BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

Identificación del Riesgo de Desarrollar Diabetes Utilizando Cómputo Suave

TESIS

Presentada para obtener el grado de:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

Presenta:

ICC. Luis Enrique Morales Márquez

Director de Tesis:

Dra. Maya Carrillo Ruíz

Asesor de Tesis:

Dr. Pedro García Juárez

Septiembre 2022

Título:

“Identificación del Riesgo de Desarrollar Diabetes Utilizando Cómputo Suave”

Estudiante:

ICC. Luis Enrique Morales Márquez

Asesores:

Dra. Maya Carrillo Ruíz

Dr. Pedro García Juárez

Agradecimientos

Quiero agradecer infinitamente a todas y cada una de las personas que me colaboraron de manera directa o indirecta durante la elaboración de este trabajo de tesis.

Primeramente, agradezco a mis padres Estela y Miguel por el enorme esfuerzo realizado a lo largo de los años para poder darme educación, vida digna y todo su apoyo sin importar la situación.

A mi hermana Maggali por estar conmigo y guiarme durante tantos años.

A mi *kitty esposa* Luz Elena por acompañarme en cada noche de trabajo, en los momentos de estrés y en los momentos de recreo, por motivarme cada día y alegrar cada momento.

A mis profesores de la maestría que resolvieron cada duda que me abordaba en clase, en particular a la Dra. Maya y al Dr. Pedro, por atenderme aún a des horas para poder llevar a cabo todas las tareas de manera satisfactoria con toda la paciencia del mundo.

A mis amigos de curso Frida y David, ¡lo logramos *amikos*! Que compartimos muchos momentos, chismes y horas de análisis sobre nuestro día a día.

Al Dr. José Gustavo por orientarme en el área de la salud, donde funge como experto.

Al Dr. Luis Enrique por guiarme en detalles de la investigación y su valiosa retroalimentación.

A la Dra. Indira, por sus atenciones y la confianza para proporcionarme datos de su investigación.

A la Dra. Josefa Somodevilla, por la orientación y luz durante el proceso de selección a la maestría.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado durante mis estudios de posgrado.

A todos aquellos que me regalaron su valioso tiempo leyendo esta tesis.

A Dios, por permitirme llegar a esta instancia.

Dedicatoria

Para mi mimor, mi pequeña Miri, te prometí esforzarme hasta el último día y superarme para ser un hombre de bien para nuestra familia. Hoy terminamos un episodio más de este camino. Gracias infinitas por recorrerlo conmigo tomados de la mano. Aún tenemos mucho más por vivir, por superar y por alcanzar.

Cada vez un poco más cerca de nuestro michi hogar <3 TE AMO, TRES MIL MILLONES.

Nunca moriré para no dejarte.

Resumen

Según la Organización Mundial de la Salud, aproximadamente el 70% de los adultos en México padecen sobrepeso u obesidad, factores determinantes en el desarrollo de diabetes mellitus tipo 2. Además, según el Instituto Nacional de Salud Pública, 10.3% de los mayores de 20 años padecen diabetes. Para facilitar las tareas de decisión o clasificación al momento de tratar a un paciente, los expertos desarrollan sistemas basados en lógica difusa, sin embargo, este diseño no suele ser infalible, por lo que es común optimizarlos para mejorar su rendimiento. El presente trabajo muestra los resultados de una comparación entre la eficiencia en la predicción de riesgo de padecimiento de diabetes tipo 2, establecida por la prueba de FINDRISC y de un sistema difuso de diseño propio optimizado por varias técnicas metaheurísticas para 295 pacientes de Acapulco, México. El algoritmo que mejores resultados proporcionó fue el Algoritmo de Polinización de Flores y la comparación frente a la prueba de FINDRISC muestra que el sistema difuso obtiene un mejor rendimiento con valores superiores de sensibilidad, especificidad y, valores predictivos positivos y negativos con mejoras generales en los intervalos de confianza, concluyendo que utilizar el sistema propuesto como auxiliar en la prevención de diabetes tipo 2 es viable y arroja resultados apegados a la realidad de los pacientes.

Abstract

According to the World Health Organization, approximately 70% of adults in Mexico are overweight or obese, determining factors in the development of type 2 diabetes mellitus. In addition, according to the National Institute of Public Health, 10.3% of those over 20 years have diabetes. To facilitate decision or classification tasks when treating a patient, experts develop systems based on fuzzy logic, however, this design is not usually infallible, so it is common to optimize them to improve their performance. The present work shows the results of a comparison between the efficiency in predicting the risk of suffering from type 2 diabetes, established by the FINDRISC test, and a fuzzy system of our own design optimized by several metaheuristic techniques for 295 patients from Acapulco, Mexico. The algorithm that obtained the best results was the Flower Pollination Algorithm and the comparison with the FINDRISC test shows that the fuzzy system obtains better performance with higher values of sensitivity, specificity, and positive and negative predictive values with general improvements in the intervals of confidence, concluding that using the proposed system as an aid in the prevention of type 2 diabetes is feasible and yields close results to the reality of the patients.

Índice general

Índice de Figuras

Índice de Tablas

Lista de Acrónimos

OMS: Organización Mundial de la Salud.

INSP: Instituto Nacional de Salud Pública.

IMC: Índice de Masa Corporal.

FIS: *Fuzzy Inference System*, Sistema de Inferencia Difuso.

ANFIS: *Adaptative NeuroFuzzy Inference System*, Sistema de Inferencia Neurodifuso Adaptativo.

IMSS: Instituto Mexicano del Seguro Social.

FINDRISC: *Finnish Diabetes Risk Score*, Puntaje Finlandés de Riesgo de Diabetes.

FINRISK: *Finland Cardiovascular Risk Study*, Estudio de Riesgo Cardiovascular de Finlandia.

DM: Diabetes Mellítus Tipo 2.

VP: Verdadero Positivo.

FP: Falso Positivo:

VN: Verdadero Negativo:

FN: Falso Negativo.

VPP: Valor Predictivo Positivo.

VPN: Valor Predictivo Negativo.

IC 95%: Intervalo de Confianza al 95%.

TSP: *Travelling Salesman Problem*, Problema del Agente Viajero.

PSO: *Particle Swarm Optimization*, Optimización por Enjambre de Partículas.

UCI: *University of California, Irvine*, Universidad de California, Irvine.

DPF: *Diabetes Pedigree Function*, Función de pedigrí de diabetes.

Capítulo 1

Generalidades

1.1 Introducción

La Organización Mundial de la Salud (OMS), en su informe mundial sobre diabetes menciona que aproximadamente el 70% de los adultos en México padecen sobrepeso u obesidad, factor importante en el desarrollo de diabetes mellitus 2, además, el Instituto Nacional de Salud Pública (INSP), en la Encuesta Nacional de Salud y Nutrición, notifica que en México 8.6 millones de personas mayores de 20 años han sido diagnosticadas con diabetes, correspondientes al 10.3% de este sector de la población [1, 2].

En el campo médico con el propósito de facilitar las tareas de decisión o clasificación al momento de tratar a un paciente, y para prevenir problemas de diabetes, se han desarrollado diversos sistemas basados en lógica difusa, no obstante, cuando se desarrolla cualquier sistema de este tipo, es de vital importancia tener una adecuada selección de las funciones y sus parámetros asociados, así como de las reglas que rigen el razonamiento del sistema, de lo contrario, el software arrojará resultados que no describen la realidad. Estos sistemas, si bien en principio suelen ser diseñados por un experto en el tema, la configuración humana no siempre resulta ser la más adecuada, por lo que no es raro hacer uso de algoritmos basados en experiencia, con el objetivo hallar configuraciones viables que ayuden al sistema a resolver el problema. Estos algoritmos, también llamados métodos heurísticos, han servido para optimizar sistemas difusos logrando describir la realidad de manera acertada. Lo expuesto anteriormente, justifica el apoyarse en sistemas computacionales que permitan conocer el riesgo de padecimiento de dicha enfermedad para llevar a cabo tareas de prevención.

El presente documento se encuentra estructurado de la siguiente manera: en el Capítulo 2 se presenta los conceptos teóricos necesarios para comprender la investigación realizada. El estado del arte que menciona algunos trabajos similares en materia de predicción de riesgo de padecimiento de diabetes utilizando lógica difusa y optimización, se muestra en el Capítulo 3, posteriormente, en el Capítulo 4 se describen las bases de datos utilizadas durante la realización de este proyecto, del que se presenta la implementación en el Capítulo 5. El Capítulo 6 muestra los experimentos llevados a cabo con sus respectivos resultados y, finalmente, las conclusiones y trabajo futuro se expresan en el Capítulo 7. Las referencias a la bibliografía consultada se encuentran en su propia sección al igual que el anexo.

1.2 Motivación

Este trabajo aporta una nueva herramienta de prevención de una enfermedad de alta prominencia en México, puesto que cada día aumenta la cantidad de diagnósticos positivos de diabetes tipo 2 en nuestro país. Si bien, dicho padecimiento es tratable, a largo plazo para la salud pública, es preferible llevar a cabo tareas de prevención en pacientes que pueden desarrollar la enfermedad y así, dirigirles hacia un estilo de vida adecuado que les permita alejarse de la diabetes. Además, se desea poner los conceptos del cómputo suave al servicio del sector salud y de la sociedad mexicana, demostrando así, una vez más, que las ciencias

computacionales pueden ser de gran ayuda para dar solución a la mayoría de los problemas existentes sin importar el área de estudio.

1.3 Objetivo general

Desarrollar un sistema difuso capaz de predecir el riesgo de padecimiento de diabetes de un paciente con base en datos clínicos, como: edad, IMC o probabilidad de padecimiento según historial familiar, entre otros, donde algunos de los parámetros de las funciones de membresía serán optimizados por técnicas metaheurísticas, buscando obtener resultados comparables a los obtenidos por los expertos, definidos en el corpus utilizado.

1.4 Objetivos específicos

- Revisar la viabilidad de uso de una red neuronal en la optimización de un sistema de lógica difusa.
- Estudiar y analizar las diferentes técnicas metaheurísticas que pueden utilizarse en la optimización de parámetros de funciones de membresía.
- Establecer los atributos para caracterizar al paciente.
- Estudiar y elegir los factores de riesgo que influyen mayormente en el diagnóstico de diabetes.
- Diseñar un sistema difuso completo para mostrar el porcentaje de riesgo de padecimiento de diabetes.
- Proponer métricas adecuadas para evaluar el rendimiento del sistema desarrollado.
- Comprobar experimentalmente el funcionamiento de las diferentes variantes del sistema propuesto.
- Comparar el rendimiento mostrado por el sistema al ser optimizado respecto a un estándar de oro.
- Llevar el sistema a los usuarios a través de una interfaz gráfica de fácil uso.

1.5 Infraestructura utilizada

- Computadora personal (Intel® Core™ i5-10300H Processor 2.5 GHZ/16GB RAM DDR4/NVIDIA® GeForce® GTX 1650 Ti 4GB GDDR6/Microsoft Windows 10 Home x64.
- MATLAB R2022a 64 bits.
- Microsoft 365 Apps for enterprise.
- Canva PRO Personal.

1.6 Impacto socieconómico y aportaciones

El sistema propuesto deberá poder ser usando en zonas rurales o de difícil acceso donde rara vez se presenta un experto para atender a pacientes potenciales, pues una sola persona con acceso al sistema puede calcular el riesgo de un grupo de personas y dirigirles directamente con un especialista en caso necesario. Incluso en las mismas zonas urbanas cualquier persona con acceso libre al sistema puede evaluar su nivel de riesgo y ser dirigido con un especialista. Las reglas difusas que se obtengan serán de utilidad para trabajos futuros relacionados, pues el conjunto que se obtenga se espera que dé un mejor resultado en la predicción, además de presentar a la sociedad un producto final en forma de software de predicción de riesgo de padecimiento de diabetes para ser utilizado como herramienta adecuada de prevención.

Capítulo 2

Marco Teórico

2.1 Lógica difusa

La lógica tradicional nos orienta al uso de dos valores al momento de definir la pertenencia de un objeto x a un determinado conjunto A, es decir, si x pertenece al conjunto A, el valor de membresía asignado es 1, de lo contrario, el valor de membresía es 0, este concepto se conoce como conjunto certero, sin embargo, este enfoque no necesariamente refleja la realidad que se intenta describir. Para reducir las limitaciones de la aproximación clásica se introduce el concepto de conjunto difuso propuesto por Zadeh en 1965 [3].

2.1.1 Conjunto difuso

Un conjunto difuso se entiende como un conjunto A de pares ordenados, cada dupla está formada por el elemento x y su valor de membresía en el conjunto A. Este valor de membresía está dado por $f_A(x)$ llamada función de membresía, el grado de pertenencia es un número real que sólo puede estar en el intervalo [0,1]. Por ejemplo:

Sea X el conjunto de todos los números x en la recta en $\mathbb{R}^1$ y sea A el conjunto difuso de los números que son mucho más grandes que 1. Note que sería muy difícil establecer un punto de corte a partir del cual un número pueda considerarse mucho más grande que 1.

Para alguna función $f_A(x)$ en $\mathbb{R}^1$, algunos valores de membresía pueden ser: $f_A(0) = 0; f_A(1) = 0;\ f_A(5) = 0.01;\ f_A(10) = 0.2;\ f_A(100) = 0.95;\ f_A(500) = 1$.

Este ejemplo muestra que para cualquier valor x menor o igual que 1, el valor de membresía es cero, pues en definitiva ninguno de estos valores cumple con ser mayor que 1, además, mientras x se aleja más de 1, el valor de membresía aumenta. En realidad, esto correspondería a un conjunto difuso infinito, sin embargo, el mismo concepto puede ser aplicado a conjuntos finitos. La función de membresía representa al conjunto difuso a lo largo de todo el intervalo que considera a los elementos de x, dicho intervalo es conocido como el *universo de discurso* y se denota como X, de modo que a partir del universo discurso y la membresía en el intervalo [0,1] se construye un segmento de plano en $\mathbb{R}^2$ donde se puede ilustrar la función de membresía.

2.1.2 Función de membresía

La función de membresía utilizada en el ejemplo anterior no está definida de alguna manera en específico, sin embargo, puesto que, en general no existen reglas sobre la forma visual de la función que se deba utilizar más allá de representar una forma convexa, se puede elegir de entre una amplia gama, a aquella que mejor se ajuste al problema abordado.

Una manera de clasificar las funciones de membresía es en funciones abiertas por la izquierda, abiertas por la derecha o cerradas [4], ver Figura 1. Por simplicidad, se utilizará la nomenclatura μ para denotar una función de membresía genérica.

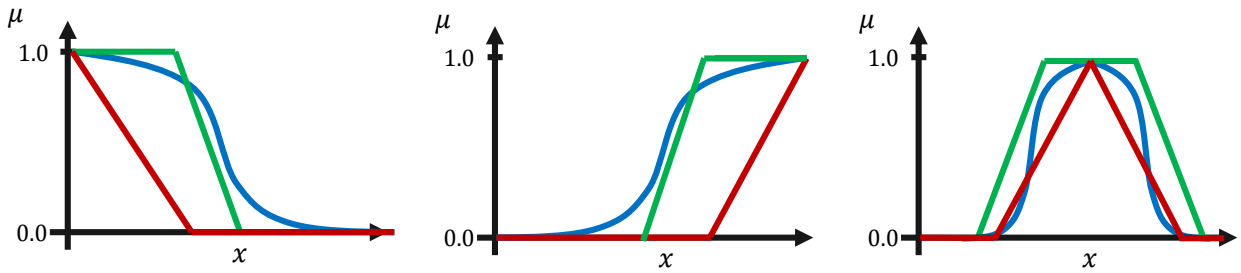

Figura 1. Funciones de membresía abiertas por la izquierda, abiertas por la derecha y cerradas.

También se pueden catalogar como simétricas, o no simétricas respecto al punto central de la función de membresía, ver Figura 2:

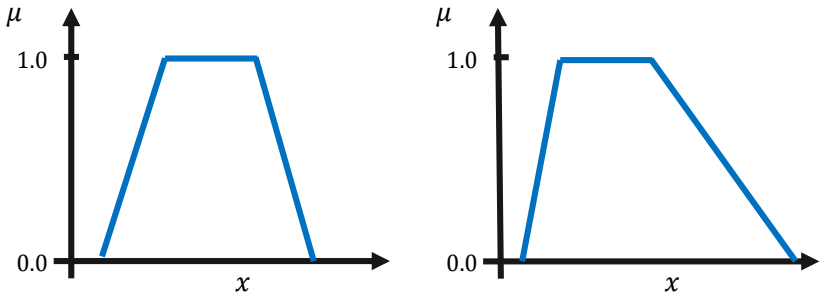

Figura 2. Funciones de membresía simétrica y no simétrica.

Para el uso en sistemas de lógica difusa, las funciones más comunes de usar son:

- Trapezoidal, como se muestra en la Figura 3, dada por la ecuación (1):

$$f(x;a,b,c,d) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \frac{d-a}{d-c} & c \le x \le d \\ 0 & x \ge d \end{cases} \qquad a \le b \le c \le d \tag{1}$$

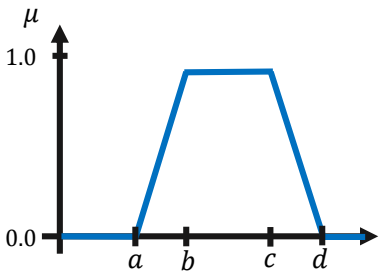

Figura 3. Función de membresía trapezoidal.

- Triangular, que es un caso particular de la función trapezoidal cuando $b = c$ en términos del trapecio. Por simplicidad, b toma el lugar de ambos parámetros, ver Figura 4, dicha forma está dada por la ecuación (2):

$$f(x;a,b,c) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ \frac{c-x}{c-b} & b \le x \le c \\ 0 & x \ge c \end{cases} \qquad a \le b \le c \tag{2}$$

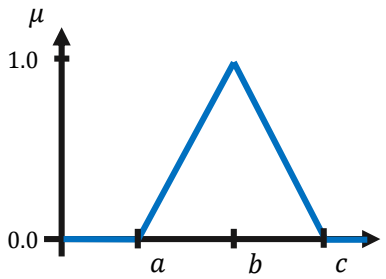

Figura 4. Función de membresía triangular.

Para generar versiones abiertas de estas dos funciones, basta con omitir los primeros o últimos dos trozos de las funciones.

- Si se desea establecer una figura derivable, se tiene la forma Campana Generalizada, donde se debe fijar el centro x_0, el ancho a y la pendiente b, véanse Figura 5 y ecuación (3):

$$f(x;a,b,x_0) = \frac{1}{1 + \left|\frac{x - x_0}{a}\right|^{2b}} \tag{3}$$

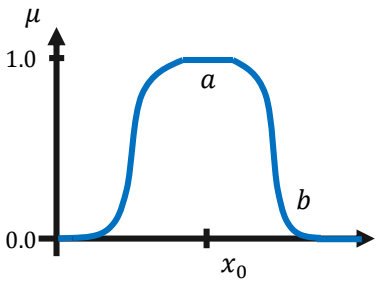

Figura 5. Función de membresía campana generalizada.

La documentación del Fuzzy Logic Toolbox de MATLAB [5] también sugiere el uso de funciones S y funciones Z, que tiene forma muy similar a las formas sigmoidales:

- Función de membresía en forma de Z, donde a determina el punto donde la función toma el valor de 1 y b donde la función toma el valor de 0, la función se determina por la ecuación (4) y puede observarse en la Figura 6:

$$f(x;a,b)=\begin{cases}1 & x\le a\\ 1-2\left(\frac{x-a}{b-a}\right)^2 & a\le x\le \frac{a+b}{2}\\ 2\left(\frac{x-b}{b-a}\right)^2 & \frac{a+b}{2}\le x\le b\\ 0 & x\ge b\end{cases} \tag{4}$$

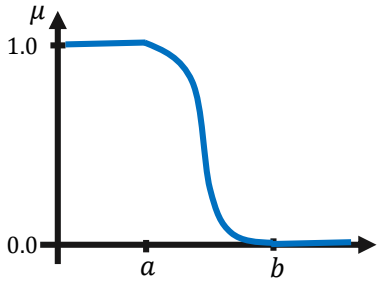

Figura 6. Función de membresía en forma de Z.

- Función de membresía en forma de S, donde a determina el punto donde la función toma el valor de 0 y b donde la función toma el valor de 1, esta función se expresa en la ecuación (5) y se puede observar en la Figura 7:

$$f(x;a,b)=\begin{cases}0 & x\le a\\ 2\left(\frac{x-a}{b-a}\right)^2 & a\le x\le \frac{a+b}{2}\\ 1-2\left(\frac{x-b}{b-a}\right)^2 & \frac{a+b}{2}\le x\le b\\ 1 & x\ge b\end{cases} \tag{5}$$

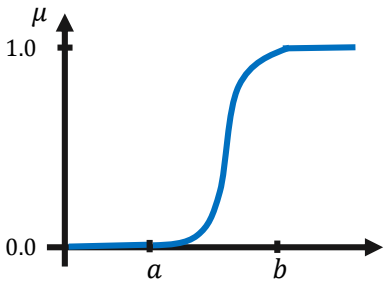

Figura 7. Función de membresía en forma de S.

El uso de cada función se queda a discreción del diseñador del sistema difuso, pues esta selección depende del juicio de un experto en el tema. Como ejemplo, considere el mapeo da las funciones de membresía para la temperatura en grados centígrados, donde las variables lingüísticas o categorías a usar

serán *frío*, *fresco*, *cálido* y *caliente*, y el universo discurso está definido entre 0° y 80° [6], la Figura 8, muestra el resultado de definir estas variables lingüísticas, note que para temperaturas de 0° e inferiores, el valor de membresía de la variable lingüística Frío es 1 mientras que para todas las demás variables, el valor de membresía es 0, para temperaturas mayores que 0° pero inferiores a los 20°, el valor de membresía para Frío disminuye a medida que el valor para Fresco asciende, a los 20°, la variable fresco obtiene un valor de membresía de 1 mientras que obtiene 0 para el resto de las variables, el comportamiento es similar a medida que se continúe analizando el universo discurso. Considere que la selección aquí mostrada es adecuada según un experto para algún problema que se abordará.

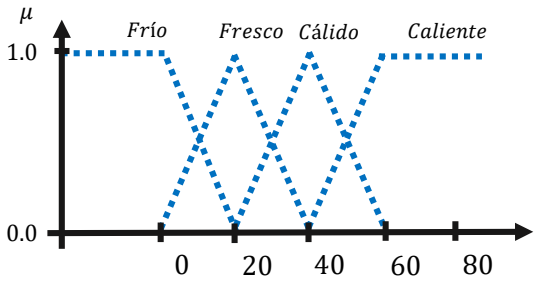

Figura 8. Funciones de membresía para Temperatura.

2.1.3 Sistema de inferencia difusa

Los humanos nos comunicamos en nuestro propio lenguaje natural con términos simples que pueden implicar ambigüedad, por ejemplo, términos como *muy lento*, *joven*, o *muy feo*. A partir de esta manera de hablar se generan modelos lingüísticos usando sistemas difusos, un sistema de este tipo para cualquier propósito incluye un módulo de Sistema de Inferencia Difuso (FIS). De manera sistemática, un FIS es una manera de mapear el espacio de entrada al espacio de salida a través de la toma de decisiones basadas en reglas. En el campo de la inteligencia artificial, existen diversas formas de representar el conocimiento, sin embargo, la más común es encapsularlo en expresiones del tipo *SI premisa (antecedente), ENTONCES conclusión (consecuente)* [6], este modelo recibe el nombre de regla *SI-ENTONCES*, y está asociado al pensamiento deductivo partiendo de un hecho o premisa para inferir un segundo hecho, también llamado concusión.

Los sistemas basados en reglas son especialmente útiles al modelar sistemas complejos apegados a la manera humana del pensamiento, pues hacen uso de variables lingüísticas, antecedentes, consecuentes y conectivos lógicos para establecer las reglas del sistema. Los conectivos lógicos *"y"* y *"o"*, tienen el propósito de establecer restricciones, además, estas reglas, así como las funciones de membresía asociadas, se encuentran almacenadas en la base de conocimiento, que es legible por la máquina, misma que es capaz de obtener razonamiento deductivo automático a partir de ella, el conocimiento almacenado debe ser lógico y consistente con la realidad y puede ser expandido [7].

Un FIS se compone fundamentalmente de tres etapas [6, 8], la primera es la fuzificación, proceso en el que una cantidad certera es convertida a una cantidad difusa, esto se logra obtenido el valor de membresía del dato de entrada x en cada una de las variables lingüísticas, es decir, la conversión de una cantidad precisa a una cantidad difusa, posteriormente se hace uso del motor de inferencia, el cual aplica las reglas de implicación *SI-ENTONCES* para transformar las variables de entrada fuzificadas en una variable difusa de salida haciendo uso de la base de conocimiento.

Finalmente se lleva a cabo el proceso de defuzificación, pues en la mayoría de las ocasiones, la salida difusa de un sistema debe ser cuantificada como un escalar único o un valor del universo discurso, el procedimiento más común es hacer el cálculo del promedio ponderado pues es de los métodos computacionalmente más eficientes, aunque solo es aplicable para funciones de membresía de salida que sean simétricas, la expresión para el promedio ponderado, obtenida con la ecuación (6):

$$z^* = \frac{\sum \mu_C(\bar{z}) \cdot \bar{z}}{\sum \mu_C(\bar{z})} \tag{6}$$

La estructura básica de un Sistema de Inferencia Difuso se muestra en la Figura 9:

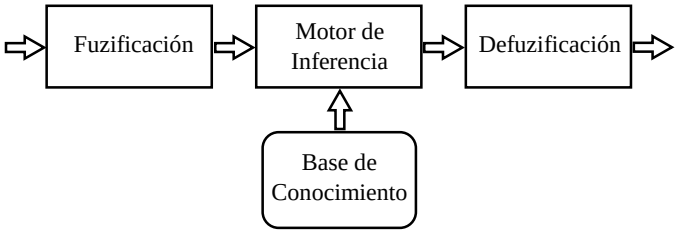

Figura 9. Estructura del Sistema Difuso de Inferencia (FIS). Fuente: Jain & Raheja [8], traducción propia.

El método de inferencia más utilizado tanto en la literatura como en la implementación es el método de Mamdani [6], desarrollado por Mamdani y Assilian en 1975, el cual, considera varias reglas con varias variables de entrada y una única variable salida.

Como ejemplo, se muestra una colección de r reglas de inferencia compuestas de 2 variables de entrada x_1, x_2 en el antecedente con un único consecuente:

$$SI\ x_1\ es\ A_1^k\ y\ x_2\ es\ A_2^k\ ENTONCES\ y^k\ es\ B^k \quad k = 1,2,\dots,r \tag{7}$$

Donde $A_1^k\ y\ A_2^k$ son conjuntos difusos que representan el antecedente de la k-ésima regla y B^k es el conjunto difuso que representa al k-ésimo consecuente.

Generalmente se utiliza el proceso *mínimo-máximo*: Para cada uno de los hechos x_1 es A_1^k y x_2 es A_2^k, se aplica el operador de intersección usando el mínimo de los valores de membresía, este será el valor al que se corte la función de salida B^k, este proceso se repite para todas las r reglas del sistema. Finalmente, se obtiene la unión con el máximo de todas las funciones de membresía de los consecuentes B^k, este último proceso recibe el nombre de agregación, pues integra todos los valores a la salida de todas las reglas del sistema. Esta técnica se muestra gráficamente en la Figura 10 y puede expresar como:

$$f_{B^k}(y) = \max_k \{min\{f_{A_1^k}(entrada(i)), f_{A_2^k}(entrada(j))\}\}, \qquad k = 1,2,\dots,r \tag{8}$$

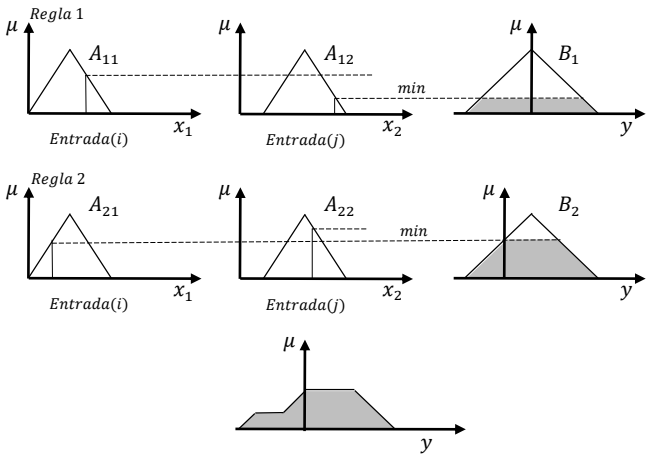

Figura 10. Inferencia mínimo-máximo para el método de Mamdani. Fuente: Ross [6], traducción propia.

2.2 Sistema de inferencia neurodifuso adaptativo

Como se ha comentado anteriormente, la configuración establecida por un experto no está exenta de fallos, con el propósito de explorar la posibilidad de omitir la participación de un experto humano, se han desarrollado sistemas difusos que se ajustan de manera automática. Estos sistemas adaptativos hacen uso de redes neuronales y técnicas de aprendizaje profundo, la mayormente usada es el Sistema de Inferencia Neurodifuso Adaptativo (ANFIS) por sus siglas en inglés para *Adaptative NeuroFuzzy Inference System*. En un intento de evitar la parte de ajuste por heurísticas para el sistema de predicción de riesgo de padecimiento de diabetes, se trae consideración el uso de un ANFIS para el proyecto.

Los sistemas ANFIS surgen como un esfuerzo por fusionar las ideas de los sistemas difusos y las redes neuronales buscando conjuntar las ventajas de ambos modelos. El objetivo es aprovechar el poder de aprendizaje de bajo nivel de las redes neuronales al mismo tiempo que el razonamiento *SI-ENTONCES* de alto nivel de los sistemas difusos, esta estructura de red neuronal tiene como consecuentes combinaciones lineales del antecedente. La forma de las reglas está dada como:

$$R^j: SI\ x_1 es\ A_1^j\ \&\ x_2 es\ A_2^j\ \&\ \dots\ \&\ x_n es\ A_n^j\ ENTONCES\ y = f_j = a_0^j + a_1^j x_1^j + a_2^j x_2^j + \cdots + a_n^j x_n^j \qquad (9)$$

Donde x_i es cada una de las variables de entrada, y es la salida del sistema, A_i^j es cada una de las variables lingüísticas necesarias en el antecedente y f_j es una combinación lineal que involucra a las variables a_i^j [9].

Una ANFIS consta de 5 capas como se muestra en la Figura 11 y sus capas se describen de la siguiente manera:

- Capa 1: Cada nodo es una variable de entrada al sistema, pasa su valor a la siguiente capa.
- Capa 2: Cada nodo es una función de membresía de cada una de las variables lingüísticas del sistema, el resultado es la fuzificación de la variable de entrada.
- Capa 3: Cada nodo es una regla de inferencia del sistema, y su salida se calcula utilizando operaciones entre conjuntos difusos.

- Capa 4: Cada nodo calcula el valor del consecuente de su respectiva regla de inferencia asociada, aplicando la combinación lineal requerida.
- Capa 5: El nodo final aplica defuzificación a través de una media ponderada entregando un dato de salida único.

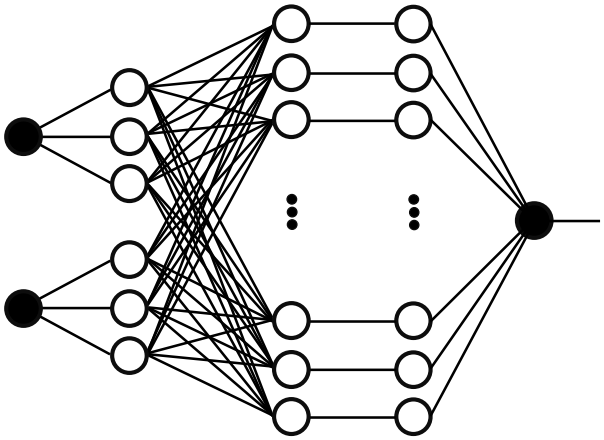

Figura 11. Estructura del ANFIS.

Un ANFIS posee propiedades de autoajuste de las redes neuronales, puede aprender en base a los ejemplos usados durante el entrenamiento aprovechando la propagación hacia atrás, los parámetros que suelen optimizarse son los requeridos por las funciones de membresía. Las formas más comunes de usar son campanas gaussianas por ser continuas y derivables, aunque también se suelen utilizar triangulares, trapezoidales, entre otras, además, se busca el valor adecuado de los coeficientes a_i^j [9].

Cuando se tiene una cantidad considerable de variables de entrada, las reglas de inferencia se suelen calcular con el método de *Substractive Clustering Method*. Esta técnica permite dividir el espacio n dimensional generado por las variables de entrada en m clústers, no se elige el valor de m pues este depende de algunos factores y se determina automáticamente siguiendo un proceso:

Primeramente, se elige el *rango de influencia* o *radio del clúster*, que tiene un valor entre 0 y 1, la recomendación es asignar 0.5 [10]. Representa la influencia que el centro del clúster tendrá sobre los demás elementos, un valor muy grande generar un clúster más grande, por lo que el conjunto de datos a agrupar tendrá un menor número de clústers al terminar la operación cuando se utiliza un valor muy grande, posteriormente se considera el *factor de depreciación* que decrementará el valor del potencial de los datos cercanos al centro del clúster, normalmente se establece con un valor de 1.5 [10].

Posteriormente se calcula el potencial o densidad de cada dato i a agrupar a través de la ecuación (10):

$$P_i = \sum_{j=1}^{m} e^{-\frac{||x_i - x_j||}{\frac{r_a^2}{2^2}}} \tag{10}$$

Donde r_a es el radio de influencia, después, el dato con mayor potencial será el centro x_{c1} del primer clúster. Para todos los puntos dentro del radio del clúster, se calcula un según potencial con la esperanza de encontrar un mejor candidato para ser el centro del clúster, esto se hace con la ecuación (11):

$$P_i = P_i - P_{c1} e^{-\frac{\left\|x_i - x_j\right\|}{\frac{r_b^2}{2^2}}} \tag{11}$$

Donde r_b es el factor de depreciación, el dato con el mejor nuevo potencial P_k ha de ser dividido entre el potencial P_{h1} del mejor centro que se obtuvo primero, este cociente se expresa en la ecuación (12):

$$ratio = \frac{P_k}{P_{h1}} \tag{12}$$

Después, debe ser evaluado respecto al *rango de aceptación* y *rango de rechazo*, los valores recomendados en [10] son 0.5 y 0.15 respectivamente. Se puede encontrar 3 posibles casos en la evaluación:

1. $ratio > rango\ de\ aceptación$: El dato es aceptado inmediatamente como centro del clúster.
2. $ratio < rango\ de\ rechazo$: El mejor potencial no pudo ser nuevo clúster, se detiene la iteración y el proceso se repite desde el inicio con los datos que no están dentro de ningún clúster.
3. $rango\ de\ rechazo \leq ratio \leq rango\ de\ aceptación$: Sólo se acepta como nuevo centro del clúster si el dato está suficientemente lejos de algún otro centro de clúster, es decir si la suma del potencial de este dato con el dato más cercano dentro de otro clúster es mayor que 1.

Un rango de aceptación demasiado grande provocará que pocos datos puedan ser nuevos clústeres, así como un rango de rechazo muy bajo provocará demasiados cambios de centro, un ejemplo de funcionamiento se puede ver en la Figura 12.

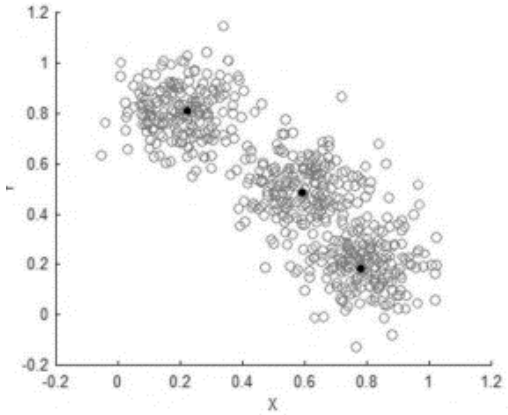

Figura 12. Agrupamiento por Substractive Clustering con 3 clústers [10].

La ANFIS es una herramienta de gran utilidad principalmente cuando no se tiene una idea clara del funcionamiento de un sistema o del problema que se aborda. Además de hacer automático el ajuste, también

los resultados son generalmente buenos con tan solo presentar un conjunto de datos de entrenamiento, sin embargo, se presentan ciertas desventajas en su uso, por ejemplo, el funcionamiento es mejor con un numero pequeño de variables de entrada, la optimización se hace solamente en las funciones de membresía y los coeficientes de las combinaciones lineales o consecuente, pero no se hace ningún cambio en las reglas de inferencia, sin mencionar que estas reglas generadas son muy poco flexibles. Como toda red neuronal, corre el riesgo de sobreajuste, es decir, el error de aproximación es mucho menor que el de generalización.

Puesto que la construcción y adaptación de las redes neuronales depende totalmente de los datos de entrada, la forma en que los datos son presentados puede afectar significativamente al resultado, se definen la estandarización de datos por puntuación Z para su uso y prueba en experimentos posteriores [11].

La puntuación Z, también llamada puntuación estándar, establece el número de desviaciones estándar a la izquierda o a la derecha a las que se encuentra un dato x respecto a la media, la conversión se muestra en la ecuación (38):

$$\boldsymbol{z = \frac{x - \mu}{\sigma}} \tag{13}$$

Donde x es el dato para puntuar, μ es la media aritmética de los datos, σ es la desviación estándar, un valor infrecuente o atípico es aquel que está más allá de 2 o -2 desviaciones estándar [11], por lo que el tratamiento de estos datos depende del propósito para el que se utilicen.

2.3 Prueba de FINDRISC

Es un instrumento de cribaje utilizado para valorar el riesgo de un individuo de desarrollar diabetes tipo 2 dentro de los próximos 10 años, es ampliamente utilizado alrededor del mundo por su simpleza de uso y basarse en un cuestionario de fácil aplicación, incluso es aplicado en el Instituto Mexicano del Seguro Social (IMSS) en unidades de medicina familiar [12].

La prueba de FINDRISC fue presentada en el año 2003 por Lindström y Tuomilehto [13], con el objetivo de disponer de una herramienta no invasiva y de fácil aplicación en unidades de medicina familiar para medir el riesgo de padecimiento de diabetes mellitus tipo 2 para personas de Finlandia. El estudio se basa en los estudios FINRISK que originalmente monitoreaban a pacientes durante 10 años en el desarrollo y tratamiento de diabetes si resultaban afectados.

A partir de 4746 individuos de prueba, se obtuvieron datos descriptivos que no requieran estudios especiales de laboratorio, además, se incluyen preguntas de carácter binario, posteriormente se aplicó regresión logística determinando los coeficientes que se utilizan para el cálculo de la probabilidad de padecimiento de diabetes. El modelo tiene un máximo de 20 puntos y es estrictamente categórico, por lo que, sin importar el valor específico de un dato, si se encuentra en algún intervalo, se usará el mismo coeficiente de toda la categoría, lo que impide establecer un porcentaje de riesgo personalizado.

La probabilidad de riesgo de padecimiento de diabetes tipo 2 en los próximos 10 años se obtiene mediante la fórmula en la ecuación (14):

$$\boldsymbol{p(diabetes)} = \frac{\boldsymbol{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots)}}}{\boldsymbol{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots)}}} \quad (14)$$

Donde β_0 es el coeficiente interceptor y β_i es el coeficiente asociado a la variable categórica x_i, naturalmente x_i solo puede adoptar valores de 1 o 0, el resultado es el porcentaje de riesgo de padecimiento dentro de los próximos 10 años. Para evadir el cálculo directo, se determinó una asignación de puntajes para cada reactivo. El modelo es validado con una sensibilidad y especificidad de 0.78 y 0.77 para el grupo original FINRISK y con un segundo grupo que obtiene 0.81 y 0.76 respectivamente.

El articulo original de FINDRISC menciona el uso de una pregunta sobre historial familiar, donde se cuestiona al individuo si tiene familiares que hayan sido detectados con diabetes tipo 2, y determina el un puntaje de 0, 3 o 5 puntos, para ningún familiar detectado, familiar cercano (padres, hermanos o hijos) y familiares menos cercanos (abuelos, tíos o primos hermanos) respectivamente. La Guía de Práctica Clínica del IMSS [12] sí toma en cuenta la pregunta sobre historial familiar además de agregar una categoría adicional para la edad, esto establece una puntuación máxima de 26 puntos. El cuestionario que se aplica en el IMSS se muestra en la Tabla 1, y el riesgo calculado con la suma de puntos se muestra en la Tabla 2.

Tabla 1. Cuestionario y puntajes de la prueba de FINDRISC del IMSS.

Variable	Puntaje
Edad <45	0
Edad 45-54	2
Edad 55-64	3
Edad >64	4
IMC <25	0
IMC 25-30	1
IMC >30	3
Circunferencia Abdominal Masculino <94 Femenino <80	0
Circunferencia Abdominal Masculino 94-101 Femenino 80-87	3
Circunferencia Abdominal Masculino >101 Femenino >87	4
¿Ha sido tratado con medicamento para la presión arterial alguna vez? SÍ	2
¿Le han detectado niveles altos de glucosa alguna vez? SÍ	5
¿Realiza menos de 4 horas de actividad física a la semana? SÍ	2
¿Su dieta carece de frutas, vegetales o bayas? SÍ	1
¿Ha habido algún diagnóstico de DM en su familia? NO	0
¿Ha habido algún diagnóstico de DM en su familia? SÍ: Abuelos, tíos o primos hermanos	3
¿Ha habido algún diagnóstico de DM en su familia? SÍ: Padres, hermano o hijos	5

Tabla 2. Posibles resultados de la prueba de FINDRISC.

Puntuación	Riesgo de desarrollar diabetes en los próximos 10 años	Interpretación
$\mathbf{P < 7}$	1%	Riesgo bajo
$\mathbf{7 \leq P \leq 11}$	4%	Riesgo ligeramente elevado
$\mathbf{12 \leq P \leq 14}$	17%	Riesgo moderado
$\mathbf{15 \leq P \leq 20}$	33%	Riesgo alto
$\mathbf{P > 20}$	50%	Riesgo muy alto

Así, la prueba de FINDRISC fue el estándar que se usará para comprobar el funcionamiento del sistema propuesto.

2.4 Herramientas para la interpretación de estudios de exactitud diagnóstica

Para describir la utilidad de los sistemas o pruebas para diagnóstico de enfermedades se definen los siguientes conceptos [14]:

- Verdadero positivo (VP): El paciente tiene la enfermedad y la prueba es positiva.
- Falso positivo (FP): El paciente no tiene la enfermedad, pero la prueba es positiva.
- Verdadero negativo (VN): El paciente no tiene la enfermedad y la prueba es negativa.
- Falso negativo (FN): El paciente tiene la enfermedad, pero la prueba es negativa.

De ellos, se desprende la tabla de contingencia [15], ver Tabla 3.

Tabla 3. Tabla de contingencia para evaluación de una prueba diagnóstica.

		Patrón de referencia +	Patrón de referencia −
Prueba diagnóstica	+	Verdaderos positivos (VP)	Falsos positivos (FP)
	−	Falsos negativos (FN)	Verdaderos negativos (VP)

Y se definen algunas proporciones utilizando valores de la matriz de contingencia

- Sensibilidad: Proporción de individuos correctamente diagnosticados con la enfermedad por la prueba diagnóstica respecto a un estándar de referencia, ver ecuación (15):

$$\boldsymbol{sensibilidad} = \frac{\boldsymbol{VP}}{\boldsymbol{VP + FN}} \tag{15}$$

- Especificidad: Proporción de individuos correctamente diagnosticados con ausencia de la enfermedad por la prueba diagnóstica respecto a un estándar de referencia, ver ecuación (16):

$$\boldsymbol{especificidad} = \frac{\boldsymbol{VN}}{\boldsymbol{VN} + \boldsymbol{FP}} \tag{16}$$

Por su parte, los valores predictivos son estimaciones de la probabilidad de que la prueba diagnóstica entre un resultado correcto si esta resulta negativa o positiva.

- Valor predictivo positivo (VPP): Probabilidad condicional de que el paciente tenga la enfermedad dado que la prueba resultó positiva, es decir, la proporción de pacientes con prueba diagnóstica positiva que realmente padecen la enfermedad, ver ecuación (17):

$$\boldsymbol{VPP} = \frac{\boldsymbol{VP}}{\boldsymbol{VP} + \boldsymbol{FP}} \tag{17}$$

- Valor predictivo negativo (VPN): Probabilidad condicional de que el paciente no tenga la enfermedad dado que la prueba resultó negativa, es decir, la proporción de pacientes con prueba diagnóstica negativa y que no padece la enfermedad, ver ecuación (18):

$$\boldsymbol{VPN} = \frac{\boldsymbol{VN}}{\boldsymbol{VN} + \boldsymbol{FN}} \tag{18}$$

Los valores predictivos son estimaciones de la probabilidad de que la prueba diagnóstica entregue un resultado correcto dado que esta resulte negativa o positiva, según sea el caso. Puesto que las proporciones mencionadas son estimaciones obtenidas con estudios llevados a cabo con muestras de población, están sujetas a variabilidad, así, para poder utilizarlas adecuadamente es necesario calcular sus intervalos de confianza, en este caso, se elige el intervalo de confianza al 95% (IC 95%). En otras palabras, se determinará el intervalo en el que el 95% de posteriores pruebas del sistema concentraran sus valores de las métricas mencionadas [15]. Se utiliza un método estándar a partir de la proporción binomial, se muestra en la ecuación (19) [14]:

$$\boldsymbol{p} \pm \boldsymbol{Z_{0.95}}^{2} \sqrt{\frac{\boldsymbol{p}(\boldsymbol{1} - \boldsymbol{p})}{\boldsymbol{N}}} \tag{19}$$

Donde p es la proporción a la que se le está estimando el intervalo de confianza, N el tamaño muestral y $Z_{0.95}$ se obtiene a partir de las tablas de la función normal, para una campana de dos colas, $Z_{0.95} = 1.96$.

Cabe resaltar que los intervalos de confianza deben ser tomados en cuenta al momento de analizar la validez de la prueba, un tamaño muestral N demasiado pequeño, menor que 30 [14], corre el riesgo de mostrar intervalos de confianza sumamente amplios, dejando limitada la validez de la prueba [15]. El tamaño óptimo del intervalo no se encuentra escrito por ninguna regla, queda sujeto al tipo de estudio y las condiciones relacionadas.

Se agrega un concepto un poco más simple, el de exactitud que indica un resultado que se acerca al valor de referencia o también llamado valor real o magnitud real. Mientras más cercano al valor real, mayor exactitud de los resultados [16].

2.5 Optimización

El interés por las tareas de optimización se puede ver diversas áreas, desde la ingeniería o la cocina hasta las disciplinas comerciales, el objetivo siempre se fija en maximizar ganancias, producción, rendimiento o eficiencia, o bien, minimizar costos, consumos de energía, o distancias de recorrido. Puesto que recursos como dinero y tiempo son limitados, la búsqueda de soluciones óptimas también debe llevarse a cabo utilizando de la mejor manera estos recursos considerando las restricciones que se impongan, y como muchos de los problemas del mundo real son complejos, el uso del cómputo se ha vuelto indispensable para hacer frente a las tareas.

Matemáticamente hablando, la mayoría de los problemas de optimización se escriben de la forma:

$$\min_{x\in\mathbb{R}^d} f_i(x), \qquad i = 1,2,\dots,M$$

Sujeto a:

$$\begin{aligned} \boldsymbol{h_j(x) = 0,}\ \ \boldsymbol{j = 1,2,\dots,J} \\ \boldsymbol{g_k(x) \le 0,} \qquad \boldsymbol{k = 1,2,..,K} \end{aligned} \tag{29}$$

Donde $f_i(x)$ son las funciones objetivo a optimizar o funciones de costo y $h_j(x)$, $g_k(x)$ son las funciones de restricción, además $x = (x_1, x_2, \dots, x_d)^T$ es el vector de variables de decisión, cuando $M = 1$, se dice que solo existe una función de objetivo única, de lo contrario, se considera un problema multiobjetivo. El espacio generado por la o las funciones objetivo donde se ha de buscar la solución se conoce como espacio de soluciones o espacio de respuesta.

En algunos casos no existe una función objetivo o no se puede expresar de manera explícita y sólo hay restricciones, esto se conoce como un problema de factibilidad y cualquier solución factible, es decir, que cumpla con las restricciones de problema es una solución óptima, en este caso recae el problema que aborda este trabajo de tesis.

Si el área donde se busca el óptimo es extremadamente grande, hallarlo puede tomar una cantidad considerable de tiempo. Si el tiempo disponible fuese infinito, en algún punto se debería poder alcanzar la solución óptima global, sin embargo, como hemos mencionado previamente, el tiempo es un recurso valioso cuyo uso debe minimizarse.

Las técnicas para optimizar suelen ser deterministas, pues se sigue una serie pasos rigurosos y los resultados siempre son repetibles, por supuesto, cada aplicación al problema en cuestión arrojará siempre los mismos resultados, pero si la solución hallada no tiene la calidad suficiente, es inútil repetir el procedimiento. Por otro lado, los algoritmos estocásticos aportan una parte aleatoria, por lo que las

diferentes soluciones generadas por el programa serán distintas en cada ejecución. Si bien, es en teoría imposible repetir la misma cadena de resultados, se puede guiar esta búsqueda con secciones o pasos determinados utilizando ciertas características aleatorias con el objetivo de buscar de manera aleatoria bajo regiones controladas.

En la mayoría de los casos no se puede saber que tan eficiente será un algoritmo para un problema específico antes de probarlo, en gran medida, la elección del algoritmo depende de la experiencia de la persona asignada a resolver el problema, los recursos disponibles y las condiciones del problema. La manera más común de elegir un algoritmo es probando una selección de ellos y comparando sus resultados, aunque el conocimiento previo puede ayudar a dirigir los algoritmos para buscar soluciones en regiones seleccionadas, el tiempo de cómputo debe ser cuidado, en este sentido, cualquier conocimiento del problema siempre es de utilidad para elegir el algoritmo más eficiente [17].

2.6 Métodos metaheurísticos

Los algoritmos metaheurísticos tratan de encontrar o descubrir a través de prueba y error soluciones de calidad en cantidades razonables de tiempo, pero no se puede garantizar el hallazgo de la solución óptima. Esta característica es especialmente útil cuando no se quiere estrictamente la mejor solución y en su lugar se solicita una solución suficientemente buena que se pueda encontrar fácilmente.

Actualmente se suele llamar metaheurísticos a aquellos algoritmos estocásticos que incluyen una parte aleatoria en conjunto a una búsqueda local en un segmento del espacio de búsqueda. La parte aleatoria es aquella que permite al algoritmo moverse de una búsqueda local hacia una búsqueda global, mientras que la búsqueda local se enfoca en explorar una región considerada como buena en busca de la mejor solución que se pueda aportar dicho subespacio. El uso de las metaheurísticas hace posible explorar una gran parte de las soluciones posibles en un tiempo aceptable, aunque no se sepa porqué funcionan para hallar la solución, si es que funcionan, así, la tarea se convierte en tener un algoritmo eficiente en la práctica y que genere buenas soluciones en la mayoría de las ocasiones.

Los algoritmos metaheurísticos son por naturaleza iterativos, pues buscan construir mejores soluciones con base en las ya obtenidas con anterioridad. De manera matemática un algoritmo A como proceso iterativo trata de generar una mejor solución x^{t+1} para un problema a partir de la solución actual x^t en el tiempo o iteración t introduciendo los parámetros p^t que requiera el algoritmo, así lo muestra la expresión (30):

$$\boldsymbol{x^{t+1} = A(x^t, p^t)} \tag{30}$$

En la metaheurísticas modernas se incluye la parte aleatoria con un conjunto de m variables aleatorias $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_m)$ que requiere el algoritmo, de modo que (30) se convierte en (31):

$$\boldsymbol{x^{t+1} = A(x^t, p^t, \epsilon^t)} \tag{31}$$

La cantidad t de iteraciones requeridas para encontrar la solución de calidad requerida determina la cantidad de cómputo que deberá utilizarse, el mejor algoritmo utiliza menos iteraciones y por consecuencia,

menos cómputo, el algoritmo ideal debería poder realizar el trabajo en una única iteración, aunque en general es inexistente.

Todos los algoritmos inspirados en la naturaleza incluyen los parámetros p mencionados previamente, la configuración de estos parámetros tiene peso en el comportamiento y desempeño del algoritmo, el mejor ajuste y control de estos parámetros es un problema desafiante. Los parámetros suelen variar al paso de las iteraciones, entonces es de suma importancia ajustar los parámetros de manera que la convergencia del algoritmo sea lo más rápida posible obteniendo los mejores resultados, los ajustes empíricos suelen dar buenos resultados [17].

Algunos de los métodos heurísticos contemplados en la obra de Yang [17] se enuncian a continuación.

2.6.1 Recocido Simulado

Diseñado originalmente para problemas de optimización combinatoria para ruteo y problema del agente viajero, en inglés *Travelling Salesman Problem* (TSP) en 1983 por Kirkpatrick, Gellat y Vecchi, el recocido simulado es una técnica de optimización global que utiliza búsqueda aleatoria en el espacio de soluciones. Trata de imitar el proceso de recocido de los materiales metálicos al enfriarlos hasta un estado de energía mínima para obtener cristales de mayor tamaño que reduzcan el defecto de sus estructuras. El recocido es un procedimiento que requiere sumo cuidado de la temperatura, tanto al calentar el metal como el ritmo de enfriamiento.

Una de sus ventajas es la capacidad de evadir mínimos locales y puede converger al óptimo local bajo condiciones apropiadas, es decir, si se utilizan buenos pasos aleatorios con un ritmo adecuado de enfriamiento. La forma de operación del recocido simulado consiste en, de manera iterativa aplicar una búsqueda aleatoria y aceptar tanto puntos que mejoren la solución como algunos otros que no lo hagan, cualquier movimiento que mejore la solución es aceptado de manera automática, sin embargo, el punto que no la mejore, puede ser aceptado con una probabilidad p, que se le conoce como *probabilidad de transición*, y está dada por la ecuación (32):

$$\boldsymbol{p = e^{-\frac{\Delta f}{T}}} \tag{32}$$

Esta probabilidad debe ser mayor que un determinado umbral de aceptación que en general es un valor aleatorio r, si esta condición se cumple, entonces la solución que no mejora el resultado es acepada. Las soluciones aceptadas en automático equivalen a la parte de explotación, pues se está trabajando sobre una pequeña región en busca el óptimo local. Una solución que no mejora, pero es aceptada aporta a la parte de exploración, pues al indagar en una región distinta, se diversifican las soluciones evadiendo los óptimos locales, note que la parte de exploración se vuelve cada vez menos relevante pues es poco probable que suceda mientras decrece la temperatura del sistema.

La selección de la temperatura inicial es muy importante, una temperatura inicial elevada, digamos $T \to \infty$, hace que $p \to 1$ por lo que, cualquier solución que no mejore el resultado sería aceptada, por el contrario, una temperatura T cercana al cero, $p \to 0$ por lo que ninguna solución que no mejore será aceptada generando una diversidad pobre de soluciones y, como consecuencia, un estancamiento prematuro en un

mínimo o máximo local. Cuidar la temperatura inicial no es el único elemento por el cual velar, el ritmo de enfriamiento también es importante.

El sistema se enfría de manera paulatina desde una temperatura elevada hasta un enfriamiento suficiente que consideraremos el estado mínimo global, el enfriamiento normalmente se programa de manera geométrica. Así, se establece una temperatura inicial T_0, además de un factor de conservación de temperatura α tal que $0 < \alpha < 1$, en cada iteración t, la temperatura está dada por la ecuación 31:

$$\boldsymbol{T_t = T_0 \alpha^t} \tag{32}$$

Esta técnica evita que se alcance $T = 0$ pues esto solo ocurre cuando $t \to \infty$ sin necesidad de establecer una cantidad de iteraciones. El factor de conservación debe ser elevado para que el enfriamiento sea lento, normalmente se usa $\alpha \in [0.7, 0.99]$. El lento descenso de la temperatura requiere de múltiples iteraciones pues una cantidad pequeña de iteraciones hará que el sistema seguramente no converja al óptimo global, sin embargo, una cantidad demasiado grande consumirá tiempo excesivo de cómputo para converger y esta situación se agrava para problemas en varias dimensiones.

La forma más común de comenzar el procedimiento es establecer una temperatura inicial muy elevada para aceptar la mayoría de los puntos visitados y reducir rápidamente la temperatura de manera que un 50% o 60% de las soluciones no mejores sean aceptadas y tomar esta temperatura como la inicial para disminuirla de manera lenta. En teoría, la temperatura final debe acercarse a cero para que no se acepten soluciones que no mejoren el resultado, por lo que se debe considerar un límite inferior cercano a cero para detener la ejecución del algoritmo.

Para determinados problemas tales como particiones de grafos, coloreo de grafos, TSP, o procesamiento de imágenes el recocido simulado se ve superado por otros algoritmos. El algoritmo general del recocido simulado se muestra en la Figura 13.

Algoritmo de Recocido Simulado

Función objetivo $f(x),\ x = (x_1, \dots, x_d)^T$
Establecer temperatura inicial T_0 y solución inicial x_0
Establecer temperatura final T_f y el número máximo de iteraciones N
Definir la programación de enfriamiento $T \mapsto \alpha T, (0 < \alpha < 1)$
mientras $(T > T_f \ \ \& \ \ t < N)$
    Elegir ϵ de una distribución Gaussian
    Moverse aleatoriamente a una nueva localización: $x_{t+1} = x_t + \epsilon$ (caminata aleatoria)
    Calcular $\Delta f = f_{t+1}(x_{t+1}) - f_t(x_t)$
    Aceptar la nueva solución si es mejor
    si no se mejoró
        Generar un número aleatorio
        Aceptar la nueva solución si $p = \exp[-\Delta f / T] > r$
    fin si
    Actualizar el mejor x_* y f_*
$t = t + 1$
fin mientras

Figura 13**.** Algoritmo de Recocido Simulado. Fuente: Yang [17], traducción propia.

2.6.2 Optimización por enjambre de partículas (PSO)

Desarrollado por Kennedy y Eberhart en 1995, la optimización por enjambre de partículas se basa en el comportamiento de los enjambres de la naturaleza, y ha sido aplicado casi en todas las áreas de optimización e inteligencia computacional. Los algoritmos de inteligencia de enjambres son ampliamente ocupados por su simplicidad y flexibilidad, utiliza movimientos aleatorios y comunicación con el resto del enjambre para continuar la búsqueda, la implementación es relativamente sencilla al no tener que establecer ningún tipo de codificación para los individuos.

Explora el espacio de búsqueda con agentes individuales llamados partículas, distribuidos a lo largo de dicho espacio, el movimiento de las partículas es individual, sin embargo, se ve influenciado por el resto del enjambre. Este movimiento consiste en una parte aleatoria y una parte determinista, cada partícula i se ve atraída hacia la mejor solución global g^* y su mejor localización individual x_i^* hasta ese momento. Cuando una partícula encuentra una mejor posición a cualquier otra encontrada por ella previamente, actualiza su mejor posición x_i^*, si esta mejor posición, también es mejor que la mejor solución global g^*, también sustituye a dicha solución y notifica al resto del enjambre. El esquema de movimiento se ilustra en la Figura 14:

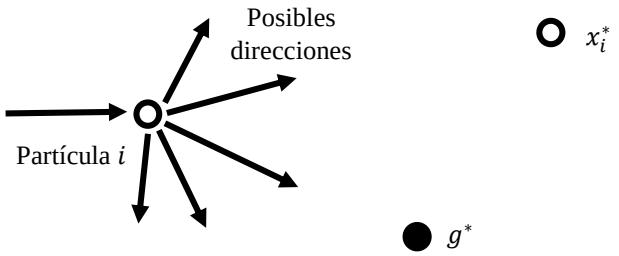

Figura 14. Movimiento de una partícula en PSO.

La actualización de la posición de la partícula en el tiempo $t+1$ se da por la ecuación (33):

$$\boldsymbol{x_i^{t+1} = x_i^t + v_i^{t+1}} \tag{33}$$

Donde la velocidad inicial es $v_i^0 = 0$, y la velocidad en el tiempo $t+1$ se obtiene a través de la ecuación (34):

$$\boldsymbol{v_i^{t+1} = \theta v_i^t + \alpha \epsilon_1 [g^* - x_i^t] + \beta \epsilon_2 [x_i^* - x_i^t]} \tag{34}$$

Aquí, θ es el coeficiente de inercia y normalmente $\theta \in [0.5,0.9]$, ese coeficiente tiene como objetivo estabilizar el proceso de movimiento y acelerar la convergencia del algoritmo, ϵ_1, ϵ_2 son números aleatorios uniformemente distribuidos con $\epsilon_1, \epsilon_2 \in [0,1]$ y α, β son parámetros de aprendizaje, normalmente $\alpha \in [0.1,0.4], \beta \in [0.1,0.7]$.

La razón de utilizar el mejor resultado de cada partícula individual es tratar de explotar los óptimos locales que se encuentren, además de explorar en busca de óptimos globales al tener influencia del mejor resultado global, aunque esta diversidad también se debe al uso de los parámetros aleatorios.

El uso de PSO se ha utilizado con amplio éxito en la mayoría de los problemas de optimización, sin embargo, la adaptación para problemas discretos es complicada debido al esquema de movimiento diseñado para exploración en superficies. El algoritmo general se describe en la Figura 15:

Algoritmo de Optimización por Enjambre de Partículas

Función objetivo $f(x),\ x = (x_1, \dots, x_d)^T$
Inicializar ubicaciones x_i y velocidad v_i para n partículas
Hallar g^* desde $\{f(x_1), \dots, f(x_n)\}$ para $t = 0$
mientras (criterio de paro)
  para todas las n partículas en todas las d dimensiones
    Generar nueva velocidad v_i^{t+1} usando la ecuación (34)
    Calcular nueva ubicación x_i^{t+1} usando la ecuación (33)
    Calcular el valor de la función objetivo en x_i^{t+1}
    Encontrar la mejor solución x_i^* de cada partícula
  fin para
  Hallar la mejor solución global g^*
  Actualizar $t = t + 1$ (pseudo tiempo o contador de iteraciones)
fin mientras
Mostrar resultados finales x_i^* y g^*

Figura 15. Algoritmo de Optimización por Enjambre de Partículas. Fuente: Yang [17], traducción propia.

2.6.3 Algoritmo de polinización de flores

Desarrollado originalmente en 2013 por Yang, el algoritmo de polinización de flores está enfocado a resolver problemas multiobjetivo, muchas veces los objetivos múltiples entran en conflicto entre sí, por lo que es imposible usar un diseño único sin comprometer la calidad de la solución. La optimización multiobjetivo supone desafíos adicionales en complejidad de tiempo de ejecución o dimensionalidad. Para optimización de un objetivo único, una solución óptima puede ser un único punto en el espacio de soluciones, pero para múltiples objetivos, las soluciones se convierten en superficies, peor aún, para dimensiones muy grandes, el problema puede desarrollarse en superficies complejas, en este documento se pretende mostrar el algoritmo básico sin extenderlo a problemas multiobjetivo.

Existen cerca de un cuarto de millón de tipos distintos de plantas en la naturaleza, de las cuales, el 80% florean, el propósito de una flor es reproducirse vía polinización, este es un proceso de transferencia de polen normalmente asociado a agentes animales. Los insectos y las flores han desarrollado una relación flor-polinizador, esta forma de reproducción se lleva a cabo de dos modos: abiótico y biótico. El 90% de las flores se reproducen de manera biótica, es decir, a través de agentes animales, mientras que el 10% restante no requiere polinizadores, llevan un proceso abiótico donde el viento y la dispersión son factores clave como el caso del césped.

Los polinizadores o vectores de polen pueden ser muy diversos, las abejas son ejemplos clásicos y tienen la particularidad de visitar especies muy específicas de flores, pues buscan maximizar la cantidad de polen a transferir hacia otras plantas de la misma especie, maximizando la reproducción además de recolectar néctar con un costo mínimo de energía y viaje pues tienen memoria limitada. La polinización también puede clasificarse como polinización hacia sí mismo, en la que el polen se transmite a flores de la misma planta, o polinización cruzada que depende de agentes bióticos para llevar polen a otras plantas, por lo que la polinización a sí mismo se puede considerar explotación de óptimos locales, mientras que la polinización cruzada puede comprenderse como la exploración hacia óptimos globales. Además, las abejas

muestran comportamiento de viajes de Lévy al momento de volar, por lo que la reproducción de las plantas se puede ver como un problema de optimización.

El algoritmo de polinización se rige bajo las siguientes reglas:

- La polinización cruzada se considera un proceso de exploración y los polinizadores obedecen los vuelos de Lévy.
- Para polinización local o explotación, utilizará polinización a sí mismo.
- Los polinizadores pueden generar favoritismo hacia ciertas especies, por lo que la reproducción es más probable si dos flores involucradas son similares.
- El cambio entre polinización local y global se lleva a cabo con una probabilidad $p \in [0,1]$ ligeramente preferente hacia la polinización local.

El paso de polinización global requiere que el polen sea llevado por insectos para viajar largas distancias, por lo que el movimiento del polen i está dado por la ecuación (35):

$$\boldsymbol{x_i^{t+1} = x_i^t + \gamma L(s,\lambda)(g_* - x_i^t)} \tag{35}$$

Donde x_i^t es la posición actual de la partícula de polen i en la iteración t, además, g_* es la mejor posición global encontrada esta ese momento. Se considera γ como un factor de escalado que controla el tamaño de paso y $L(s,\lambda)$ es el tamaño de paso basado en vuelo de Lévy expresado en la ecuación (36):

$$\boldsymbol{L(s,\lambda) = \frac{\lambda \Gamma(\lambda) \sin\left(\frac{\pi\lambda}{2}\right)}{\pi s^{1+\lambda}}} \tag{36}$$

En la práctica, s puede ser tan pequeño como 0.1.

Para la etapa de polinización local, el movimiento del polen se hace a través de la ecuación (37):

$$\boldsymbol{x_i^{t+1} = x_i^t + \epsilon(x_j^t - x_k^t)} \tag{37}$$

Donde x_j^t, x_k^t son polen de diferentes flores de la misma especie, representando la constancia de las flores en un vecindario limitado, esta parte utiliza caminata aleatoria tradicional al usar $\epsilon \in [0,1]$ como un número aleatorio distribuido uniformemente.

En principio, la polinización se da tanto a escala local como a escala global, sin embargo, en la práctica las flores tienen mayor tendencia a ser polinizadas de manera local, aquí resalta la importancia del parámetro p, estudios empíricos arrojan que un valor $p = 0.8$ debería funcionar para la mayoría de las aplicaciones. El algoritmo se muestra en la Figura 16:

Algoritmo de Polinización de Flores (Algoritmo Simple de Flores)

Función objetivo $f(x),\ x = (x_1, \ldots, x_d)^T$
Inicializar la población de n gametos de polen con soluciones aleatorias
Hallar la mejor solución g^* en la población inicial
Definir la probabilidad de transición $p \in [0,1]$
mientras (t <Máximo de iteraciones)
  para todas las n flores en la población
    si $rand < p$
      Generar un vector de movimiento L de una distribución de Lévy
      Polinizar de manera global usando la ecuación (35)
    si no
      Elegir un valor ϵ de una distribución uniforme entre 0 y 1
      Polinizar de manera local usando la ecuación (37)
    fin si
    Evaluar las nuevas soluciones
    Si son mejores, actualizarlas en la población
  fin para
  Encontrar la mejor solución g^*
fin mientras
Mostrar la mejor solución encontrada

Figura 16. Algoritmo de Polinización de Flores. Fuente: Yang [17], traducción propia.

2.6.4 Búsqueda armónica

Desarrollado por Geem en 2001, es un algoritmo inspirado en la música, una pieza musical busca un perfecto estado de armonía, que es equivalente a encontrar el óptimo en un proceso de optimización. Se puede comparar al proceso de improvisación de los músicos, y la armonía se determina por el estándar de estética del audio.

La calidad estética de un instrumento se determina por la tono, timbre y volumen. Diferentes notas tienen diferentes tonos, la medida de la armonía cuando diferentes tonos ocurren de manera simultánea como cualquier cualidad estética, es meramente subjetiva, pero aun así es posible hacer algunas estimaciones estándar de armonía.

Cuando un músico improvisa, tiene tres opciones:

1. Tocar cualquier pieza famosa, es decir una serie de tonos en armonía exactamente iguales desde su memoria.
2. Tocar algo similar a una pieza famosa conocida, ajustando un poco el tono.
3. Componer notas nuevas o de manera aleatoria.

En resumen, se puede usar la memoria armónica, ajustar tonos o recurrir a la aleatoriedad. La memoria armónica es similar a elegir los mejores individuos de los algoritmos genéticos, pues los mejores armónicos serán llevados a la nueva memoria armónica.

Se puede elegir un parámetro $r_{accept} \in [0,1]$ para determinar la aceptación de la memoria armónica, si es demasiado bajo, solo unas pocas de las mejores armonías serán elegidas y la convergencia será lenta, por el contrario, si es muy alto, casi todas las armonías de la memoria serán usadas por lo que se estará haciendo poca exploración y se estancará en un mínimo local, típicamente $r_{accept} \in [0.7,0.95]$.

Para ajustar el tono, se requiere de un método para ajustar la frecuencia de manera eficiente, el nuevo tono o solución se determina por la ecuación (38):

$$\boldsymbol{x_{new} = x_{old} + b_p(2rand - 1)} \tag{38}$$

Donde $rand$ es un número aleatorio entre 0 y 1 uniformemente distribuido y b_p es el ancho de banda que controla el rango local de ajuste de tono, este ajuste es en realidad una caminata aleatoria. Se puede definir una proporción de ajuste de tono r_{pa} para controlar el grado de ajuste, si es demasiado bajo, rara vez se notarán cambios, pero si es muy alto, el algoritmo a lo mejor no convergerá, normalmente se utiliza $r_{pa} \in [0.1,0.5]$ en la mayoría de los casos.

Finalmente, la opción de aleatoriedad corresponde a una búsqueda para diversificar las soluciones, esto permite explorar varias regiones para incrementar la probabilidad de hallar el óptimo global. El algoritmo se especifica en la Figura 17.

Algoritmo de Búsqueda Armónica

Función objetivo $f(x),\ x = (x_1, \dots, x_d)^T$
Generar armonías iniciales (arreglos de números reales)
Definir el rango del ajuste de tono r_{pa} y los límites de ajuste
Definir el rango de aceptación r_{accept} de la memoria armónica
mientras (t <Máximo de iteraciones)
    Generar nuevas armonías aceptando las mejores armonías
    Ajustar el tono para obtener nuevas armonías (soluciones)
    si ($rand > r_{accept}$)
        Elegir una armonía existente de maneara aleatoria
    si no, si ($rand > r_{pa}$)
        Ajustar el tono aleatoriamente usando la ecuación (38)
    si no
        Generar una nueva armonía de manera aleatoria
    fin si
    Aceptar las nuevas armonías (soluciones) si son mejores
fin mientras
Hallar las mejores estimaciones

Figura 17. Algoritmo de Búsqueda Armónica. Fuente: Yang [17], traducción propia.

Capítulo 3

Estado del Arte

En esta sección se describen algunos trabajos involucrados en la detección y clasificación de pacientes con diabetes a través de sistemas basados en lógica difusa y optimizados por métodos heurísticos, así como su eficiencia obtenida.

En 2015, Jain y Raheja [8], toman la base de datos de Pima Indians Diabetes Database [18], un experto selecciona los seis datos más relevantes y diseña funciones de membresía triangulares para los datos de entrada, y la predicción de riesgo de padecimiento como salida del sistema, además de seis reglas que dictaron el comportamiento del sistema difuso construido con el método de inferencia de Mamdani. El sistema no es optimizado por alguna heurística, en su lugar se opta por normalizar los datos de entrada para establecerlos entre 0 y 1, finalmente, se prueba la efectividad del sistema calculando el porcentaje de precisión respecto a la hecha en el corpus, el sistema es generalmente mejor con un 87.2% de precisión que otros sistemas hechos por computadora que alcanzan 85.03% de aciertos, incluso superando sistemas médicos de especialistas que oscilan entre los 77.6% y 81.7%.

En 2017, Reddy y Khare [19], publican un estudio que establece una predicción para problemas cardiacos y riesgo de padecimiento de diabetes. Usando teoría de conjuntos aproximados y binarización de los datos provenientes de la base de datos de la UCI, de un total de 13 factores de riesgo como edad, sexo, IMC, entre otros, se seleccionan 6 para después normalizarlos y optimizarlos usando la heurística de Búsqueda Cuckoo, posteriormente un experto diseña funciones de membresía triangulares que no son optimizadas. Para evaluar el desempeño, se revisa tanto la precisión respecto al corpus, como la sensibilidad y especificidad, la puntuación fue de 91%, 94% y 90% respectivamente. Al establecer una comparación contra un sistema basado en búsqueda por luciérnagas y murciélagos, se nota que los 3 criterios son mejores, superando el 68%, 79% y 84% de dicho estudio, y el 72.6%, 100% y 0% de un sistema ya existente en la industria, aunque con otros conjuntos de datos se puede percibir un rendimiento inferior esperado, enfatizando la reducción de los datos en más del 50%.

También en 2017, Sahu, Verma y Reddy [20], llevan a cabo un estudio donde se obtuvieron una serie de datos no especificados de un laboratorio patológico local, dichos datos se redujeron utilizando Reducción por Preservación Local de Proyecciones y las reglas del sistema se calcularon utilizando Búsqueda por Colonia de Hormigas, estas reglas fueron optimizadas usando la técnica de Búsqueda Cuckoo, las funciones de membresía se diseñaron con forma triangular por un experto y no fueron optimizadas. Se utiliza la precisión de los resultados respecto al corpus consultado, así como la sensibilidad y especificidad para medir la eficacia del sistema en 10 ejecuciones, luego se compara contra otra implementación del mismo sistema optimizado mediante un híbrido de Búsqueda por Luciérnagas y Murciélagos. Se concluye que el uso de Búsqueda Cuckoo muestra un mejor rendimiento que la variante Luciernaga-Murciélago, sin embargo, los resultados en cada ejecución son variables en el porcentaje de precisión del sistema oscilando entre 64% y 74%.

En el 2020, Bressan, Flamia de Azevedo y Souza [21], desarrollan un clasificador difuso para catalogar a un grupo de personas como pacientes que requieren atenciones básicas o que requieren atenciones

especializadas. Partiendo de los datos clínicos de edad, triglicéridos, tiempo de evolución de la enfermedad, IMC, ingresos per cápita, circunferencia abdominal y tiempo de escolaridad proporcionados por el Unified Health System de Brasil, un experto descartó los no esenciales y, usando un árbol de decisión con el algoritmo C4.5 se generaron todas las reglas requeridas por el sistema. Se hizo una implementación para cada grupo, las funciones de membresía se eligieron con forma triangular y se ajustaron haciendo uso de un Sistema de Inferencia Neurodifuso Adaptativo, evadiendo posterior optimización. Finalmente se simularon 10000 pacientes para ser clasificados por el sistema, los resultados se compararon contra un sistema experto. La clasificación fue consistente con la del sistema experto por lo que el sistema ahorra tiempo y dinero al sistema de salud al proporcionar tratamiento de especialidad solo en casos necesarios.

Finalmente, en 2020, Risqy et al. [22], implementan un sistema basado en el método de inferencia de Tsukamoto para calcular el riesgo de padecimiento de diabetes con muy poca información de los pacientes, el corpus utilizado consta solamente de 2 pacientes voluntarios, los datos proporcionados son edad, IMC y presión sanguínea. Un experto diseña funciones de membresía triangulares que posteriormente son optimizadas por el algoritmo de Enjambre de Partículas, además, proporcionó 6 reglas para el control del sistema. La eficiencia se obtiene calculando el error cuadrático medio respecto al porcentaje de riesgo predicho por un estudio especializado, un error de 0.012% se obtiene después de la optimización y se compara con el 25.9% de error obtenido por el sistema sin optimización. Los autores no descartan la posibilidad de utilizar otros corpus disponibles en línea para continuar con el desarrollo y refinamiento del sistema.

Un resumen de los trabajos presentados se muestra a continuación en las Tablas 4 y 5:

Tabla 4. Resumen de estado del arte.

	Autores	**Año**	**Tipo de reducción de datos**	**Método de obtención de reglas de inferencia**
Fuzzy Tsukamoto Membership Function Optimization Using PSO to Predict Diabetes Mellitus Risk Level	Risy S, Cantika N, & Fitra A	2020	Ninguna	Diseñadas por un experto.
A Fuzzy Approach for Diabetes Mellitus Type 2 Classification	Bressan, Flamia de Azevedo, & Molina de Souza	2020	Datos seleccionados por un experto	Obtenidas a través de árbol de decisión con algoritmo C4.5
Diabetes Classification using Fuzzy Logic and Adaptive Cuckoo Search Optimization Techniques	Nikkita Sahu, Toran Verma & Thippa Reddy	2017	Reducción por Preservación Local de Proyecciones	Obtención por Colonia de Hormigas
Improving the Prediction Rate of Diabetes using Fuzzy Expert System	Vaishali Jain, & Supriya Raheja	2015	Datos seleccionados por un experto	Diseñadas por un experto
Cuckoo Search Optimized Reduction and Fuzzy Logic Classifier for Heart Disease and Diabetes Prediction	Thippa Reddy & Neelu Khare	2017	Reducción por Conjuntos aproximados	Diseñadas por un experto

Tabla 5. Resumen de estado del arte.

	Método de obtención de funciones de membresía	**Criterios para valuación de efectividad**	**Atributos y Corpus utilizado**	**Tipo de optimización**
Fuzzy Tsukamoto Membership Function Optimization Using PSO to Predict Diabetes Mellitus Risk Level	Triangulares diseñadas por experto.	Error cuadrático medio.	***Edad, IMC, Presión sanguínea*** Datos de dos pacientes reales.	**Funciones de membresía:** Optimización por Enjambre de Partículas.
A Fuzzy Approach for Diabetes Mellitus Type 2 Classification	Triangulares obtenidas por Sistema de Inferencia Neuro Difuso Adaptativo.	Comparación respecto a la predicción indicada en el corpus.	***Edad, Trigliceridos, Tiempo de evolución de la enfermedad, IMC, Ingresos per cápita, Circunferencia abdominal, Tiempo de escolaridad*** Corpus del Unified Health System de Brasil.	Ninguna.
Diabetes Classification using Fuzzy Logic and Adaptive Cuckoo Search Optimization Techniques	Triangulares diseñadas por un experto.	Precisión en 10 corridas. Sensibilidad y Especificidad respecto a un sistema similar basado en optimización por luciérnagas-murciélago.	Características no especificadas provenientes de un laboratorio patológico local.	**Reglas de inferencia:** Optimización por Búsqueda Cuckoo.
Improving the Prediction Rate of Diabetes using Fuzzy Expert System	Triangulares diseñadas por un experto..	Precisión respecto a la predicción indicada en el corpus.	***Glucosa, Presión arterial, IMC, Riesgo de padecimiento por herencia, Edad, Cantidad de orina expulsada*** Corpus del Pima Indians Diabetes Database	**Datos de entrada:** Normalización binaria 0,1.
Fuzzy Membership Function Generation using DMS-PSO for the Diagnosis of Heart Disease	Triangulares calculadas con base en literatura consultada.	Precisión respecto a la predicción indicada en el corpus, una ejecución con todos los datos sin optimización, contra datos seleccionados con optimización y contra datos seleccionados sin optimización.	***Edad, Género, Tipo de dolor en el pecho, Presión sanguínea, Índice de colesterol*** Corpus del repositorio de UCI.	**Funciones de membresía:** Optimización del vértice superior de los triángulos por Enjambre de Partículas Multi enjambre Dinámico **Reglas de inferencia**: Ponderación y selección de las más importantes.

Capítulo 4

Conjuntos de Datos

Es complicado encontrar bases de datos con información clínica de acceso libre, una de las más utilizadas para estudios computacionales sobre diabetes es la Pima Indians Diabetes Database, que se encuentra accesible en línea [23].

4.1 Pima Indians Diabetes Database

Esta base de datos es una aportación de Smith, J.W. et al., del año de 1998, a raíz de un trabajo llamado *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus* [24]. El estudio surge como una manera de lidiar con los problemas que acarrean las técnicas de análisis estadístico como las regresiones al momento de clasificar o tomar decisiones, tales inconvenientes se relacionan con la existencia de relaciones entre los datos que pueden resultar no evidentes y, al no ser consideradas, pueden no generar resultados satisfactorios cuando se tiene muestras de tamaño pequeño. Los autores utilizan un algoritmo de red neuronal de aprendizaje adaptativo, similar al ANFIS.

El sector de población que cubre el estudio consta de indios Pima establecidos cerca de Phoenix, Arizona, la elección de este grupo se debe al gran interés que el National Institute of Diabetes and Digestive and Kidney Diseases ha mostrado desde 1965 por su alta incidencia en casos de diabetes de manera histórica. Se tomaron los datos de 768 mujeres Pima, que incluyen cantidad de embarazos, glucosa en ayunas, presión sanguínea diastólica, grosor de la piel en los bíceps, insulina en ayunas, índice de masa corporal, posibilidad de padecimiento de diabetes según historial familiar y edad, para un total de 8 variables. Al paso de 5 años, se evalúa a las personas para saber si han desarrollado diabetes tipo 2 y establecer la salida mostrada en la novena columna, los datos de 576 de estas mujeres se utilizan para el entrenamiento, y los 192 casos restantes se utilizan para evaluar el funcionamiento del sistema, resultando con un 76% de aciertos.

Un dato que resalta dentro de la base de datos es la variable de posibilidad de padecimiento por historial familiar (DPF), pues no existe una formulación o modelo oficial que pueda calcular este valor. Es sabido que la herencia juega un papel importante en el desarrollo de diabetes tipo 2, por lo que los autores de este trabajo proponen la siguiente manera para obtener dicho dato. El objetivo es computar una medida de la influencia genética de parientes afectados y no afectados del sujeto para su predicción de riesgo usando la ecuación (39):

$$\boldsymbol{DPF} = \frac{\sum_i \boldsymbol{K_i}(\boldsymbol{88} - \boldsymbol{ADM_i}) + \boldsymbol{20}}{\sum_j \boldsymbol{K_j}\left(\boldsymbol{ALC_j} - \boldsymbol{14}\right) + \boldsymbol{50}} \tag{39}$$

Donde i es el i-ésimo pariente que ha desarrollado diabetes el día de examinación del sujeto, j es el j-ésimo pariente que no ha desarrollado diabetes el día de examinación del sujeto y K_x es el porcentaje de genes compartidos con el x-ésimo pariente y puede adoptar valores de:

- 0.5 si el pariente es padre o hermano directo
- 0.25 si el pariente es medio hermano, abuelo, tío o tía.
- 0.125 si es medio tío, media tía o primo hermano.

Además, ADM_i es la edad en años del i-ésimo pariente al ser diagnosticado, ACL_j es la edad en años del j-ésimo pariente en su último diagnóstico negativo de diabetes previo al examen del sujeto.

Las constantes 88 y 14 representan, salvo raras excepciones la edad máxima y mínima en la que los parientes de los sujetos de estudio desarrollaron diabetes.

Las constantes 20 y 50 se eligen de manera que

1. Un sujeto sin parientes por evaluar obtenga un DPF ligeramente menos que el promedio
2. El DPF decrezca tan relativamente lento como haya parientes sin diabetes a la base de datos.
3. El DPF incremente relativamente rápido si hay parientes con diabetes desarrollada.

Note que el valor de DPF incrementa a mayor número de parientes diagnosticados con diabetes, menor edad de desarrollo de la enfermedad de los parientes y mayor porcentaje de genes compartidos. Análogamente, el DPF se reduce cuando hay más parientes libres de la enfermedad, sus edades de último examen negativo crecen y la cantidad de genes compartidos aumenta. Este modelo ha sido propuesto para considerar al historial familiar y asignarle un valor numérico, sin embargo, depende de la cantidad de parientes de cada individuo que hayan acudido al estudio, así pues, una persona que no haya presentado familiares tendrá una posibilidad de riesgo de padecimiento intermedio, el cual no necesariamente refleja su realidad.

Se procedió a buscar todos los trabajos relacionados con diabetes de todos los autores desde el 1998 hasta el 2021 con la intención de indagar sobre una actualización del estudio, validación del modelo mostrado o una reconstrucción, desafortunadamente no se han encontrado más información al relacionada por lo que el modelo sigue en estado no validado. Aunado a este inconveniente, la base de datos representa a un sector de población ajeno a nuestro país, con condiciones especiales que los hacen tener una mayor incidencia de padecimiento de esta enfermad, además, la columna de resultado de la base de datos es binaria, es decir, indica si al paso de 5 años el individuo desarrollo o no diabetes tipo 2, no indica un valor de riesgo de padecimiento, además la prueba de FINDRISC solicita datos que no incluye esta base de datos.

En principio se consideró su utilización debido a su fácil acceso en línea pues el autor principal ha decidido compartirla para su uso libre, además de ser ampliamente citada y utilizada por una gran cantidad de estudios relacionados a diabetes, sin embargo, se ha determinado que por las razones previamente mencionadas se busque una base de datos que represente mejor a la comunidad mexicana y que tenga los datos necesarios para usar la prueba de FINDRISC como estándar de oro. Eso no dejó fuera del estudio a la base de datos mencionada, fue utilizada para las pruebas requeridas para la comparación de rendimiento entre prueba de FINDRISC, ANFIS y sistema difuso diseñado manualmente. Para las pruebas del sistema definitivo se encontró el trabajo “Evaluación del desempeño del Finnish Diabetes Risk Score (findrisc) como prueba de tamizaje para diabetes mellitus tipo 2” por Mendiola, I. et. al.[25].

4.2 Base de datos de Mendiola et. al.

El trabajo dirigido por la Dra. Mendiola [25] en el 2017 tuvo por objetivo evaluar el desempeño de la prueba de FINDRISC, con pacientes libres de diabetes tipo 2 adscritos a la Unidad de Medicina Familiar No. 9 del IMSS de Acapulco Guerrero, se seleccionan de manera aleatoria 295 participantes para aplicar el cuestionario FINDRISC, así como tomas orales de glucosa en ayuno para determinar si padecen o no diabetes, donde se indica la cantidad de pacientes puntuados con alto y bajo riesgo según la prueba de FINDRISC, y el diagnostico posterior a la consulta después de prueba de glucosa de laboratorio, de acuerdo a la American Diabetes Association, clasifica a los pacientes como glucosa normal $< (100mg/dl)$, prediabetes $(100\ a\ 125mg/dl)$ y diabetes tipo 2 $(> 125mg/dl)$, además de clasificar a los pacientes con najo riesgo a los individuos con 14 puntos o menos en la prueba de FINDRISC, y de alto riesgo a aquellos con 15 o más puntos. Los resultados mostraron que la prueba de FINDRISC tiene un desempeño adecuado para el tamizaje de diabetes tipo 2 en población mexicana, con sensibilidad mayor a 80%, pacientes con diabetes tipo 2 no diagnosticada, a los que se aplique el cuestionario tendrán 15 puntos o más. Mientras que detecta un gran número de falsos positivos, tiene la capacidad de discriminar a aquellos que no están enfermos. La prueba de FINDRISC resultó ser adecuada y útil como prueba de tamizaje además de ser de fácil uso y no invasiva. Se decidió contactar a la Dra. Mendiola y muy amablemente accedió a proporcionar la base de datos utilizada para su estudio. Los resultados tabulados relevantes para nuestro propósito se muestran en la sección 6 al comparar su rendimiento con el de sistema ANFIS y difuso diseñado manualmente.

Capítulo 5

Implementación

Este capítulo menciona las implementaciones del sistema ANFIS y el sistema difuso diseñado manualmente, así como el análisis que justifica la selección de las metaheurísticas aplicadas.

5.1 Construcción del ANFIS

Se sometió a consideración el uso de un sistema ANFIS para reducir considerablemente el tiempo de diseño y optimización del sistema difuso propuesto, para el entrenamiento de la red neuronal y las pruebas realizadas se hace uso de la base de datos mexicana [25].

Las ANFIS se generaron y entrenaeron con la herramienta *anfisedit* disponible en MATLAB, al usar substractive clustering, el software en automático elige funciones de membresía con forma de campana gaussiana, solamente se requiere ajustar los parámetros solicitados por la herramienta. Se decidió utilizar el 80% del conjunto de datos para el entrenamiento y el 20% restante para pruebas. Mayores detalles se mencionan en la sección 6.1.

5.2 Diseño del sistema difuso

El FIS que se diseña hace uso del método de inferencia de Mamdani, se elige puesto que no se conocen las salidas deseadas en el sistema, si bien la prueba de FINDRISC nos da un porcentaje de riesgo de padecimiento de diabetes, el objetivo del hacer un sistema propio es indicar riesgos personalizados, por lo que la prueba de FINDRISC funge como indicador para verificar la exactitud de cada prueba pero no puede ser utilizado como la salida deseada, además, diseñar un sistema basado en otro método, por ejemplo el de Takagi-Sugeno-Kang significaría generar reglas de inferencia de manera automática dejando de lado el conocimiento que puede aportar la base de datos de Mendiola [25], además, se utiliza la técnica de inferencia máximo mínimo. Uno de los objetivos de este trabajo es establecer un sistema capaz de indicar el riesgo de padecimiento de un paciente de manera personalizada, este resultado se diferenciará del obtenido por la prueba de FINDRISC al indicar un resultado entre 1% y 50% en lugar de indicar uno de los 5 posibles resultados de dicha prueba, además de solicitar al usuario una menor cantidad de datos de entrada.

La prueba de FINDRISC solicita edad, IMC, circunferencia abdominal, realización de actividad física, consumo de frutas y verduras, si se ha tenido tratamiento previo para tensión arterial alta, si se ha tenido un estudio con índices altos de glucosa, también se pregunta si algún familiar cercano o no cercano ha sido diagnosticado con diabetes. Sin embargo, Lindström & Toumilehto resaltan que las preguntas sobre actividad física y consumo de frutas y vegetales no portan mucho a su modelo predictivo, pero se tomaron en cuenta para enfatizar la importancia del ejercicio y la dieta sana [13], por lo que para nuestro propósito se eliminan de la información requerida por el sistema difuso, además, Mendiola et. al. Mencionan que la edad, IMC catalogado como sobre peso u obesidad, y, por ende, circunferencias abdominales elevadas

aumentan el riesgo de padecimiento de diabetes, también es bien sabido que la diabetes puede ser heredada y que estudios previos de glucosa alta, sumados con la hipertensión son factores de peso en el desarrollo de la diabetes, por lo que el resto de los datos solicitados por la prueba de FINDRISC se mantienen, salvo los excluidos que se mencionaron con anterioridad. Las reglas para el sistema difuso se obtienen utilizando la base de datos de Mediola et. al. [25] y el razonamiento explicado a continuación:

La prueba de FINDRISC puntúa la respuesta a cada pregunta en una escala de 0 a 5 puntos, por lo que se usó cada una de estas opciones como una variable lingüística:

- 0 puntos – Factor de riesgo muy bajo.
- 1 punto – Factor de riesgo bajo.
- 2 puntos – Factor de riesgo ligeramente elevado.
- 3 puntos – Factor de riesgo moderado.
- 4 puntos – Factor de riesgo alto.
- 5 puntos – Factor de riesgo muy alto.

Además, FINDRISC establece 5 categorías de riesgo mencionadas en la Tabla 2, de modo que, para cada uno de los 295 casos dentro de la base de datos, se obtuvo la puntuación requerida por cada pregunta y su resultado de riesgo, generando una regla cuyo antecedente se compone de los factores en la base de datos y el consecuente es el riesgo calculado por FINDRSIC, como ejemplo considere el siguiente caso femenino:

- Edad: 55 años – 3 puntos.
- Circunferencia abdominal: 87 cm – 3 puntos.
- Índice de masa corporal: 26.75 – 1 punto.
- Tratamiento previo para presión arterial: 2 puntos.
- Estudio previo de glucosa alta: 0 puntos.
- Historial familiar de diabetes: 0 puntos.
- Riesgo de FINDRISC: 4% - Riesgo ligeramente elevado

Cuya regla difusa está dada por:

SI **edad** es de riesgo moderado y **circunferencia abdominal** es de riesgo moderado y **índice de masa corporal** es de riesgo bajo y **tratamiento para presión arterial** es de riesgo ligeramente elevado y **estudio previo de glucosa** es de riesgo muy bajo y **historial familiar de diabetes** es de riesgo muy bajo, $ENTONCES$ **riesgo de padecimiento** es ligeramente elevado.

Si un antecedente ya existía en el sistema, entonces era omitido, tanto el sistema femenino como el masculino están gobernados por 102 reglas difusas obtenida en total. El sistema difuso diseñado de manera manual utiliza el método de inferencia de Mamdani con método AND y de implicación el mínimo, método OR y de agregación el máximo y como método de defuzificación, centroide de gravedad. Las funciones de membresía para el género femenino y masculino son las mismas, salvo para la variable de medida de circunferencia abdominal.

5.3 Selección de metaheurísticas

En la sección 2.6 se abordaron las técnicas metaheurísticas de Recocido Simulado, PSO, Polinización de Flores y Búsqueda Armónica, sin embargo, se revisaron los métodos de Algoritmos Genéticos, Evolución Diferencial, Algoritmo de Luciérnagas, Búsqueda Cuckoo, Algoritmo de Murciélagos, Colonia de Hormigas y Colonia de Abejas para ser consideradas. Las Tablas 6 y 7 muestran el cuadro comparativo de todas las metaheurísticas estudiadas, las seleccionadas y presentadas en la sección 2.6 se encuentran resaltadas en *letras negritas*.

Tabla 6. Cuadro comparativo de técnicas metaheurísticas.

	Tipo de población	**Forma de exploración**	**Forma de explotación**	**Tipos de problemas**	**Estrategia**	**Cambios en tiempo de ejecución**	**Tipo de caminata**
Recocido simulado.	**Individuo único.**	**Aceptar algunas soluciones peores.**	**Aceptar cualquier solución mejor.**	**Continuos.**	**Exploración a altas temperaturas, explotación a bajas temperaturas.**	**Probabilidad de transición.**	**Aleatoria.**
Algoritmos genéticos	Población de genes.	Operador de mutación.	Operador de cruza.	Combinatorios y continuos.	Explotación con alta ocurrencia, exploración con baja ocurrencia.	Selección por elitismo.	Aleatoria.
Evolución diferencial.	Población de genes.	Operador de mutación.	Operador de cruza.	Continuos.	Mutar a partir de 3 vectores y cruzar con dicho vector con cierta probabilidad combinando exploración y explotación.	Selección por elitismo.	Dirigida por 3 vectores.
PSO.	**Enjambre de individuos.**	**Influencia del mejor resultado global durante el movimiento.**	**Influencia del mejor resultado de la partícula durante el movimiento.**	**Continuos.**	**Influenciar el movimiento de la partícula con óptimos local y global combinando exploración y explotación.**	**Notificar al resto del enjambre una mejor solución global.**	**Aleatoria dirigida por los óptimos local y global.**
Algoritmos de luciérnagas.	Enjambre de individuos.	División de enjambre en grupos a lo largo del espacio de soluciones	Atracción hacia la luciérnaga con mejor.	Continuos.	Dispersar luciérnagas explorando el espacio de soluciones y las atractivas formarán grupos de explotación local.	Reducir el elemento aleatorio del movimiento al acercarse a un óptimo.	Aleatoria dirigida por el óptimo local.

Tabla 7. Cuadro comparativo de técnicas metaheurísticas.

	Tipo de población	**Forma de exploración**	**Forma de explotación**	**Tipos de problemas**	**Estrategia**	**Cambios en tiempo de ejecución**	**Tipo de caminata**
Búsqueda Cuckoo.	Población de nidos.	Búsqueda de nuevas soluciones aleatorias.	A agregar al intruso si es mejor.	Continuos.	Explorar nuevas soluciones abandonado una porción de los nidos y conservar las mejores soluciones.	Selección por elitismo.	Aleatoria o vuelo de Lévy.
Algoritmos de murciélago.	Población de murciélagos.	Influencia del mejor resultado global.	Movimiento aleatorio partiendo de la mejor solución local.	Continuos.	Explotar nuevas soluciones locales y globales, aceptar si son mejores que las seleccionadas previamente.	Explotar soluciones locales para explorar globales	Aleatoria.
Algoritmo de polinización de flores.	**Población de partículas de polen.**	**Influencia del polen de dos flores de la misma planta en el movimiento al explotar los óptimos locales.**	**Influencia del óptimo global en el movimiento del polen al explorar por óptimos globales.**	**Continuos.**	**Elegir el tipo de polinización con base en probabilidad preferencia hacia polinización local.**	**Notificar al resto de la población de una mejor solución global.**	**Aleatoria o vuelo de Lévy.**
Algoritmo de colonia de hormigas	Colonia de hormigas.	Influencia de la feromona en la elección de caminos.	Aleatoriedad en la selección cuando la feromona es tenue.	Discretos.	Elegir caminos de manera probabilística la selección se ve guía por feromona, esperar autoorganización.	Disipación de la feromona en el ambiente.	Aleatoria influenciada por la feromona.
Algoritmo de colonia de abejas	Colonia de abejas.	Sustitución de peores soluciones después de un tiempo determinado.	Caminata aleatoria local desde las mejores soluciones.	Continuos.	Elegir aleatoriamente la fuente de alimento a explorar, iniciar explotación y exploración.	Selección por elitismo.	Aleatoria.
Búsqueda armónica	**Individuo único**	**Composición de notas de manera aleatoria.**	**Ajuste de tono variando ligeramente la memoria armónica.**	**Continuos.**	**Generar nuevas composiciones con las mejores armonías, ajustar el tono de una solución o generar una nueva.**	**Selección por elitismo.**	**Aleatoria.**

Puesto que se tiene una idea concreta de la configuración de los parámetros de las funciones de membresía, y, además, solo se requieren ajustes pequeños, se opta por elegir las heurísticas que tengan cierta especialización y preferencia por la etapa de explotación.

El recocido simulado es elegido pues a temperaturas bajas, el enfoque es casi exclusivo de explotación, además de presentar un único agente para desarrollar la tarea y ser de fácil implementación. Los algoritmos genéticos, si bien tienen un alto grado de preferencia hacia la etapa de explotación, esta se ve limitada por la cantidad y calidad de individuos de la primera generación, se requeriría de un cuidadoso diseño de la etapa de exploración aún para encontrar mejores candidatos locales, además de requerir una cantidad considerable de agentes independientes para poder tener variedad de soluciones locales, estas condiciones son similares a las presentadas por la evolución diferencial, por lo que ninguna de estas opciones es elegida.

PSO es una opción prometedora, pues se puede considerar al primer óptimo global a la configuración actual, además de establecer una cantidad disminuida de partículas en la región señalada, mejorando la convergencia y aprovechando la explotación del área, por lo que se considera para las pruebas de optimización.

El algoritmo de luciérnagas tiene opciones que permiten centralizar la búsqueda en manera local, no obstante, la división del enjambre puede no ser la mejor opción pues el objetivo es encontrar una mejor configuración muy cercana a la que ya se tiene al inicio, por lo que no es elegida.

Búsqueda cuckoo, se encuentra enfocada a la exploración al ingresar un factor de abandono de nidos, si bien, este puede variarse para preferir explotación, la generación de intrusos es un procedo complejo y demasiado elaborado para generar una solución que requiera poco ajuste a la ya obtenida, así, no fue utilizada en la etapa de optimización.

Los algoritmos de murciélago permiten movimiento enfocado a la mejor solución local, sin embargo, al hallar una, el algoritmo optará por diversificar mientras se encuentra una nueva presa, por lo que este algoritmo tomara tiempo en una etapa que no es de nuestro interés, entonces, no se considera para la optimización.

El algoritmo de polinización puede ser manipulado para centrarse de manera exclusiva en polinización local, y no se requiere de profundizar en vuelos de Lévy por la nula utilización que se le dará a la exploración, por esta razón es utilizada en la optimización.

Las colonias de hormigas quedan completamente descartadas al ser enfocadas en problemas discretos, además, la adaptación al espacio continuo, así como las reglas de actualización de feromona es demasiado compleja para este tipo de problemas.

La opción de colonia de abejas requiere de una etapa de exploración para hacer variaciones a las soluciones, similar a lo ocurrido con los algoritmos genéticos, por lo que se requiere de utilizar tiempo en una parte que no es de interés, así, queda descartara del proceso de optimización.

Finalmente, la búsqueda armónica, utiliza un único agente, además de poder manipular los umbrales de probabilidad para enfocar al algoritmo a ajustar tonos de la memoria armónica actual, el objetivo de hacer ajustes mínimos a la configuración actual se ve satisfecho, por lo que se utilizó en la etapa de optimización.

Capítulo 6

Experimentos y Resultados

A continuación, se presentan los resultados de los experimentos realizados con el sistema ANFIS, el sistema difuso de diseño propio y la comparación de los resultados de estos con la prueba de FINDRISC. Así mismo, se muestran los resultados de la optimización de los métodos metaheurísticos y el rendimiento del sistema seleccionado para el producto final.

6.1 Rendimiento del ANFIS

Puesto que las ANFIS hacen uso del aprendizaje, la forma de presentar los datos de entrada puede influir en el aprendizaje, por lo que se experimentó con los siguientes casos para decidir el tipo de estandarización:

- Datos femeninos sin estandarización utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.
- Datos femeninos sin estandarización utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.
- Datos femeninos estandarizados con puntuación Z utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.
- Datos femeninos estandarizados con puntuación Z utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.
- Datos masculinos sin estandarización utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.
- Datos masculinos sin estandarización utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.
- Datos masculinos estandarizados con puntuación Z utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.
- Datos masculinos estandarizados con puntuación Z utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.

El motivo de construir ANFIS por cada género es que, la medida de circunferencia abdominal femenina y masculina no consideran los mismos intervalos para determinar las variables lingüísticas. Por cada caso, se entrenaron 10 ANFIS, es decir, se realizaron 10 pliegues ordenando los datos de manera aleatoria y utilizando el 80% de los datos para entrenamiento con el 20% restante para pruebas. La base de datos contiene 205 datos femeninos y 90 masculino, por lo que, para el entrenamiento se utilizan 164 y 72 datos respectivamente, mientras que, para las pruebas, se utilizan conjuntos de 41 y 18 datos. Las funciones de membresía se generan con el método de substractive clustering, los parámetros de rango de aceptación y de rechazo se mantienen similares a los recomendados por la literatura buscando generar exactamente 4 funciones de membresía para cada variable de entrada.

Cada caso mencionado previamente generó una tabla que muestra para cada una de las 10 ANFIS asociadas, el error de entrenamiento alcanzado, el porcentaje de datos de entrenamiento que alcanzaron la exactitud requerida, este criterio, para ser aceptado como un acierto, el paciente evaluado debe obtener una diferencia no mayor a 15% de riesgo respecto al indicado por la prueba de FINDRISC, el porcentaje de aciertos del conjunto de prueba, los parámetros de radio y factor de depreciación requeridos para que cada variable tenga 4 funciones de membresía y la cantidad de épocas necesarias para alcanzar una diferencia menor a 10^{-5} de los errores de entrenamiento entre cada época. Las Tablas 8 a 15 muestran los resultados obtenidos.

Tabla 8. Datos femeninos sin estandarización utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.

Femenino con puntuación FINDRSIC sin estandarización						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	2.34667	93.29%	78.05%	0.95	2.75	75
2	2.94254	92.68%	85.37%	0.9	2.5	24
3	3.27462	92.07%	70.73%	0.9	2.75	18
4	3.05869	85.90%	78.05%	0.9	2.75	240
5	3.39924	90.24%	87.80%	0.9	2.6	29
6	2.91833	86.59%	75.61%	0.9	2.5	60
7	3.18571	92.07%	85.37%	0.9	2.75	16
8	2.9753	88.41%	85.37%	0.95	2.5	70
9	3.33144	85.37%	73.17%	0.9	2.6	14
10	3.22446	87.80%	73.17%	0.85	2.5	31
PROMEDIO	3.0657	89.44%	79.27%			

Tabla 9. Datos femeninos sin estandarización utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.

Femenino con puntuación binaria o triaría sin estandarización						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	2.49579	93.90%	75.61%	0.95	2.75	23
2	3.84809	90.24%	90.24%	0.95	2.5	15
3	3.15643	87.80%	78.05%	0.9	2.5	187
4	2.80733	90.85%	85.37%	0.85	2.6	33
5	3.22427	92.07%	85.37%	0.8	2.6	7
6	3.54791	89.63%	82.93%	0.9	2.5	13
7	2.63155	90.85%	75.61%	0.9	2.75	32
8	3.25196	85.98%	87.80%	0.9	2.75	12
9	3.37425	94.51%	87.80%	0.9	2.5	6
10	3.09011	88.41%	82.93%	0.9	2.5	15
PROMEDIO	3.142769	90.42%	83.17%			

Tabla 10. Datos femeninos estandarizados con puntuación Z utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.

Femenino con puntuación FINDRSIC estandarizados con puntuación Z						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	0.14134	93.90%	75.61%	0.95	2.75	35
2	0.17716	86.59%	87.80%	0.9	2.6	28
3	0.19859	88.41%	92.68%	0.9	2.75	69
4	0.17911	92.68%	82.93%	0.9	2.5	290
5	0.21552	90.24%	75.61%	0.9	2.7	36
6	0.09664	94.51%	95.12%	0.95	2.6	1000
7	0.17276	90.85%	87.80%	0.9	2.75	22
8	**0.142563**	**97.56%**	**100.00%**	**0.9**	**2.6**	**291**
9	0.2297	89.02%	75.61%	0.9	2.5	18
10	0.19885	92.68%	90.24%	0.85	2.6	252
PROMEDIO	0.1752233	91.64%	86.34%			

Tabla 11. Datos femeninos estandarizados con puntuación Z utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.

Femenino con puntuación binaria o triaría estandarizados con puntuación Z						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	0.14041	92.07%	78.05%	0.95	2.75	72
2	0.16813	89.63%	80.49%	0.95	2.75	37
3	0.12058	96.95%	82.93%	0.9	2.75	378
4	0.17063	86.59%	90.24%	0.9	2.5	110
5	0.1951	88.41%	87.80%	0.9	2.75	112
6	0.15522	92.07%	95.12%	0.8	2.55	44
7	0.20285	92.07%	85.37%	0.9	2.5	53
8	0.20276	89.63%	80.49%	0.85	2.6	35
9	0.20694	87.20%	82.93%	0.9	2.5	31
10	0.20863	87.20%	92.68%	0.9	2.75	28
PROMEDIO	0.177125	90.18%	85.61%			

Tabla 12. Datos masculinos sin estandarización utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.

Masculino con puntuación FINDRSIC sin estandarización						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	2.05291	94.44%	88.89%	0.95	2.75	40
2	1.20882	97.22%	83.33%	0.95	2.85	245
3	2.39303	93.06%	66.67%	0.95	2.75	28
4	1.53665	97.22%	88.89%	0.9	2.75	67
5	2.05158	94.44%	77.78%	0.95	2.75	125
6	2.75135	91.67%	88.89%	0.95	2.75	16
7	2.08011	90.28%	94.44%	0.9	2.5	129
8	1.08886	100.00%	77.78%	0.95	2.75	360
9	2.97198	88.89%	88.89%	0.9	2.5	81
10	1.72038	94.44%	72.22%	0.95	2.6	72
PROMEDIO	1.985567	94.17%	82.78%			

Tabla 13. Datos masculinos sin estandarización utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.

Masculino con puntuación binaria o triaría sin estandarización						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	1.43682	95.83%	94.44%	0.95	2.75	30
2	2.61132	91.67%	77.78%	0.9	2.5	8
3	1.44616	97.22%	77.78%	0.9	2.75	33
4	2.00594	95.83%	94.44%	0.9	2.5	34
5	3.25837	88.89%	72.22%	0.9	2.75	15
6	1.94792	91.67%	88.89%	0.95	2.75	64
7	2.85817	88.89%	66.67%	0.95	2.75	7
8	2.54268	91.67%	77.78%	0.95	2.75	12
9	1.45914	95.83%	88.89%	0.9	2.75	44
10	0.95842	97.22%	100.00%	0.9	2.5	53
PROMEDIO	2.052494	93.47%	83.89%			

Tabla 14. Datos masculinos estandarizados con puntuación Z utilizando la puntuación de FINDRISC para las preguntas binarias y triarías.

Masculino con puntuación FINDRSIC estandarizados con puntuación Z						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	0.12797	91.67%	88.89%	0.95	2.75	32
2	0.13997	97.22%	77.78%	0.9	2.5	44
3	0.10611	95.83%	83.33%	0.9	2.75	121
4	0.06775	100.00%	88.89%	0.9	2.5	84
5	0.0466	97.22%	100.00%	0.95	2.75	119
6	0.08684	95.83%	83.33%	0.9	2.75	58
7	**0.04739**	**100.00%**	**100.00%**	**0.9**	**2.5**	**140**
8	0.1127	95.83%	83.33%	0.9	2.5	76
9	0.04696	97.22%	94.44%	0.9	2.6	115
10	0.18264	86.11%	72.22%	0.95	2.6	43
PROMEDIO	0.096493	95.69%	87.22%			

Tabla 15. Datos masculinos estandarizados con puntuación Z utilizando valores binarios y triarios para las preguntas binarias y triarías respectivamente.

Masculino con puntuación binaria o triaria estandarizados con puntuación Z						
No.	**ERROR ENTRENAMIENTO**	**ACEPTACIÓN ENTRENAMIENTO**	**ACEPTACIÓN PRUEBAS**	**RADIO**	**FAC DEP**	**EPOCAS**
1	0.08439	97.22%	100.00%	0.95	2.75	58
2	0.11784	95.83%	83.33%	0.9	2.75	67
3	0.02831	98.61%	72.22%	0.95	2.8	105
4	0.17789	87.50%	88.89%	0.9	2.75	45
5	0.19325	87.50%	66.67%	0.95	2.85	47
6	0.08316	95.83%	88.89%	0.9	2.5	108
7	0.14371	93.06%	66.67%	0.95	2.6	21
8	0.02377	100.00%	77.78%	0.95	2.9	118
9	0.1212	94.44%	88.89%	0.9	2.5	16
10	0.12328	90.28%	83.33%	0.9	2.75	91
PROMEDIO	0.10968	94.03%	81.67%			

Se destaca que las ANFIS entrenadas con datos estandarizados con puntuación Z, arrojan errores de entrenamiento considerablemente menores a las ANFIS que trabajan con datos sin estandarizar. El porcentaje es superior en 1% al utilizar estandarización, fenómeno que se repite casi siempre al revisar la aceptación para los conjuntos de prueba. La única excepción es el último caso donde se obtiene la menor aceptación aun utilizando puntuación Z. Nótese que, sin estandarizar los datos, se obtienen mejores resultados al utilizar puntuaciones binarias o triarías. Mientras que, al estandarizar los datos, los mejores resultados se obtienen al utilizar la puntuación establecida por la prueba de FINDRISC. Tanto para

masculinos como para femeninos, el mejor resultado se obtiene al estandarizar los datos y utilizar la puntuación de FINDRISC, pues se obtienen mejores promedios de error y aceptación, además, de que la ANFIS femenina número 8 de este caso, es la que obtiene los mejores resultados con un error de entrenamiento de 0.1425 y porcentajes de aceptación de 97.56% para el conjunto de entrenamiento y 100% en el conjunto de pruebas. En el caso masculino, con la ANFIS número 7 el error de entrenamiento obtenido es de 0.0473 con aceptación para entrenamiento y pruebas de 100%.

Es por ello que se eligen estas ANFIS para comparar su rendimiento con el de la prueba de FINDRISC y el sistema difuso diseñado. Las ANFIS seleccionadas se han resaltado en *letras negritas* en las Tablas 10 y 14.

6.2 Rendimiento del sistema difuso

Como ya se ha mencionado, se tiene una amplia variedad de formas de funciones de membresía a utilizar, la cantidad de variables lingüísticas es igual a las categorías de la variable de entrada establecida en la prueba de FINDRISC. La posición de las funciones se establece de manera que el punto central de la función objetivo corresponda al punto central de la categoría en la prueba de FINDRISC. La pendiente y anchura de las funciones, se ajusta para abarcar de la mejor manera el resto del universo de discurso. Se desprenden únicamente 3 variantes de configuración por género, pues los valores de circunferencia abdominal varían el riesgo según el género. Los casos evaluados para la elección de la mejor configuración se describen a continuación:

- Datos femeninos utilizando funciones de membresía trapezoidales.
- Datos femeninos utilizando funciones de membresía trapezoidales en los extremos del universo discurso y triangulares en la región interna.
- Datos femeninos utilizando funciones de membresía Z y S en los extremos del universo de discurso y campanas generalizadas en la región interna.
- Datos masculinos utilizando funciones de membresía trapezoidales.
- Datos masculinos utilizando funciones de membresía trapezoidales en los extremos del universo de discurso y triangulares en la región interna.
- Datos masculinos utilizando funciones de membresía Z y S en los extremos del universo de discurso y campanas generalizadas en la región interna.

La exactitud establecida para contabilizar la aceptación en este experimento es considerar solamente a aquellas evaluaciones cuya diferencia promedio no sea mayor al 5% en cuanto al riesgo reportado por la prueba de FNIDRISC, se evalúan las 295 entradas de la base de datos en cada caso. Los resultados se condensan en la Tabla 16.

Tabla 16. Resultados de distintas configuraciones de funciones de membresía para el sistema difuso.

GÉNERO	FORMA DE LAS FUNCIONES	PROMEDIO DE DIFERENCIA	MÍNIMA DIFERENCIA	MÁXIMA DIFERENCIA	PORCENTAJE DE ACEPTACIÓN
Femenino	TRP	1.96885	0.28757	8.21046	81.46%
Femenino	TRI-TRP	2.97574	0.27193	11.67042	77.07%
Femenino	**ZS-BELL**	**3.06305**	**0.00664**	**10.93380**	**91.22%**
Masculino	TRP	4.2811	0.58743	17.59969	74.44%
Masculino	TRI-TRP	3.1831	0.43423	11.67318	78.89%
Masculino	**ZS-BELL**	**2.40929**	**0.30096**	**12.38788**	**78.89%**

Para ambos géneros, el porcentaje de aceptación es superior al utilizar funciones S, Z y campanas generalizadas, si bien, la máxima diferencia es la mayor para el sistema masculino y la segunda mayor para el femenino, esta diferencia mayor a 10% de riesgo de padecimiento respecto a FINDRISC no necesariamente sugiere que el sistema difuso este dando resultados erróneos, algunos de estos casos con diferencias relativamente grandes se analizan más adelante. Además, los promedios de diferencia de la configuración de S, Z y campanas generalizadas no son mayores al 3%. Dados estos resultados, se seleccionan funciones de membresía S, Z y campanas generalizadas para establecer una comparación de rendimiento con el obtenido por el sistema ANFIS, la selección se ha resaltado en *letras negritas* en la Tabla 16.

Como se mencionó en el párrafo anterior, el sistema difuso arrojó diferencias de 10% de riesgo o más respecto al resultado definido por la prueba de FINDRISC, en la Tabla 17 se revisan algunos casos para verificar la validez del resultado del sistema difuso.

Tabla 17. Algunos casos con al menos 10% de diferencia respecto a la prueba de FINDRISC.

ID	Edad	IMC	Circ. Abdominal	Act física	Frutas y verduras	Tratamiento previo presión sanguínea	Estudio previo glucosa alta	Hist familiar	Riesgo FINDRISC	Puntos FINDRISC	Riesgo Difuso
11	28	30.457656	87	2	0	0	0	3	4	11	14
	MUY BAJO	MODERADO	MODERADO	LIG. ELEV	MUY BAJO	MUY BAJO	MUY BAJO	MODERADO	LIG. ELEV		
70	35	29.131695	87	2	0	0	0	5	4	11	14
	MUY BAJO	BAJO	MODERADO	LIG. ELEV	MUY BAJO	MUY BAJO	MUY BAJO	MUY ALTO	LIG. ELEV		
79	35	29.131695	87	2	0	0	0	5	4	11	14
	MUY BAJO	BAJO	MODERADO	LIG. ELEV	MUY BAJO	MUY BAJO	MUY BAJO	MUY ALTO	LIG. ELEV		
89	27	29.516249	91	2	1	0	0	5	17	14	27
	MUY BAJO	BAJO	ALTO	LIG. ELEV	BAJO	MUY BAJO	MUY BAJO	MUY ALTO	MODERADO		
98	43	27.434842	88	2	0	0	0	5	4	11	14
	MUY BAJO	BAJO	MODERADO	LIG. ELEV	MUY BAJO	MUY BAJO	MUY BAJO	MUY ALTO	LIG. ELEV		
127	27	29.516249	91	2	1	0	0	5	17	14	27
	MUY BAJO	BAJO	ALTO	LIG. ELEV	BAJO	MUY BAJO	MUY BAJO	MUY ALTO	MODERADO		
176	28	30.457656	87	2	0	0	0	3	4	11	14
	MUY BAJO	MODERADO	MODERADO	LIG. ELEV	MUY BAJO	MUY BAJO	MUY BAJO	MODERADO	LIG. ELEV		
186	43	27.434842	88	2	0	0	0	5	4	11	14
	MUY BAJO	BAJO	MODERADO	LIG. ELEV	MUY BAJO	MUY BAJO	MUY BAJO	MUY ALTO	LIG. ELEV		

Para las personas 11, 70, 79, 98, 176, 186, se tiene un riesgo por FINDRISC de 4%, mientras que el sistema difuso arroja 14% de riesgo. Sin embargo, la puntuación de FINDRISC de 11 puntos, se encuentra a solamente 1 punto de entrar a la categoría de riesgo de 17%, ver Tabla 2, por lo que se considera correcto el resultado del sistema difuso. Los casos 89, 127. Se tiene un riesgo por FINDRISC de 17%, mientras que el sistema difuso arroja 27% de riesgo. Similarmente, la puntuación de FINDRISC de 14 puntos, que se encuentra a solamente 1 punto de entrar a la categoría de riesgo de 33%, por lo que se considera correcto el resultado del sistema difuso.

6.3 Comparación de rendimiento de ANFIS, FIS y FINDRISC

Se revisaron los sistemas ANFIS para decidir su posible inclusión en el sistema final, y de esta manera ahorrar tiempo de cómputo y optimización. Para comparar el rendimiento del sistema ANFIS y el sistema difuso diseñado, se utilizan las proporciones indicadas por Mendiola et al. en la Tabla 4 de la publicación asociada [25], por simplicidad, el resultado de ambos sistemas para cada paciente es redondeado al entero más cercano.

La Tabla 18 muestra la comparación de rendimiento entre la prueba FINDRISC, el sistema difuso y el sistema ANFIS, utilizando los datos de prueba y el conjunto completo de datos, se revisa el valor predictivo positivo (VPP), valor predictivo negativo (VPN), así como sensibilidad y especificidad con sus respectivos índices de confianza al 95%. La Tabla 19 muestra las diferencias mínima y máxima que obtiene cada sistema respecto al resultado dado por la prueba de FINDRSIC, así como el promedio de diferencias, las filas correspondientes a FINDRISC se toman directamente del estudio de Mendiola et. al [25].

Para dar una mayor validación a cada sistema, se utilizó la base de datos de PIMA Indians Diabetes Database [23], pero para determinar los valores ausentes que se requieren para la prueba de FINDRISC, se utilizó el proceso descrito a continuación. Se calcularon los datos faltantes a partir de interpolación lineal para la variable de circunferencia abdominal

- Si $IMC < 25$: circunferencia = $\frac{(IMC-15)(79-60)}{24-15} + 60$
- Si $25 \leq IMC \leq 30$: circunferencia = $\frac{(IMC-25)(87-80)}{30-25} + 80$
- Si $IMC > 30$: circunferencia = $\frac{(IMC-31)(140-88)}{45-31} + 88$

Donde se han utilizado los valores de corte de circunferencia de la prueba de FINDRISC y los valores máximos y mínimos de IMC y circunferencia considerados en sus respectivos universos de discurso del sistema difuso.

Las variables de tratamiento previo para presión alta y resultados previos de glucosa se determinaron a partir de los datos existentes, las variables de actividad física y consumo regular de frutas y verduras se establecieron de la siguiente manera:

- Si $IMC < 25$: Consumo de frutas = 0, Actividad física = 0
- Si $25 \leq IMC \leq 30$: Consumo de frutas = a$aleatorio\{0,1\}$, Actividad física = $aleatorio\{0,2\}$
- Si $IMC > 30$: Consumo de frutas = 1, Actividad física = 2

La puntuación para historial familiar se determina según los casos:

- Si $DPF < 0.33$: Historial familiar = 0
- Si $0.33 \leq DPF < 0.50$: Historial familiar = 3
- Si $DPF \geq 0.50$: Historial familiar = 5

Este proceso se llevó a cabo con el objetivo de complementar la información faltante de manera que los datos inexistentes requeridos para la prueba de FINDRISC se obtuvieran manteniendo coherencia respecto a los datos existentes en la base de datos.

Tabla 18. Proporciones con sus respectivos IC 95% respecto a la prueba de FINDRISC.

MENDIOLA ET.AL.								
	VPP	**VPP IC 95%**	**VPN**	**VPN IC 95%**	**SENSIB**	**SENSIB IC 95%**	**ESPECIF**	**ESPECIF IC 95%**
FINDRISC	0.2244	0.1615-0.2908	0.9640	0.9181-0.9882	0.8750	0.7320-0.9591	0.5255	0.4623-0.5881
FUZZY	0.3039	0.2146-0.3931	0.9533	0.9236-0.9831	0.7750	0.6455-0.9044	0.7215	0.6665-0.7765
ANFIS	0.1923	0.0408-0.3438	0.9696	0.9112-1.0000	0.8333	0.5351-1.0000	0.6037	0.4720-0.7354
PIMA								
	VPP	**VPP IC 95%**	**VPN**	**VPN IC 95%**	**SENSIB**	**SENSIB IC 95%**	**ESPECIF**	**ESPECIF IC 95%**
FINDRISC	0.4989	0.4519-0.5457	0.8500	0.8108-0.8891	0.8195	0.7733-0.8657	0.5540	0.5100-0.5979
FUZZY	0.6015	0.5421-0.6609	0.7802	0.7438-0.8167	0.6353	0.5774-0.6931	0.7882	0.7520-0.8243
ANFIS	0.5249	0.4761-0.5737	0.8451	0.8074-0.8827	0.7923	0.7456-0.8419	0.6111	0.5679-0.6541

Tabla 19. Diferencias respecto a la prueba de FINDRISC.

MEDNIOLA ET. AL.			
	MÁXIMA DIFERENCIA	**MÍNIMA DIFERENCIA**	**PROMEDIO DE DIFERENCIA**
FINDRISC	NA	NA	NA
FUZZY	10	0	2.8635
ANFIS	82	0	4.89
PIMA			
	MÁXIMA DIFERENCIA	**MÍNIMA DIFERENCIA**	**PROMEDIO DE DIFERENCIA**
FINDRISC	NA	NA	NA
FUZZY	26	0	8.29
ANFIS	527	0	10.9

Al utilizar la base de datos de Mendiola et. al., se observa que el sistema difuso obtiene resultados mejores para VPP con un IC 95% de casi 6% mayor, mientras que el VPN es ligeramente menor con un IC 95% de apenas 0.5% menor. Respecto a la sensibilidad, esta es menor en un 10%, el IC 95% se mantiene casi 4% más grande al obtenido por FINDRISC. Por otro lado, la especificidad es muy superior por casi 20% para el sistema difuso además de mostrar un IC 95% ligeramente más reducido.

Se resalta que el promedio de diferencia es de menos de 3 unidades y una diferencia máxima de 10 unidades, sin embargo, este último dato no significa un peor desempeño, al no ser una diferencia considerable y analizar casos específicos, se puede concluir que es un resultado coherente, pues como se comentó en la sección 6.2, se pueden presentar situaciones en las que el resultado de FINDRISC clasifique a una persona en una categoría aún cuando se encuentre en el límite de pertenencia y la realidad indique que está mas cercana a una nueva categoría.

El ANFIS revela un VPP casi 3% inferior con un IC 95% de cerca de 30%, no obstante, el VPN es superior pero muy similar con un IC 95% de tamaño cercano a 9%, la sensibilidad es menor por un poco más de 4% pero con un IC 95% más grande de casi 47%, además la especificidad resulta superior en casi un 10% con un IC 95% de casi 27%. El promedio de diferencia es de 4.89 unidades, pero la diferencia máxima es de más de 82 unidades, para los casos similares, esta diferencia es preocupante por un riesgo distinto en 80 unidades, que es claramente un resultado erróneo por la magnitud alcanzada.

Probando con la base de datos de PIMA, la prueba de FINDRISC regresa un VPP de casi 50% con una IC 95% de tamaño cercano a 10%, un VPN de 0.85 con IC 95% de tamaño cercano a 9, sensibilidad de 0.8195 con IC 95% ligeramente mayor a 9%, especificidad mayor al 50% con IC 95% de casi 9 unidades.

El sistema difuso mejora el VPP en casi 11% con un IC 95% reducido en casi 2%, un VPN inferior en poco menos de 7%, y un IC 95% muy similar, la sensibilidad es considerablemente inferior por casi 18% con un IC 95% de apenas 2% mayor la especificidad crece en casi un 13% manteniendo un IC 95% de menor tamaño pero muy similar. El promedio de diferencia es de 8.29% de riesgo con una diferencia máxima de 26 unidades, una magnitud considerablemente grande por lo que casos similares puedes considerarse errores del sistema.

El sistema ANFIS obtiene un VPP ligeramente superior con un IC 95% de tamaño muy similar, un VPN inferior pero muy parecido con un IC 95% también muy acercado, la sensibilidad continúa siendo ligeramente menor con un IC 95% más grande por casi 1%, la especificidad es caso 6% mejor con un IC 95% muy similar. Finalmente, el promedio de diferencia es de 10.9% de riesgo que, como se comentó anteriormente no es un indicador de un error, es una magnitud razonable, por otro lado, una diferencia máxima de más de 527 unidades representa claramente un error del sistema al menos para ese caso en específico, considere que este caso se trata de un paciente real, por lo que pacientes con un perfil igual o similar, seguramente obtendrá un error de la misma magnitud, situación que no es permisible. De este modo se decide dejar de lado el uso de las ANFIS y continuar con la etapa de optimización del sistema difuso diseñado, con el objetivo de mejorar los resultados obtenidos por la prueba de FINDRISC.

6.4 Rendimiento de las metaheurísticas

Se tomaron los sistemas difusos femenino y masculino utilizados en la comparación de rendimiento de la sección anterior y fueron optimizados en 40 ocasiones por cada técnica metaheurística, el objetivo es

maximizar los indicadores de sensibilidad, especificidad y valores predictivos positivo y negativo, aunque la prioridad es aumentar los valores de sensibilidad [15].

La forma de optimización consiste en tomar los parámetros $a\ y\ b$ de las funciones de campana generalizada, correspondientes a la anchura y pendiente para las variables cuyos datos son números y no puntuación de FINDRISC, es decir, las dos funciones centrales de las variables de edad, una función central para índice de masa corporal y una función central para circunferencia abdominal. El problema de optimización tiene un total de 8 dimensiones. Se muestra el máximo valor obtenido de cada métrica, así como su mínimo, varianza, desviación estándar y promedio de cada género y algoritmo. Los resultados para el sistema femenino se muestran en las Tablas 20 a 23, mientras que los resultados para sistema masculino están plasmados en las Tablas 24 a 27.

Tabla 20. Resultados de búsqueda armónica para el FIS femenino.

Búsqueda armónica (Femenino)				
	VPN	**VPP**	**SENSIBILIDAD**	**ESPECIFICIDAD**
Mínimo	0.969	0.3289	0.862	0.7102
Máximo	0.9854	0.4386	0.931	0.8182
Promedio	0.971575	0.381715	0.86382	0.76803
Varianza	5.61E-06	0.000578	0.000116	0.000537
Desviación estándar	0.002369	0.024032	0.010757	0.02318

Tabla 21. Resultados de la polinización de flores para el FIS femenino.

Polinización de flores (Femenino)				
	VPN	**VPP**	**SENSIBILIDAD**	**ESPECIFICIDAD**
Mínimo	0.9714	0.3462	0.8621	0.7102
Máximo	0.986	0.4355	0.931	0.8125
Promedio	0.983893	0.391433	0.92411	0.761638
Varianza	1.61E-05	0.000593	0.000427	0.000663
Desviación estándar	0.00401	0.024357	0.02067	0.025755

Tabla 22. Resultados de PSO para el FIS femenino.

PSO (Femenino)				
	VPN	**VPP**	**SENSIBILIDAD**	**ESPECIFICIDAD**
Mínimo	0.9701	0.3521	0.8621	0.7386
Máximo	0.9848	0.4464	0.931	0.8239
Promedio	0.971905	0.390423	0.863823	0.776848
Varianza	4.68E-06	0.000382	0.000116	0.000359
Desviación estándar	0.002162	0.019539	0.010757	0.018948

Tabla 23. Resultados de recocido simulado para el FIS femenino.

Recocido simulado (Femenino)				
	VPN	**VPP**	**Sensibilidad**	**Especificidad**
Mínimo	0.9574	0.3521	0.7931	0.7386
Máximo	0.9706	0.3623	0.8621	0.767
Promedio	0.970208	0.360953	0.860375	0.749
Varianza	4.23E-06	1.14E-05	0.000116	2.25E-05
Desviación estándar	0.002057	0.003376	0.010773	0.00474

Tabla 24. Resultados de búsqueda armónica para el FIS masculino.

Búsqueda Armónica (Masculino)				
	VPN	**VPP**	**Sensibilidad**	**Especificidad**
Mínimo	0.9818	0.2857	0.9091	0.6835
Máximo	0.9846	0.4	0.9091	0.8101
Promedio	0.98322	0.33207	0.9091	0.742418
Varianza	5.07E-07	0.000913	4.93E-32	0.001101
Desviación estándar	0.000712	0.030211	2.22E-16	0.033183

Tabla 25. Resultados de la polinización de flores para el FIS masculino.

Polinización de flores (Masculino)				
	VPN	**VPP**	**Sensibilidad**	**Especificidad**
Mínimo	0.9818	0.2857	0.9091	0.6835
Máximo	0.9846	0.4	0.9091	0.8101
Promedio	0.983463	0.34305	0.9091	0.75443
Varianza	5.83E-07	0.000972	4.93E-32	0.001224
Desviación estándar	0.000764	0.031181	2.22E-16	0.034986

Tabla 26. Resultados de PSO para el FIS masculino.

PSO (Masculino)				
	VPN	**VPP**	**Sensibilidad**	**Especificidad**
Mínimo	0.9818	0.2857	0.9091	0.6835
Máximo	0.9846	0.4	0.9091	0.8101
Promedio	0.983868	0.36176	0.9091	0.774043
Varianza	4.7E-07	0.000891	4.93E-32	0.001021
Desviación estándar	0.000685	0.02985	2.22E-16	0.03196

Tabla 27. Resultados de recocido simulado para el FIS masculino.

Recocido simulado (Masculino)				
	VPN	**VPP**	**Sensibilidad**	**Especificidad**
Mínimo	0.9524	0.2963	0.7273	0.7089
Máximo	0.9836	0.3448	0.9091	0.7595
Promedio	0.981615	0.324385	0.90001	0.738628
Varianza	4.5E-05	0.000203	0.00157	0.000252
Desviación estándar	0.00671	0.014235	0.039622	0.015888

Para el sistema femenino, los mejores resultados de sensibilidad son obtenidos con búsqueda armónica, polinización de flores y PSO con 0.9310, que es mayor al 0.86297 obtenido por la prueba de FINDRISC, con especificades de 0.8182, 0.8125 y 0.8239, superiores al 0.53977 de FINDRISC. Si bien el algoritmo que muestra mejores resultados máximos es PSO, se elige como mejor heurística al algoritmo de polinización pues la técnica que tiene el promedio mayor de sensibilidad de las 40 ejecuciones con 0.9241, además, los valores predictivos y positivos de los 3 métodos son muy similares, también, las varianzas y desviaciones estándar entre todos los resultados son minúsculas, esto significa que las 40 ejecuciones entregan resultados similares, entonces los algoritmos son estables, por esto, el sistema optimizado por polinización de flores con mejor sensibilidad y especificidad será usado en el sistema para procesar datos femeninos.

Por otro lado, para el sistema masculino la sensibilidad de 0.9091 obtenida por FINDRISC no se ve superada, pero si igualada por los 4 algoritmos, no obstante, el algoritmo de recocido simulado obtiene el peor resultado de especificad con 0.7595, que no deja de ser superior al 0.49367 obtenido por la prueba de FINDRISC, pero está por debajo del 0.8101 de las otres 3 heurísticas. Aunque el promedio de especificidad de 0.774 de PSO es mayor al 0.7544 del algoritmo de polinización, este último es elegido al tener una diferencia menor a 0.02, además de corresponder a la elección de la heurística para el sistema femenino, también es la heurística más novedosa seleccionada y, seguramente que puede mejorar su rendimiento cuando se disponga de mayor cantidad de datos, del mismo modo, las desviaciones estándar y varianzas de todos los resultados son minúsculas por lo que se considera que los algoritmos se comportar de manera estable en cada ejecución.

6.5 Sistema optimizado

En la sección 5.2 se mencionó que las funciones de membresía para el género femenino y masculino son las mismas, salvo para la variable de medida de circunferencia abdominal, estas se muestran a continuación, la configuración de las variables de edad, índice de masa corporal y circunferencia abdominal del sistema sin optimizar en conjunto a las versiones optimizadas seleccionadas en la sección 6.4 para las variables de entrada ajustadas en las Figuras 18 a 27:

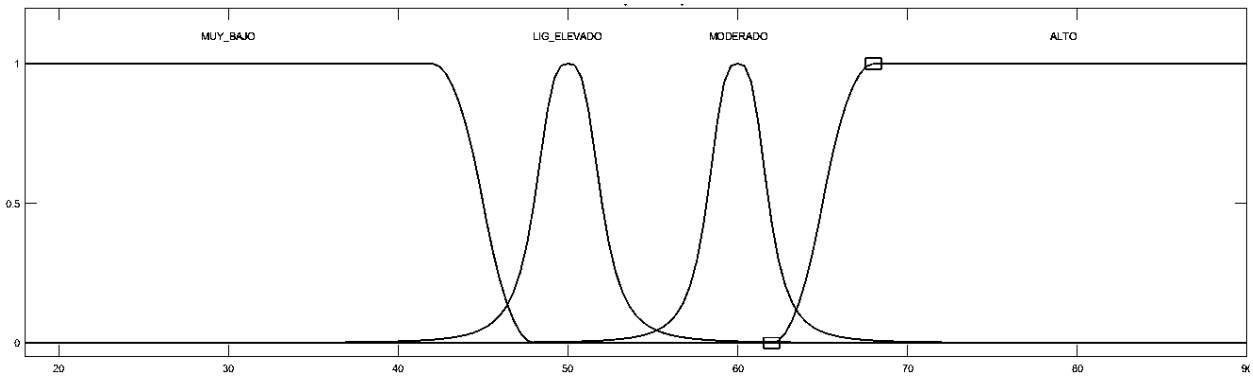

Figura 18. Funciones de membresía sin optimizar de la variable edad.

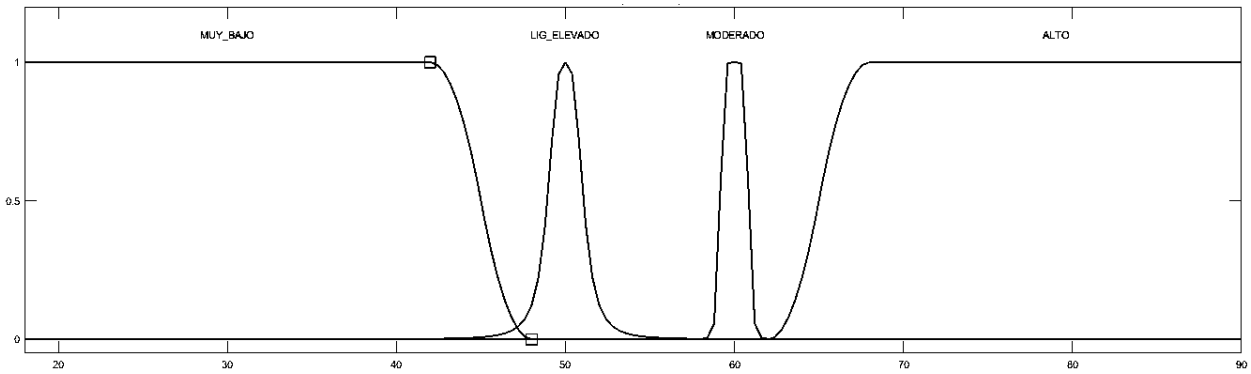

Figura 19. Funciones de membresía optimizadas para el género femenino de la variable edad.

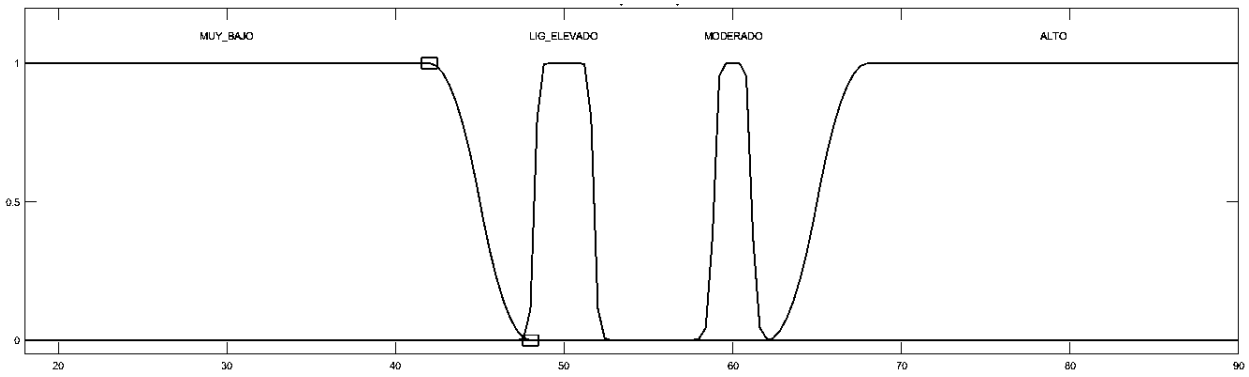

Figura 20. Funciones de membresía optimizadas para el género masculino de la variable edad.

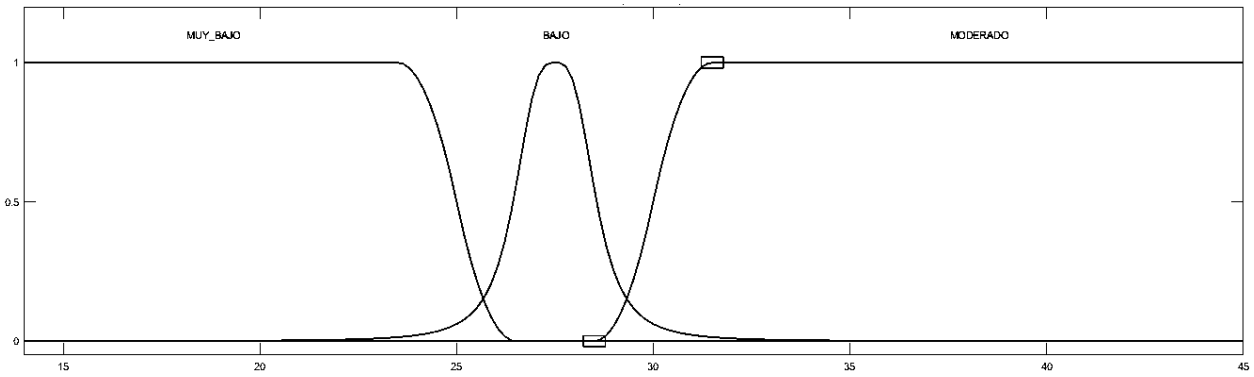

Figura 21. Funciones de membresía sin optimizar de la variable índice de masa corporal.

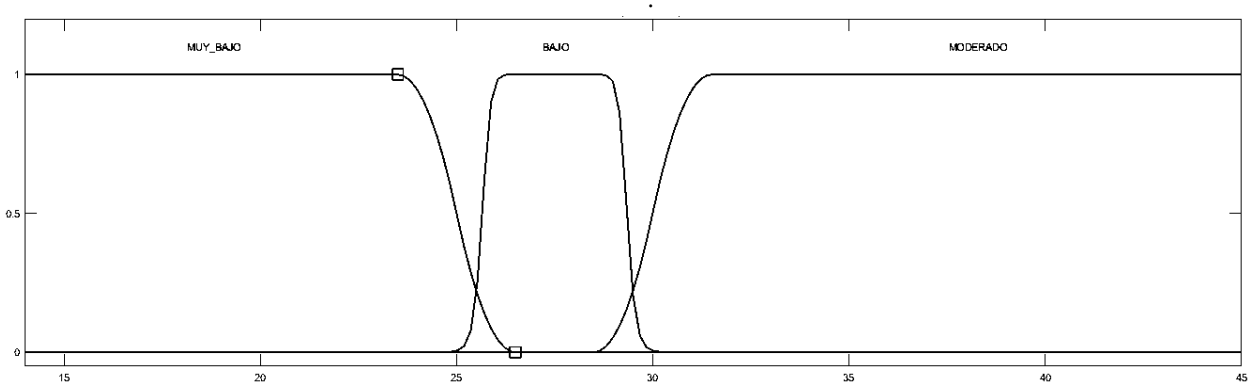

Figura 22. Funciones de membresía optimizadas para el género femenino de la variable índice de masa corporal.

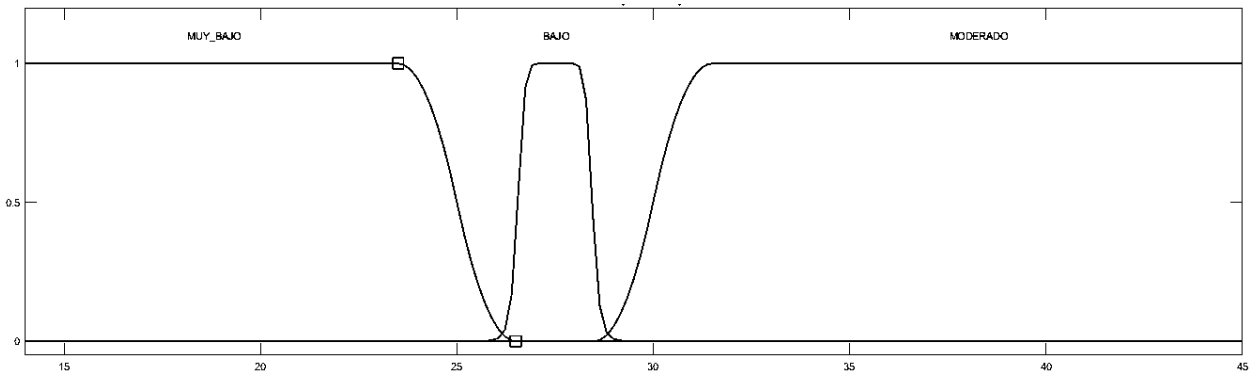

Figura 23. Funciones de membresía optimizadas para el género masculino de la variable índice de masa corporal.

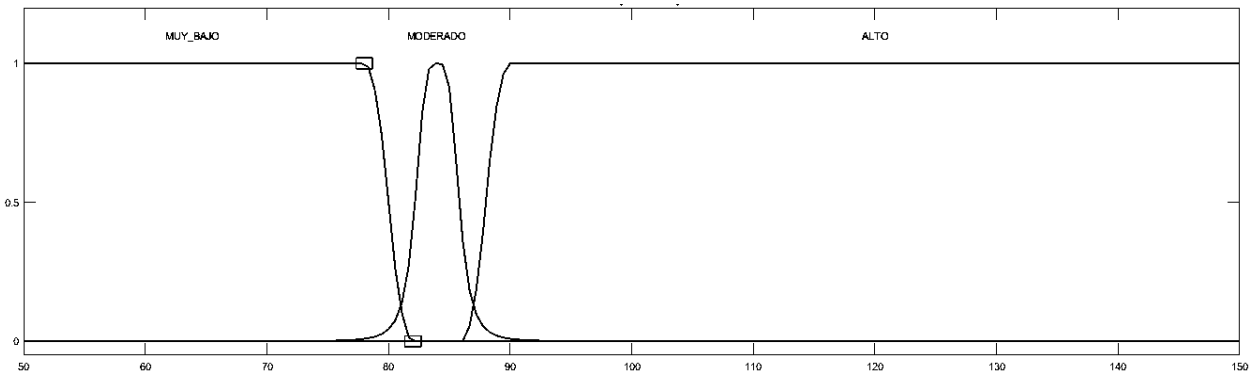

Figura 24. Funciones de membresía sin optimizar para el género femenino de la variable circunferencia abdominal.

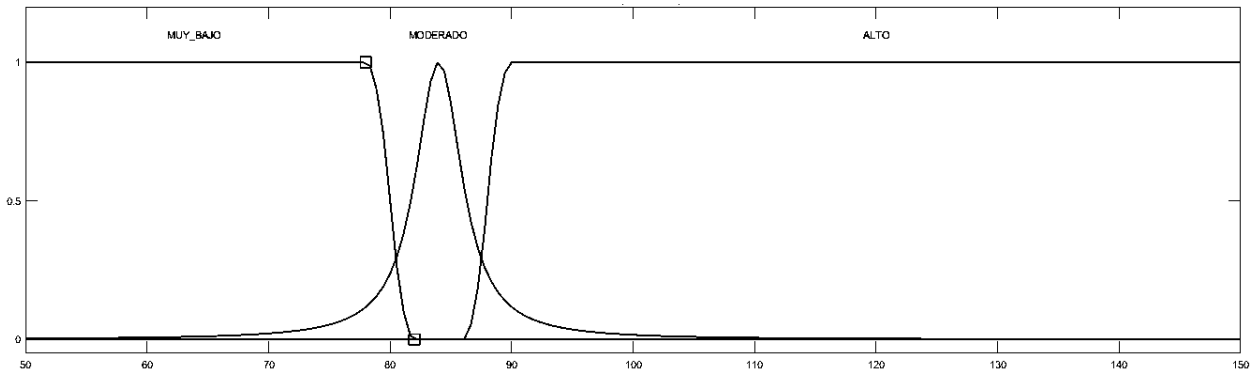

Figura 25. Funciones de membresía optimizadas para el género femenino de la variable circunferencia abdominal.

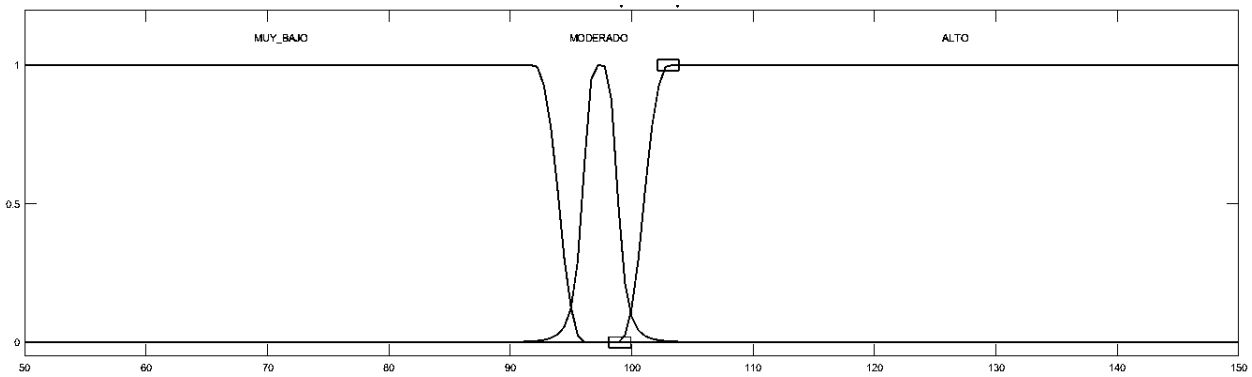

Figura 26. Funciones de membresía sin optimizar para el género masculino de la variable circunferencia abdominal.

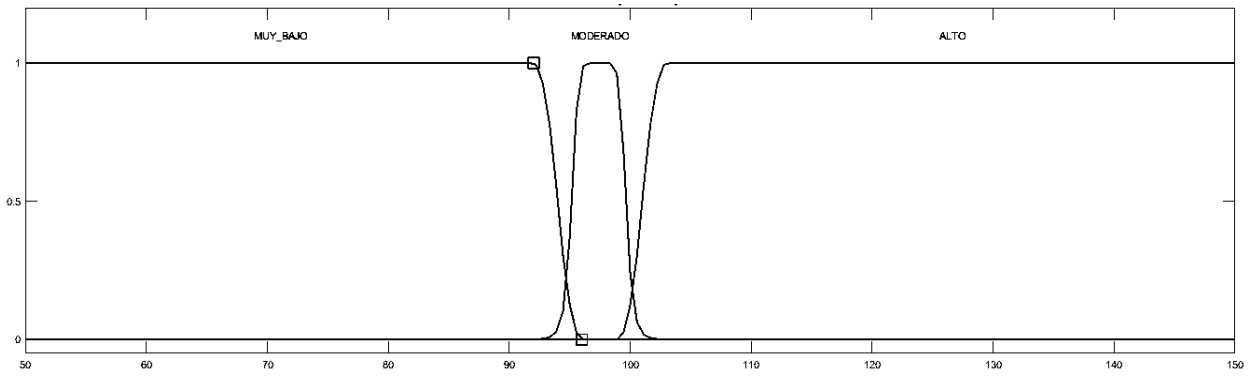

Figura 27. Funciones de membresía optimizadas para el género masculino de la variable circunferencia abdominal.

Las preguntas binarias y triarias de tratamiento previo de presión arterial, estudio previo de glucosa alta e historial familiar no requirieron optimización, se muestran en las Figuras 28 a 30:

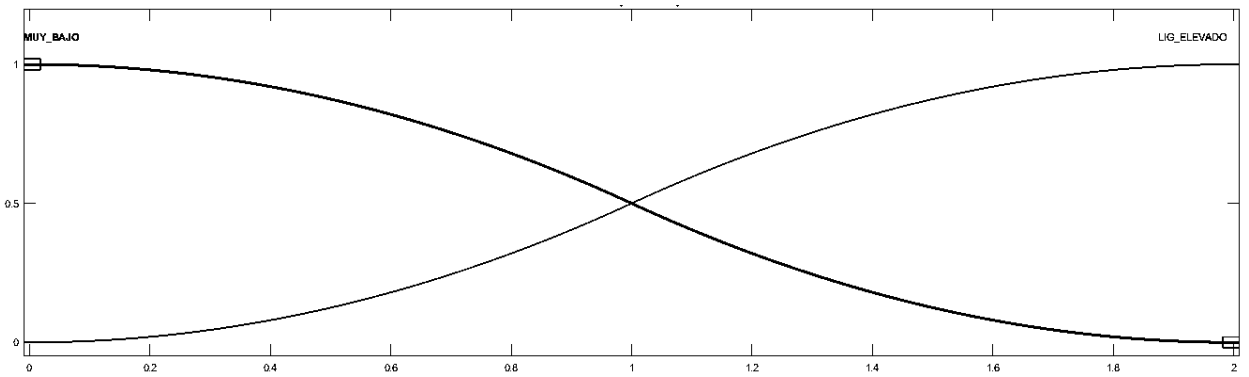

Figura 28. Funciones de membresía de la variable tratamiento previo de presión arterial.

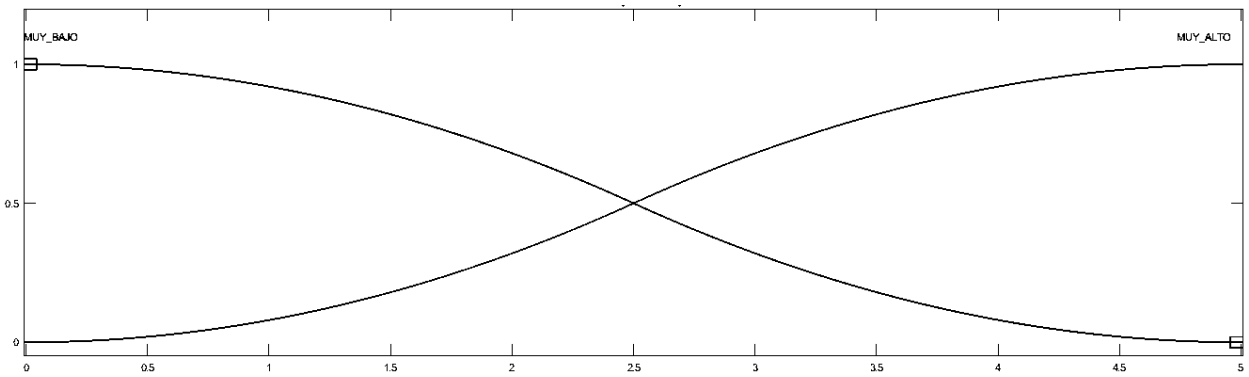

Figura 29. Funciones de membresía de la variable estudio previo de glucosa alta.

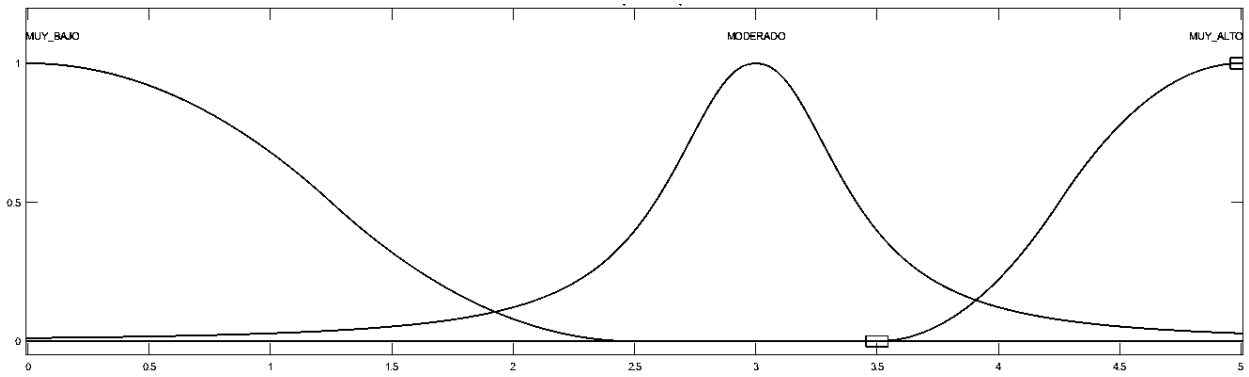

Figura 30. Funciones de membresía de la variable historial familiar de diabetes.

Las funciones de membresía para la salida de ambos sistemas se aprecian en la Figura 31:

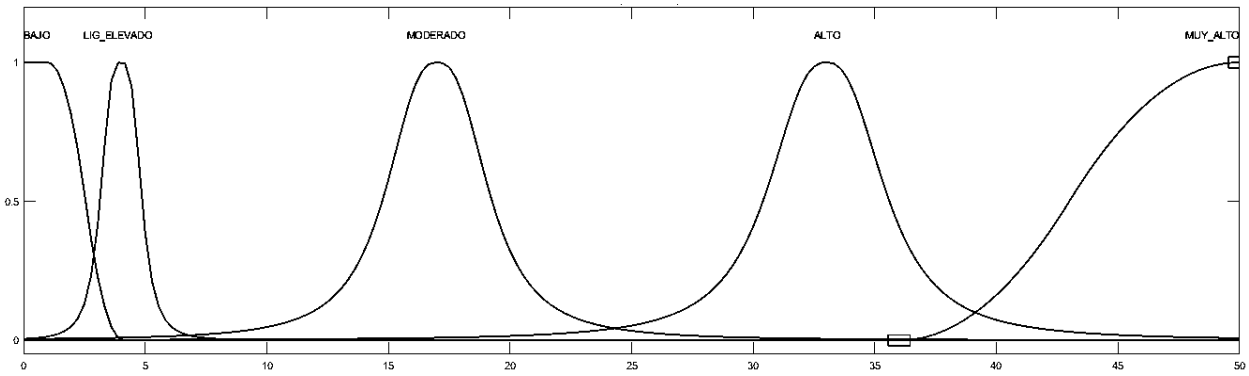

Figura 31. Funciones de membresía de la salida o riesgo de padecimiento de diabetes.

El diagrama del sistema definitivo se muestra en la Figura 32:

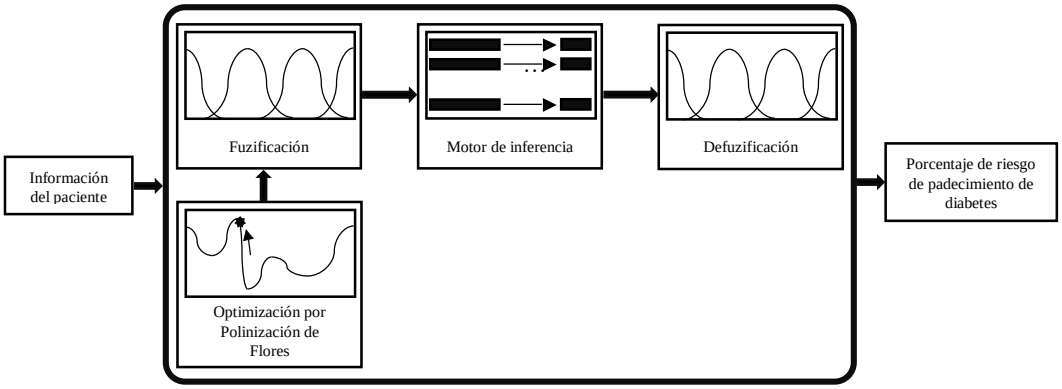

Figura 32. Diagrama del sistema difuso definitivo.

6.6 Rendimiento del sistema definitivo

Se eligieron los resultados difusos optimizados por el algoritmo de polinización de flores para ambos géneros. Los resultados, de las métricas seleccionadas para la medición del rendimiento generados se comparan con los resultados obtenidos por Mendiola et. al. para la puebla de FINDRISC [25] y se muestran en la Tabla 28.

Tabla 28. Comparación de la prueba de FINDRISC y el sistema difuso optimizado.

	TOTALES							
	VPN FINDRISC	**VPN DIFUSO**	**VPP FINDRISC**	**VPP DIFUSO**	**SENSIB FINDRISC**	**SENSIB DIFUSO**	**ESPECIF FINDRISC**	**ESPECIF DIFUSO**
	0.9640	0.9856	0.2244	0.4253	0.8750	0.9250	0.5255	0.8039
IC 95%	0.9181-0.9882	0.9694-1.0000	0.1615-0.2908	0.3214-0.5292	0.732-0.9591	0.8434-1.0000	0.4623-0.5881	0.7552-0.8527
TAMAÑO IC95%	0.0701	0.0324	0.1293	0.2078	0.2271	0.1633	0.1258	0.0975

Capítulo 7

Conclusiones y Trabajo Futuro

Se puede observar que el valor predictivo negativo aumento en 0.0216 en el sistema difuso optimizado, y el IC95% redujo su tamaño en más de la mitad. También, el valor predictivo positivo, casi se duplica en el sistema difuso al aumentar 0.2009 unidades al costo de incrementar hasta 0.2070 el tamaño del IC95% que, no es un tamaño preocupante. Por otro lado, la sensibilidad en el sistema difuso es mayor por 0.05 unidades con un intervalo de confianza reducido en casi 0.06 unidades respecto a la prueba de FINDRISC, finalmente, el sistema difuso tiene una mayor especificidad por 0.2784 unidad que FINDRISC con un IC95% también reducido. Estos datos muestran una clara mejora en el poder predictivo del sistema difuso optimizado sobre la prueba de FINDRISC, por lo que se concluye que se ha mejorado el desempeño del estándar actual. El sistema difuso optimizado se utilizó para construir una herramienta de usuario final.

Dados los resultados, se han cumplido los objetivos planteados al inicio del presente trabajo. Adicionalmente, se ha desarrollado una herramienta de prevención cuyo uso como auxiliar en materia de prevención de diabetes mellitus 2 pude ser de gran utilidad al brindar resultados que describen adecuadamente la realidad, comparados contra un estándar de oro. Lo anterior se logra porque la herramienta está construida con reglas provenientes de pacientes reales y está optimizada por metaheurísticas. Además, se ha desarrollado una versión de escritorio de la herramienta, que permite a cualquier persona conocer el riesgo de padecimiento con valor no categórico, y de ser necesario, dirigirse oportunamente con un especialista para la realización de estudios y tratamiento especializado. La descripción de dicha aplicación se muestra en el Anexo 1.

Como trabajo futuro se pretende realizar una etapa nueva de optimización para las funciones de membresía con datos más recientes, de más entidades federativas de México y en mayor cantidad si es posible hallarlos, esto también cumpliría con el objetivo de buscar una mejor diversidad de reglas para el sistema difuso e, incrementar su poder predictivo para la población mexicana. También, se desea exportar la aplicación de escritorio hacia dispositivos móviles incrementando la portabilidad y facilidad de aplicación en zonas de difícil acceso.

Anexo 1

Aplicación de escritorio

Utilizando el compilador de aplicaciones de escritorio de MATLAB y la herramienta *App Designer* del mismo software, se diseñó la interfaz de usuario y se programaron las funcionalidades. Se pensó en una distribución de los elementos que resultase amigable y fácil de entender, para evitar valores fuera de rango, los elementos gráficos se encuentran limitados al intervalo válido.

Se dispone de un botón para limpiar los datos y colocarse en el estado inicial. También se tiene un botón para calcular el riesgo usando los datos ingresados, el indicador púrpura muestra el riesgo de padecimiento de diabetes acompañado de un indicador semáforo que ayuda como guía visual del estado del paciente. Además, se dispone de un botón de ayuda para la correcta medición de la circunferencia abdominal.

Una vez calculado el riesgo, se habilita un botón que permite ver un resumen de los datos ingresados, junto con el riesgo calculado y la opción de guardar el resumen como imagen. La interfaz a la que el usuario tiene acceso está explicada en las Figuras 33 y 34, y muestra los siguientes componentes:

1. Selector de Edad, comprende valores entre 18 y 90 años.
2. Selector de Altura en centímetros, comprende valores entre 140 y 200.
3. Selector de Peso en kilogramos, comprende valores entre 40 y 150.
4. Selector de Altura en centímetros, comprende valores entre 50 y 150.
5. Selector de género, tiene como opciones: *Femenino* y *Masculino*.
6. Selector para informar sobre tratamiento previo para presión arterial alta, tiene como opciones: *No* y *Sí*.
7. Selector para informar sobre estudios previos de glucosa alta, tiene como opciones: *No* y *Sí*.
8. Selector para informar sobre parientes con diagnósticos de diabetes tipo 2, tiene como opciones: *No*, *Padre/Hermano/Hijo* y *Abuelo/Tío/Primo*.
9. Botón de ayuda que muestra una ventana con las indicaciones del Instituto Mexicano del Seguro Social (IMSS) para tomar la medida de la circunferencia abdominal, ver Figura 35.
10. Indicador en modo de espera del riesgo de padecimiento de diabetes tipo 2.
11. Indicador semáforo en modo de espera del riesgo de padecimiento de diabetes tipo 2.
12. Botón para reiniciar los valores iniciales del selector y colocar en modo de espera a los indicadores de riesgo de padecimiento de diabetes tipo 2.
13. Botón para calcular el riesgo de padecimiento de diabetes tipo 2 utilizando los datos en los selectores.
14. Botón para mostrar una ventana resumen de los datos ingresados y el riesgo calculado, y guardar el contenido de dicha ventana en formato de imagen, sólo se encuentra disponible mientras se esté mostrado un resultado, ver Figuras 36 y 37.

15. Indicador con el riesgo de padecimiento de diabetes tipo 2 resultante con los datos en los selectores, el color del indicador es el mismo que el correspondiente en el indicador semáforo.
16. Indicador semáforo con la posición del riesgo de padecimiento de diabetes tipo 2 en una escala de 1% hasta 50%.

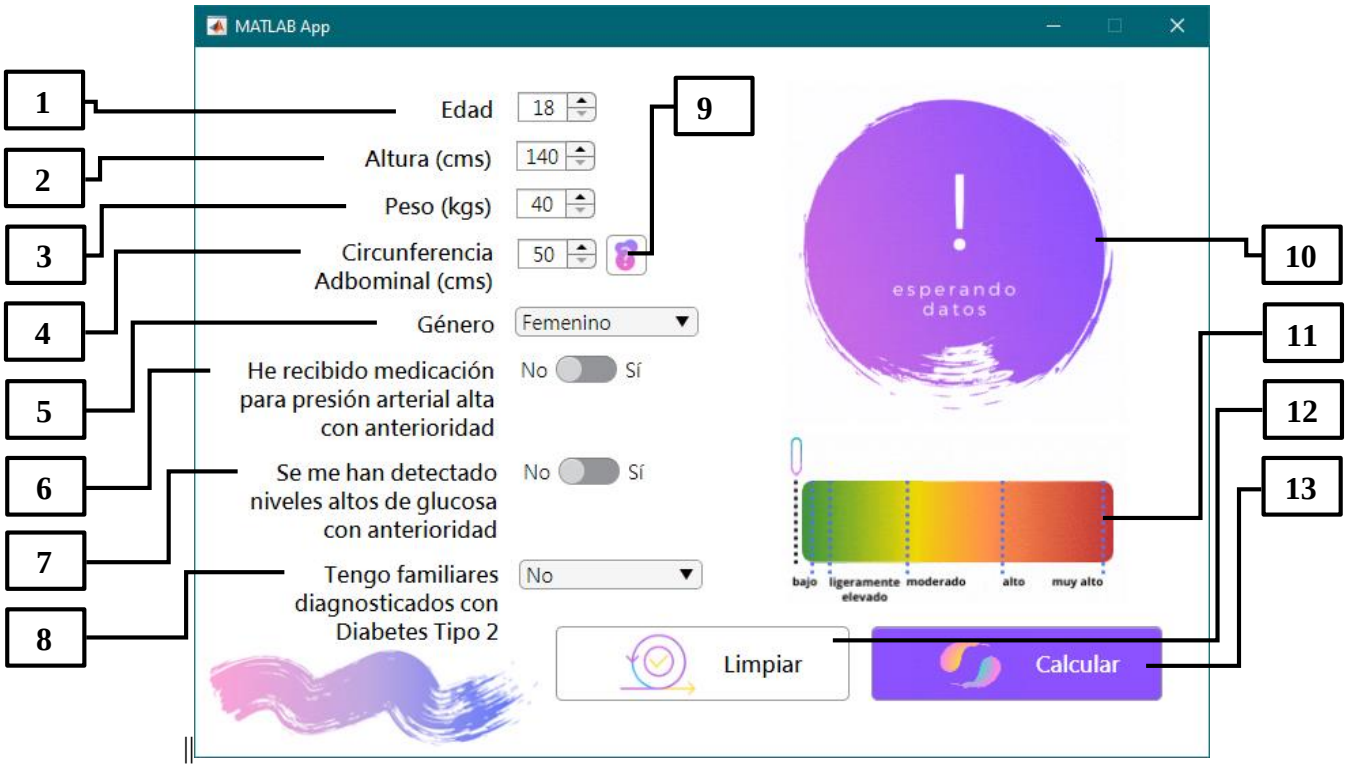

Figura 33. Interfaz de usuario en modo de espera.

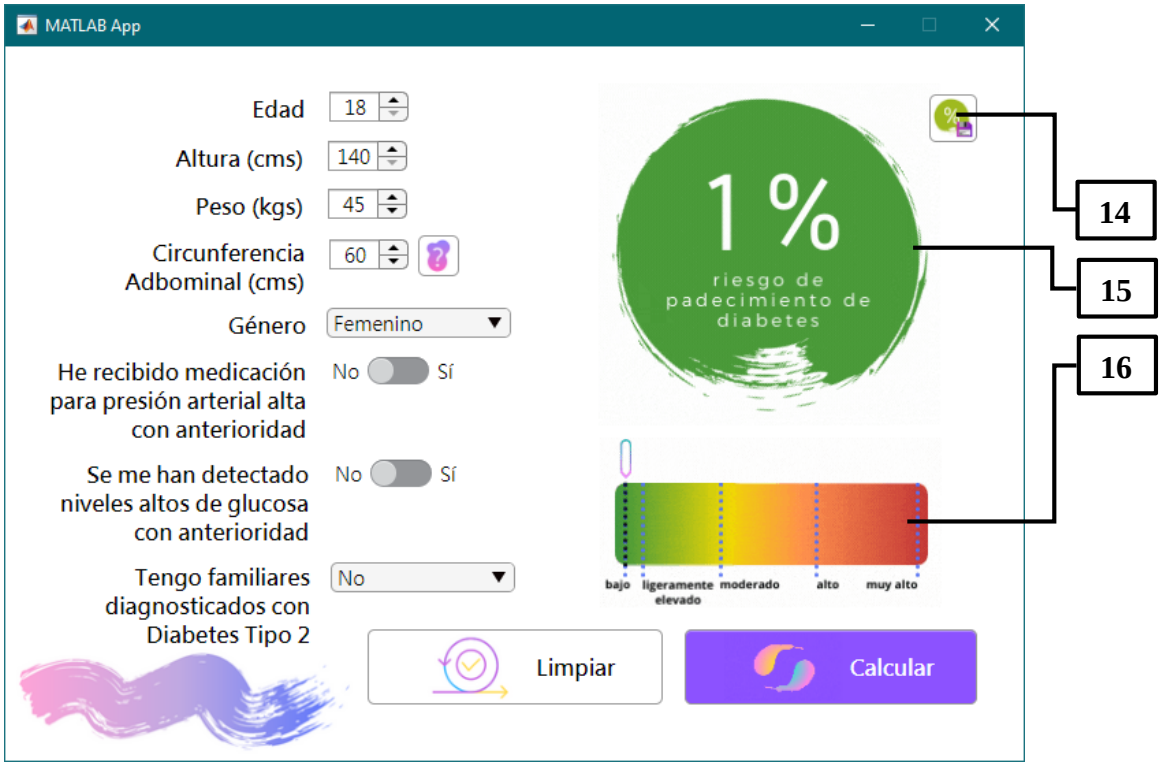

Figura 34. Interfaz de usuario mostrando un riesgo calculado.

Figura 35. Ventana de ayuda para la toma de medida de circunferencia abdominal.

Figura 36. Ventana con el riesgo calculado y resumen de los datos seleccionados que se guardará como imagen.

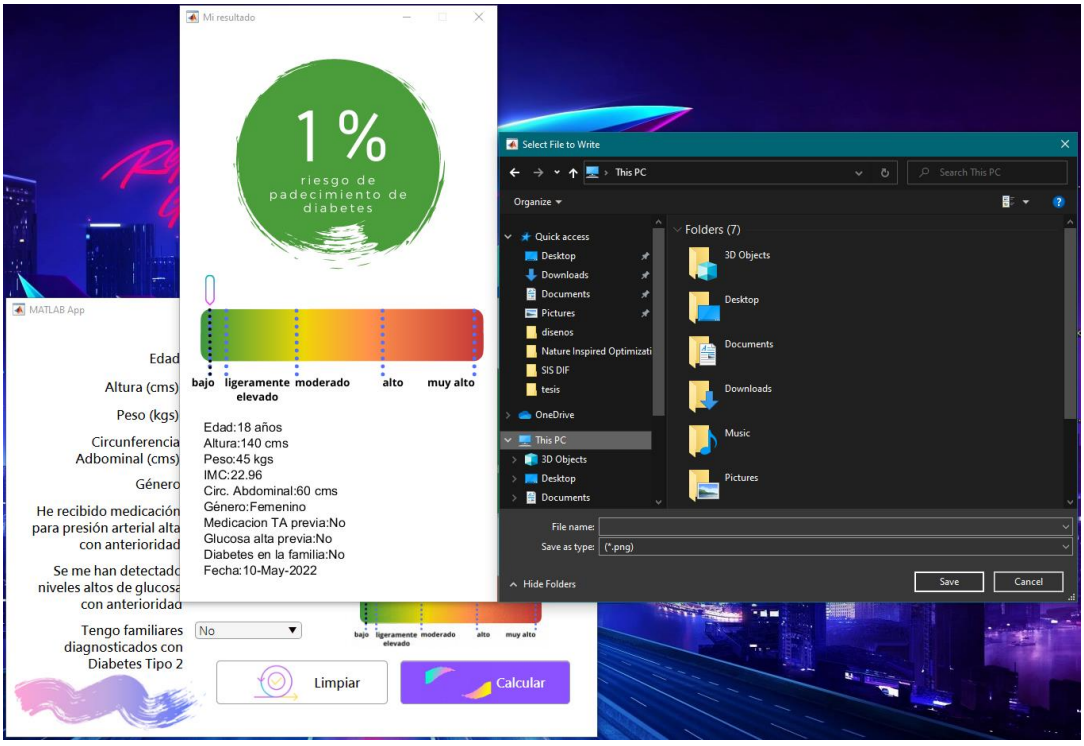

Figura 37. Almacenamiento de la ventana resumen como imagen.

Referencias

[1] Organización Mundial de la Salud. (2016). Informe Mundial Sobre la Diabetes.
[2] Instituto Nacional de Salud Pública. (2018). Encuesta Nacional de Salud y Nutrición.
[3] Zadeh, L. A.: Fuzzy Sets. In: Information and Control, pp. 8(3), 338-353, (1965).
[4] Zamora, Curso de Lógica Difusa - Hackeando Tec, https://www.youtube.com/playlist?list=PLIyIZGa1sAZoWAeT_tL7zCv3wi1ISrBa0, accesado el 8 de febrero de 2022.
[5] Fuzzy Logic Toolbox Documentatión, https://la.mathworks.com/help/fuzzy/, accesado el 10 de febrero de 2022.
[6] Ross, T. J.: Fuzzy Logic with Engineering Applications, John Wiley & Sons, 3rd. Ed., (2010).
[7] "Base de conocimiento", https://es.wikipedia.org/wiki/Base_de_conocimiento, accesado el 14 de febrero de 2002.
[8] Jain, V. & Raheja, S.: Improving the Prediction Rate of Diabetes using Fuzzy Expert System. In: International Journal of Information Technology and Computer Science (IJITCS), p. 7(10), 84, (2015).
[9] M. EROL KESKIN, DILEK TAYLAN & ÖZLEM TERZI (2006) Adaptive neural-based fuzzy inference system (ANFIS) approach for modelling hydrological time series, Hydrological Sciences Journal, 51:4, 588-598, DOI: 10.1623/hysj.51.4.588.
[10] FARIZAL HAKIKI & ARIS TRISTIANTO WIBOWO (2014), FORMULATION OF ROCK TYPE PREDICTION IN CORED WELL USING FUZZY SUBTRACTIVE CLUSTERING ALGORITHM, PROCEEDINGS, INDONESIAN PETROLEUM ASSOCIATION Thirty-Eighth Annual Convention & Exhibition, DOI: 10.29118/IPA.46.14.SE.118.
[11] Triola, M. Estadística, 9na Ed. Pearson Eduaction, pp. 92-93, (2004).
[12] Diagnóstico y Tratamiento Farmacológico de la Diabetes Mellitus Tipo 2 en el Primer Nivel de Atención, Catálogo Maestro de Guías de Práctica Clínica, INSTITUTO MEXICANO DEL SEGURO SOCIAL (2018), accesado el 12 de octubre del 2021.
[13] Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003 Mar;26(3):725-31. doi: 10.2337/diacare.26.3.725. PMID: 12610029.
[14] Bravo, S & Cruz, J. Estudios de exactitud diagnóstica: Herramientas para su Interpretación. Revista Chilena de Radiología. Vol. 21 Nº 4. 2015; 158-164.
[15] Ochoa, C & Orejas, G. Epidemiología y metodología científica aplicada a la pediatría (IV): Pruebas diagnósticas. An Esp Pediatr 1999; 50:301-314.
[16] "Exactitud", https://www.significados.com/exactitud/, accesado el 10 de febrero de 2022.
[17] Yang, X-S., Nature-Inspired Optimization Algorithms, Elsevier, 1st. Ed., (2014), ISBN: 9780124167438.
[18] Pima Indians Diabetes Database, https://www.kaggle.com/uciml/pima-indians-diabetes-database, accesado el 20 de agosto de 2021.
[19] Reddy, G.T. & Khare, N. (2017). Cuckoo Search Optimized Reduction and Fuzzy Logic Classifier for Heart Disease and Diabetes Prediction. *International Journal of Fuzzy System Applications (IJFSA)*, 6 (2), 25-42. doi: 10.4018/IJFSA.2017040102.
[20] Sahu, N., Verma, T. & Reddy, G.T. (2017). Diabetes classification using fuzzy logic and adaptive cuckoo search optimization techniques. *International Journal on Future Revolution in Computer Science & Communication Engineering (IJFRCSCE)*, 3 (9), 252-255.
[21] Bressan, G., Flamia Azevedo, B. & Souza, R. (2020). A Fuzzy Approach for Diabetes Mellitus Type 2 Classification. *Brazilian Archives of Biology and Technology*, 63. doi: 10.1590/1678-4324-2020180742.
[22] Risqy, F., et al. (2020). Fuzzy Tsukamoto Membership Function Optimization Using PSO to Predict Diabetes Mellitus Risk Level. *Pro-ceedings of SIET '20: 5th International Conference on Sustainable Information Engi-neering and Technology*, 101-106. doi: 10.1145/3427423.3427451.
[23] Pima Indians Diabetes Database, https://www.kaggle.com/uciml/pima-indians-diabetes-database, accesado el 28 de agosto del 2021.
[24] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.
[25] Mendiola, I., et al. Evaluación del desempeño del Finnish Diabetes Risk Score findrisc como prueba de tamizaje para diabetes mellitus tipo 2. Aten Fam. 2018;25(1):22-26.