CENTRO DE CIENCIAS BÁSICAS

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

IDENTIFICACIÓN DE FACTORES DE RIESGO EN PATOLOGÍAS MÉDICAS MEDIANTE MÉTODOS DE SELECCIÓN DE SUBCONJUNTOS DE CARACTERÍSTICAS

TESIS QUE PRESENTA

Alexis Edmundo Gallegos Acosta

PARA OBTAR POR EL GRADO DE

Maestría en Ciencias con Opción a Computación

TUTORES

Dr. Francisco Javier Álvarez Rodríguez

Dra. María Dolores Torres Soto

INTEGRANTES DEL COMITÉ TUTORAL

Dra. Aurora Torres Soto

Aguascalientes, Ags, 19 de mayo de 2018

Director de Tesis:

Dr. Francisco Javier Álvarez Rodríguez

Universidad Autónoma de Aguascalientes –México

Centro de Ciencias Básicas

Departamento de Ciencia de la Computación

Codirector de Tesis

Dra. María Dolores Torres Soto

Universidad Autónoma de Aguascalientes –México

Centro de Ciencias Básicas

Departamento de Sistemas de Información

Asesor:

Dra. Aurora Torres Soto

Universidad Autónoma de Aguascalientes –México

Centro de Ciencias Básicas

Departamento de Ciencias de la Computación

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

ALEXIS EDMUNDO GALLEGOS ACOSTA
MAESTRÍA EN CIENCIAS CON OPCIÓN A LA COMPUTACIÓN Y MATEMÁTICAS APLICADAS
P R E S E N T E.

Estimado alumno:

Por medio de este conducto me permito comunicar a Usted que habiendo recibido los votos aprobatorios de los revisores de su trabajo de tesis y/o caso práctico titulado: **"IDENTIFICACIÓN DE FACTORES DE RIESGO EN PATOLOGÍAS MÉDICAS MEDIANTE MÉTODOS DE SELECCIÓN DE SUBCONJUNTOS DE CARACTERÍSTICAS"**, hago de su conocimiento que puede imprimir dicho documento y continuar con los trámites para la presentación de su examen de grado.

Sin otro particular me permito saludarle muy afectuosamente.

A T E N T A M E N T E
Aguascalientes, Ags., a 22 de mayo de 2018
"Se lumen proferre"
EL DECANO

M. en C. JOSÉ DE JESÚS RUÍZ GALLEGOS

c.c.p.- Archivo.

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

M. en C. JOSÉ DE JESÚS RUIZ GALLEGOS:
DECANO DEL CENTRO DE CIENCAS BÁSICAS
P R E S E N T E

Por medio del presente como Tutor designado del estudiante **ALEXIS EDMUNDO GALLEGOS ACOSTA** con ID **139484** quien realizó la tesis titulada: **IDENTIFICACIÓN DE FACTORES DE RIESGO EN PATOLOGÍAS MÉDICAS MEDIANTE MÉTODOS DE SELECCIÓN DE SUBCONJUNTOS DE CARACTERÍSTICAS**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 21 de mayo de 2018

Dr. Francisco Javier Álvarez Rodríguez
Director de Tesis

c.c.p.- Interesado
c.c.p.- Secretaría de Investigación y Posgrado
c.c.p.- Jefatura del Depto. De Ciencias de la Computación
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

M. en C. JOSÉ DE JESÚS RUIZ GALLEGOS:
DECANO DEL CENTRO DE CIENCAS BÁSICAS
P R E S E N T E

Por medio del presente como Tutor designado del estudiante **ALEXIS EDMUNDO GALLEGOS ACOSTA** con ID **139484** quien realizó la tesis titulada: **IDENTIFICACIÓN DE FACTORES DE RIESGO EN PATOLOGÍAS MÉDICAS MEDIANTE MÉTODOS DE SELECCIÓN DE SUBCONJUNTOS DE CARACTERÍSTICAS**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 21 de mayo de 2018

Dra. María Dolores Torres Soto
Co-Director de Tesis

c.c.p.- Interesado
c.c.p.- Secretaría de Investigación y Posgrado
c.c.p.- Jefatura del Depto. De Ciencias de la Computación
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

M. en C. JOSÉ DE JESÚS RUIZ GALLEGOS:
DECANO DEL CENTRO DE CIENCAS BÁSICAS
P R E S E N T E

Por medio del presente como Tutor designado del estudiante **ALEXIS EDMUNDO GALLEGOS ACOSTA** con ID **139484** quien realizó la tesis titulada: **IDENTIFICACIÓN DE FACTORES DE RIESGO EN PATOLOGÍAS MÉDICAS MEDIANTE MÉTODOS DE SELECCIÓN DE SUBCONJUNTOS DE CARACTERÍSTICAS**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla, así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 21 de mayo de 2018

Dra. Aurora Torres Soto
Asesor de Tesis

c.c.p.- Interesado
c.c.p.- Secretaría de Investigación y Posgrado
c.c.p.- Jefatura del Depto. De Ciencias de la Computación
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

Agradecimientos

A cada una de las personas que, de alguna u otra manera, se involucraron en la realización de este proyecto, así como aquellos que me alentaron y creyeron en mí en momentos en los que, incluso, yo mismo no lo hacía. Muchas gracias.

A mis padres Edmundo y Tere; y mis hermanos Luis y Amairani por estar conmigo incondicionalmente y siempre tener un consejo o una palabra de aliento que me permite seguir adelante.

A mis amigos Paulina, Gabriela y Fernando por su apoyo, su amistad y por ser una parte muy importante en esta etapa profesional y personal.

A mis tutores la Dra. Dolores, Dra. Aurora y el Dr. Francisco por el apoyo y dedicación; por todo lo que he aprendido de cada uno de ustedes. Son un ejemplo como personas y como profesionistas. Muchas Gracias.

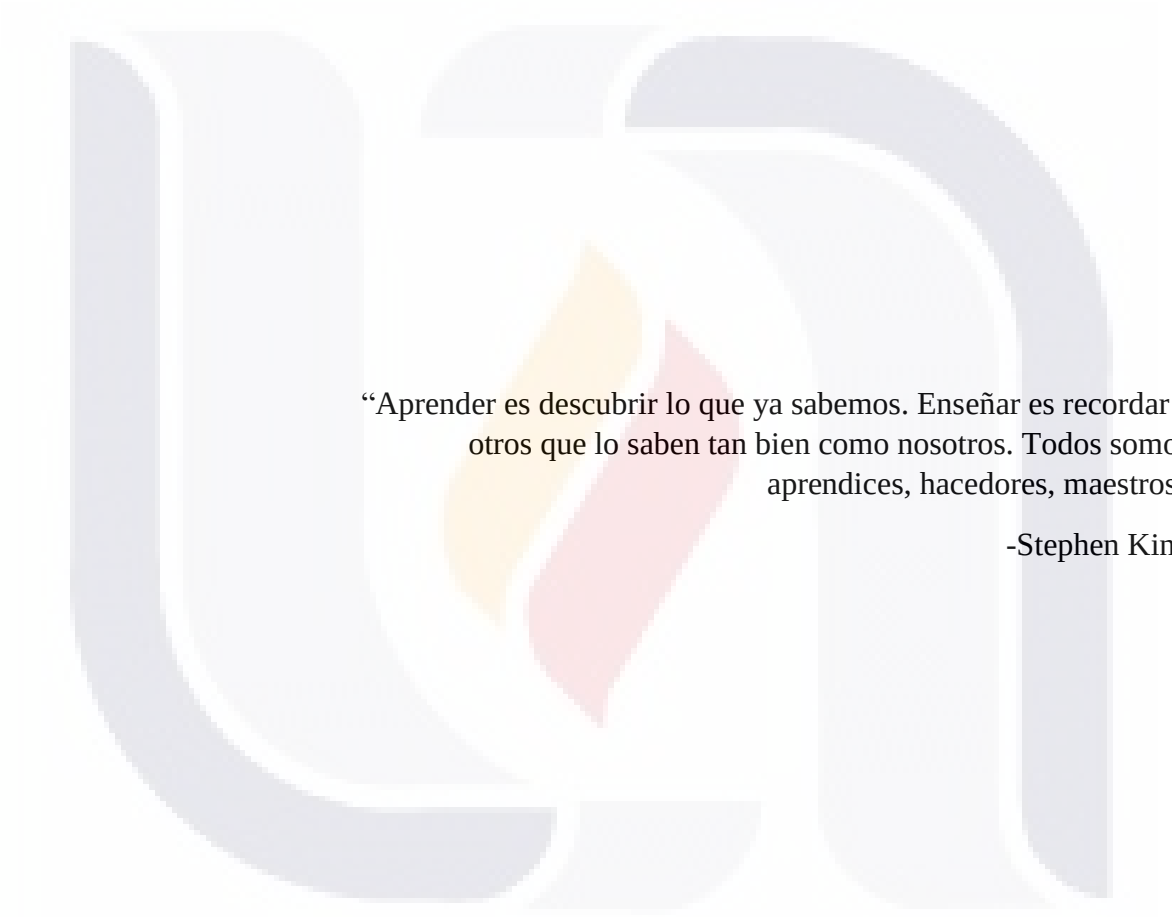
A Betty Morquecho y Martín Cuellar, agradezco infinitamente su guía y sus consejos.

A Conacyt y a la Universidad Autónoma de Aguascalientes por su apoyo a lo largo de este proyecto.

Dedicatoria

A mi abuelita María de Jesús, fuiste un ejemplo incansable de humildad y fortaleza. Te extraño mucho.

A mis ahijados Oscar e Israel, estoy muy orgulloso de ustedes.



“Aprender es descubrir lo que ya sabemos. Enseñar es recordar a otros que lo saben tan bien como nosotros. Todos somos aprendices, hacedores, maestros”

-Stephen King

Tabla de contenido

Índice de Tablas

Índice de Figuras

Índice de Ecuaciones

Resumen

En la actualidad, existen bases de datos cada vez más robustas como consecuencia de la digitalización de la información. Tal es el caso de la *medicina*, cuyos avances en el análisis, tratamiento y diagnóstico de patologías, proveen grandes cantidades de información. Por esta razón, la interacción de la *medicina* con las *ciencias computacionales* resulta en una novedosa perspectiva que reduce costos, tiempo y errores.

El presente documento describe la aplicación la *selección de subconjuntos de características* y la *teoría de testores* en patologías médicas, encontrando aquellas características que inciden de forma determinante en cada enfermedad. En concreto, se analizó una base de datos de células malignas y benignas de cáncer de mama, así como una base de datos de casos leves, moderados y graves de hemofilia. Para ello, se hizo uso de la *teoría de testores* que permite identificar conjuntos de *testores típicos*, los cuales representan la información mínima necesaria para distinguir objetos en sus respectivas clases.

La identificación de *testores típicos* es un problema exponencial respecto al número de características involucradas. Por esta razón, se hibridaron dos *metaheurísticas: el algoritmo genético (AG) y el algoritmo de estimación de la distribución (EDA)* por medio de la inclusión de un nuevo operador denominado como *operador de alteración*, el cual altera un porcentaje de la solución en búsqueda de mejores soluciones haciendo que la búsqueda en el espacio de soluciones sea más rápida.

La implementación de las *metaheurísticas* fue apoyada por la adaptación de modelos *arquitectónicos de software* y modelos de *patrones de diseño* propuestos por la *ingeniería de software*. Estos modelos proveen de diferentes niveles de abstracción de cada sistema permitiendo la identificación de sus componentes y sus interacciones para lograr su objetivo.

Al final de esta investigación se contó con dos metaheurísticas hibridas y afinadas para cada problema, es decir, se determinaron los valores de parámetros adecuados para mejorar su desempeño; además de la identificación de los testores típicos y su respectiva interpretación para cada una de las patologías.

Abstract

At the present, there are more and more robust databases as a result of the digitalization of information. Such is the case of *medicine*, whose advances in the analysis, treatment and diagnosis of pathologies, provide a large wealth of information. For this reason, the interaction of medicine with *computer sciences* results in a novel perspective that reduces costs, time and errors.

This document describes the application of the *Feature Subset Selección* and *Typical Testors Theory* to medical pathologists, finding those features that have a determining influence in each disease. Specifically, a database of malignant and benign breast cancer cells was analyzed, as well as a data base of mild, moderate and severe cases of hemophilia. To do this, the Typical Testors theory was applied, allowing the identification of sets of typical testors, which represent the minimun information needed to distinguish objects in their respective classes.

The identification of the typical testors is an exponential problem regarding the the number of the features involved. For this reason, two metaheuristics were hybridized: the genetic algorithm (GA) and the estimation of the distribution algorithm (EDA) by means of the inclusión of a new operator named as alteration operator, which modufies a percentage of the solution in search for better solutions making the search in the space of solutions faster.

The implementation of the metaheuristics was supported by the adaptation of architertural software and design patterns models proposed by software engineering. These models provide different levels of abstraction of each system allowing the identification of their components and their interactions to achieve its objective.

At the end of this investigation, there were two hybrid and refined metaheuristics for each problem, i.e., the appropiate parameter values were determined to improve their performance; in addition to the identification of typical testors and their respective interpretation for each pathology.

CAPÍTULO 1

Introducción

El presente documento de tesis aplica el enfoque *lógico-combinatorio* del *reconocimiento de patrones* a *patologías médicas*, como *cáncer de mama* y *hemofilia*, por medio de la *selección de subconjuntos de características (SSC)* y de la aplicación de la *teoría de testores* haciendo uso del *algoritmo exhaustivo* y algoritmos metaheurísticos hibridados: *algoritmo genético* y *algoritmo de estimación de la distribución*. Estos tres algoritmos o modelos se diseñaron aplicando herramientas de construcción de software, como las *arquitecturas de software* y los *patrones de diseño*, proporcionados por la *ingeniería de software* facilitando su implementación, evaluación y gestión. Estos algoritmos permitieron analizar un conjunto de características que describen las *patologías* ya mencionadas determinando cuáles son más importantes y apoyar así, el diagnóstico médico.

Este trabajo se compone de cuatro apartados más, de los cuales el apartado 2 aborda el marco teórico con los conceptos básicos que fueron necesarios para la realización de este documento. En el apartado 3 se describe la metodología que se siguió y las características de su aplicación. Por su parte, el apartado 4 expone los resultados obtenidos desde el punto de vista computacional y desde el punto de vista médico. Finalmente, las conclusiones se abordan en el apartado 5, de igual forma, desde el punto de vista de las ciencias computacionales y de la medicina.

Dentro de la *teoría de testores* se encuentran conceptos como *testor* y *testor típico*, los cuales serán descritos en el apartado 2. Sin embargo, ambos *son subconjuntos de características* de un problema dado y su diferencia radica en que el *testor típico* es un *testor irreducible*, es decir, es imposible prescindir de una característica sin que el subconjunto pierda su estado de *testor*. Estos subconjuntos son suficientes para describir objetos e identificarlos como miembro de una clase de un conjunto mayor de clases. El *algoritmo exhaustivo* para encontrar *testores* y *testores típicos* tiene una complejidad exponencial en base al número de características de dicho problema. Por esta razón se aplicaron *algoritmos metaheurísticos* como el *algoritmo genético* y el *algoritmo de estimación de la distribución* con el objetivo de encontrar *testores* y *testores típicos* con mayor eficiencia. Para ello se diseñaron hibridaciones a los algoritmos que apoyan en la exploración del espacio de soluciones y la obtención de *testores* con mayor eficiencia. En este documento se reporta la aplicación de estos algoritmos en el ámbito médico, en bases de datos que describen patologías.

La *medicina* es una de las áreas del conocimiento más beneficiadas con la interacción cercana con las *ciencias computacionales* y las *matemáticas* debido a que la digitalización de la información a producido el reto de lidiar con bases de datos cada vez más complejas[1]. Dicha complejidad de los datos hace que exista un área de oportunidad para el aprendizaje automático y la minería de datos, como campos de las ciencias computacionales[2].

De acuerdo con Lugo-Reyes et al. [1], el diagnóstico es una de las tareas más importantes en la medicina por lo que requiere de capacitación, reconocimiento de patrones, cálculo de probabilidad condicional y experiencia para el desarrollo de la intuición por parte del médico

para el análisis de datos que no pueden ser tomados en cuenta de manera aislada. Debido a la gran cantidad de información que puede involucrar la realización de un diagnóstico médico, se realizó un análisis de patologías médicas (cáncer de mama y hemofilia) con la intención de apoyar el proceso de diagnóstico por medio del enfoque lógico-combinatorio del reconocimiento de patrones.

1.1 Justificación

Los algoritmos para la selección de características suelen tener altas complejidades, tal como es el caso de la *búsqueda exhaustiva de testores típicos*. Este problema es *no polinómico (NP) de complejidad exponencial* (2^n), siendo *n* el número de características de problema a tratar[3]. Por este motivo se diseñaron, por medio de *herramientas de construcción de software*, operadores que hibridan las *metaheurísticas* poblacionales utilizadas: *algoritmo genético* y *algoritmo de estimación de la distribución* con el objetivo de facilitar su construcción e implementación.

El uso de *metaheurísticas* y sus *hibridaciones* tiene el objetivo de evitar la exploración exhaustiva del espacio de soluciones para encontrar un óptimo o el óptimo [4], siendo en este caso, el conjunto de *testores* de las patologías *cáncer de mama* y *hemofilia*.

De acuerdo con Ávila y Lugo-Reyes en [1, 5], la *medicina* ha tenido un rápido crecimiento gracias a la introducción de la *computación* y la digitalización de la *información*. Gracias a ello, existe un crecimiento en el volumen de información que se maneja haciendo, a su vez, que las bases de datos se vuelvan más complejas y robustas. Así pues, la interacción entre la computación y la medicina supone *"una novedosa perspectiva que reduce costos, tiempo, errores médicos y potencia el uso de recursos humanos en ramas médicos con mayores requerimientos"*, menciona María del Carmen Expósito en [6].

Una de las tareas más importantes de la *medicina* es el *diagnóstico*, cuyo error representa el error más común en el área cuando se considera atrasado, errado o incompleto [7]; siendo además, muy difíciles de identificar[8]. En [1] Lugo-Reyes menciona una estimación de 150

de cada mil pacientes son mal *diagnosticados*; mientras que Monestel y Samaha en [7] hacen referencia al Hardvard Medical Practice Study que estima que el error en el *diagnóstico* provocó el 17% de errores prevenibles en pacientes hospitalizados, además, menciona que este error es el más costoso al sumar 38.8 billones de dólares entre 1986 y 2010. Por otro lado, José Ceriani [9] reporta que los errores no son inferiores al 25% de los *diagnósticos* y, en adición, reporta un estudio en autopsias de 1966 a 2002 de los cuales el 23.5% fue la tasa media de muerte por posible error en el *diagnóstico*.

El error en el *diagnóstico médico* es parte de lo que se conoce como error médico que conforma un "*tema complejo, polémico y difícil de estudiar*"[10], además es el "más importante factor causal de eventos adversos o consecuencias indeseadas del proceso de la atención médica" especialmente cuando se trata de patologías que requieren de un *diagnóstico* oportuno y de un tratamiento prolongado como lo son el *cáncer* y la *hemofilia*.

El término *cáncer* designa alrededor de 200 entidades distintas [11] que contribuyen un serio problema debido a las altas incidencias y mortalidad en el mundo[12], además de problemas de orden psicológico, familiar, económico, entre otros[13]. Específicamente en este trabajo de tesis se trabajó con *cáncer de mama*, el cual, es el *tumor* más común a nivel mundial y la primera causa de muerte en mujeres que fallecen por *tumor maligno*. De acuerdo con el *Consenso Mexicano sobre diagnóstico y tratamiento del cáncer mamario* [14], se estiman alrededor de 1.67 millones de diagnósticos de mujeres con cáncer de mama y cerca de 522 mil pacientes fallecidos por esta enfermedad. Además, la *Organización Panamericana de Salud* (PAHO) estima 596 mil nuevos casos y 142,100 muertes en América para el año 2030, principalmente en Latinoamérica y el Caribe [15].

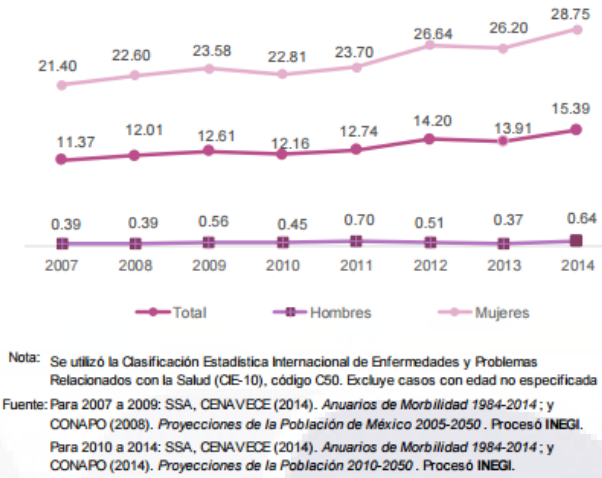


Figura 1 *Incidencia de tumor maligno de mama en población de 20 años o más de 2007 a 2014 por cada 100,000 habitantes de cada sexo [15]*

Como se puede observar en la Figura 1, se hace un análisis de la incidencia de cáncer de mama entre 2007 y 2014 que muestra que los hombres mantienen una incidencia baja y relativamente estable, pues por cada caso diagnosticado en hombres se detectan 29 en mujeres [16] cuya tendencia se mantiene al alza alcanzando 28.75 casos nuevos por cada 100 mil mujeres mayores a 20 años.

En México, el promedio de edad en el diagnóstico de cáncer de mama es de 53 años, la cual es casi una década menos comparado con Estados Unidos de América, Canadá y algunos países de Europa, donde el promedio está alrededor de los 60 años. Además, el riesgo para una mujer mexicana de padecer cáncer de mama durante su vida es del 2.9%, mientras que a nivel mundial es del 4.27% y finalmente, de 7.14% en países desarrollados [17].

La segunda patología tratada para este documento de tesis fue la *hemofilia*, la cual es un desorden genético, que afecta a uno de cada 10 mil varones nacidos vivos, en la que los pacientes tienen deficiencias en algún factor de coagulación ocasionando sangrados prolongados y hemorragias en órganos [18, 19]. De acuerdo con la World Federation of Hemophilia, presente en 134 países incluido México, se tenía registro para 2017 de 184,723 pacientes con *hemofilia* [20]. Específicamente en México, se tiene un registro de 6,022 casos al 22 de mayo de 2017 [19].

En la Figura 2 a continuación, se muestra el comportamiento a nivel global de la *hemofilia* de 1999 a 2016 obtenido del Annual Global Survey 2016 realizado por la World Federation of Hemophilia [20]. A pesar de dicho incremento, existe falta de conocimiento sobre la enfermedad y pocos pacientes reciben tratamiento adecuado [21], estimando que solo el 25% de las personas con *hemofilia* reciben un tratamiento adecuado en todo el mundo [22]. Con el *diagnóstico* oportuno los pacientes pueden llevar una vida normal, sin embargo, si no se trata correctamente se pueden presentar daños severos en articulaciones, discapacidad y muerte temprana [19].

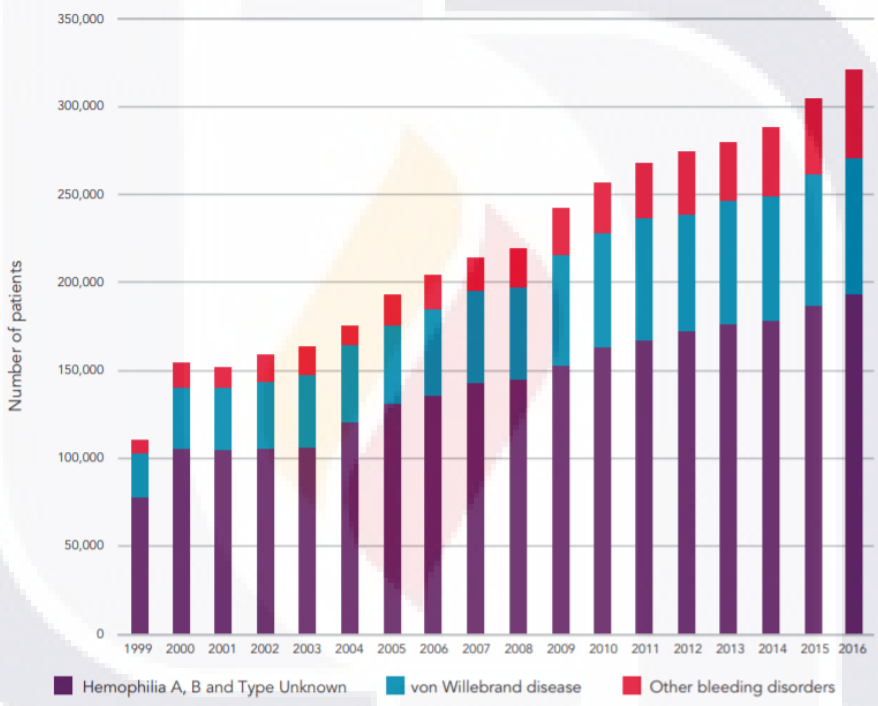


Figura 2 *Pacientes identificados a lo largo del tiempo por tipo de desorden en la sangre[20]*

Una vez identificada la importancia de un *diagnóstico* oportuno en medicina para enfermedades como el *cáncer* y la *hemofilia*, así como la gran cantidad de información a la que se puede tener acceso en un mundo digital; se procede a exponer en este documento de tesis la aplicación de la *teoría de testores* para la *selección de las características* que mejor describen a las patologías mencionadas por medio del método exhaustivo y dos metaheurísticas construidas por medio de herramientas de *ingeniería de software*.

1.2 Objetivos

1.2.1 Objetivo General

Identificar de factores de riesgo en cáncer de mama y hemofilia mediante el uso de métodos metaheurísticos híbridos de Selección de Subconjuntos de Características.

1.2.2 Objetivos Particulares

1. Programar un mecanismo de exhaustivo para encontrar el 100% de los testores típicos de las patologías: cáncer de mama y hemofilia.

2. Identificar factores en las patologías de cáncer de mama y hemofilia, así como su ponderación mediante el peso informacional.

3. Diseñar, implementar y ajustar una metaheurística tipos AG y una metaheurística tipo EDA.

4. Implementar algoritmos basándose en procesos de construcción de software.

5. Diseñar operador para la reducción del espacio de búsqueda en las metaheurísticas AG y EDA.

6. Validación médica de los factores encontrados en las patologías: cáncer de mama y hemofilia.

CAPÍTULO 2

Marco Teórico

2.1 Introducción

La realización de este documento de tesis involucró cuatro áreas del conocimiento: la *selección de características*, las *metaheurísticas*, *herramientas de construcción de software* y la *medicina* (ver Figura 3). En este apartado se describen cada uno de los conceptos fundamentales que respaldaron la realización de este documento de tesis como los son: el *reconocimiento de patrones*, la *selección de características* y las *metaheurísticas* como parte central del trabajo; *arquitecturas de software* y *patrones de diseño* utilizados para facilitar la implementación de los algoritmos *metaheurísticos*; y finalmente, se describen las patologías médicas utilizadas en este trabajo: el *cáncer de mama* y la *hemofilia*.

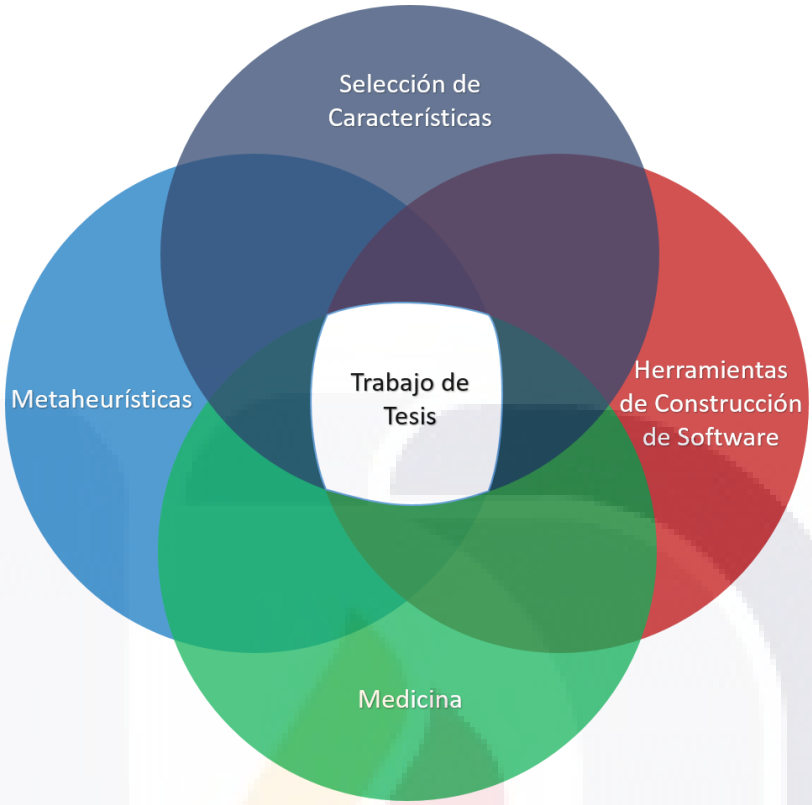


Figura 3 *Áreas involucradas*

2.2 Reconocimiento de Patrones

La *selección de características* forma parte del enfoque *lógico-combinatorio* del *reconocimiento de patrones*, razón por la cual se opta por comenzar a definirla y así, tener una visión de la posición en la que se trabajó para este documento de tesis.

A la fecha existen diferentes maneras de definir "*reconocimiento de patrones*" debido a que enmarcar una definición de una disciplina no es sencillo [23]. Sin embargo, Ruíz Shulcloper hace una aproximación de la siguiente manera:

"Zona del conocimiento (de caracter interdisciplinario) que se ocupa del desarrollo de teorías, métodos, técnicas y dispositivos para la realización de procesos ingenieriles, computacionales y/o matemáticos, relacionados con objetos físicos y/o abstractos, que tienen el propósito de extraer información que permita establecer propiedades y/o vínculos de entre

un conjunto de dichos objetos sobre la base de los cuales se realiza una tarea de identificación o clasificación" [23].

En otras palabras, es el área multidisciplinaria de la ciencia, específicamente en *inteligencia artificial*, que se ocupa de procesos en ingeniería, computación y matemáticas relacionados con objetos, cuyo propósito es extraer información que permita establecer propiedades de o entre conjuntos de dichos objetos [24]. Así pues, el objetivo del *reconocimiento de patrones* es la clasificación y la recuperación de características únicas que identifiquen un sujeto de la misma especie o clase [25].

Por sí mismo, el *reconocimiento de patrones* representa un desafío dentro de la *inteligencia artificial* debido al interés y alta demanda de aplicaciones especializadas en áreas como *minería de datos*, *biometría*, *herramientas de toma de decisiones*, entre otras.

2.2.1 Enfoques del Reconocimiento de Patrones

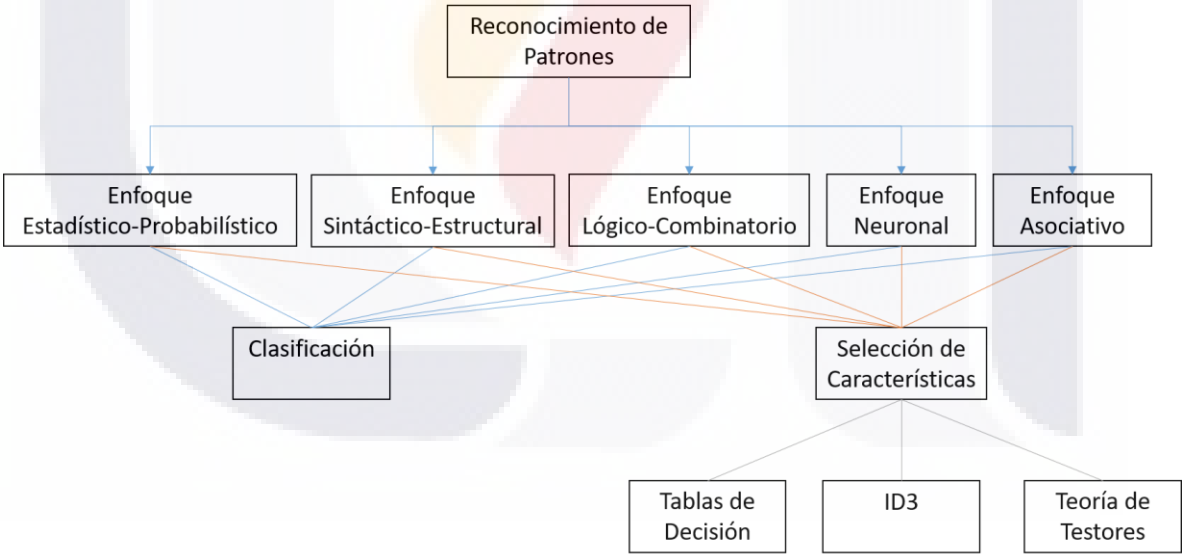


Figura 4 *Enfoques del Reconocimiento de Patrones / Ubicación de la Selección de Características [24, 26]*

En la actualidad, el *reconocimiento de patrones* es una de las líneas de investigación teórica y aplicada con alto auge que ha propiciado la existencia de diferentes enfoques entre los cuales se encuentra el *estadístico-probabilístico*, el *sintáctico estructural*, el *lógico-combinatorio*, el *neuronal* y el *asociativo* [24, 27]. Estos enfoques no son necesariamente independientes y los problemas atacados pueden ser interpretados desde diferentes enfoques, como se puede observar en la Figura 4. A continuación, se abordan brevemente el objetivo y las características de cada uno de los enfoques mencionados anteriormente descritos por Ariel Carrasco Ochoa y Francisco Martínez Trinidad en su artículo "Reconocimiento de Patrones" para la revista mexicana de inteligencia artificial Komputer Sapiens [24] :

- **Enfoque estadístico-probabilístico**
    - Basado en teoría de probabilidad y estadística.
    - Supone un conjunto de medidas numéricas con distribuciones de probabilidad conocidas o estimables, a partir de las cuales se hace el reconocimiento.
    - Históricamente, el primer enfoque abordado [28].
- **Enfoque sintáctico-estructural:**
    - Basado en teoría de Autómatas y lenguajes formales [28].
    - Estudio de objetos descritos como cadenas de símbolos, grafos, etc.
    - Enfocado en la búsqueda de relaciones estructurales de los objetos de estudio.
- **Redes Neuronales:**
    - Resolución de problemas por medio de modelo matemáticos de las redes neurológicas biológicas [28].
    - Estas redes se entrenan para dar cierta respuesta cuando se presentan determinados valores en sus entradas.
    - Una red neuronal puede dar una respuesta similar cuando recibe entradas parecidas a la usadas en su entrenamiento.
- **Enfoque asociativo[29]:**
    - Utiliza modelos de memorias asociativas para crear clasificadores robustos.
    - El propósito de una memoria asociativa es recuperar patrones completos a partir de patrones de entrada con posible ruido.
    - Creado en el Centro de Investigación en Computación del IPN en 2002.

- **Lógico-Combinatorio**
    - Alternativa a los enfoques anteriores.
    - Basado en la idea de que el modelo de un problema debe ser lo más cercano a la realidad del mismo.
    - Los atributos que describen a los objetos de estudio son tratados cuidadosamente para no resultar en operaciones antinaturales.
    - Acepta atributos cualitativos y cuantitativos e incluso con ausencia de información.

2.2.2 Problemas del Reconocimiento de Patrones

En este apartado, se busca describir las características de los problemas que el *reconocimiento de patrones* puede resolver por medio de sus diferentes enfoques entre los cuales se encuentra la s*elección de características*, que determina el conjunto más adecuado para describir objetos; la c*lasificación supervisada*, para la clasificación de objetos por medio de una muestra ya clasificada; y la c*lasificación no supervisada*, donde la clasificación no requiere información de clasificación previa de los objetos e incluso, sin definición de clases [24, 27].

- **Selección de Características [28, 30]:**
    - Determina el conjunto de características/variables más adecuando para describir objetos.
    - Por tanto, remueve variables poco relevantes (ruido o distractores).
    - Mejoramiento del desempeño de los clasificadores.
    - Representaciones estables de los objetos.
    - Facilidad de visualización y comprensión de datos.
- **Clasificación [24, 31]:**
    - Asignación de un objeto/fenómeno a una de las categorías o clases que se especifiquen dependiendo de las características que comparta con una clase definida.
    - Ampliamente utilizada en el *reconocimiento de patrones*.

- **Clasificación Supervisada [31]:**
  - Clasificación de objetos por medio de una muestra ya clasificada (conjunto de entrenamiento).
  - Compuesta de dos fases: *entrenamiento* y *clasificación*.
  - En la fase de entrenamiento se cuenta propiamente con el conjunto de entrenamiento y uno adicional para validación. Ambos constituyen un *modelo de clasificación*.
- **Clasificación no Supervisada [31]:**
  - No cuenta con conocimiento de clasificación previa.
  - Algunas veces, no hay definición previa de clases.
  - También conocida como *clustering*.
  - Se basa en el descubrimiento de grupos de objetos cuyas características similares de por resultado la separación en clases diferentes.

2.3 Selección de Subconjuntos de Características

Como ya se mencionó anteriormente, el problema de *reconocimiento de patrones* a abordar es la *selección de características* desde el enfoque *lógico-combinatorio* [32] (ver Figura 4). Dicho enfoque fue creado por la ex-Unión Soviética, destacando por ofrecer la posibilidad de trabajar con cualquier tipo de variable a pesar de que los algoritmos sean de alta complejidad [24, 26].

La *selección de características,* término usado habitualmente en *minería de datos* para describir herramientas y técnicas encargadas de *"la reducción eficiente del número de variables o características con las cuales se deben describir objetos y para encontrar las variables que inciden de forma determinante en un problema"* [33, 34] con el propósito de remover características poco relevantes consideradas como ruido o distractores para ahorro de tiempo y memoria. Además, mejora el rendimiento de un modelo clasificador al recibir representaciones estables, disminuyendo la posibilidad de sobre-entrenamiento facilitando la visualización y comprensión de datos [30].

De acuerdo con Microsoft [35], la *selección de características* es esencial para generar un análisis eficiente debido a que los conjuntos de datos suelen contener más información de la necesaria para la generación de modelos. Es decir, si un *modelo clasificador* se construye a partir de todo el conjunto de *n* columnas, dará como resultado la necesidad de mayor capacidad de procesamiento y memoria en el proceso de entrenamiento y en el almacenamiento del modelo completo [34]. Así pues, la *selección de características* debe su importancia a beneficios como la reducción de costos, eficiencia y precisión en clasificación o descripción de fenómenos con gran número de variables [4].

Existen diferentes métodos de *selección de características*, de los cuales, los convencionales evalúan diferentes combinaciones de *subconjuntos de características* usando algún índice y seleccionan el mejor de ellos. Este índice mide la capacidad del subconjunto en clasificación dependiendo del tipo (selección supervisada o no supervisada). La complejidad de este proceso crece cuando los conjuntos de datos son mayores, siendo de complejidad exponencial en base a la dimensión de los datos en búsquedas exhaustivas [36]. Para dar solución a este problema se han desarrollado algunas *heurísticas* como ramificación y poda, algoritmos Greedy, etc.

Más adelante en este documento, se analizarán algunos algoritmos creados para la *selección de características* por medio de la obtención del conjunto de *testores típicos*.

2.4 Teoría de Testores

Existen diferentes herramientas o metodologías para llevar a cabo la *selección de características* (Ver Figura 4), de las cuales, en este documento, se aborda la *teoría de testores* formulada como dirección científica de *Cibernética Matemática* en los años 60 en la *Unión de Repúblicas Socialistas Soviéticas (URSS)*, cuyo origen está vinculado con el uso de *lógica matemática* para localizar fallas en circuitos electrónicos que realizan funciones booleanas [32].

El enfoque de la *teoría de testores* a utilizar, es el de Dimitriev, Zhuravlev y Krendeleiev para problemas clásicos del *reconocimiento de patrones* (clasificación con aprendizaje y selección de características) creado en 1965 donde *"las clases son conjuntos disjuntos, el criterio de comparación entre rasgos es booleano y el criterio de semejanza entre objetos asume que dos objetos son diferentes si al menos uno de sus rasgos también lo es"* [37].

Uno de los conceptos que define ésta teoría es el de *testor*, el cual es un *subconjunto de características* que distingue objetos de diferentes clases [3], es decir, ningún objeto perteneciente a una clase T_0 se confunde con algún objeto de la clase T_1 mirados a través de los valores en sus características [38].

De acuerdo a Shulcloper et al. en [32], la definición de *testor* se puede extender a más de dos clases como lo hace Zhuravlev, de manera que, dado un subconjunto de columnas $\tau = \{i_1, \ldots, i_s\}$ de columnas de la tabla T y sus rasgos ($\alpha_{i1}, \ldots, \alpha_{is}$), se considera testor para $(T_0, \ldots, T_r)=T$, si después de eliminar de T todas las columnas de τ no existe alguna fila en T_0 igual a una en cualquiera del resto de las r clases en T.

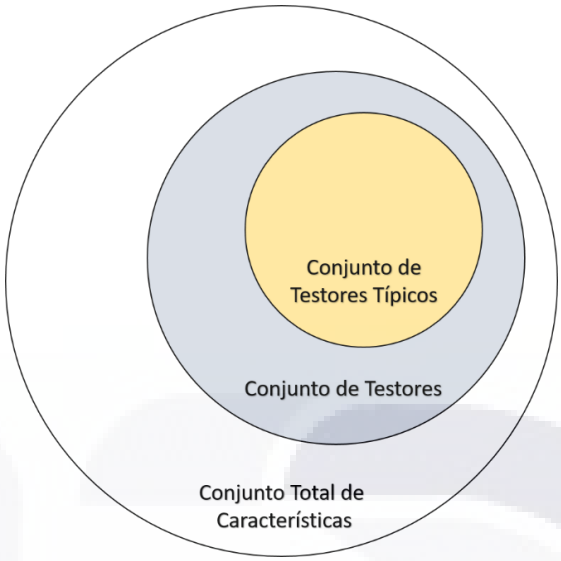


Figura 5 *Conjuntos de Características*

Dentro del conjunto de *testores* se encuentran los llamados *testores típicos* o *testores irreductibles* (Ver Figura 5), los cuales son *testores* que si perdieran cualquiera de sus características perderían a su vez el estado de *testor* [32, 33]. Por tanto, un *testor típico* es el subconjunto de características mínimo necesario para diferenciar objetos de diferentes clases [3].

Así pues, la importancia del *conjunto de testores* radica propiamente en la reducción del espacio de representación de objetos (*selección de características*), como apoyo en la evaluación de peso informacional de características y como apoyo a sistemas de clasificación [33].

En el apartado 2.4.1 se explican conceptos básicos para entender el proceso de obtención del conjunto de *testores típicos* tanto en el *modelo exhaustivo* como en los *modelos metaheurísticos* generados a lo largo del trabajo de tesis.

2.4.1 Conceptos Básicos

El proceso para la obtención de *testores típicos* comienza con una *matriz de aprendizaje (MA)* que contiene las descripciones del conjunto de objetos $\Omega = \{O_1, O_2, \ldots, O_m\}$ en términos de sus *n* rasgos $\Re = \{X_1, X_2, \ldots, X_n\}$ divididos en *r* clases y un conjunto de *criterios de comparación*, donde cada criterio C_k se asocia a un rasgo X_k del conjunto $\Re$ [33]. Como ejemplo de criterios se tienen, el *criterio de comparación con igualdad estricta* y el *criterio de comparación con error admisible* cuyas representaciones se ilustran a continuación:

1) $$C_k\left(X_k(O_i), X_k(O_j)\right) = \begin{cases} 0 \text{ si } X_k(O_i) = X_k(O_j) \\ 1 \text{ e. o. c} \end{cases}$$ "Criterio de comparación de igualdad estricta"

2) $$C_k\left(X_k(O_i), X_k(O_j)\right) = \begin{cases} 0 \text{ si } \left|X_k(O_i) - X_k(O_j)\right| \leq \varepsilon \\ 1 \text{ e. o. c} \end{cases}$$ "Criterio de comparación con error admisible"

A partir de una *MA* y un *conjunto de criterios de comparación* $\{C_1, C_2, \ldots, C_n\}$, se obtiene una *matriz de diferencias (MD)* realizando comparaciones booleanas dos a dos entre las descripciones de objetos que pertenecen a distintas clases [32, 33]. [32, 33]. La *MD* es un conjunto binario, que muestra la igualdad (0) o diferencia (1) característica a característica a partir de un *criterio de comparación* entre pares de objetos pertenecientes a clases distintas, compuesto por *m'* filas definidas de la siguiente manera [32]: suponiendo que existen *r* clases $(m_1, m_2, \ldots m_r)$, donde cada m_i representa el número de objetos pertenecientes a la clase *i, la MD tendrá:*

$$m' = m_1(m_2 + \cdots + m_r) + m_2(m_3 + \cdots + m_r) + \cdots + (m_{r-1})m_r \text{ renglones}$$

Ecuación 1

El siguiente concepto es la *matriz básica (MB)* que, de manera sencilla, contiene las filas básicas de la *MD* [33]. La *MB* sintetiza la información contenida en la *MD* eliminando información redundante eliminando renglones que son superconjuntos de los demás. Formalmente, suponiendo i_p e i_t filas de una MD, se dice que i_p es subfila de i_t (i_t es superfila de i_p) si y solo si [32]:

a) $$\forall(a_{i_t j} = 0 \Rightarrow a_{i_p j} = 0)$$

b) $$\exists j_o \left(a_{i_t j_o} = 1 \wedge a_{i_p j_o} = 0\right)$$

Esto es i_t, fila de *MD*, es básica si y solo si en la *MD* no existe fila i_p alguna que sea subfila de i_t. Así pues, como ya se mencionó, la *MB* se compone de las filas básicas de la *MB* excluyendo filas repetidas.

Tanto la *MA*, como la *MD* y la *MB* aportan la misma información desde el punto de vista de la diferenciación de objetos con la diferencia de que a cada una se le da una representación más compacta a medida que se van obteniendo.

El siguiente elemento a considerar en el proceso es el c*onjunto potencia (CP)* utilizado por *algoritmos de escala exterior* [39]. Este conjunto representa los potenciales *testores*, que conforma una matriz de 2^n filas de cadenas binarias de n elementos, donde n es la cantidad de características que supone el problema analizado.

Cada fila del *CP* representa un subconjunto diferente de rasgos $\tau = \{X_{i1}, \ldots, X_{is}\}$ de una *MA* que puede ser *testor* (Ver Figura 6) si al eliminar de su *MB* todas las columnas, excepto aquellas que corresponden a los elementos de τ, no existen filas completas de ceros [33].

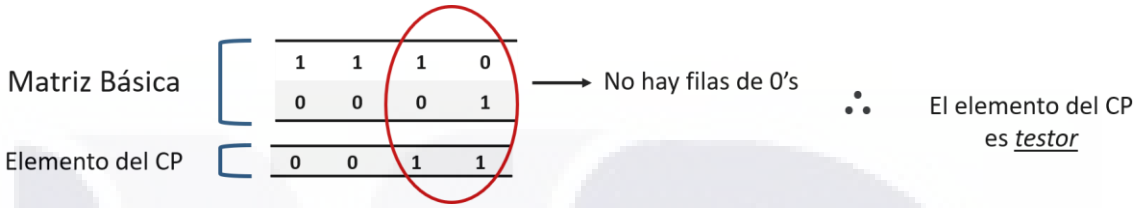


Figura 6 *Obtención de testores a partir del conjunto potencia [33]*

Una vez obtenido el *conjunto de testores típicos* se puede computar el *peso informacional*, el cual, es una puntuación determinada a cada característica a partir de dicho conjunto de testores típicos. Esta puntuación es obtenida a partir del cálculo de la frecuencia relativa de cada una de las características como se puede observar en la Ecuación 2. Sea τ el número de testores típicos obtenidos en el problema y $\tau(i)$ el número de testores típicos en el que aparece la característica x_i, el *peso informacional (P)* de x_i estará dado por [40]:

$$P(x_i) = \frac{\tau(i)}{\tau} \qquad \textit{Ecuación 2}$$

Para i=1, …, n, para $x_i \in R$

Ésta puntuación representa una medida de significancia para cada característica para el proceso de clasificación [41]. Es decir, si una variable o característica aparece varias veces en diferentes testores típicos es más complicado descartarla, por tanto, es más importante para diferenciar clases [40]. Finalmente, el uso de *peso informacional* es una buena herramienta para la *selección de características* que muestra resultados tangibles de la obtención de *testores típicos*.

2.4.2 Algoritmos para la Búsqueda de Testores Típicos

En la literatura existen diferentes algoritmos para el cálculo de *testores típicos* que de acuerdo con la estrategia utilizada pueden clasificarse como algoritmos de *escala exterior* y *escala interior*. Donde los *algoritmos de escala exterior* hacen uso de un *conjunto potencia* del conjunto de columnas de la *MB* en orden determinado. Por otra parte, *los algoritmos de escala interior* realizan el cálculo de *testores* basándose en la estructura interna la matriz, buscando condiciones que garanticen el estado de *testor* en las columnas asociadas a las posiciones unitarias de la matriz [33]. A continuación, en la Tabla1 se muestran algunos de los algoritmos existentes para el cálculo de *testores típicos*.

Algoritmo	Escala	Características	Desventajas
BT	Exterior	Su orden es el generado por los números naturales en forma ascendente en notación binaria mediante un n-uplo booleano.	Genera vectores que son supraconjuntos de testores típicos que son evaluados innecesariamente.
		Este orden constituye todos los elementos no nulos del *CP* de características/variables de la *MA*.	La comprobación de cada vector se realiza con toda la *MB* como si no se hubiese hecho en iteraciones anteriores.
		Prueba cada vector del conjunto es *testor* o *testor típico*.	
		El algoritmo termina al llegar al n-uplo (1…11).	
CT	Interior	Basado en conceptos de *conjunto de filas independientes* y *conjunto completo*.	Encuentra *testores típicos repetidos*. Un mismo test se puede originar a partir de más de un par de *conjunto*

		Conjunto completo → testor	*completo-conjunto de filas independientes*.
		Primero encuentra todos los *testores* y después comprueba si son *típicos* o no.	
		Explora todas las posiciones unitarias de la *MB* encontrando conjuntos completos.	
REC	Exterior	Trabaja directamente con la MA.	Enorme manejo de información.
CER	Exterior	Creado como mejora del algoritmo REC.	
		Introduce un nuevo orden de CP.	
LEX	Exterior	Implanta otra alternativa de orden del *CP*.	No asegura el salto de algún n-uplo cuando la *MB* es compleja.
		Define saltos para obviar algunos conjuntos.	
		Los vectores construidos deben poseer la propiedad de tipicidad.	

Tabla 1 *Algoritmos para el cálculo de testores típicos [33]*

Como se puede observar en la Tabla 1, los *algoritmos de escala exterior* hacen uso del *conjunto potencia* y analizan cada uno de sus vectores con la *matriz básica* para después determinar si dicho vector es un *testor* o un *testor típicos*. La diferencia en entre estos algoritmos es el orden del *conjunto potencia*, el cual es considerado primordial en el desempeño de los algoritmos. A continuación, en la Tabla 2 se presenta un ejemplo utilizado por Santiesteban y Pons [33] para mostrar el ordenamiento sobre P(ℜ) de cada uno de los *algoritmos de escala exterior*. Este ejemplo representa las características por sus índices en ℜ y su representación binaria.

REC		CER		BT		LEX	
$X \subseteq \Re$	Bin	$X \subseteq \Re$	Bin	$X \subseteq \Re$	Bin	$X \subseteq \Re$	Bin
123	111	∅	000	∅	000	∅	000
12	110	1	100	3	001	1	100
13	101	2	010	2	010	12	110
1	100	3	001	23	011	123	111
23	011	12	110	1	100	13	101
2	010	13	101	13	101	2	010
3	001	23	011	12	110	23	011
∅	000	123	111	123	111	3	001

Tabla 2 *Ordenamiento de conjunto potencia para algoritmos de escala exterior*

2.5 Heurísticas y Metaheurísticas

Como ya se pudo observar en el apartado 2.3, los algoritmos para la *selección de características* suelen tener altas complejidades, específicamente la *búsqueda exhaustiva de testores típicos* en un problema *no polinómico (NP)* de *complejidad exponencial* (2^n) [3], por tanto, se aborda el tema de *metaheurísticas* que evitan la exploración exhaustiva en todo el espacio de soluciones y obtener resultados cercanos al óptimo o el óptimo mismo [4]. Primeramente, se expone el término *heurística* como antesala a lo que son las *metaheurísticas* y su beneficio para la solución de problemas.

2.5.1 Heurísticas

Al igual que la *búsqueda de testores típicos*, existen otros problemas de interés en la actualidad que no tienen un algoritmo exacto con complejidad polinómica que encuentre solución óptima con espacio de búsqueda enorme y con un tiempo aceptable para encontrar una solución, por lo que se cuenta con métodos de aproximación o *heurísticas* que permiten obtener soluciones en tiempo razonable [42] sin llevar a cabo un análisis riguroso del problema. El término *heurística* se remonta a Arquímedes y es definido *"Técnica de la indagación y del descubrimiento"* mientras que Zanakis et al [43] definen *heurística*, desde el punto de vista *computo-matemático*, como un *"procedimiento simple, a menudo basado en sentido común, que supone la obtención de una buena solución no necesariamente óptima a problemas difíciles de modo sencillo y rápido"*.

Las motivaciones principales para la aplicación de *heurísticas* destacan las siguientes [44]:

- No existe un método exacto para encontrar una solución.
- Excesivo uso de tiempo y/o memoria.
- No hay requerimiento de un óptimo global, un óptimo local es suficiente.
- Baja fiabilidad de datos, solo se requiere un resultado aproximado.

De acuerdo con Duarte [42], la principal desventaja de la aplicación de *heurísticas* es su incapacidad de escapar de óptimos locales, posiblemente de baja calidad, debido a que no cuentan con algún mecanismo que les permita continuar con la búsqueda aún después de encontrar un óptimo local. Para ello, las investigaciones en las últimas décadas han diseñado técnicas aún más inteligentes que evitan este problema de manera que son procedimientos de alto nivel que guían *métodos heurísticos* para evitar estancamientos en óptimos locales [45]. Estos métodos son conocidos como *metaheurísticas* cuyo concepto se profundiza un poco más en el apartado siguiente.

2.5.2 Metaheurística

Las *metaheurísticas*, conocidas en algunas fuentes como *meta-heurísticas* o *heurística moderna* [46], es un término acuñado a Fred Glover en 1986. Por su artículo sobre *búsqueda tabú*, refiriéndose a *heurísticas* de nivel superior (prefijo griego "*meta*", "*más allá*", "*de alto nivel*"). Así pues, de acuerdo con Duarte [42], Glover definía el término como "*procedimientos de alto nivel que guían y modifican otras heurísticas para explorar soluciones más allá del simple óptimo local"*.

Desde entonces se han creado gran cantidad de propuestas de diseño de procedimientos para la solución de problemas por lo que las *metaheurísticas* pueden ser específicas a un problema, para mejorar el rendimiento con mayor especialización; o generales con el objetivo de ser sencillas, adaptables y robustas [44].

La lógica general de las *metaheurísticas* [47] es similar en cada una: tienen como punto de partida una solución, o un conjunto de ellas, que normalmente no es óptima. A partir de esta solución se obtienen varias similares, de entre las cuales se elige la que cumpla con algún criterio definido para comenzar el proceso con la nueva solución. La metahuerística se detiene bajo un criterio de paro definido previamente.

Según Rodríguez-Piñero en [48] todas las *metaheurísticas* cumplen con las siguientes características:

- Ciegas: No saben cuándo encontraron la solución óptima por lo que necesitan de una indicación para detenerse. Al igual que las *heurísticas*, no se asegura una respuesta óptima.
- Algoritmos aproximados: Como ya se mencionó, no garantizan la solución óptima.
- Aceptan malos movimientos: Ocasionalmente aceptan una solución mala como paso intermedio para acceder a regiones no exploradas del espacio de soluciones.
- Relativamente sencillos: Tiene como requerimiento una representación del espacio de soluciones, una solución inicial/conjunto de soluciones inicial y un mecanismo para explorar el espacio de soluciones.

- Son generales: Son aplicables a cualquier problema de optimización combinatoria. Suelen ser mejores si las operaciones tienen más relación con el problema considerado.
- La regla de selección depende del instante del proceso y del historial de dicho momento: Suponiendo que dos iteraciones obtienen la misma solución, la siguiente no necesariamente obtendrá esa solución.

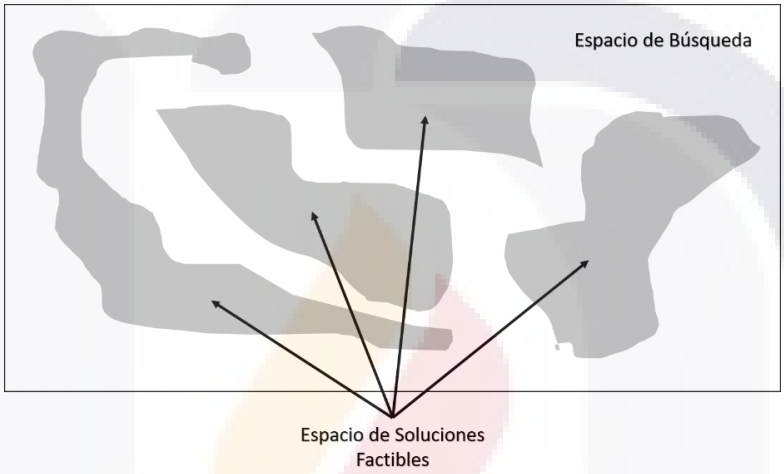


Figura 7 *Espacio de búsqueda y sub espacio de soluciones factibles [42]*

Como en cualquier resolución algorítmica de casi cualquier problema de optimización combinatoria es necesario especificar tres conceptos básicos [42]:

- Representación: Codifica las soluciones factibles para su manipulación determinando el tamaño del espacio de búsqueda del problema (cardinalidad). Ver Figura 4.
- Objetivo: El propósito que debe tener una representación dada. Predicado matemático que expresa la tara a realizar.
- Función de evaluación: Asocia a cada solución factible un valor que determine su calidad. Correspondencia f entre los puntos del espacio de soluciones y $\mathbb{R}$.

Las *metaheurísticas* cuentan con mecanismos para alejarse de óptimos locales orientando su búsqueda dependiendo de la evolución del proceso [47] incorporando conceptos de diversas áreas del conocimiento como la *genética*, la *biología*, la int*eligencia artificial*, *matemáticas*, *física*, etc.[45]. Entre las *metaheurísticas* más conocidas se tienen [42, 45]

- Optimización por Colonia de Hormigas (ACO).
- Algoritmos Evolutivos (EA):
  - Algoritmos Genéticos (GA)
  - Algoritmos Meméticos (MA)
- Procedimientos de Búsqueda Miope, Aleatorizados y Adaptativos (GRASP)
- Búsqueda Local Iterativa (ILS)
- Búsqueda Tabú (TS)
- Re-encadenamiento de trayectorias (PR)
- Recocido Simulado (SA)
- Búsqueda Dispersa (SS)

Para este trabajo se decidió usar el *Algoritmo Genético (GA)* y el *Algoritmo de Estimación de la Distribución (EDA)* con el fin de dar solución a la *búsqueda de testores típicos* para la *selección de características.* En los apartados 2.5.3 y 2.5.4 se profundiza en la teoría base acerca de dichos algoritmos para después presentar su implementación específica a lo largo del documento de tesis.

2.5.3 Algoritmo Genético

El primer algoritmo con el que se trabajó es el *Algoritmo Genético (AG),* el cual es una *metaheurística* basada en *poblaciones* debido a que emplea un conjunto de soluciones (población) en cada iteración. Éste tipo de *metaheurísticas* proporciona un mecanismo de exploración paralelo del espacio de soluciones y su eficiencia depende de cómo se manipule dicha población [42].

A su vez, el *AG* pertenece a *los algoritmos evolutivos (EA),* los cuales se basan en la idea neo-darwiniana de la evolución de las especies en la cual se establece que los individuos mejor adaptados tienen más probabilidad de sobrevivir, tener descendencia y heredar sus características [42, 49].

A lo largo de la historia, han existido diferentes trabajos sobre los *algoritmos genéticos,* pero es John Holland (1975) quien es considerado el creador de dichos algoritmos, cuyo objetivo era el estudio formal de la adaptación natural en sistemas computacionales preguntándose la forma en que la naturaleza creaba seres cada vez más perfectos [48, 50].

Por medio de ese estudio, Holland logra abstraer la evolución biológica y proporciona las bases para la adaptación del Algoritmo Genético bajo la nomenclatura de "*población*", para nombrar el conjunto de soluciones, y la definición de operadores de selección, cruce y mutación.

De acuerdo con Goldberg [51], discípulo de Holland, una definición formal para un Algoritmo Genético (AG) sería: *"Algoritmo de búsqueda basado en mecanismos de selección natural y genética natural. Combina la supervivencia de los más compatibles entre las estructuras de cadenas, con una estructura de información ya aleatorizada, intercambiada para construir un algoritmo de búsqueda con algunas de las capacidades de innovación de la búsqueda humana."*

El *AG* debe su proliferación a su robustez que le permite abordar problemas en diferentes áreas del conocimiento. A pesar de no garantizar soluciones óptimas, los *AGs* tienen la certeza empírica de que ofrece soluciones de alto nivel en tiempo competitivo. Los *AGs* no son la única solución a problemas, pues existen métodos más concretos y efectivos, pero permiten la hibridación con técnicas específicas para mejorar su desempeño [52].

En base a las características establecidas, en el siguiente subapartado, se explicará más detalladamente el funcionamiento del llamado *Algoritmo Genético Simple o Canónico* por medio de los operadores ya mencionados.

Algoritmo Genético Simple

Como se dijo anteriormente, el AG es un *algoritmo evolutivo (AE)* y por tanto, de acuerdo con Duarte [42], debe constar de los siguientes elementos:

- **Población:** Conjunto de *soluciones* candidatas de un problema dado. Cada uno de sus elementos se conoce como *individuo (Cromosoma).*
- **Selección:** Mecanismo sesgado de selección de individuos en el que sean más probable seleccionar a los mejores para transmitir sus características a la siguiente población.
- **Alteración:** Mecanismo para generar nuevos individuos mediante modificación estocástica de los individuos anteriores. Según el número de individuos, éste elemento puede ser unaria (mutación) o de orden superior (cruce).

La Figura 8 muestra en forma de pseudocódigo el funcionamiento de un a*lgoritmo genético simple*, el cual parte de una población inicial evaluada y continúa generando nuevas poblaciones hasta que se cumpla la *convergencia* o una *condición de parada* determinada.

```
BEGIN /* Algoritmo Genético Simple */
    Generar una población inicial.
    Computar la función de evaluación de cada individuo
    WHILE NOT Terminado DO
    BEGIN /* Producir nueva generación */
        FOR Tamaño población/2 DO
        BEGIN /*Ciclo Reproductivo */
            Seleccionar dos individuos de la anterior generación, para el cruce
            (probabilidad de selección proporcional a la función de evaluación del
            individuo).
            Cruzar con cierta probabilidad los dos individuos obteniendo dos
            descendientes
            Mutar los dos descendientes con cierta probabilidad.
            Computar la función de evaluación de los dos descendientes mutados.
            Insertar los dos descendientes mutados en la nueva generación.
        END
        IF la población ha convergido THEN
            Terminado := TRUE
    END
END
```

Figura 8 *Algoritmo Genético Simple [51]*

El algoritmo hereda las características de un *AE* y hace uso de sus propias características que lo hacen singular y con entidad propia como *metaheurística* [42].

Para la construcción de un AG es necesaria una r*epresentación* (propiedad de una metaheurística, ver apartado 2.5.2), la cual constituye una correspondencia entre soluciones factibles (fenotipo) y la codificación de variables (genotipo). Originalmente, la *representación* usada consiste en cadenas binarias, pero pueden utilizarse otras representaciones dependiendo del problema [42]. En la Figura 9 se muestra un ejemplo sencillo de representación binaria [53] de una solución para el problema *OneMax* que busca el mayor número de 1’s en una cadena.

Una vez definida la *representación*, se puede comenzar el algoritmo. El *AG simple* comienza con la generación de una *población inicial* de individuos, la cual consiste en un conjunto de *soluciones candidatas* que representan un punto en el espacio de búsqueda [48] creadas, generalmente, de forma *aleatoria* [42, 52]. Cada uno de los individuos, se define

como *cromosoma* compuesto por un conjunto de *genes* que representan las *variables* del problema [52].

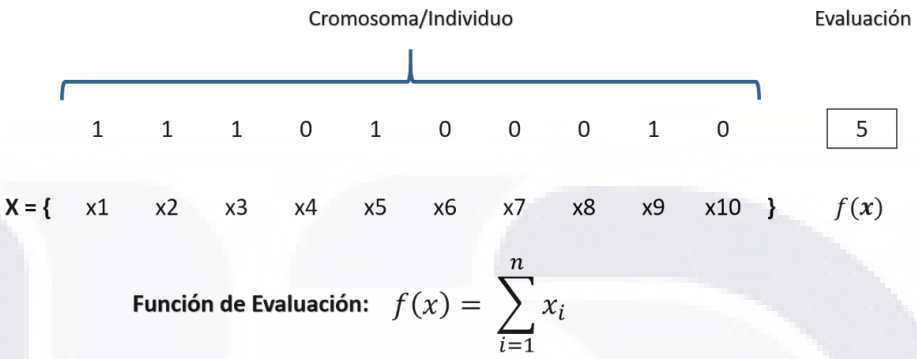


Figura 9 *Ejemplo de representación y evaluación [53]*

Una vez generada la *población inicial*, se evalúa cada individuo por medio de una *función de evaluación*, también conocida como *función de adaptación* diseñada específicamente para el problema dado que determina la calidad de cada solución por medio de la asignación de un valor que crece según mejor sea el individuo [42, 52, 54].

Una vez evaluada la población, se somete a un proceso de s*elección* que permite elegir pseudo-aleatoriamente individuos de la población. Dicho proceso suele favorecer a los mejor calificados por la *función de evaluación*, es decir, los más fuertes o mejor adaptados. Existen diferentes mecanismos para realizar *selección*, siendo el más popular el de *la ruleta*, el cual se observa en la Figura 10, que asigna probabilidad proporcionalmente a la evaluación obtenida. Otros mecanismos son *Ranking* y *Selección por Torneo* [42, 54, 55].

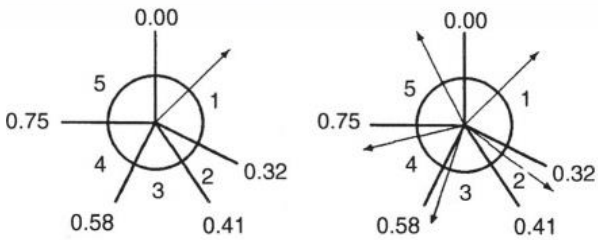


Figura 10 *Ejemplo de ruleta para selección de individuos [55]*

Lo siguiente en el algoritmo, es la aplicación de *operadores genéticos*, los cuales son métodos probabilísticos que obtienen nuevos individuos [42]. El primer operador es el *cruce*, que consiste en reemplazar algunos genes en una parte de un padre por los genes correspondientes a otro (Ver Figura 11) dando lugar a nuevos individuos diferentes a sus padres, pero con algunas similitudes [55]. Este tipo de *cruce* es conocido como *cruce basado en un punto* del que se esperan dos descendientes por cada cruce. Normalmente el *cruce* no es aplicado a todos los elementos seleccionados, sino que se decide aleatoriamente con una probabilidad entre 0.5 y 1.0 [42].

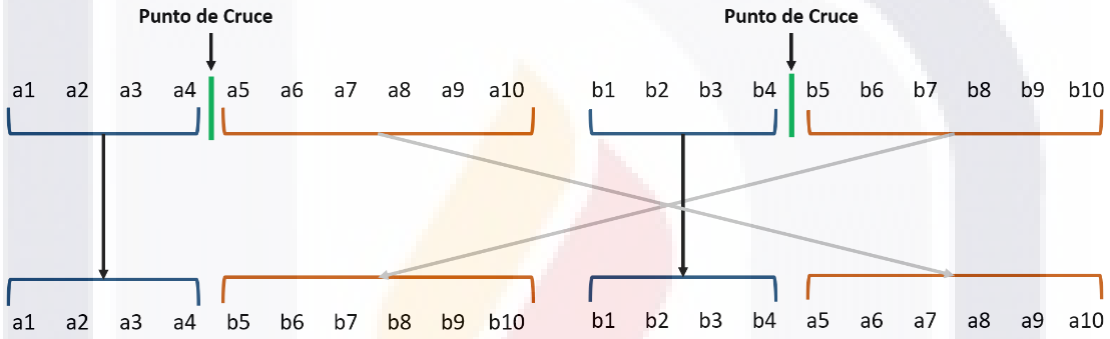


Figura 11 *Cruce basado en un punto [54, 55]*

El segundo *operador genético* es la *mutación*, que funciona como complemento al *cruce* implementando cambios inesperados en los individuos hijos [52]. Este operador se aplica a cada individuo normalmente con una probabilidad muy pequeña de que ocurra [48]. El caso más sencillo es seleccionar un elemento del individuo y cambiar su valor por uno contrario, en caso de ser cadenas binarias cambiar 0 por 1 y viceversa, tal como se observa en la Figura 9. Otra alternativa es la *permutación* en la que se seleccionan dos elementos del individuo y se intercambian sus posiciones [52]. El objetivo primordial de este operador es preservar la diversidad y explorar nuevas zonas del espacio de soluciones [42].

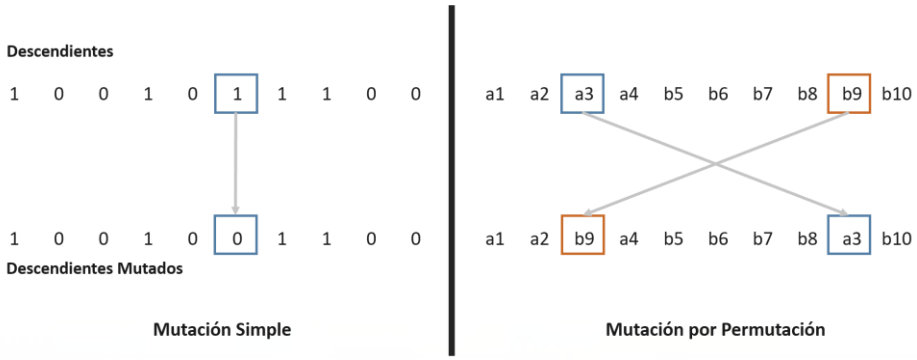


Figura 12 *Alternativas de mutación [52, 54]*

Al finalizar el *cruce* y la *mutación* se contará con nuevos individuos que son sometidos al *operador de evaluación* para conocer su nivel de adaptación y luego ser insertados en una *población descendiente* o la *nueva generación de soluciones*. Una vez obtenida la nueva población, el algoritmo itera de nuevo hasta cumplir un *criterio de parada*.

El *criterio* más simple, es la definición de un número máximo de iteraciones cuya desventaja es el posible encuentro del óptimo antes de cumplir el número de iteraciones [52]. Otra alternativa es limitar el tiempo de resolución, aunque los métodos más eficientes son los relacionados con los indicadores del estado de evolución de la población [56]. Por ejemplo, el *criterio de paro* relacionado con la progresión a la uniformidad introducida por De Jong, establece que una variable ha convergido cuando al menos el 95% de la población comparte el mismo valor para la misma variable [57]. Por tanto, el criterio es detener el *AG* cuando un determinado número de variables haya convergido [52]. Finalmente, la recomendación de la literatura consultada es combinar dos *criterios de parada* ya que la convergencia puede llevar muchas generaciones y mucho tiempo de ejecución [52, 54, 56].

El *AG Simple*, es una visión general de un AG, por tanto, es posible agregar gran variedad de estrategias distintas en cada uno de los operadores mencionados e incluso combinarlas. Es decir, el AG es muy flexible para implementar múltiples configuraciones dependiendo del problema a solucionar, hasta elementos de otras metodologías [56].

2.5.4 Algoritmo de Estimación de la Distribución

La segunda *metaheurística* con la que se trabajó en este documento de tesis, es el *Algoritmo de Estimación de la Distribución (EDA)*, el cual, al igual que el *AG*, forma parte de los *algoritmos evolutivos (EA)* y, a su vez, se trata de una *metaheurística* basada en poblaciones, usando una colección de soluciones candidatas para explorar trayectorias de búsqueda evitando los óptimos locales [42]. La *metaheurística EDA* fue propuesta por Mühlenvein y Paasss en 1996, aunque existen trabajos previos como el de Holland que ya consideraba esta alternativa para incorporarla en el contexto del AG.

El EDA fue motivado por su número reducido de parámetros asociados [42] en comparación al AG que depende de las probabilidades de *cruce* y *mutación*, tamaño de la población, tasa de reproducción, el número de generaciones, etc., por lo que requiere de buena inversión de tiempo para determinar los valores adecuados al problema, además de que el procedimiento de *cruce* puede destruir agrupaciones de componentes de alta calidad [42, 58].

Como se puede observar, los *EDAs* pueden verse como una alternativa o una generalización a los *AGs* [42] que pueden identificar las interrelaciones entre las variables involucradas, y evitar la necesidad de ajustar un gran número de parámetros [59]. Es decir, los *EDAs* muestrean de una *distribución de probabilidad* estimada de los datos contendidos en un conjunto de individuos seleccionados de la generación anterior [58].

Por tanto, el *AG* y en *EDA* representan dos enfoques diferentes, ver Figura 13. Los *AG*s representan el primer enfoque, que se basa en la manipulación de la representación de las soluciones previniendo trastornos en las variables que interactúan. Mientras que los *EDAs*, se basan en cambiar el proceso de variación aprendiendo y aprovechando las interacciones entre variables mediante *estimación de distribución de la población* con la que se realiza un *muestro de descendientes* [53].

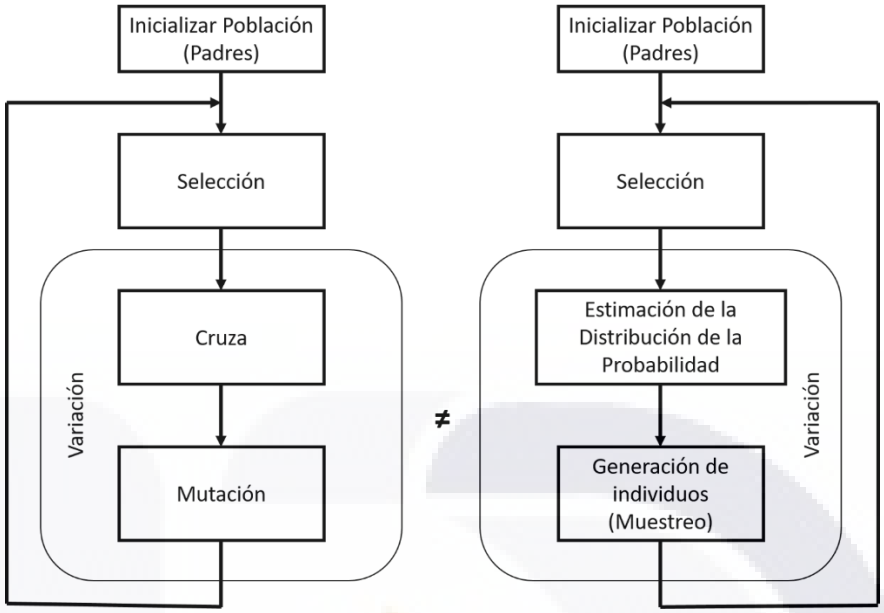


Figura 13 *Diferencia de enfoques, EDA y AG [55]*

A continuación, se explica una aproximación general de un *algoritmo de estimación de la distribución* y algunas alternativas existentes en la literatura.

Aproximación General del Algoritmo de Estimación de la Distribución

Al igual que el AG, el *Algoritmo de Estimación de Probabilidad (EDA)* tiene una base general, a partir de la cual se pueden hacer adaptaciones o hibridaciones dependiendo del problema que se pretenda resolver. Por tanto, en este apartado se ahonda un poco en la aproximación general del *EDA*. Entre la amplia variedad de *heurísticas* y *metaheurísticas*, el *EDA* permite una búsqueda *heurística probabilística* que se fundamenta en tres pasos u operadores: la *selección* de individuos, una *estimación* de distribución de los individuos seleccionados y el *muestreo* en base a la probabilidad aprendida [58].

BEGIN /* Algoritmo de Estimación de la Distribución*/

$D_0 \leftarrow$ Generar una Población Inicial (m individuos)

Evaluar la Población D_0

REPEAT for $l = 1, 2, \ldots$ hasta verificación de criterio de parada

$D_{l-1}^{Se} \leftarrow$ Seleccionar $n \leq m$ individuos de D_{l-1} acorde con el método de selección.

$p_l(x) = p(x|D_{l-1}^{Se}) \leftarrow$ Estimar la distribución de probabilidad de que un individuo se encuentre en los individuos seleccionados.

$D_l \leftarrow$ Muestrear M individuos (nueva Población) de $p_l(x)$

END REPEAT

END

Figura 14 *Aproximación general del EDA [60]*

De acuerdo con el pseudocódigo ilustrado en la Figura 11, la aproximación general de un *EDA* comienza de un conjunto de *m* individuos generados al azar que conforman la *población inicial* D_0 [58]. Una vez obtenida la *población inicial*, cada uno de los individuos es evaluado por medio de la *función de evaluación* que determina la calidad de las soluciones. A continuación, se comienza con los tres operadores:

- **Seleccionar:** Se elige un conjunto $n<=m$ de la población anterior. Regularmente, se espera que los seleccionados sean los que obtuvieron mejor evaluación o mayor peso de acuerdo con la *función objetivo* [42, 58].
- **Estimar** la *distribución de probabilidad* conjunta $p_l(x)$ partiendo de los individuos seleccionados [59]. La distribución de un dato o un conjunto de ellos se refiere a la probabilidad de que se presente un dato o un conjunto en el universo del que se extrae [53].
- **Muestrear:** Generar una población con *m* individuos nuevos a partir de la función de distribución $p_l(x)$ obtenida por el operador de *estimación.*

En los *EDA*, la estimación de la *función de distribución* es la fase más complicada. Como ya se mencionó, se realiza a partir del subconjunto de individuos seleccionados de la población inicial.

De acuerdo con Duarte [42], el modelado de la *función de distribución* se vuelve más difícil cuando la dependencia entre los componentes se vuelve más complicada. En casos

sencillos, el individuo se compone de *n* atributos independientes; mientras que, en casos más generales, $p_l(x)$ es una *función de distribución conjunta* de dimensión *n* que no se puede estimar de forma exacta, teniendo como desventaja que si *n* es suficientemente grande el problema se vuelve inabordable. Para estos casos, se necesita suponer que la *función de distribución de probabilidad* se puede factorizar en un modelo simplificado. Algunos de los modelos existentes se tienen los mencionados a continuación.

Funciones de Distribución de Probabilidad sin Dependencias

Se asume que no existen interacciones entre variables, por tanto, *la función de probabilidad* $p(X=x)$ se factoriza como el producto de funciones de probabilidad univariantes e independientes. Estos modelos son los más simples y computacionalmente económicos usados en problemas donde la interacción entre variables no es importante [42, 53, 59].

Entre los modelos sin dependencias se tienen los siguientes:

- UMDA (Univariate Marginal Distribution Algorithm),
- PBIL (Population Based Incremental Learning),
- cGA (Compact Genetic Algotithm).

Funciones de Distribución con Dependencias Bi-variadas

Considera las interacciones entre pares de variables, por lo que se consideran estadísticos de orden 2. Este tipo de modelos se comportan mejor en problemas donde existe relación entre pares de variables [42, 53, 59]. Ejemplos:

- MIMIC (Mutual Information Maximization for Input Clustering),
- COMIT (Combining Optimizers with Mutal Information Trees).
- BMDA (Bivariate Marginal Distribution Algorithm).

Funciones de Distribución con Dependencias Multivariadas:

Es el caso más general de todos, considera las interacciones entre variables de orden mayor a dos. Este modelo, llega a ser más complejo que los modelos anteriores llegando a incrementar exponencialmente por las posibles combinaciones de las interacciones haciendo infactible buscar todas las posibilidades [42, 53, 59]. Ejemplos:

- ECGA (Extended Compact Genetic Algorithm)
- FDA (Factorised Distribution Algorithm)
- BOA (Bayesian Optimization Algorithm)
- LFDA (Learning Factorised Distribution Algorithm)
- EBNA (Estimation of Bayesian Network Algorithm)

Finalmente, los tres operadores son ejecutados iterativamente, hasta que se cumpla una o más *condiciones de parada*, los cuales pueden ser: determinación de un número máximo de iteraciones, número máximo de individuos evaluados, uniformidad en la población generada, no existencia de mejora con respecto a los individuos obtenidos en iteraciones previas, etc.[53].

En términos generales, el modelo de distribución de probabilidad seleccionado según el problema a tratar para aprender las interrelaciones entre variables influirá notablemente en el comportamiento del algoritmo, tanto en tiempo como en resultados [61].

2.6 Herramientas de Construcción

Como ya se ha mencionado, este trabajo de tesis implicó la construcción de aplicaciones para la *búsqueda exhaustiva y metaheurística de testores*. Dichas aplicaciones tienen cierto grado de complejidad, por lo que se hizo uso de herramientas de *Ingeniería de Software* para facilitar la construcción de dichas aplicaciones. Las herramientas consultadas en la literatura son *arquitecturas de diseño* y *patrones de diseño*, las cuales son descritas a continuación.

2.6.1 Arquitectura de Software

La *Arquitectura de Software* es un área relativamente nueva, tal como los sistemas de software se han vuelto más distribuidos y complejos de forma que el diseño de una arquitectura se ha vuelto un paso importante para la construcción de *sistemas de software* [62, 63]. Así pues, la tarea de diseño de un *modelo arquitectónico de software* representa la primera etapa de diseño de software [64].

De Acuerdo con Jalote [62], una *arquitectura de software* particiona un sistema en partes lógicas de manera que cada una se comprenda independientemente y así, describir el sistema completo en términos de sus partes y sus relaciones formando una vista de alto nivel del funcionamiento de dicho sistema. Así pues, una arquitectura de software pues ser definida como: *"La estructura o estructuras del sistema, lo que comprende a los componentes del software, sus propiedades externas visibles y las relaciones entre ellas"* [65]. Dichos componentes pueden ser algo simple como un módulo de software o una clase orientada a objetos, cuyas propiedades son necesarias para entender las interacciones dentro de un sistema de software.

En resumen, una *arquitectura de software* no supone una aplicación de software operativa, pero en cambio, establece un marco estructural básico que identifica los principales componentes de un sistema y las comunicaciones entre dichos componentes [64] que permite [65]:

- Analizar la efectividad del diseño para el cumplimiento de requerimientos.
- Considerar alternativas en un punto donde los cambios aún son relativamente fáciles.
- Reducir riesgos asociados con la construcción de software.

Las ventajas que se encuentran al invertir en una arquitectura de software se tienen [66]:

- Facilidad de evaluación
- Construcción estable
- Construcción escalable

- Facilidad de gestión
- Disminución en tiempo de pruebas
- Optimización de recursos
- Producción flexible

Como producto del proceso del diseño de una *arquitectura de software* es un documento de *diseño arquitectónico* que, de acuerdo con Sommerville [64], puede incluir varias representaciones gráficas del sistema junto con un texto descriptivo asociado. Esta descripción debe cubrir la *estructura del sistema*, la aproximación adoptada y cómo se estructura cada subsistema en módulos. Los modelos que se incluyeron en este proyecto de tesis son los siguientes:

- **Modelo Estático:**
  El modelo estático muestra "*los subsistemas o componentes que se desarrollarán como unidades separadas*", siendo cada uno de dichos componentes un bloque de construcción de software que puede ser desplegable y sustituible [64, 65].
- **Modelo Dinámico:**
  Encargado de mostrar "*cómo se organiza el sistema en tiempo de ejecución*". Específicamente, un modelo dinámico aborda la estructura de la arquitectura en indica la manera en cambia la configuración en función de los eventos [64, 65].
- **Modelo de Organización del Sistema:**
  Representa la organización del sistema bajo una estrategia básica de estructuración de dicho sistema. La organización del sistema puede reflejarse directamente en la estructura de los subsistemas aún que en modelos posteriores se incluyan más detalles [64].
- **Modelo de Control:**
  Consiste en un plan de control para la forma en que un sistema se descompone en subsistemas. De esta manera, existe un mecanismo de control para los subsistemas reciban entradas y entreguen las salidas correspondientes en el momento preciso [64].

En el apartado 2.6.2 se explica la segunda herramienta de la *Ingeniería de Software* que apoya la construcción de los sistemas que se han mencionado en este capítulo.

2.6.2 Patrón de Diseño

En la etapa de *diseño* siempre se encuentran problemas recurrentes reconocibles que son fácilmente solucionados con la aplicación de formatos estándar. Estos modelos son llamados *patrones de diseño* creados por el arquitecto constructor Christopher Alexander quien reconoció que al construir edificios siempre se presentan problemas recurrentes. Alexander define estas soluciones como patrones, los cuales *"describen un problema recurrente en el ambiente, así como el núcleo de la solución de forma tal que es posible usarla millones de veces sin tener que elaborarla dos veces de la misma forma"* [65, 67, 68].

Un *patrón de diseño* no es una invención, más bien es una expresión documentada de la mejor manera de resolver un problema que fue observada o descubierta durante un estudio o la construcción de numerosos sistemas de software [67]. En general, un *patrón de diseño* se *"caracteriza como una regla de tres partes que expresa una relación entre cierto contexto, un problema y una solución"* [65]. Sin embargo, Gamma [68] generaliza el *patrón de diseño* en cuatro elementos esenciales:

1. **Nombre de Patrón,** que describe un problema de diseño, sus soluciones y consecuencias en pocas palabras dando en alto nivel de abstracción en contexto con el que se creó el documento.
2. **El problema,** que describe cuando puede ser aplicado el problema. Explica con mayor profundidad el contexto del problema y sus problemas de diseño. En esta parte se menciona la manera en que se representan los algoritmos como objetos. En algunas ocasiones se presentan las condiciones que deben conocerse antes de aplicar el patrón.
3. **La Solución,** describe los elementos que conforman el diseño, sus relaciones, responsabilidades y colaboraciones. Este apartado no debe describir una solución en particular, sino desarrollarse como un modelo o plantilla que pueda ser aplicado en

diferentes situaciones. Así, el patrón es una descripción abstracta de un problema de diseño.

4. **Las consecuencias,** se refieren a los resultados y consecuencias de la aplicación del patrón. Son críticos para la evaluación de alternativas de diseño y para entender los costos y beneficios de aplicar un patrón.

Como se puede observar, un patrón de diseño no provee soluciones a cada problema encontrado en el diseño y desarrollo de software del mundo real. En lugar de ello, un patrón provee la mejor solución a un problema en un contexto particular por lo que no producirá una solución efectiva al mismo problema en un contexto diferente [67].

En base a Coplien en [69] un patrón de diseño es eficaz si cumple con las siguientes características:

- **Resuelve un problema:** Contienen la solución a un problema, además de estrategias abstractas.
- **Concepto probado:** Las soluciones cuentan con un historial, no teorías.
- **La solución no es obvia:** Los mejores patrones de diseño generan indirectamente una solución a un problema, enfoque necesario para los problemas más difíciles del diseño.
- **Describe una relación:** Los patrones describen módulos, estructuras y mecanismos más profundos.
- **Tiene un componente humano significativo:** Los mejores patrones recurren explícitamente a la estética y a la utilidad debido que todo software sirve para el confort humano o a la calidad humana.

Así pues, un *patrón de diseño* incorpora conocimiento pragmático ganado con dificultad que permite ser reutilizado sin tener que elaborarlo dos veces de la misma forma evitando la reinvención de la rueda o, incluso, evitar crear una nueva rueda que no sea suficiente para el contexto establecido [65].

2.7 Interacción Ciencias Computacionales y Medicina

De acuerdo con Ávila [5], la *medicina* es "*prácticamente contemporánea con la humanidad*" y, al igual que otras ciencias, ha manejado *información* para su desarrollo. En la actualidad, la *información* que se genera en el ámbito *médico* ha tenido un crecimiento acelerado gracias a la introducción del mundo digital.

Gracias al crecimiento en la información que se maneja, las bases de datos se vuelven complejas y robustas representando así, un área de oportunidad para que la *medicina* sea una de las áreas del conocimiento que más beneficio obtienen de su interacción cercana con las *ciencias computacionales* y las *matemáticas* [1]. Esta interacción supone, como menciona María del Carmen Expósito en [6], una "*novedosa perspectiva*" que reduce costos, tiempo, errores médicos; así como un potenciador del uso de recursos humanos en ramas médicas con mayores requerimientos.

La *medicina* y las *ciencias computacionales* han *"logrado agilizar y mejorar procesos de apoyo médico"* tendiendo gran influencia especialmente con la introducción de la *inteligencia artificial* [5], la cual es una combinación de ciencias con múltiples ramas que tiene importantes aplicaciones *médicas* [70] que pueden vigilar pacientes con equipos biomédicos, realizan procesamiento de grandes cantidades de información y toman decisiones, entre otras aplicaciones [5].

La *inteligencia artificial* ha tenido un impacto importante a través de *sistemas expertos* aplicados especialmente al *diagnóstico médico* [70] que almacenan gran cantidad de conocimiento, simulan el razonamiento especialista, desarrollan hipótesis, resuelven conflictos y proporcionan el *diagnóstico* probable y un posible tratamiento justificando sus conclusiones [70-72]. El *diagnóstico médico*, por su parte, implica una tarea compleja que requiere de mucha capacitación por parte del especialista, debido a la diversidad de enfermedades y síntomas que pueden ser confusos que pueden evitar la obtención de un *diagnóstico* oportuno [1, 72]. Así, la aplicación de *sistemas expertos* busca apoyar a los especialistas en este complejo proceso y reducir errores pues, según lo descrito por Lugo-Reyes et al.[1], se estiman 150 de cada 1,000 pacientes mal diagnosticados.

Hoy en día, las aplicaciones y dispositivos enfocados en salud se han visto beneficiados por la existencia de los smartphones. Dichas aplicaciones reciben gran cantidad de información y, a través de sistemas expertos, pueden predecir padecimientos y mejorar las condiciones de salud del usuario [73]. Estas aplicaciones van desde el monitoreo de estilo de vida como la calidad de sueño, actividad física, etc. [73, 74]; .[73, 74]; hasta aplicaciones capaces de diagnosticar depresión a partir de las fotografías que los pacientes comparten en redes sociales de Harvard y la Universidad de Vermont [75].

Finalmente, en base a Lugo-Reyes et al.[1], gracias al creciente volumen en la información que se genera en la actualidad y a las características de complejidad de la medicina, el uso de las ciencias computacionales se vuelve vital en la búsqueda de sistemas de salud más eficientes simplificando en gran medida el trabajo de los médicos.

2.8 Cáncer de Mama

De manera general y lejos de lo que se puede pensar, el término *cáncer* no es una enfermedad, sino muchas de ellas. *Cáncer* es un término que designa alrededor de 200 entidades distintas [11] que constituyen un serio problema debido a las altas incidencia y mortalidad en el mundo [12], además de problemas de orden psicológico, familiar, económico, entre otros [13].

Todos los tipos de *cáncer* se caracterizan por un desequilibrio en la proliferación *celular* y los mecanismos de muerte *celular* [12], es decir, algunas *células* del cuerpo comienzan a dividirse sin control y se extienden en tejidos circundantes, es decir, existe un desequilibrio en la proliferación celular y los mecanismos normales de muerte celular [11]. Como se puede observar en la Figura 15, las *células* sanas se multiplican cuando el cuerpo las necesita y muere (apoptosis) cuando se dañan o el cuerpo ya no las requiere, de manera que, cuando el material genético de la *célula* cambia activando ciertos genes y desactivando otros, provoca que crezcan y se dividan sin ningún control u orden, cuya consecuencia final es la generación de tumores [11, 76, 77].

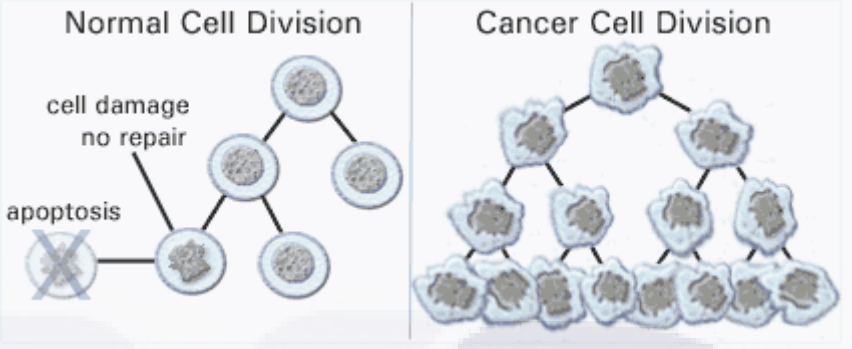


Figura 15 *División celular normal y anormal [78]*

Este desorden en el proceso *celular* puede generarse en casi cualquier parte del cuerpo humano, el cual se compone de trillones de *células* [77]. La diferencia entre los diferentes tipos de cáncer radica en la velocidad de crecimiento y propagación, así como su respuesta al tratamiento [79]. La mayoría de los tipos de cáncer forman una masa anormal de tejido corporal sin funciones fisiológicas conocidas como *tumor*, *masa* o *neoplasia* [79].

Es importante remarcar que la presencia de un *tumor* en el cuerpo no significa el desarrollo de un *cáncer*, pues algunas células que mutan desarrollando cambios leves que desaparecen sin tratamiento, mientras que otras desarrollan anomalías genéticas creando nuevas *células* cada vez más anormales hasta convertirse en *cáncer* [80, 81].

De acuerdo con el comportamiento clínico del *tumor* se puede catalogar como *tumor benigno* o *maligno*. El *benigno* es un aquel *tumor* no grave, es decir, no implican *cáncer*, que se encuentra bien localizado con tendencia de crecimiento lento. Este tipo de *tumores* rara vez causa problemas más graves, una vez extirpados, no suelen reaparecer [80, 82]. Mientras que los *tumores malignos* pueden dar paso a cualquier tipo de *cáncer* cuyas *células* son completamente deformes y desorganizadas provocando un crecimiento descontrolando que puede invadir tejidos cercanos interfiriendo en funciones del cuerpo y causar la muerte. Al igual que los *tumores benignos*, lo *malignos* pueden desaparecer o retirarse, pero tienen la posibilidad de reaparecer en algún momento [82, 83].

El tipo de cáncer en el que se centrará este proyecto es el *cáncer de mama*, uno de los cánceres tumorales más antiguos (Egipto, 1600 a.C. aproximadamente) [84]. El *cáncer de mama consiste* en un *tumor maligno* desarrollado a partir de células mamarias. En la mayoría de los casos el cáncer se genera en los conductos que llevan leche al pezón (cáncer ductal). En otras ocasiones se originan en las glándulas (lobulillos) que producen leche (cánceres lobulillares) [85, 86]. En menor incidencia se tienen cánceres de seno que comienzan en otros tejidos, tal es el caso de los sarcomas y linfomas, que no son considerados como *cáncer de mama* [86].

Al igual que otros tipos de cáncer, el origen del *cáncer de mama* es multifactorial, no hay una razón única para desencadenar su aparición y desarrollo posterior [15]. Si el problema no es tratado a tiempo las *células malignas* pueden invadir el tejido mamario circundante y llegar a los ganglios linfáticos de las axilas, encargados de eliminar sustancias extrañas del cuerpo. Si esto sucede, las *células malignas* tendrán acceso al resto del cuerpo [85].

Con una visión general del problema de cáncer y, específicamente, el *cáncer de mama* se puede abordar el problema de *selección de subconjuntos de características* y *teoría de testores* aplicado a las características que describen células de *cáncer de mama*, cuyos valores fueron obtenidos por un sistema de visión artificial reportados en [87-89]. La aplicación de *selección de características* se reporta más adelante en los apartados de Metodología y Resultados.

2.9 Hemofilia

Como ya se ha mencionado, la segunda patología abordada fue *hemofilia*, la cual consiste en un desorden genético, recesiva y ligada al cromosoma X que ocasiona que los pacientes tengan deficiencias en algún *factor de coagulación* ya sea del *factor VIII* para el caso de la *hemofilia A* o del *factor IX* para el caso de la *hemofilia B* [22, 90]. Este déficit provoca que el cuerpo tarde más tiempo en formar coagulos cuando hay una hemorragia [91], es decir, provoca sangrados más prolongados que pueden poner en peligro la vida del paciente[18].

Se estima que la *hemofilia* es diagnosticada en 1 de cada 10,000 nacimientos, siendo la *hemofilia A* la más frecuente con un estimado de un 80 a 85% de los casos [90]. De acuerdo con lo descrito por García-Chávez y Majluf-Cruz en [22], *"la hemofilia se manifiesta clínicamente solo en pacientes varones; mientras que las mujeres son portadoras, aunque pueden padecerla bajo condiciones muy específicas"* (ver Figura 16). Como se mencionó anteriormente, en el cromosoma X se encuentran los genes que determinan los factores de coagulación VIII y IX. Las mujeres tienen dos cromosomas X, mientras que los hombres solo tienen uno. De esta manera si un hombre tiene alteraciones en alguno de los dos factores, desarrollará hemofilia, mientras que las mujeres necesitan que las alteraciones se presenten en ambos cromosomas X para desarrollar la enfermedad [18].

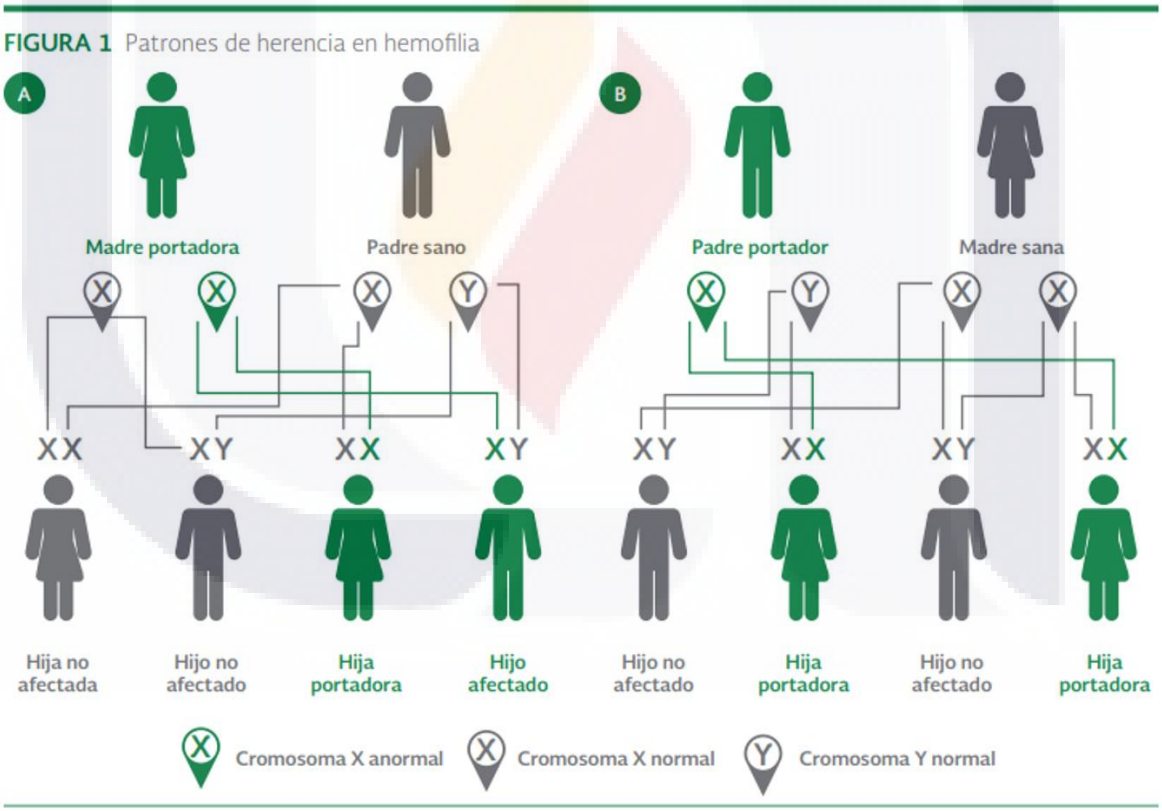


Figura 16 *Patrones de herencia en hemofilia [18]*

Las dificultades por enfrentar, tanto en hemofilia A como en B, son las mismas; aunque el tratamiento es diferente en cada una. De acuerdo con Bergés et al en [18], *"la clasificación de la hemofilia se basa en los niveles de actividad plasmática del factor VIII o IX"*, es decir, la gravedad de la enfermedad depende de los niveles del factor de coagulación correspondiente en la sangre. Estos niveles determinan cierto patrón en el nivel del sangrado y, por lo regular, son consistentes en la vida del paciente [91]. En la Tabla 3 a continuación, se muestra la clasificación de la hemofilia de acuerdo a su gravedad según el nivel del factor de coagulación.

GRAVEDAD	NIVEL DE FACTOR DE COAGULACIÓN	EPISODIOS HEMORRÁGICOS
Severa	<1%	• Hemorragias espontáneas en articulaciones y músculos.
Moderada	1 a 5%	• Hemorragias espontaneas ocasionales. • Hemorragias prolongadas ante traumatismos o cirugías menores.
Leve	5 a 40%	• Hemorragias en traumatismos o cirugías importantes. • Hemorragias espontáneas poco frecuentes.

Tabla 3 *Clasificación de hemofilia según los niveles del factor de coagulación [18, 90, 91]*

Los pacientes con *hemofilia* pueden llevar una vida normal llevando el tratamiento adecuado [92], para ello es necesario un *diagnóstico* oportuno desde la infancia pues, en caso de no detectarse a tiempo, puede derivar en costos más altos de tratamiento y cuidados [21]. Sin embargo, los tratamientos actuales permiten que los pacientes lleven un control adecuado de su enfermedad permitiéndoles, incluso, practicar deportes especialmente aquellos que fortalecen articulaciones, permitan un desarrollo neuromuscular normal, fortalecimiento de músculos y la coordinación como natación y tenis [21, 90].

Hasta este punto, se han definido los objetivos a alcanzar y los conceptos involucrados en el desarrollo de este reporte de tesis. Además, se tuvieron como escenarios de

experimentación la aplicación de los modelos: *exhaustivo* y *metaheurístico* (*AG* y *EDA*) en las patologías de *cáncer de mama* y *hemofilia*.

En el apartado 3 a continuación, se describe la metodología a partir de la cual se obtienen el conjunto de testores y el conjunto de testores típicos, así como el peso informacional de cada una de las características involucradas en las patologías descritas y los parámetros que mejor desempeño generan en la ejecución de los modelos metaheurísticos.

CAPÍTULO 3

Metodología

3.1 Introducción

A continuación, se describe la metodología que se siguió para la construcción de este reporte de tesis, en la cual se ha implementado el *algoritmo exhaustivo* y se han adaptado *metaheurísticas* (*AG* y *EDA*) que permitieron aplicar la *teoría de testores* para la *selección de características*. Lo anterior con el objetivo de analizar conjuntos de datos y reducir el número de variables o características que describen objetos y determinar cuáles inciden de forma determinante en un problema (ver apartados 2.3 y 2.4).

Como ya se mencionó anteriormente, se implementó el *algoritmo exhaustivo* para la búsqueda de testores, el cual asegura el 100% de ellos con el inconveniente de que es un algoritmo de complejidad exponencial (2^n) según el número de características a analizar. Por esta razón, se construyeron dos aplicaciones más que hibridan las aproximaciones generales

de las *metaheurísticas*: *algoritmo genético (AG)* y el *algoritmo de estimación de la distribución (EDA)* descritas en los apartados 2.5.3 y 2.5.4 respectivamente.

En ambas metaheurísticas se implementó un nuevo operador al que se le llamó *operador de alteración* que busca mejores soluciones a partir de soluciones ya evaluadas; todo con el objetivo mejorar la exploración del espacio de soluciones y obtener el conjunto de testores en los escenarios de investigación: *cáncer de mama* y *hemofilia*.

La implementación de las tres aplicaciones se realizó con el apoyo de modelos existentes en la literatura de la ingeniería de software para la construcción de herramientas que facilitaron su diseño estable (ver apartado 2.6).

Finalmente, se hace un estudio de los parámetros recibidos por los algoritmos metaheurísticos, el *AG* y el *EDA* para obtener aquellos que permiten un mejor desempeño en la búsqueda del conjunto de *testores* y *testores típicos*.

3.2 Descripción de la Metodología

En este aparato se describe a detalle la metodología que se siguió para la investigación reportada en este trabajo de tesis. Dicha metodología es representada a nivel general en la Figura 17 a continuación.

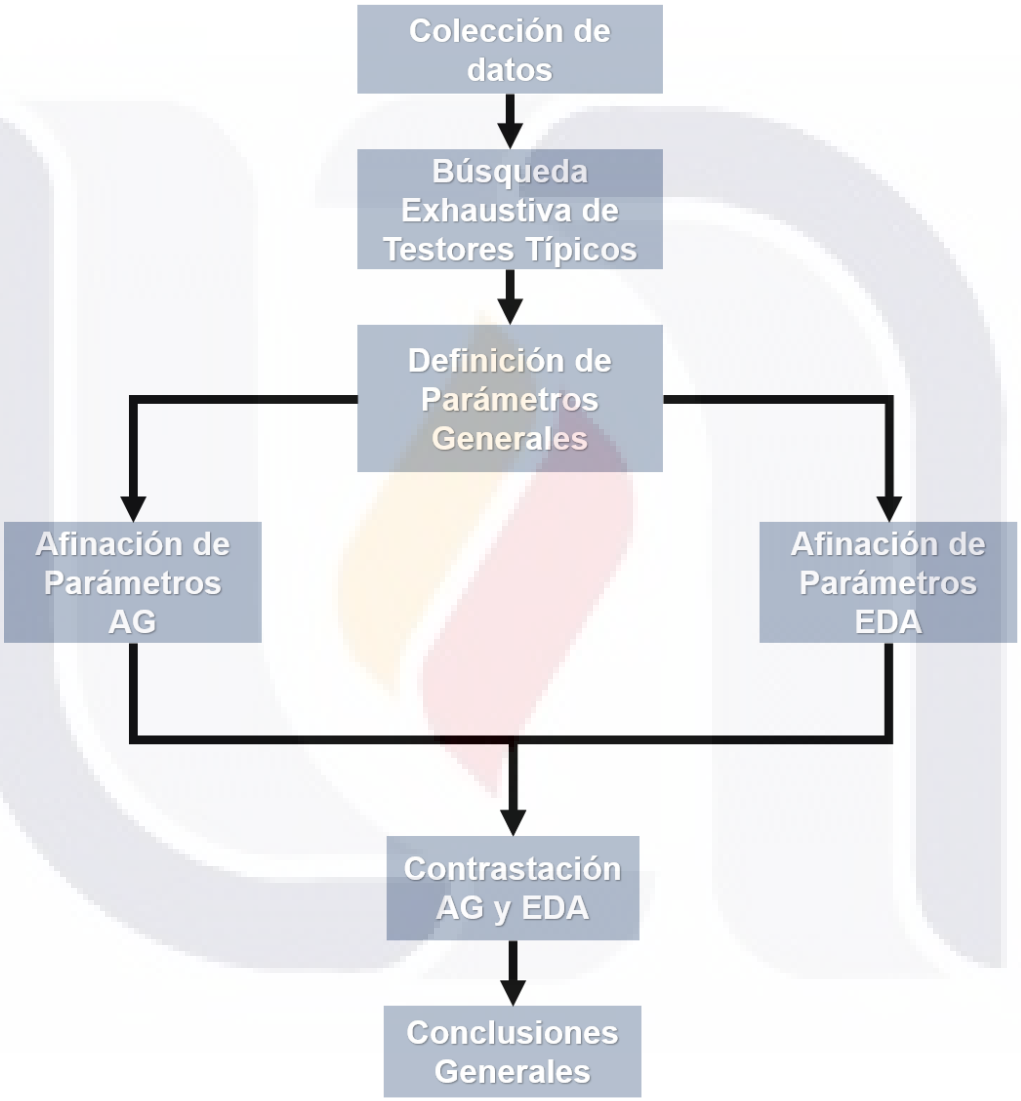


Figura 17 *Metodología*

En el subapartado a continuación, se describen los elementos que conforman la metodología presentada.

3.2.1 Colección de datos

La metodología comienza con la *colección de datos* (ver Figura 17), que representa un elemento importante de la misma, pues conforma su insumo principal. De acuerdo con la teoría del apartado 2.4, los conjuntos de datos pueden pertenecer a cualquier área del conocimiento. Para el caso de este trabajo de tesis los datos pertenecen al área médica, específicamente a dos patologías: el *cáncer de mama* y la *hemofilia*. Estos conjuntos de datos describen instancias u objetos por medio de *n* características, donde cada objeto pertenece a una clase determinada de un conjunto mayor. Sin embargo, no se debe perder de vista que entre mayor sea el número de características más pesado será el procesamiento exhaustivo debido a su complejidad exponencial (Ver apartado 2.4). Los conjuntos de datos utilizados recibieron un preprocesamiento en el cual cada característica fue discretizada en base a la literatura de la patología y la opinión de un especialista, con el objetivo de facilitar su procesamiento en la metodología.

3.2.2 Búsqueda Exhaustiva de Testores Típicos

El segundo paso de la metodología es la *búsqueda exhaustiva de testores típicos,* ejemplificada en la Figura 18. Como su nombre lo indica, obtiene el conjunto de *testores* y *testores típicos,* además de calcular el *peso informacional* de cada característica involucrada. El *análisis exhaustivo* es descrito a detalle en el apartado 2.4. Esta parte de la metodología recibió como parámetro un conjunto de *k* archivos, que constituyen la *matriz de aprendizaje* obtenida en el paso anterior, donde *k* es el número de clases que conformaron el problema a analizar.

Conforme se procesa la *matriz de aprendizaje,* cada módulo de la aplicación genera un archivo de texto con su salida correspondiente, es decir, al final se contó con un archivo para la *matriz de diferencias*, un segundo archivo para la *matriz básica,* el tercero para el *conjunto potencia*, el cuarto para el conjunto de *testores*, el quinto para los *testores típicos* y finalmente, uno para el cómputo del *peso informacional*.

Con la información obtenida en esta parte de la metodología se tiene un punto de comparación para evaluar el desempeño de las metaheurísticas aplicadas posteriormente dentro de la metodología.

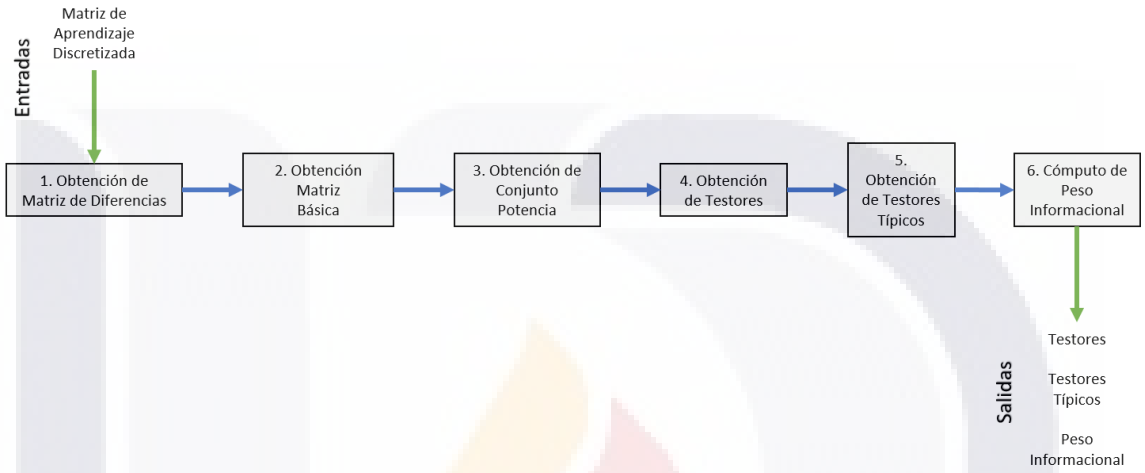


Figura 18 *Análisis exhaustivo*

La construcción de esta aplicación requirió de un diseño de software, el cual se realizó con el apoyo de herramientas de construcción de software como los *modelos arquitectónicos* y los *patrones de diseño* abordados en el apartado 2.6 aportados por el área de ing*eniería de software*. Los modelos que integran la *arquitectura de software* para el *modelo exhaustivo* están expuestos en el Anexo A.1, en el cual se incluye varias representaciones gráficas en las que se divide el modelo en partes lógicas independientes que permiten diferentes visualizaciones de alto nivel del sistema. Los modelos presentados son el *modelo estático* para identificar los componentes como unidades del sistema, un *modelo dinámico* para conocer la interacción entre los componentes, un *modelo de organización* con una estrategia de estructuración y un *modelo de control* que permita organizar la ejecución del sistema. Para mayor información estos modelos son descritos con más detalle en el apartado 2.6.1.

Por su parte, el anexo A.2 contiene la adaptación de un *patrón de diseño* que describe la solución de diseño del *modelo exhaustivo* en un formato estándar, el contexto en el que se

implementó dicho modelo, además de una representación abstracta de la estructura del sistema. Para más información consultar el apartado 2.6.2.

3.2.3 Definición de Parámetros Generales

A lo largo de este documento se ha mencionado la implementación del *algoritmo genético* y del *algoritmo de estimación de la distribución*. Ambas metaheurísticas son poblacionales evolutivas, por lo que su principio es similar. Por esta razón en el punto de *definición de parámetros generales* de la metodología de la Figura 17, define aquellos parámetros comunes para ambas metaheurísticas. Ambas reciben los parámetros generales:

1. **Tamaño de la población/conjunto**: Numero de cadenas/individuos por conjunto/población.
2. **Porcentaje de alteración**: Este porcentaje permite buscar posibles buenas soluciones en la vecindad de un *individuo* evaluado. De esta manera se buscan mejores soluciones a las ya obtenidas.
3. **Número de iteraciones**: Se trata del máximo número de iteraciones que ser realizan por cada experimento. En otras palabras, es una de las condiciones de paro.

Además de los parámetros descritos, ambas *metaheurísticas* reciben los archivos de la matriz básica, testores y testores típicos obtenidos por la *búsqueda exhaustiva*. El primer archivo es utilizado como evaluación de cada individuo del conjunto de soluciones generados por las *metaheurísticas*. Por su parte, los archivos de *testores* y *testores típicos* sirven a las *metaheurísticas* como punto de evaluación de desempeño en la búsqueda del espacio de soluciones.

3.2.4 Afinación de Parámetros AG

La *afinación de parámetros AG* realiza la búsqueda de *testores típicos* por medio de una hibridación del *algoritmo genético simple* descrito en el apartado 2.5.3. Esta hibridación fue

diseñada y construida en base a una arquitectura y un patrón de diseño basados en los modelos existentes en la literatura de la *ingeniería de software*. En el apartado A.3 se expone el *modelo estático* que determina los componentes del sistema como unidades lógicas, un *modelo dinámico* que visualiza la interacción entre las unidades del sistema, un *modelo de organización* con una estrategia básica de estructuración, así como un *modelo de control* que organiza la activación de cada componente en el momento requerido. Por su parte, en el Anexo A.4 se expone un *patrón de diseño* que expone la solución de diseño que permitió la construcción de la aplicación que ejecuta el *AG hibridado* en el contexto de la búsqueda de testores y testores típicos.

Además de la búsqueda de *testores típicos* de los problemas descritos, la aplicación permite la *afinación de parámetros* con el objetivo de obtener aquellos que permitan un mejor desempeño por parte del algoritmo. Para ello, se tiene un *diseño de experimentos factorial* con el cual cada parámetro recibe una serie de valores como se muestra a continuación en la Tabla 4.

PARAMETRO	VALORES
1. Tamaño de población	700, 800, 900
2. Porcentaje de alteración	10, 15, 20, 25
3. Probabilidad de mutación	2, 5, 7
4. Número de iteraciones	10, 20, 30

Tabla 4 *Definición de parámetros para análisis AG*

Como se puede observar, además de los parámetros definidos en el paso anterior de la metodología, la búsqueda de *testores típicos* con AG también necesita del parámetro de *probabilidad de mutación*, con el cual, se decide por cada solución de la población si será mutada o no.

Con los valores de los parámetros definidos para ambos escenarios de experimentación propuestos para este trabajo de tesis (*cáncer de mama* y *hemofilia*) se realizó la búsqueda de *testores típicos* bajo el diseño de experimento factorial en el que se exploran todas las posibles combinaciones de valores como se muestra en la Figura 19. Además, cada una de

las combinaciones fue replicada 30 veces, aunque la aplicación permite definir un valor distinto, con el objetivo de realizar un análisis estadístico y así, determinar aquellas combinaciones que permiten un mejor desempeño de la metaheurística en base a:

- Porcentaje de testores encontrados.
- Porcentaje de testores típicos encontrados
- Tiempo utilizado
- Numero de iteraciones realizadas

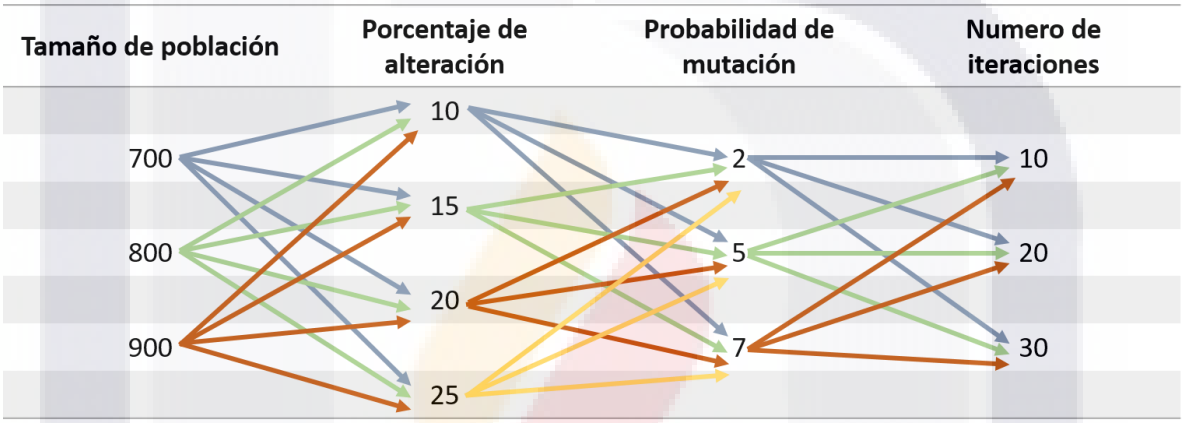


Figura 19 *Ejemplo de combinaciones de parámetros esperadas en el experimento con AG*

3.2.5 Afinación de Parámetros EDA

Por su parte, la afinación de parámetros *EDA* se encarga la búsqueda de *testores típicos* por medio de la hibridación del *algoritmo de estimación de la distribución* descrito en el apartado 2.5.4, además de determinar los valores para los parámetros que mejor desempeño tienen en la búsqueda.

Para la implementación del *EDA* se seleccionó la función de distribución *UMDA* (Univariate Marginal Distribution Algorithm) para el operador de estimación, el cual fue seleccionado por su simpleza y a que es computacionalmente económico (ver apartado 2.5.4). Por otro lado, al igual que con el AG, se incorpora el *operador de alteración* para evaluar la

vecindad de cada individuo del conjunto con la intención de mejorar las soluciones ya obtenidas en la iteración en curso (ver apartado 4).

La hibridación del *EDA* para la *búsqueda de testores típicos* fue diseñada en base a la adaptación de modelos de *arquitectura de software*, los cuales son expuestos en el Anexo A.5. Al igual en la hibridación del AG, se incluye un *modelo estático* con los componentes del sistema como bloques de construcción, un *modelo dinámico* que muestra el comportamiento del sistema es estado de ejecución, un *modelo de organización* que muestra la estructura y finalmente, un *modelo de control* que permite la inclusión de un componente adicional que organice al resto en tiempo de ejecución. Por su parte, el Anexo A.6 expone la adaptación de un modelo de *patrón de diseño* que muestra la solución encontrada para el diseño de la aplicación encarga de la ejecución del *EDA híbrido* para el contexto de la búsqueda de *testores típicos*.

PARAMETRO	VALORES
1. Tamaño de población	700, 800, 900
2. Porcentaje de alteración	10, 15, 20, 25
3. Porcentaje de selección	20, 30, 40
4. Número de iteraciones	10, 20, 30

Tabla 5 *Definición de parámetros para el análisis EDA*

En la Tabla 5 se presentan los parámetros que requiere para la búsqueda de *testores típicos*. Como se puede observar, además de los parámetros generales, este análisis difiere en un nuevo parámetro que define el *porcentaje de selección*. Este parámetro, como su nombre lo indica, permite seleccionar un porcentaje del conjunto de soluciones para realizar la *estimación de la distribución* y, posteriormente, el *muestreo* de un nuevo conjunto.

Para la realizar la afinación del algoritmo se hizo un diseño de experimentos factorial (ver Figura 20) agregando la posibilidad de asignar una serie de valores a cada uno de los parámetros, como se muestra en la Tabla 5. Por tanto, se corren todas las posibles combinaciones de valores replicadas 30 veces cada uno para realizar su estudio estadístico, para así, determinar aquellos con los que el EDA obtiene mejores resultados.

Al final de este punto de la metodología, se tiene el conjunto de *testores* y *testores típicos* encontrados por la metaheurística híbrida y los valores para los parámetros que obtuvieron mejor desempeño en la búsqueda.

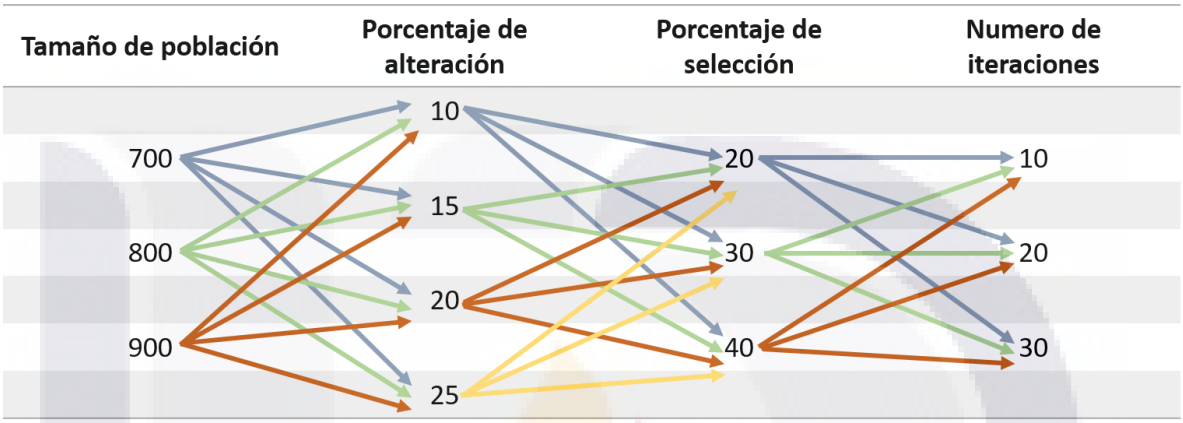


Figura 20 *Ejemplo de combinaciones de parámetros esperadas en el experimento con AG*

3.2.6 Contrastación de Resultados AG y EDA

Una vez realizado el análisis de resultados sobre el desempeño del *AG* y del *EDA* se procedió a contrastar su desempeño en cada escenario de experimentación. Es decir, se realizó un estudio estadístico que permitió conocer cuál de las dos *metaheurísticas* seleccionadas fue mejor en la búsqueda de *testores típicos* en base a su naturaleza y la influencia de la hibridación en cada uno de los modelos en un problema dado.

3.2.7 Conclusiones Generales

Para finalizar la metodología se describen las conclusiones en dos sentidos importantes: desde el punto de vista del *problema médico* y desde el punto de vista del desempeño de las *metaheurísticas*. El primer punto de vista describe la información obtenida a partir del conjunto de testores típicos y su apego a la realidad de la patología. Desde el segundo punto

de vista se describe el desempeño de la hibridación en cada una de las *metaheurísticas* para encontrar el conjunto de *testores típicos* con respecto al resultado del *análisis exhaustivo*.

A continuación, se describen los escenarios de experimentación con los que se pusieron a prueba los modelos desarrollados para este trabajo de tesis.

3.3 Escenarios de Experimentación

Como se ha mencionado a lo largo de este documento, se contaron con dos escenarios de experimentación en los que se aplicaron los 3 modelos desarrollados: el *algoritmo exhaustivo* y las *metaheurísticas AG* y *EDA* para realizar *selección de subconjuntos de características* por medio de la *teoría de testores*. Los escenarios comprenden dos patologías medicas: *cáncer de mama* y *hemofilia*. Donde la primera proviene de un repositorio de la Universidad de California, Irvine; mientras que la segunda es una base de datos de uso personal de la Dra. Cardiel del IMSS N° 1 en Aguascalientes, México. A continuación, se describen en detalle ambos escenarios.

3.3.1 Escenario de Experimentación 1: Cáncer de Mama

La *matriz de aprendizaje* de este primer escenario de experimentación comprende un conjunto de datos de la Universidad de California, Irvine; específicamente, del Machine Learning Repository. La fuente está denominada como Diagnóstico de Cáncer de Mama de Wisconsin (Wisconsin Diagnostic Breast Cancer) [93].

Este conjunto de datos contiene el *diagnóstico* de *células de mama* y 10 características obtenidas por medio de un sistema de visión por computadora a partir de una imagen digitalizada de una muestra de tejido de mama (ver Figura 21). Las características a analizar fueron resultado de este estudio que fue realizado por Wolberg, Street y Mangasarian reportado en [87-89].

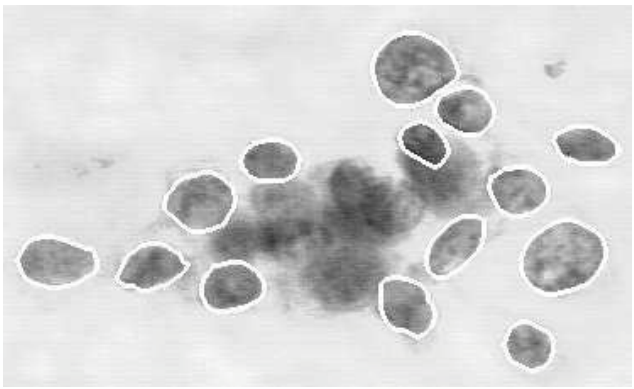


Figura 21 *Ejemplo de imagen tomada por un sistema de visión por computadora y el contorno de la célula [87]*

Las características evaluadas en este escenario se describen a continuación de acuerdo a lo consultado en [87, 93]:

A. Diagnóstico (M=maligno, B=benigno):

El *diagnóstico* es el resultado final de la evaluación de las características del *núcleo celular* con un sistema de *diagnóstico* con visión por computadora [87-89]. Cada *célula* en la base de datos tiene uno de dos posibles *diagnósticos*, puede ser *célula maligna* registrada con la letra "M" o *benigna* registrado con la letra "B". De acuerdo con lo descrito en el apartado 4.3, las *células benignas* no suponen *cáncer* mientras que las malignas sí suponen un caso de *cáncer* con la posibilidad de extenderse en todo el cuerpo.

B. Radio:

El *radio* de la *célula* fue medido promediando la longitud de los segmentos de las líneas radiales definidos por el centroide de la *célula* y los puntos individuales en el límite de la célula. Las líneas radiales fueron definidas pro Street, Wolberg y Mangasarian en [87, 89] se muestran a continuación en la Figura 22.

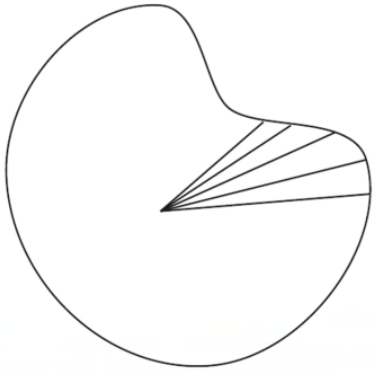


Figura 22 *Líneas radiales medidas en una célula*

C. Textura:

Como se mencionó anteriormente, cada característica fue extraída de un sistema de un sistema de visión por computadora, por lo que la textura de la *célula* fue medida encontrando la varianza en intensidades de escala de grises en los pixeles de la computadora [87, 89]. Ejemplo de lo anterior se puede observar en la Figura 21.

D. Perímetro:

El perímetro es definido como la distancia total entre puntos individuales llamados "serpiente" en [87]. Estos puntos individuales comprenden las líneas blancas en el perímetro de las *células*, como se observa en el ejemplo de la Figura 21.

E. Área:

El área de la *célula* se obtiene contando el número de pixeles en el interior de la línea blanca o conjunto de puntos "serpiente" (Figura 21) añadiendo los pixeles en el perímetro [87].

F. Lisura:

La lisura del contorno nuclear es calculada por la diferencia entre la longitud de una línea radial y la longitud media de las líneas que lo rodean [87]. Básicamente, la lisura de la *célula* es la variación local en las longitudes del radio. Un ejemplo de líneas radiales se puede observar en la Figura 22.

G. Compacidad:

Para calcular la compacidad del núcleo *celular*, el perímetro y el área son combinados usando la fórmula:

$$\text{compacidad} = \frac{\text{perímetro}^2}{\text{área}} \qquad \textit{Ecuación 3}$$

Este valor dimensional o medida de forma minimiza con un disco circular e incrementa con la irregularidad del límite de la célula. Además, esta medida aumenta con *núcleos celulares* alargados, lo cual puede indicar mayor probabilidad de malignidad [87].

H. Concavidad:

La concavidad analiza las irregularidades de forma en el *núcleo de la célula*. Street, Wolberg y Mangasarian miden el número y la severidad de las concavidades y hendiduras en el *núcleo de la célula*. Los autores dibujan líneas entre cada punto blanco no adyacente y miden hasta qué punto el límite real del núcleo se encuentra en el interior de cada línea. Este parámetro es afectado por la longitud de estas líneas, entre más pequeñas sean las líneas capturan concavidades más pequeñas [87, 89]. Un ejemplo de las líneas trazadas para medir la concavidad se observa en la Figura 23.

Figura 23 *Líneas usadas para calcular concavidad [89]*

I. Puntos cóncavos:

Los puntos cóncavos es una medida similar a la concavidad, cuya diferencia es que ésta característica mide el número de concavidades en el contorno *celular*, en lugar de la magnitud [87].

J. Simetría:

La simetría se obtiene a partir de la línea más larga que pase por el centro *celular*. Entonces, de acuerdo a [87], se trazan líneas perpendiculares a dicha línea, como se muestra en la Figura 24, para medir la diferencia de longitudes en las dos direcciones de la línea central.

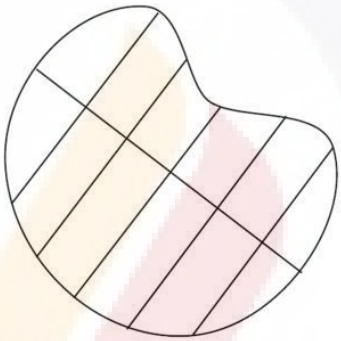


Figura 24 *Segmentos usados para el cálculo de simetría [87]*

K. Dimensión fractal:

La dimensión fractal es una característica de forma [89], es decir, a mayor valor corresponde a un menor contorno y por tanto, a una mayor probabilidad de malignidad [87]. La dimensión fractal se aproxima usando la aproximación de costa de Mandelbrot [87, 94]. El perímetro del *núcleo* es medido usando "reglas" cada vez más grandes. Esto es, a medida que aumenta el tamaño de la regla, decrece la precisión de la medición, el perímetro observado disminuye. Ahora, trazando estos valores a una escala logarítmica y medir la pendiente descendente da el negativo de una aproximación de la dimensión fractal [87].

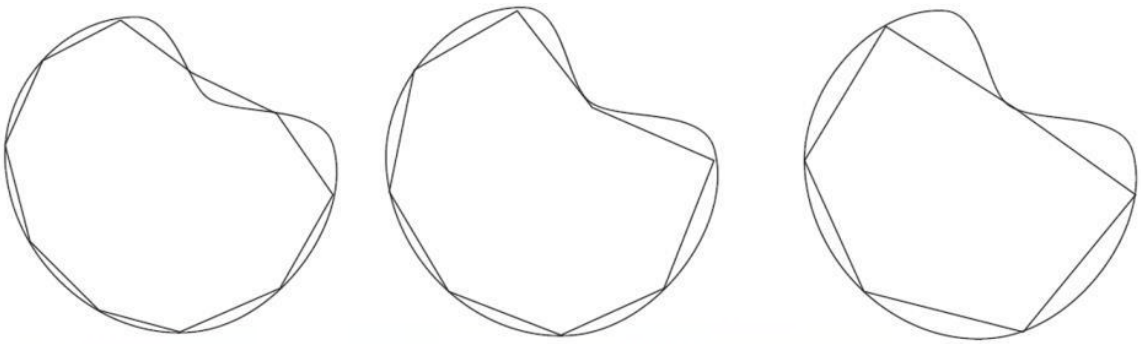


Figura 25 *Secuencia de medidas para el cálculo de dimensión fractal [87]*

Discretización de Características:

Como se describe en la metodología de la Figura 17, se comienza con la colección de datos que, además implica un preprocesamiento, el cual incluye una discretización de datos. Para el caso del escenario de experimentación de cáncer de mama se realizó de la siguiente manera:

CATEGORÍA	RADIO	TEXTURA
1. Totalmente benigno	6.9 – 9.9	9.7 – 17
2. Localmente agresivo	10 – 15.9	17.1 – 24.1
3. Bajo grado de malignidad	16 – 22.9	24.2 – 31.2
4. Alto grado de malignidad	23 – 28.11	31.3 – 39.78

Tabla 6 *Discretización de datos para cáncer de mama: radio y textura*

CATEGORÍA	PERIMETRO	AREA
1. Totalmente benigno	43.79 – 72.17	143.5 – 732.5
2. Localmente agresivo	72.18 – 108.7	732.6 – 1321.5
3. Bajo grado de malignidad	108.18 – 144.17	1321.6 – 1901.5
4. Alto grado de malignidad	144.18 – 188.5	1901.6 - 2501

Tabla 7 *Discretización de datos para cáncer de mama: perímetro y área*

CATEGORÍA	LISURA	COMPACIDAD
1. Totalmente benigno	0.05263 – 0.08263	0.01938 – 0.09938
2. Localmente agresivo	0.08264 – 0.11263	0.09939 – 0.17938
3. Bajo grado de malignidad	0.11264 – 0.14263	0.17939 – 0.25938
4. Alto grado de malignidad	0.14264 – 0.1634	0.25939 – 0.3454

Tabla 8 *Discretización de datos para cáncer de mama: lisura y compacidad*

CATEGORÍA	CONCAVIDAD	PUNTOS CONCAVOS
1. Totalmente benigno	0 – 0.1	0.0 – 0.05
2. Localmente agresivo	0.2 – 0.2067	0.06 – 0.1003
3. Bajo grado de malignidad	0.2068 – 0.3067	0.1004 – 0.1503
4. Alto grado de malignidad	0.3068 – 0.4268	0.1504 – 0.2012

Tabla 9 *Discretización de datos para cáncer de mama: concavidad y puntos cóncavos*

CATEGORÍA	SIMETRIA	DIMENSION FRACTAL
1. Totalmente benigno	0.106 – 0.146	0.04996 – 0.05996
2. Localmente agresivo	0.147 – 0.186	0.05997 – 0.06996
3. Bajo grado de malignidad	0.187 – 0.226	0.06997 – 0.07996
4. Alto grado de malignidad	0.227 – 0.304	0.07997 – 0.09744

Tabla 10 *Discretización de datos para cáncer de mama: Simetría y Dimensión Fractal*

En este caso, la discretización de las características se realizó en base a la experiencia de un especialista, el Dr. Luis Muñoz Fernández, patólogo del Centenario Hospital Miguel Hidalgo en Aguascalientes, Ags.

3.3.2 Escenario de Experimentación 2: Hemofilia

En el segundo escenario de investigación comprende una *matriz de aprendizaje* proveída por la Dra. Cardiel del Hospital Zona N°1 del Instituto Mexicano del Seguro Social (IMSS) en el Estado de Aguascalientes. El cual consiste en una base de datos de uso personal que contiene la información sobre el estado actual de pacientes diagnosticados con *hemofilia* en la entidad.

Para este escenario se realizaron experimentos de acuerdo a las clases que dicta el *nivel de gravedad* de la enfermedad pues, como se describe en el apartado 2.9, la clasificación de la *hemofilia* se basa en los niveles de actividad de los *factores VIII* o *IX*, es decir, de acuerdo a su gravedad. Además, se cuenta con 11 características más descritas a continuación:

A. Gravedad (S=severa, M=moderada, L=leve):

De acuerdo con la Tabla 3 del apartado 2.9 la *hemofilia severa* presenta hemorragias espontaneas sin razón aparente. Por su parte, la *hemofilia moderada* presenta hemorragias ocasionales y prolongadas en traumatismos o cirugías. Finalmente, la *hemofilia leve* presenta hemorragias espontáneas poco frecuentes e importantes en cirugías.

B. Edad:

Edad en años cumplidos del paciente pues es de gran importancia la detección de este trastorno hemorrágico en edad temprana, pues de no detectarlo puede derivar en costos altos de tratamiento y cuidados (ver apartado 2.9).

C. IMC:

La obesidad de un paciente es uno de los principales determinantes de la salud [95]. En este caso, se realiza por medio del cálculo de índice de masa corporal (IMC), el cual es un indicador internacional de la OMS que corresponde a la relación entre el peso y la talla[96].

D. Tipo de Enfermedad:

La hemofilia puede presentarse en dos tipos, la hemofilia A en la que existe deficiencia en el factor de coagulación VIII; y la hemofilia B en la existe deficiencia en el factor IX. En ambos casos se producen sangrados prolongados que pueden poner en peligro la vida del paciente. Para más información consultar apartado 2.9.

E. Artropatía:

La artropatía es una afección en una articulación. De acuerdo con el Instituto Mexicano del Seguro Social en [97] *"la artropatía constituye la morbilidad más importante del paciente con hemofilia, ya que lo convierte en un discapacitado físico al destruir las articulaciones más importantes como las rodillas, tobillos, caderas, codos por las hemartrosis recurrentes".*

F. Articulaciones con Daño:

En caso de que en la característica anterior resulte positiva, es decir, el paciente presenta artropatía, en la variable 6 se registró la cantidad de articulaciones dañadas.

G. VIH:

Esta es una característica que registra si el paciente presenta VIH o no. El virus de inmunodeficiencia humana o VIH infecta a las células del sistema inmunitario que altera o anula su función. El VIH es un problema de salud pública que se diagnostica por análisis de sangre, del cual no existe cura para la infección [98].

H. VHC:

Al igual que la característica anterior, registra la presencia de VHC o hepatitis C en el paciente con hemofilia. El VHC es una enfermedad hepática, causada por un virus RNA que provoca infecciones tanto agudas como crónicas. Por su parte, la infección aguda es asintomática con una duración máxima de seis meses. Por otro lado, la infección crónica que puede desarrollar insuficiencia hepática y es la causa principal de cirrosis y trasplante hepático [99].

I. VHB:

Registra la presencia o ausencia del virus de la hepatitis B (VHB) en el paciente con hemofilia. De acuerdo con la Organización Mundial de la Salud [100], la hepatitis B es "*una infección hepática potencialmente mortal que constituye un problema de salud mundial*". Al igual que la VHC conlleva un alto riesgo de muerte por cirrosis y cáncer hepático.

J. Inhibidores:

Los inhibidores son anticuerpos que presentan los pacientes de hemofilia al medicamento de tratamiento concentrado del factor de coagulación correspondiente. Cuando se producen inhibidores, el cuerpo deja de aceptar del factor suministrado como parte normal de la sangre provocando deficiencia en el tratamiento. Esta característica mide los títulos del inhibidor en unidades Bethesda (UB), que mientras más altos sean mayor presencia existe del inhibidor [101].

K. Número de Hemorragias al Año:

Como su nombre lo indica, esta característica registra la cantidad de hemorragias que presentó cada paciente con hemofilia en el último año al momento de su registro en la base de datos.

L. Modalidad de Tratamiento:

En los tratamientos actuales se suministra preventivamente (profilaxis) el factor de coagulación del cual el paciente es deficiente con el objetivo de mejorar la calidad de vida del paciente [102]. De acuerdo con el Reporte sobre Hemofilia en México [18], la profilaxis pueden ser primaria, secundaria y a demanda. La primera se aplica de una a tres veces por semana antes de los 30 meses de vida, sin artropatías. Por su parte la secundaria inicia después de los 3º meses de vida cuando existen hemorragias en articulaciones. Finalmente, el tratamiento a demanda se aplica únicamente cuando se presenta un evento hemorrágico o después de un traumatismo [18].

Discretización de Características

Para llevar a cabo la discretización de las características de este escenario de experimentación se combinó la investigación de la literatura de la patología y se tuvo a disposición la opinión del experto. A continuación, se presenta la discretización de las características descritas en este apartado sobre hemofilia como parte del proceso de colección de datos de la Figura 17.

CATEGORÍA	EDAD
1. Infancia	0 a 4 años
2. Niñez	5 a 13 años
3. Adolescencia	14 a 18 años
4. Adulto Joven	19 a 44 años
5. Adulto Maduro	Mayor a 45 años

Tabla 11 *Clasificación de edad para hemofilia (con opinión del experto)*

CATEGORÍA	IMC
1. Bajo	<18.5
2. Saludable	18.5 a 24.9
3. Sobrepeso	25 a 29.9
4. Obeso	30 a 39.9
6. Obesidad Extrema	7. >40

Tabla 12 *Discretización del Índice de Masa Corporal (IMC) [96]*

CATEGORÍA	TIPO DE HEMOFILIA
1. A	Hemofilia A
2. B	Hemofilia B

Tabla 13 *Discretización de tipo de hemofilia (ver apartado 2.9)*

CATEGORÍA	ARTROPATÍA
1. Si	Positivo
2. No	Negativo
3. SD	Sin Determinar

Tabla 14 *Discretización de la presencia de artropatías (con opinión de experto)*

CATEGORÍA	ARTICULACIONES CON DAÑO
1. 3	3 o más
2. 2	2
3. 1	1
4. 0	0
5. SD	Sin Determinar

Tabla 15 *Discretización para en número de articulaciones con daño (con opinión de experto)*

CATEGORÍA	VIH
1. Si	Positivo
2. No	Negativo
3. SD	Sin Determinar

Tabla 16 *Discretización de la presencia de VIH*

CATEGORÍA	VHC
1. Si	Positivo
2. No	Negativo
3. SD	Sin Determinar

Tabla 17 *Discretización de la presencia de VHC*

CATEGORÍA	VHB
1. Si	Positivo
2. No	Negativo
3. SD	Sin Determinar

Tabla 18 *Discretización de la presencia de VHB*

CATEGORÍA	INHIBIDORES
1. 1	>5UB
2. 2	<5UB
3. 3	Negativo
4. SD	Sin Determinar

Tabla 19 *Discretización de inhibidores en el paciente*

CATEGORÍA	N° DE HEMORRAGIAS AL AÑO
1. Grave	24 a 48
2. Moderada	4 a 6
3. Leve	<3

Tabla 20 *Discretización de numero de hemorragias al año [103]*

CATEGORÍA	MODALIDAD DE TRATAMIENTO
1. PP	Profilaxis Primaria
2. PS	Profilaxis Secundaria
3. TD	Tratamiento a Demanda
4. SD	Sin Determinar

Tabla 21 *Discretización de la modalidad de tratamiento del paciente con hemofilia*

Una vez descrita la metodología utilizada y los escenarios de investigación en los que se aplicó se procede a describir los resultados obtenidos de dicha aplicación en el apartado siguiente.

CAPÍTULO 4

Resultados

4.1 Introducción

A continuación, se muestran los resultados obtenidos con el desarrollo de este trabajo de tesis. Dichos resultados se presentan desde dos puntos de vista: el computacional y el médico. En el primero se exponen las propuestas de hibridación para las metaheurísticas AG y EDA, ya descritas en el apartado 2.5, que incluye la creación de una población inicial, la adición de un *operador de alteración*, así como la creación de una nueva población de soluciones utilizada en cada iteración de la metaheurística. Además, este punto de vista incluye el resultado de la afinación, descrita en el apartado 3, con las combinaciones de valores parámetro que mejor desempeño obtuvieron para cada uno de los escenarios de experimentación utilizados.

Por otra parte, se tiene el punto de vista médico en el que se describe el significado del conjunto de testores típicos referente al contexto de cada una de las patologías que constituyen los escenarios de experimentación.

4.2 Resultados Computacionales

El primer punto de vista a expones es el de las *ciencias computaciones*, desde el cual se describe la incorporación de nuevos mecanismos que hibridan las *metaheurísticas* seleccionadas en beneficio del desempeño en el recorrido del espacio de soluciones.

El primer mecanismo a describir es la *producción de poblaciones iniciales*, el cual se compone de tres ruletas diferentes. Cada una de estas ruletas puede seleccionar uno de dos posibles valores 1 y 0, donde la probabilidad de que sea uno u otro valor es distinta en cada ruleta, lo cual posibilita tener mayor diversidad de individuos en la población.

El segundo mecanismo trata de un nuevo operador para las dos *metaheurísticas* con las que se trabajó. Este nuevo operador se le nombró *operador de alteración* el cual, como su nombre lo dice, se encarga de alterar un porcentaje de cada uno de los individuos de una población evaluada con el objetivo de obtener mejores soluciones en cada iteración.

Una vez descritos los mecanismos implementados, se procede a describir las metaheurísticas, tanto *EDA* y *AG*, hibridados con la incorporación de los mecanismos descritos, así como la descripción de las condiciones de paro.

Finalmente, se presentan los mejores valores para los parámetros definidos en el apartado anterior. Dichos valores resultan de la experimentación factorial y de un análisis estadístico como se describe en la metodología.

4.2.1 Producción de Poblaciones Iniciales

Como se mencionó en el apartado 2.5 tanto el *algoritmo genético (AG)* como el *algoritmo de estimación de la distribución (EDA)* son *metaheurísticas* poblacionales evolutivas, por tanto, se diseñó un mecanismo u operador que permitió generar una *población inicial aleatoria* a partir de la cual, comienza la búsqueda en el espacio de soluciones por parte de la metaheurística.

Para la realización de este proyecto de tesis se determinó el uso de 3 ruletas como se muestra en la Figura 26, donde cada una de las ruletas tiene diferentes probabilidades de asignar un 1 o 0 en un elemento de la cadena que representa el individuo en una población.

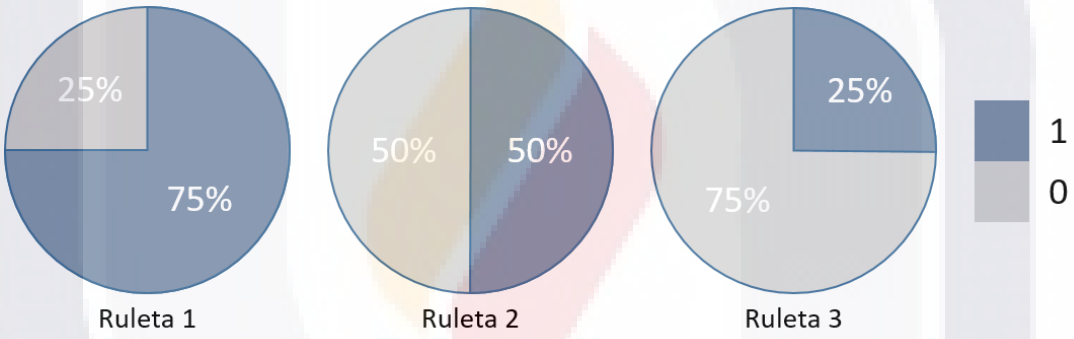


Figura 26 *Ruleta para la creación de poblaciones iniciales*

Cada ruleta es seleccionada al azar con la misma probabilidad para cada elemento del individuo, es decir, cada elemento *i* de una cada es determinado por una de las tres ruletas. Por lo tanto, dos elementos continuos *i* e *i+1* no necesariamente se asignaron con la misma ruleta. Por ejemplo, suponiendo que se selecciona la Ruleta 1 (ver Figura 26), se genera un número aleatorio entre 1 y 100. Si este valor es menor o igual a 25, se asigna 0 al elemento *i* del individuo y se asigna 1 en caso contrario.

Este operador fue diseñado con el objetivo de obtener mayor diversidad de individuos dentro del espacio de soluciones de un problema específico, en este caso, la búsqueda de *testores típicos*.

4.2.2 Operador de Alteración

Con el objetivo de mejorar el desempeño del recorrido del espacio de soluciones por parte de las metaheurísticas, se incorporó un nuevo operador al que se denominó *operador de alteración*. El cual, trabaja sobre la población evaluada en cada iteración de la *metaheurística*, donde cada individuo representa un *subconjunto de características* que posiblemente pertenezca al *conjunto de testores* (ver apartado 2.4) por medio de valores binarios, siendo 1 para la presencia de una característica y 0 para la ausencia.

Este operador analiza la vecindad de cada individuo de la población alterando un porcentaje de este, agregando características si el individuo original no es *testor* y eliminando características en caso contrario. El número de cambios o alteraciones depende de un *porcentaje de alteración* establecido como parámetro en las dos *metaheurísticas* con las que se trabajaron (AG y EDA).

En la Figura 27 se describe el algoritmo del *operador de alteración* implementado, el cual, como ya se mencionó, recibe un *porcentaje alteración* que determina el número de alteraciones que se aplicarán al *subconjunto de características* o solución. Una vez verificada la existencia o no existencia de *testores* en la vecindad de cada una de las soluciones, se obtiene un nuevo conjunto compuesto de soluciones mejoradas, nombradas como *conjunto sobreviviente.*

```
BEGIN /* Operador de Alteración*/
    Sobrevivientes[ ] = NULL
    i=1
    REPEAT for l = 1, 2, ... hasta l = Número de soluciones en el conjunto
        Pivote = Solución l
        IF Solución l es testor THEN
            número de pruebas = (número bits 1 en Solución l * porcentaje de alteración) / 100
            REPEAT for m = 1, 2, ... hasta m = número de pruebas
                Seleccionar aleatoriamente un bit 1 del Pivote sin repetir
                Solución l' = Pivote con bit 1 seleccionado cambiado a bit 0.
                IF Solución l' es testor THEN
                    Pivote = Solución l'
                END IF
            END REPEAT
        ELSE
            número de pruebas = (número bits 0 en Solución l * porcentaje de alteración) / 100
            REPEAT for m = 1, 2, ... hasta m = número de pruebas
                Seleccionar aleatoriamente un bit 0 del Pivote sin repetir
                Solución l' = Pivote con bit 0 seleccionado cambiado a bit 1.
                IF Solución l' es testor THEN
                    Pivote = Solución l'
                END IF
            END REPEAT
        END IF
        Sobrevivientes [i] = Pivote
        i = i + 1
    END REPEAT
END
```

Figura 27 *Algoritmo del operador de alteración*

Como se puede observar, cuando el operador tiene como pivote un individuo *testor* trabaja sobre los bits 1, es decir, sobre las características presentes. En cambio, si el pivote tiene un individuo *no testor* trabaja sobre los bits 0 o sobre las características ausentes en el subconjunto.

Por ejemplo, en la Figura 28 se tiene un individuo de 10 características $(x_1, \ldots, x_{10})$ representando un *subconjunto de características* de ocho bits 1 en estado de *testor* (T=1) y un *porcentaje de alteración* de 20%. A partir de esta información, el operador calcula el

número de alteraciones (N) con el cociente del producto del *porcentaje de alteración (p)* y el *número de bits 1 (b)* en el individuo, sobre 100 (ver Ecuación 4).

$$N = \frac{(p * b)}{100} \qquad \textit{Ecuación 4}$$

De acuerdo a lo anterior, se determinan dos pruebas a realizar para el ejemplo.

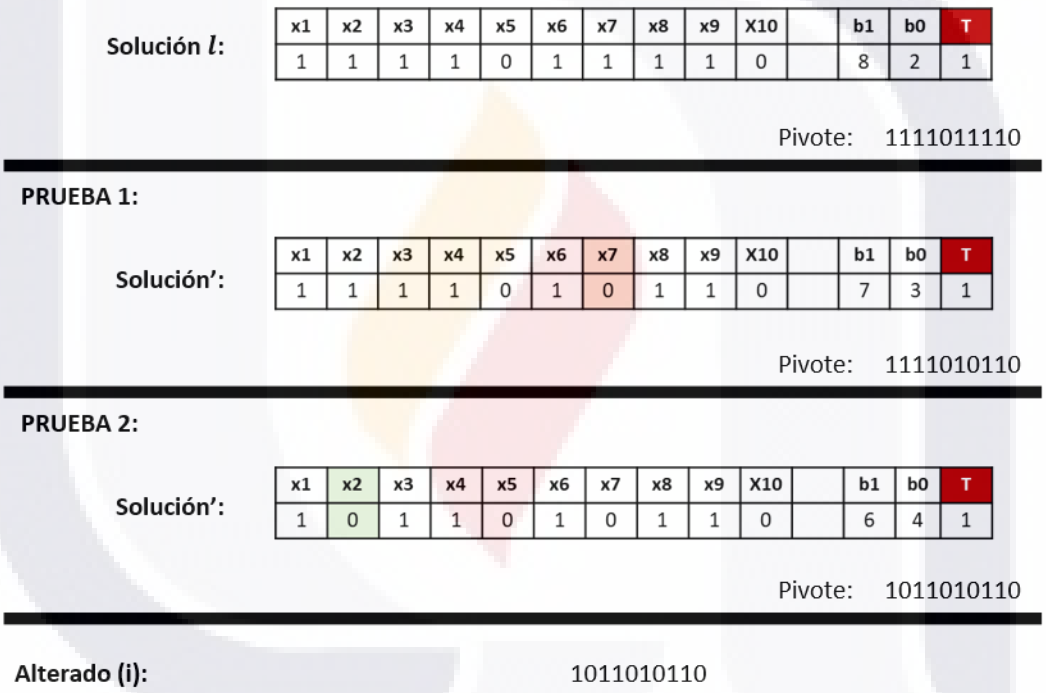


Figura 28 *Ejemplo de alteración 1*

Una vez determinado el número de alteraciones, el operador procede propiamente a cumplir su objetivo. En el ejemplo la *solución l* es asignada como pivote inicial en el cual, el operador selecciona aleatoriamente un bit 1 y lo sustituye con un bit 0, es decir, se elimina una característica del *subconjunto* generando una *solución'*. Lo anterior con la intención de verificar que la *solución'* mantiene el estado de *testor* mediante el proceso de evaluación

descrito en el apartado 2.4 haciendo uso de la *matriz básica* de problema abordado. De acuerdo con el ejemplo de la Figura 28, la *solución'* mantiene el estado de *testor* por lo que el pivote cambia por esta nueva solución para que el operador haga una segunda prueba, esta vez, con este nuevo pivote.

En la segunda prueba, el operador selecciona nuevamente un bit 1 al azar del conjunto restantes de bits 1 y lo sustituye de un bit 0 obteniendo una nueva *solución'*. Como se muestra en la Figura 28, la *solución'* de la segunda prueba fue evaluada nuevamente como testor, por lo que se convierte en el nuevo pivote. Al tratarse de la última prueba prevista por el *porcentaje de alteración*, el pivote actual es la salida del operador y es asignado al *conjunto sobreviviente*.

Como ya se mencionó, el operador de alteración también analiza soluciones no testor, cuya diferencia radica en que se trabaja sobre los bits 0. Por tanto, el número de alteraciones estará dado por el cociente del producto del *porcentaje de alteración* y el número de bits 0 de la solución sobre 100.

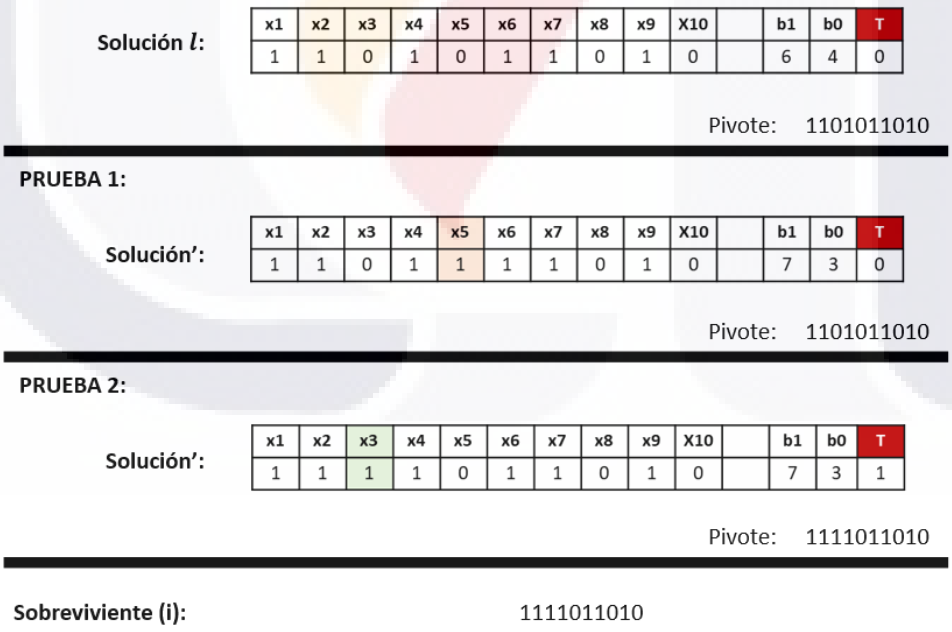


Figura 29 *Ejemplo de alteración 2*

El ejemplo de la Figura 29 supone una *solución no testor* (T=0) de 10 características a evaluar *(x_1, …, x_{10})* con dos pruebas a realizar. Al iniciar el proceso, la *solución l* es asignada como pivote y se selecciona un bit 0 aleatoriamente para reemplazarla por un bit 1, agregando la característica a la solución. Como se observa en la figura, la primera prueba resulta negativa manteniendo el pivote con la *solución l* ignorando por completo la nueva solución. Para la segunda prueba, se selecciona un nuevo bit 0 sin repetir el bit seleccionado anteriormente y se repite una nueva prueba. En este caso, la nueva solución resulta en un *testor*, por tanto, se asigna como pivote. Al ser la última prueba, la solución del pivote es considerada como *sobreviviente* y forma parte del conjunto de salida del operador. En caso de existir una tercera prueba, ésta se realizaría sobre los bits 1 al tener un pivote testor.

En resumen, se considera sobreviviente al último pivote en estado de testor encontrado por el operador, el cual puede ser incluso, el individuo original. Al final del recorrido de la población, la iteración de la metaheurística mostró un conjunto de soluciones sobrevivientes, las cuales tienen un estado de testor. A partir de este conjunto, las *metaheurísticas* continúan su procesamiento por medio de sus operadores habituales como se describe en los apartados siguientes.

4.2.3 Algoritmo Genético Híbrido

La primera *metaheurística* con la que se trabajo fue el *algoritmo genético*, el cual basa su estructura en el *AG simple* descrito en el apartado 2.5.3 por lo que cuenta con los operadores básicos de evaluación, selección, cruce y mutación. Dichos operadores fueron encaminados a la búsqueda de *testores típicos* con el apoyo de los mecanismos descritos en los apartados 4.2.1 y 4.2.2 que hibridan la *metaheurística* mejorando la exploración del espacio de soluciones.

El *AG híbrido* implementado comienza con la creación de una *población inicial* al igual que la aproximación general como se muestra en la Figura 30. En este caso se crea aleatoriamente por medio del mecanismo de *producción de poblaciones iniciales* del apartado 4.2.1 donde cada individuo es un subconjunto de características representado por

medio de cadenas binarias, recordando que 1 significa la presencia de la característica *i*, mientras que el 0 representa la ausencia.

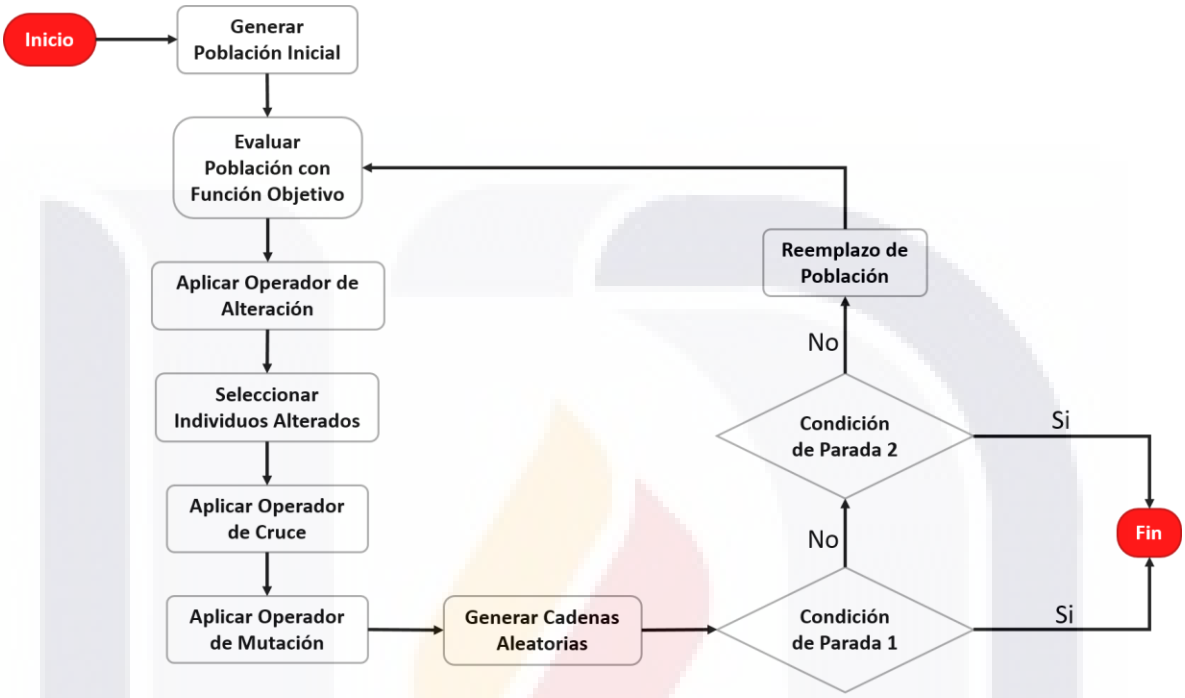


Figura 30 *Algoritmo genético híbrido para la búsqueda de testores típicos*

De acuerdo con la Figura 30, una vez que se cuenta con una población inicial de soluciones, el algoritmo continúa con la evaluación por medio de una función objetivo. En este caso la evaluación es realizada por medio de la *matriz básica* del problema que se está abordando a partir de la cual, se determina qué individuos son *testores*. Esta evaluación es descrita en el apartado 2.4.1 específicamente en la Figura 6. Si la solución es *testor* el *fitness* es alto, mientras que en caso contrario el *fitness* permanece en cero.

Una vez que se conocen los individuos *testores* y *no testores* entra en acción el *operador de alteración* descrito en el apartado 4.2.2. Al finalizar el procesamiento de alteración se contó con un conjunto de sobrevivientes, los cuales tienen estado de *testor*. A partir de este conjunto se procede a la aplicación de los operadores habituales de selección, cruce y mutación.

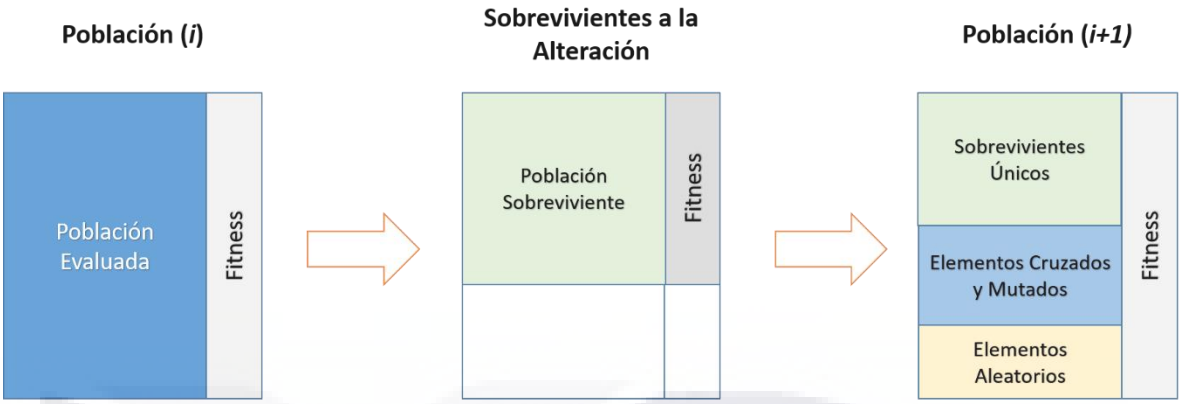


Figura 31 *Creación de nuevas poblaciones en el AG híbrido*

De acuerdo con lo descrito en la teoría base del *algoritmo genético* en el apartado 2.5.3, los operadores de selección, cruce y mutación tienen el objetivo de crear una nueva población que tendrá soluciones aún más cercanas al óptimo. En este caso, una población *i+1* se conforma de tres partes; el conjunto de *sobrevivientes únicos*, un conjunto de soluciones obtenidas a partir selección, cruce y mutación, y un conjunto de soluciones aleatorias.

La presencia de *sobrevivientes únicos* es una implementación de *elitismo*, el cual consiste en copiar directamente la(s) mejor(es) solución(es) a la nueva población para evitar que la mejor solución se pierda entre las generaciones permitiendo la convergencia del algoritmo. Lo anterior fue demostrado por Günter Rudolph en 1994 en [104].

Como ya se mencionó, el segundo conjunto que conforma una nueva población es el generado por los operadores canónicos del AG. Primeramente, se seleccionan los individuos a cruzar por medio de una ruleta en la que se da preferencia a aquellas soluciones mejor evaluadas dándoles mayor probabilidad de ser seleccionadas. En cuanto al cruce, se utilizó el método de cruce basado en un punto descrito en el apartado 2.5.3 a partir del cual, se obtienen dos descendientes por cada par de individuos seleccionados. En este caso, no se manejó una probabilidad de cruzamiento, por lo que cada par de individuos seleccionados fue cruzado para obtener descendencia. Para la aplicación de la mutación se seleccionaba aleatoriamente un bit del individuo y se cambiaba el bit contrario, por ejemplo, si el bit seleccionado era un 1, el proceso de mutación le asignaba un bit 0. Este proceso decidía por

ruleta que podía tomar valores entre 1 y 100, si el valor resultante era menor o igual al *porcentaje de mutación*, este proceso se aplicaba.

La tercera parte de la población se generó por medio del mecanismo de producción de poblaciones iniciales descrito en el apartado 4.2.1. Este conjunto de individuos aleatorios permitió al algoritmo asegurar mayor diversidad en la exploración del espacio de soluciones.

De acuerdo con el diagrama de la Figura 30, el AG híbrido implementado continua con la evaluación de las dos *condiciones de paro*. La primera *condición de paro* es el número de *testores típicos* encontrados, es decir, si el algoritmo ha encontrado el total de los *testores típicos* esperados éste se detiene. Para determinar si la condición se cumple, el algoritmo hace uso del archivo de *testores típicos* encontrados por el *método exhaustivo* con el cual, el algoritmo conoce el número de *testores típicos* que debe encontrar y, además, le permite comparar cada individuo del conjunto de *sobrevivientes únicos* con los que se encuentran en dicho archivo para determinar si es típico o no. Cuando un *individuo testor* resulta ser *típico* lo almacena en su propio archivo de *testores típicos*. Cuando el número de testores típicos alcanza el esperado, el algoritmo se detiene.

La segunda condición consiste simplemente en completar un número máximo de iteraciones, el cual es otorgado al algoritmo como parámetro como se describe en la metodología del apartado 3. Entonces, cuando el algoritmo ha cubierto este número de iteraciones se detiene. Si ninguna de las condiciones de paro es satisfecha la población nueva reemplaza a la anterior y continua la exploración del espacio de soluciones.

Como se recordará este algoritmo es ejecutado un número determinado de ocasiones según el diseño de experimento expuesto en el apartado 3.2.4, donde cada combinación de valores de parámetros fue replicada 30 veces. Al final de cada réplica se registra un resumen de desempeño el cual contiene un identificador, el porcentaje de testores encontrados, el porcentaje de testores, el número de iteraciones realizadas y el tiempo que le llevo terminar. Como también se describe en el apartado 3.2.4, estos datos fueron utilizados en un estudio estadístico que permitió conocer aquellos valores de parámetros que permitieron un mejor desempeño en la búsqueda de testores típicos.

4.3.4 Algoritmo Híbrido de Estimación de la Distribución

La segunda metaheurística que fue hibridada fue el *algoritmo de estimación de la distribución (EDA)*, la cual, al igual que el *AG*, es una *metaheurística* evolutiva que trabaja con poblaciones de soluciones, cuya aproximación general es descrita en el apartado 2.5.4. A su vez, es una *metaheurística* muy flexible, pues permite la integración de nuevos mecanismos que benefician la exploración del espacio de soluciones.

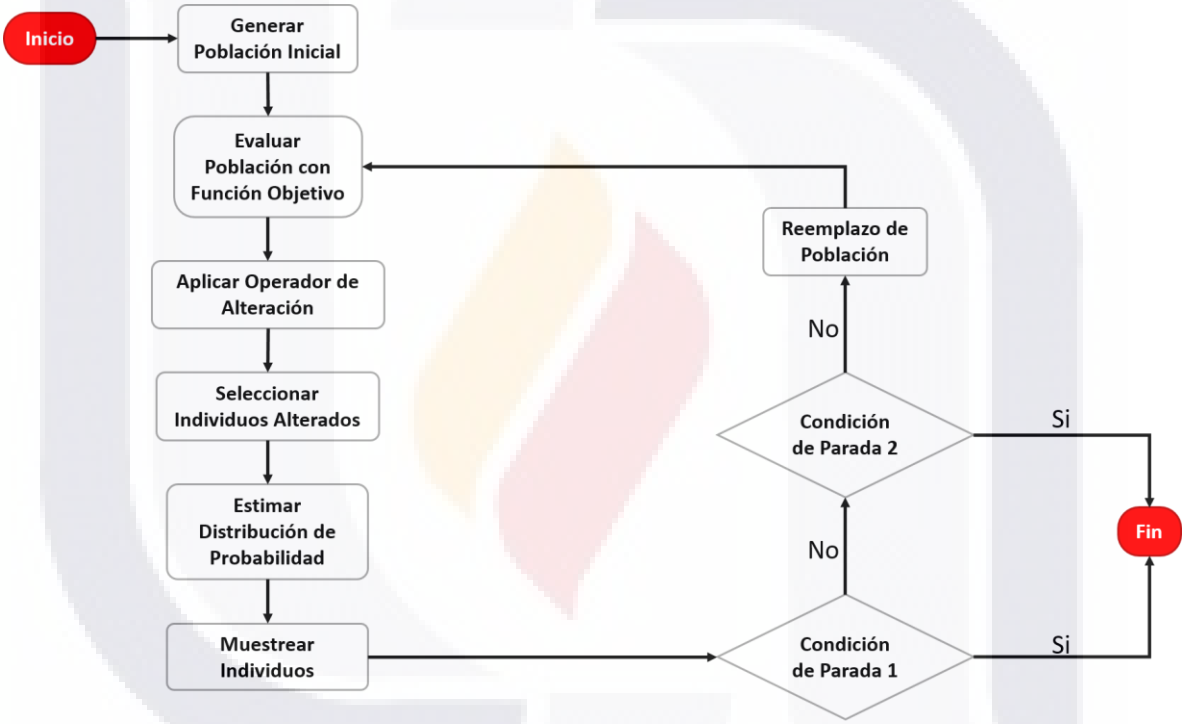


Figura 32 *Algoritmo de estimación de distribución para la búsqueda de testores*

En la Figura 32 muestra el diagrama de flujo del *EDA híbrido* orientado a la búsqueda de testores típicos. Al igual que el *AG*, inicial con una población inicial que, en este caso, se utiliza el mecanismo de producción de poblaciones inicial descrito en el apartado 4.2.1. La cantidad de individuos a generar es determinada como parámetro de acuerdo a los descrito en el apartado 3.2.5 sobre la afinación de la *metaheurística* como parte de la metodología que se siguió.

Una vez que se cuenta con la población inicial, el algoritmo la evalúa con la función objetivo. En este caso es necesario el uso de la *matriz básica* del problema abordado, con la cual se determina qué individuos presentan un estado de *testor*. El proceso de evaluación se describe en la Figura 6 del apartado 2.4.1. Si la solución es *testor* el fitness del individuo es alto, mientras que en otro caso el fitness permanece en cero.

A continuación, el algoritmo aplica el *operador de alteración* del apartado 4.4.2 como se muestra en la Figura 33, el cual provee un conjunto de *individuos sobrevivientes* en estado de *testor* provenientes tanto de soluciones *testor* como de soluciones *no testor*.

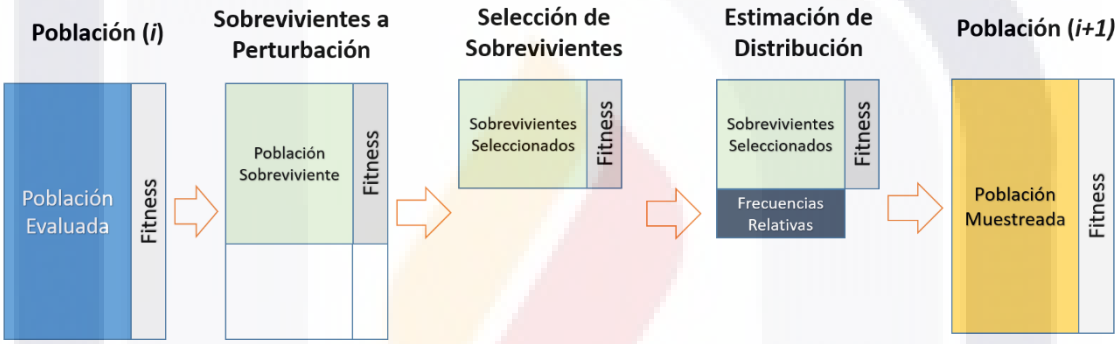


Figura 33 *Creación de nuevas poblaciones en el EDA híbrido*

Como se muestra en la Figura 33, A partir de este conjunto, se continua con la *selección de individuos*. Para ello, se utilizó una ruleta aleatoria, la cual provee de mayor probabilidad de selección a aquellos individuos con mayor fitness, es decir, los mejor evaluados. La cantidad de individuos a seleccionar está dada por el *porcentaje de selección* obtenido como parámetro, de acuerdo con lo descrito en el apartado 3.2.5. A partir de los *individuos seleccionados*, se realizó propiamente la *estimación de distribución* $p_l(x)$ calculando la probabilidad de obtener un bit 1 en cada característica por medio de la *función de distribución* sin dependencias *UMDA*. Para ello, bastó con calcular las frecuencias relativas para cada característica para determinar la probabilidad de obtener un bit 1 en la característica *i*. A partir de la función de distribución $p_l(x)$ obtenida se realiza un proceso de *muestro* donde se obtienen nuevos individuos que conforman la nueva población $(i+1)$.

Para el *EDA híbrido* también se contó con dos *condiciones de paro* como se muestra en la Figura 32. La primera de ellas se encargaba de comprobar que el algoritmo había encontrado el total de *testores típicos* con apoyo del archivo de *testores típicos* encontrados por el *algoritmo exhaustivo* comparando cada individuo del *conjunto de sobrevivientes* con los elementos de dicho archivo. Cada vez que se determina que un *testor* era *típico*, éste fue almacenado por la metaheurística en su propio archivo sin duplicados. Una vez que el archivo de *testores típicos* de la *metaheurística* y el del *algoritmo exhaustivo* coincidían, el algoritmo detenía su búsqueda.

La segunda condición se encarga de comparar el número de iteraciones realizadas hasta un determinado tiempo. En caso de que se hubiera alcanzado un número máximo de iteraciones, la *metaheurística* se detenía independientemente del estado de la búsqueda.

En el caso de que ninguna de las condiciones se cumpliera, la población nueva obtenida reemplazaba a la anterior para comenzar una nueva iteración hasta lograr el cumplimiento de alguna de las *condiciones de paro*.

Cada configuración de valores en los parámetros fue replicada 30 veces (ver apartado 3.2.5), cuyos resultados fueron analizados estadísticamente para determinar cuál de las configuraciones se desempeñó mejor. Para ello, la aplicación contó un registro con el resumen de cada experimento/réplica en el que se visualiza un identificados, el porcentaje de testores encontrados, el porcentaje de testores típicos encontrados, el número de iteraciones realizadas y el tiempo que llevo terminar el experimento.

Como se determinó en la metodología, se llevó a cabo un diseño de experimentos factorial para las *metaheurísticas hibridadas*. Los resultados obtenidos de este diseño fueron sometidos a un *análisis estadístico* que permitió conocer las configuraciones para las cuales, cada una de las *metaheurísticas*, presentaban mejor desempeño en cada uno de los escenarios de experimentación, los cuales son, como se recordará, cáncer de mama y hemofilia. A continuación, se exponen dichos resultados para cada experimento realizado.

4.2.4 Afinación de Metaheurísticas para Cáncer de Mama

Como se mencionó en los apartados en el apartado 3.2.4 y 3.2.5, la *metodología* que se siguió involucró un proceso de afinación de las *metaheurísticas* empleadas: *AG* y *EDA*, por lo que en este apartado se exponen los resultados de dicho proceso orientado al escenario de experimentación del *cáncer de mama* descrito en el apartado 3.3.1 bajo las siguientes variables de salida:

- Porcentaje de testores encontrados (variable continua)
- Porcentaje de testores típicos encontrados (variable continua)
- Número de iteraciones realizadas (variable continua)
- Tiempo utilizado (variable continua)

Por otra parte, se exponen los resultados de la contratación de metaheurísticas (ver apartado 3.2.6) a partir de la cual, se determinó cuál metaheurística presentó mejor desempeño en el contexto del escenario de experimentación basado en su naturaleza y la influencia de las hibridaciones realizadas.

Afinación de Algoritmo Genético

Como se recordará, el apartado 3.2.4 se expuso el diseño de experimentos factorial a partir del cual cada parámetro del *algoritmo genético* recibió un conjunto de valores con los cuales se ejecutaron todas las posibles combinaciones de los mismos. Para el caso de cáncer de mama con el AG híbrido, se definieron los valores de la Tabla 22 que resultaron en la realización de 400 experimentos diferentes.

PARAMETRO	VALORES
1. Tamaño de población	10, 20, 40, 60, 80
2. Porcentaje de alteración	10, 15, 20, 25, 30
3. Probabilidad de mutación	5, 8, 10, 12
4. Número de iteraciones	50, 75, 100, 150

Tabla 22 *Parámetros utilizados para la Afinación del AG en el contexto de cáncer de mama*

De acuerdo con la metodología, cada uno de los experimentos fue replicado 30 veces formando 400 grupos independientes, los cuales fueron sometidos a una prueba estadística como se describe a continuación.

Para comenzar, se aplicó una prueba de Levene que permitió determinar que el conjunto de experimentos son *no homocedásticas*, existiendo diferencia en la varianza de las variables continuas (ver Anexo B.1). Por otro lado, los datos no siguen una *distribución normal* de acuerdo con una prueba de Kolmogorov Smirnov como se observa en el Anexo B.2 por lo que se siguió con pruebas no paramétricas.

De acuerdo con lo anterior, se realizó una prueba de Kruskal Wallis la cual determinó la existencia de diferencia significativa entre los 400 grupos estudiados en las cuatro variables continuas (ver Anexo B.3).

La variable de salida con mayor prioridad para este estudio fue el *porcentaje de testores típicos*. Variable en la cual, los 400 experimentos estudiados encontraron en promedio, un mínimo del 96.66% y un máximo del 100% haciéndolos estadísticamente iguales respecto a esta variable, de acuerdo con la prueba de U-Mann Whitney. Por lo tanto, se concluye que el algoritmo es inmune a los valores definidos en la Tabla 22 para el contexto del cáncer de mama. En otras palabras, los valores definidos no influyen estadísticamente en el desempeño del algoritmo. En la Tabla 23 se muestran los mejores resultados de experimentos con porcentajes de testores típicos mayores al 96.66%, mientras que en la 24 se muestran sus parámetros respectivamente.

Debido a la igualdad estadística existente, se realizó una nueva prueba que determinó que los experimentos que encontraron en promedio entre el 96.6% y el 100% de los *testores* son estadísticamente similares. A partir de este resultado se observó que para obtener mejores resultados el algoritmo trabaja mejor con poblaciones de 10, 40, 60 y 80 individuos, mientras mantiene su inmunidad a los valores del resto de los parámetros.

De manera más estricta, se extraen los experimentos 4, 9, 350, 395, 183, 170 y 354 de la Tabla 23, cuyos parámetros se exponen en la Tabla 24, a partir de los cuales se observó que el algoritmo se ve afectado por el *tamaño de población* al mejorar su desempeño con tamaños

de 10, 40 y 80 individuos. En cuanto al *porcentaje de alteración*, se tienen solo los valores 10, 15, 20 y 30 para beneficiar el desempeño del AG. Por otro lado, la proba*bilidad de mutación* y el *número de iteraciones* no influyen en el algoritmo al presentarse todos los valores posibles manteniendo sobresalientes a los experimentos citados.

	Variables de Salida			
#Exp	% Testores	%Testores Típicos	Iteraciones Realizadas	Tiempo (s)
4	100	100	11.27	0.1687
9	98.89	100	12.33	0.1505
350	98.89	100	2.33	0.2006
395	97.77	100	1.9	0.1734
183	97.77	100	2.93	0.1114
170	96.66	100	3.37	0.1235
354	96.66	100	2.066	0.1743
65	85.56	98.33	14.7	0.1875
50	92.22	98.33	16.27	0.2484
17	86.67	96.67	15.9	0.2443
118	88.89	100	4.3	0.0917

Tabla 23 *Salidas de los estadísticamente mejores experimentos AG aplicados al contexto de cáncer de mama*

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	Prob. de Mutación	Número de Iteraciones
4	10	10	5	150
9	10	10	10	50
350	80	15	12	75
395	80	30	10	100
183	40	15	8	100
170	40	10	10	75
354	80	20	5	75
65	10	30	5	100
50	10	25	5	75
17	10	15	5	50
118	20	20	8	75

Tabla 24 *Parámetros de los mejores experimentos AG aplicados al contexto de cáncer de mama*

A partir del grupo de experimentos que resultaron iguales respecto a las variables de porcentaje de testores típicos y porcentaje de testores, los experimentos 170, 183 y 395 difieren estadísticamente del resto respecto al *tiempo* y al *número de iteraciones* realizadas. Siendo los grupos 170 y 183 iguales entre ellos reduciendo el *tiempo* de ejecución y, a su vez, distintos al 395 el cual, minimiza el *número de iteraciones* a realizar.

Empíricamente hablando, los experimentos con mejor desempeño cuyos parámetros se muestran en la Tabla 24, son:

- El experimento 4, al maximizar el *porcentaje de testores típicos* y el *porcentaje de testores*.
- El experimento 183, al minimizar el *tiempo* de ejecución.
- El experimento 395, al minimizar el *número de iteraciones* realizadas.
- El experimento 118, al minimizar el *tiempo* de ejecución, pero difiere estadística y empíricamente de los grupos anteriores respecto al *porcentaje de testores* encontrados.

Por el contrario, el experimento 17 obtuvo el desempeño menos deseado al encontrar menor *porcentaje de testores típicos* con mayor cantidad de *iteraciones* realizadas, cuyos parámetros se pueden observar en la Tabla 24.

A continuación, se muestra la afinación del algoritmo de estimación de la distribución hibridado a partir del cual, y con apoyo de la afinación del algoritmo genético de este apartado, se describan los resultados obtenidos de un proceso de contrastación con el que se determinó qué algoritmo responde con mejor desempeño en el contexto del escenario de experimentación del cáncer de mama.

Afinación de Algoritmo de Estimación de la Distribución

De acuerdo con la metodología del apartado 3, se realizó un proceso de afinación del *algoritmo de estimación de la distribución hibridado* (descrito en el apartado 4.3.4) para lo cual, se aplicó un diseño de experimentos factorial definiendo los valores de parámetros de la Tabla 25, los cuales resultaron en la realización de 400 experimentos diferentes.

PARAMETRO	VALORES
1. Tamaño de población	10, 30, 40, 60, 80
2. Porcentaje de alteración	10, 15, 20, 25, 30
3. Porcentaje de selección	20, 30, 40, 50
4. Número de iteraciones	50, 75, 100, 150

Tabla 25 *Parámetros utilizados para la afinación del EDA en el contexto de cáncer de mama*

De acuerdo con la *metodología* y el proceso de *afinación*, cada uno de los experimentos fue replicado 30 veces, de manera que formaron 400 grupos independientes a partir de los cuales se realizaron pruebas para afinar la *metaheurística* y encontrar aquella combinación de valores que permiten un mejor desempeño en la *búsqueda de testores típicos*.

La primera prueba realizada fue una prueba de Levene para determinar la *homogeneidad de varianzas*, a partir de la cual se concluyó el conjunto de experimentos no es *homocedástico* existiendo diferencia significativa en la varianza de las variables de entrada (ver Anexo C.1). A continuación, se determinó que los datos no siguen una *distribución normal* de acuerdo con la aplicación de una prueba de Kolmogorov Smirnov expuesta en el anexo C.2. Por lo tanto, se aplicó la prueba de Kruskal Wallis que concluyó en que no todos los grupos estudiados son iguales estadísticamente (ver Anexo C.3).

A diferencia del AG, los experimentos EDA encontraron entre el 88.33% y el 100% de los *testores típicos* de los cuales, aquellos que encontraron entre el 95% y el 100% resultaron ser estadísticamente iguales de acuerdo con la aplicación de pruebas U de Mann Whitney. A partir de este resultado y al observar la presencia de todos los valores de cada parámetro de

la Tabla 25 se concluyó que dichos valores no afectan el desempeño del algoritmo aplicado al contexto del cáncer de mama, tal como sucede con el AG. Por otro lado, extrayendo

Debido a lo anterior se procedió a realizar nuevas pruebas U de Mann Whitney respecto al *porcentaje de testores* encontrados en los mejores experimentos de la prueba anterior. Dichas pruebas concluyeron en que los experimentos que encontraron entre el 96.67% y el 100% de los *testores* presentan igualdad estadística, en los que se observa de manera general que el algoritmo mantiene su inmunidad a los valores definidos en los parámetros. En cambio, al observar empíricamente los experimentos numerados del 203 al 349 de la Tabla 26 y 27, se determina el algoritmo se ve influenciado por el tamaño de población al mejorar su desempeño con 40, 60 y 80 individuos. Por su parte el porcentaje de alteración afecta positivamente el resultado con 10, 15 y 20%. De igual manera, el EDA se desempeña mejor con 50, 100 y 150 iteraciones. En cambio, el algoritmo se mantiene inmune al número de iteraciones definidas.

	Variables de Salida			
#Exp	% Testores	%Testores Típicos	Iteraciones Realizadas	Tiempo (s)
203	100	100	4.6	0.1287
255	100	100	3.96	0.1479
256	98.89	100	3.16	0.1077
343	98.89	100	2.3	0.1229
325	97.77	100	2.73	0.1214
336	97.77	100	2.67	0.1390
243	96.66	100	3.36	0.1224
349	96.66	100	2.67	0.1406
38	96.66	98.33	13.6	0.1625
105	96.66	98.33	11.5	0.2937
10	96.66	96.67	19.96	0.2243
17	90	88.33	22.03	0.2503

Tabla 26 *Resultados de la experimentación EDA aplicado a cáncer de mama*

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	% de Selección	Número de Iteraciones
203	40	20	40	100
255	60	10	50	100
256	60	10	50	150
343	80	15	30	100
325	80	10	30	50
336	80	10	50	150
243	60	10	20	100
349	80	15	50	50
38	10	20	30	75
105	30	15	40	50
10	10	10	40	75
17	10	15	20	50

Tabla 27 *Experimentos EDA que encontraron entre el 96.67% y el 100% de testores típicos*

La siguiente variable en prioridad es el *tiempo* que le tomó a los experimentos terminar el algoritmo. En esta variable los experimentos 256, 325, 243 se distinguen estadísticamente del resto. Como se puede observar en la Tabla 26, estos experimentos minimizan el *tiempo* de ejecución. Por tanto, el algoritmo se nota más influenciado por los parámetros ingresados pues al priorizar el *tiempo*, el *tamaño de población* se reduce a 60 u 80 individuos; un *porcentaje de alteración* de 10%, significando que el algoritmo se ve altamente afectado por esta variable; el *porcentaje de selección* se reduce a los valores 20, 30 y 50%; y, finalmente, el *número de iteraciones* queda en 50, 100 y 150.

A su vez, los experimentos 256, 325 y 243 mantienen su igualdad estadística respecto al *número de iteraciones* realizadas. Por lo tanto, el experimento 256 cuyos parámetros se pueden consultar en la Tabla 27, es el que destaca estadísticamente al maximizar el *porcentaje de testores* y el *porcentaje de testores típicos*, así como de minimizar el *tiempo* y el *número de iteraciones* realizadas.

De acuerdo con la Tabla 26 y desde el punto de vista empírico los experimentos con mejor desempeño cuyos parámetros se muestran en la Tabla 27, son:

- El experimento 203, al maximizar el porcentaje de *testores típicos* y el *porcentaje de testores* encontrados en un *tiempo* de 0.1287 segundos.
- El experimento 256, al maximizar el *porcentaje de testores típicos* encontrados y minimizando el *tiempo* de ejecución requerido.
- El experimento 343, al maximizar el *porcentaje de testores típicos* encontrados y minimizando el *número de iteraciones* a realizar.

Finalmente, se detectó un grupo 17 cuya combinación de valores de parámetros solo logró encontrar en promedio el 88.33% de los *testores típicos*, con un promedio del 90% de los *testores* y, además, le tomó una cantidad considerable de iteraciones en promedio como se observa en la Tabla 26. Debido a que es el grupo con el desempeño más bajo se exponen sus parámetros en la Tabla 28 con la intención de evitar su uso en el contexto del escenario del cáncer de mama.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	% de Selección	Número de Iteraciones
17	10	15	20	50

Tabla 28 *Parámetros con bajo desempeño en la aplicación EDA en el contexto de cáncer de mama*

Contrastación de Metaheurísticas

En este apartado se describen los resultados de un proceso en el que se contrastan las dos metaheurísticas aplicadas al escenario de investigación del *cáncer de mama* a partir de los mejores parámetros encontrados en sus respectivas afinaciones descritas dentro de este apartado 4.3.4, a partir del punto de vista empírico.

En la Tabla 29 a continuación, se exponen los parámetros del experimento 4 cuyo desempeño fue el mejor en cuanto a la ejecución del algoritmo genético en el contexto del cáncer de mama.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	Prob. De Mutación	Número de Iteraciones
4	10	10	5	150

Tabla 29 *Mejores parámetros obtenidos empíricamente de la afinación del AG aplicado a cáncer de mama*

Para el caso del *algoritmo de estimación de la distribución* se seleccionó el conjunto de parámetros del experimento 203, el cual tuvo empíricamente el mejor desempeño en la ejecución del diseño de experimentos factorial. Los parámetros de este experimento se presentan a continuación en la Tabla 30.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	% de Selección	Número de Iteraciones
203	40	20	40	100

Tabla 30 *Mejores parámetros obtenidos empíricamente de la afinación del EDA aplicado a cáncer de mama*

Para contrastar ambas *metaheurísticas* se realizaron nuevas ejecuciones independientes de los experimentos de las Tablas 29 y 30 para después, realizar pruebas estadísticas que determinaron cuál de éstas *metaheurísticas* responde mejor a la *búsqueda de testores típicos* en el escenario de investigación del *cáncer de mama*.

Para encontrar la mejor *metaheurística* cada experimento fue replicado 30 veces formado dos grupos independientes los cuales fueron sometidos a una prueba de Levene, que puede consultarse en el Anexo D.1, cuyo resultado indica que el AG y el EDA son homocedásticos respecto al *porcentaje de testores encontrados*, mientras que ambos encuentran el 100% de los *testores típicos* en cada una de sus réplicas. En cambio, las *metaheurísticas* son homocedáticos respecto al *tiempo utilizado* y el *número de iteraciones realizadas*.

A continuación, se determinó que los algoritmos no siguen una distribución normal por medio de una prueba de Kolmogorov Smirnov (Anexo D.2) por lo tanto, se continuó con pruebas estadísticas no paramétricas.

En el Anexo D.3 se puede consultar la prueba Kruskal Wallis en la que se descubrió que los experimentos presentan diferencias significativas en respecto al *tiempo utilizado* y el *número de iteraciones realizadas*. Por el contrario, no existe diferencia significativa respecto en los grupos de réplicas respecto al *porcentaje de testores encontrados* y el *porcentaje de testores típicos encontrados*.

		Variables de Salida			
Meta	**#Exp**	% Testores	%Testores Típicos	Iteraciones Realizadas	Tiempo (s)
AG	**4**	94.45	100	14.03	0.9909
EDA	**203**	97.78	100	4.23	0.2281

Tabla 31 *Resultados promedio en los experimentos seleccionados para contrastación en el contexto de cáncer de mama*

De acuerdo con una Prueba de U de Mann Whitney ambos experimentos son estadísticamente iguales respecto al porcentaje de *testores* y respecto al por*centaje de testores típicos*. En cambio, ambos experimentos difieren estadísticamente respecto al *tiempo* promedio en el que terminaron su respectivo algoritmo y el *número de iteraciones* que realizaron, destacando así el EDA, al minimizar dichas variables. Finalmente, destaca el EDA sobre el AG tanto del punto de vista estadístico y desde el punto de vista empírico como se muestra en la Tabla 31, al encontrar un porcentaje mayor de *testores*, en el menor *tiempo* y con el menor número de *iteraciones*.

4.2.5 Afinación de Metaheurísticas para Hemofilia

Según con lo establecido por la metodología en el apartado 3, se realizó un proceso de afinación de las metaheurísticas hibridadas: AG (apartado 4.2.3) y EDA (apartado 4.2.4), específicamente para el contexto del escenario de experimentación de hemofilia, el cual es descrito en el apartado 3.3.2. Como se ha mencionado, el proceso de afinación involucró un diseño factorial de experimentos del cual, se obtuvieron las siguientes variables de salida:

- Porcentaje de testores encontrados (variable continua)
- Porcentaje de testores típicos encontrados (variable continua)
- Número de iteraciones realizadas (variable continua)
- Tiempo utilizado (variable continua)

A partir de estas variables se determinaron aquellos valores que permitieron obtener un mejor desempeño por parte de las metaheurísticas por medio de una prueba estadística. Además, se realizó una contrastación entre el desempeño del AG y el EDA para determinar aquel que tiene mejor comportamiento en el contexto de hemofilia.

Afinación de Algoritmo Genético

El diseño factorial de experimentos descrito en el apartado 3.24 permitió que cada parámetro se le asignara un conjunto de valores con los cuales, se ejecutaron múltiples experimentos cubriendo todas las posibles combinaciones de valores. Para este caso, se definieron los valores de la Tabla 32 que permitieron la realización de 400 experimentos. Dichos experimentos fueron replicados un total de 30 veces formando así, grupos independientes con los que se realizó la afinación como se describe a continuación.

PARAMETRO	VALORES
1. Tamaño de población	700, 750, 800, 850, 900
2. Porcentaje de alteración	10, 15, 20, 25, 30
3. Probabilidad de mutación	5, 8, 10, 12
4. Número de iteraciones	10, 15 ,20, 25

Tabla 32 *Parámetros para afinación del AG en el contexto de hemofilia*

Los grupos de experimentos fueron estudiados por medio de una prueba de Levene mediante la cual, se determinó que *no son homocedásticos* al existir diferencia significativa en la varianza de las variables de entrada (ver Anexo E.1). Además, se concluyó que tampoco siguen una distribución normal de acuerdo con la aplicación de una prueba de Kolmogorov Smirnov, la cual puede ser consultada en el Anexo E.2.

De acuerdo con lo anterior, se optó por una prueba no paramétrica de Kruskal Wallis con la cual se descubrió que la existencia de diferencia significativa desde el punto de vista del *porcentaje de testores encontrados*, el *tiempo utilizado* y el *número de iteraciones realizadas*. En cambio, no existe diferencia significativa en el *porcentaje de testores típicos encontrados* (ver Anexo E.3) debido a que los 400 grupos de réplicas de experimentos logran encontrar el conjunto completo de *testores típicos*.

Debido a lo anterior, se aplicaron pruebas U de Mann Whiney respecto a la capacidad de los experimentos para encontrar testores, los cuales fueron capaces de encontrar entre el 41.7 y el 76.12% de los testores. Dicha prueba concluyó en que aquellos experimentos que localizaron entre el 73.68 y el 76.12% de los testores son estadísticamente iguales, los cuales se muestran en la Tabla 33. De acuerdo con este resultado se observó que el *tamaño de población* tiene efecto en el desempeño del algoritmo al presentar mejores resultados con 800, 850 y 900 individuos. Por su parte, el *porcentaje de perturbación* tiene especial influencia al solo optar por un 10% de perturbación para asegurar un buen desempeño. En caso contrario, el EDA es inmune a las variables *probabilidad de mutación* y el *número de iteraciones* al mantener un buen desempeño independientemente del valor seleccionado en dichas variables. Ver Tabla 34.

Al aplicar una nueva prueba U de Mann Whitney respecto al tiempo a los experimentos con mejor desempeño en porcentaje de testores, se encontró que aquellos experimentos con un tiempo entre 2.7210 y 2.9668 segundos son estadísticamente iguales. A pesar de lo anterior, las variables de entrada mantienen su influencia sobre el algoritmo.

En cuanto al número de iteraciones, una prueba de U de Mann Whitney encontró que les tomó entre 2.33 y 2.66 presentan igualdad estadística. De esta manera, el algoritmo se vuelve más sensible al *tamaño de población* desempeñándose mejor con 850 y 900 individuos. En

cuando al *porcentaje de alteración*, el algoritmo se mantiene especialmente sensible al solo requerir de un 10% para mejorar su desempeño. Además, el AG no requiere más de 20 *iteraciones*. Por su parte, la *probabilidad de mutación* no logra influir en el algoritmo al aceptar cualquier valor definido sin afectar los resultados.

	Variables de Salida			
#Exp	% Testores	%Testores Típicos	Iteraciones Realizadas	Tiempo (s)
253	76.12	100	2.73	3.0471
325	75.77	100	2.5	2.9502
254	74.99	100	2.66	2.9366
321	75.43	100	2.46	2.8564
323	75.12	100	2.5	2.9513
334	74.95	100	2.43	3.0310
330	74.56	100	2.33	2.7210
161	74.21	100	2.86	2.9668
247	73.89	100	2.53	2.8240
162	73.75	100	2.76	2.8619
168	73.68	100	2.83	2.9376
242	73.57	100	2.5	2.7521

Tabla 33 *Resultados de experimentos EDA aplicado a hemofilia*

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	Prob. De Mutación	Número de Iteraciones
253	850	10	12	10
325	900	10	8	10
254	850	10	12	15
321	900	10	5	10
323	900	10	5	20
334	900	10	12	15
330	900	10	10	15
161	800	10	5	10
247	850	10	8	20
162	800	10	5	15
168	800	10	8	25
242	850	10	5	15

Tabla 34 *Experimentos AG que encontraron entre el 73.57 y el 76.12% de los testores*

Empíricamente, los experimentos con mejor desempeño se muestran a continuación (los parámetros correspondientes pueden consultarse en la tabla 34):

- El experimento 253, al maximizar el *porcentaje de testores* encontrados.
- El experimento 330, al minimizar el *tiempo* requerido y el *número de iteraciones* realizadas, aunque con un *porcentaje de testores* menor al 253 recordando que ambos son estadísticamente iguales respecto a dicha variable.

Finalmente, se expone el experimento 68 con el desempeño más bajo respecto al *porcentaje de testores* encontrados con un promedio del 41.7% a pasar de haber encontrado el 100% de los *testores típicos*. Los parámetros asignados a este experimento se presentan en la Tabla 35 a continuación.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	Prob. De Mutación	Número de Iteraciones
68	700	30	5	25

Tabla 35 *Experimento con el desempeño más bajo en la ejecución del AG en Hemofilia*

A continuación, se describe la afinación realizada al algoritmo de estimación de la distribución aplicado al contexto de la hemofilia. Una vez explicada, se procede a describir la contrastación entre ambas metaheurísticas determinado cuál se desempeña mejor en dicho contexto.

Afinación de Algoritmo de Estimación de la Distribución

Para la realización de la afinación del algoritmo de estimación de la distribución se definieron los conjuntos de valores en cada parámetro los cuales se presentan en la Tabla 36 a continuación. Como ya se ha mencionado, se siguió un diseño factorial de experimentos de manera que se ejecutaron todas las posibles combinaciones de experimentos. Así pues, en

este caso resultó en la realización de 400 experimentos los cuales fueron replicados 30 veces que fueron utilizados para determinar estadísticamente los grupos con mejor desempeño.

PARAMETRO	VALORES
1. Tamaño de población	700, 750, 800, 850, 900
2. Porcentaje de alteración	10, 15, 20, 25, 30
3. Porcentaje de selección	20, 30, 40, 50
4. Número de iteraciones	10, 15 ,20, 25

Tabla 36 *Parámetros para afinación del EDA en el contexto de hemofilia*

A partir de una prueba de Levene (Anexo F.1) resultan ser no homocedásticos existiendo diferencia entre las varianzas de los grupos respecto al *porcentaje de testores*, el *tiempo requerido* y el *número de iteraciones* realizadas. Por su parte, no hay diferencias en el *porcentaje de testores típicos* debido a que el 100% de los grupos encuentra el conjunto en su totalidad. Por su parte, una prueba de Kolmogorov Smirnov, expuesta en el Anexo F.2, los grupos no siguen una distribución normal por lo que se prosiguió con estadística no paramétrica como se describe a continuación.

De acuerdo con lo anterior, se determinó que existe diferencia significativa en algunos grupos de experimentos respecto al *porcentaje de testores*, el *tiempo* y el *número de iteraciones realizadas* por medio de una prueba de Kruskal Wallis, la cual puede consultarse en el Anexo F.3.

Para determinar estadísticamente los grupos con el mejor desempeño se realizaron pruebas U de Mann Whitney dando prioridad al *porcentaje de testores* debido a que, como ya se mencionó, el algoritmo fue capaz de encontrar el 100% de los *testores típicos* en los 400 casos observados. De acuerdo con dichas pruebas, y como se muestra en la Tabla 33, los grupos que obtuvieron entre 80.85% y el 83.3% son estadísticamente iguales en su desempeño en *búsqueda de testores*. A partir de lo anterior, se observa que el EDA es sensible al *tamaño de la población* (ver Tabla 38), pues para obtener un mejor desempeño requiere de 800, 850 y 900 individuos. Por su parte, el algoritmo es especialmente sensible al *porcentaje de selección* pues es estrictamente necesario de un 10% para asegurar un buen

desempeño. Por el contrario, el *porcentaje de selección* y *numero de iteraciones* no influyen en el desempeño del algoritmo al poder asignarse cualquier valor del conjunto correspondiente.

Con la aplicación de una nueva prueba U de Mann Whitney respecto al tiempo requerido para terminar el algoritmo, se encontró que los experimento con igualdad estadística respecto al porcentaje de testores típicos mantienen su igualdad respecto al tiempo. Por tanto, las variables de entrada también mantienen su influencia descrita en el apartado anterior.

En cambio, respecto al número de iteraciones realizadas los experimentos que realizaron como mínimo 2.33 y como máximo 2.6 iteraciones son estadísticamente iguales (ver Tabla 37). Provocando que el EDA aplicado a hemofilia se vuelva más sensible al tamaño de la población al responder mejor con 900 individuos con un porcentaje de alteración del 10%. Mientras que para el porcentaje de selección y el número de iteraciones se mantiene inmune a los valores que pueden tomar (ver Tabla 38).

	Variables de Salida			
#Exp	% Testores	%Testores Típicos	Iteraciones Realizadas	Tiempo (s)
327	83.30	100	2.53	2.478
335	82.70	100	2.5	2.457
256	82.40	100	2.63	2.454
328	82.26	100	2.43	2.385
326	82.17	100	2.6	2.532
321	81.76	100	2.43	2.401
166	81.51	100	2.7	2.335
333	81.15	100	2.33	2.402
323	81.08	100	2.43	2.366
330	80.96	100	2.43	2.458
334	80.85	100	2.4	2.328

Tabla 37 *Desempeño de los mejores experimentos EDA en el contexto de hemofilia*

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	% de Selección	Número de Iteraciones
327	900	10	30	20
335	900	10	50	20
256	850	10	50	25
328	900	10	30	25
326	900	10	30	15
321	900	10	20	10
166	800	10	30	15
333	900	10	50	10
323	900	10	20	20
330	900	10	40	15
334	900	10	50	15

Tabla 38 *Parámetros de los mejores experimentos EDA en el contexto de hemofilia*

Desde el punto de vista empírico destacan los siguientes experimentos, cuyos parámetros se reportan en la Tabla 38:

- El experimento 327, al maximizar el *porcentaje de testores* encontrados.
- El experimento 334, al minimizar el *tiempo* utilizado con 2.328 segundos y es estadísticamente igual al experimento 327 de acuerdo con el *porcentaje de testores* encontrados.
- El experimento 333, al minimizar el *número de iteraciones* realizadas con un promedio de 2.33 iteraciones. Este experimento es estadísticamente similar al 327 respecto al *porcentaje de testores* encontrados.

Finalmente, en la Tabla 39 se expone los parámetros del experimento que empíricamente, genero el desempeño más bajo al encontrar solamente un promedio 43.36% de los *testores* esperado a pesar de encontrar el 100% de los *testores típicos* y haber realizado un promedio de 1.43 *iteraciones* en un promedio de 1.072 segundos.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	% de Selección	Número de Iteraciones
76	700	30	40	25

Tabla 39 *Parametros del experimento EDA con desempeño más bajo aplicado a hemofilia*

Una vez identificados los mejores parámetros para las metaheurísticas AG y EDA, se procede en el siguiente subapartado a exponer el proceso de contraste en el que se identificó la metaheurística que posee mejor respuesta el escenario de experimentación de hemofilia.

Contrastación de Resultados

A continuación, se describe el proceso en el que se contrastan las metaheurísticas AG y EDA con el cual, se determinó aquella que tienen mejor desempeño en la búsqueda de testores típicos en el contexto de la hemofilia descrita en el apartado 3.3.2. Dicho proceso parte de la selección empírica de los mejores experimentos tanto del AG como del EDA, en base a la afinación descrita en este apartado.

En la Tabla 40 se muestran los parámetros del experimento 253 perteneciente a la ejecución del *AG* en el contexto de hemofilia. Este experimento fue seleccionado al maximizar el *porcentaje de testores* encontrados y obteniendo, a su vez, el 100% de los *testores típicos* esperados.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	Prob. De Mutación	Número de Iteraciones
253	850	10	12	10

Tabla 40 *Experimento AG con mejor desempeño en el contexto del problema de hemofilia*

Para el caso del EDA se seleccionó el experimento 327 cuyos parámetros son expuestos a continuación en la Tabla 41. Desde el punto de vista empírico y, al igual que el experimento 253 del AG, obtuvo el mayor porcentaje de testores en promedio asegurando el conjunto completo de testores típicos.

	Variables de Entrada			
#Exp	Tamaño de Población	% de Alteración	% de Selección	Número de Iteraciones
327	900	10	30	20

Tabla 41 *Experimento EDA con mejor desempeño aplicado al contexto de la hemofilia*

Para conocer la metaheurística con mejor desempeño en la *búsqueda de testores típicos* en el contexto del escenario de experimentación de la *hemofilia* se realizaron nuevas ejecuciones independientes de los experimentos de las Tablas 40 y 41, los cuales fueron replicados 30 veces formando dos grupos de experimentos analizados de la siguiente manera.

Mediante una prueba de Levene expuesta en el Anexo G.1, se determinó que ambos grupos de experimentos homocedásticos respecto al *tiempo utilizado* y el *número de iteraciones* realizadas, mientras que respecto al *porcentaje de testores* no lo son. Para el caso del *porcentaje de testores típicos*, ambos experimentos lograr encontrar el 100% en el total de las réplicas realizadas.

A continuación, se determinó que ambos experimentos no siguen una distribución normal respecto al *porcentaje de testores encontrados*, mientras que no existe una distribución normal respecto al *tiempo* utilizado y el *número de iteraciones realizadas*. Lo anterior, se base en los resultados de una prueba de Kolmogorov Smirnov expuesta en el Anexo G.2

Por su parte, en el Anexo G.3 se puede consultar la prueba Kruskal Wallis con la que se concluyó que existe diferencia significativa respecto al *porcentaje de testores encontrados*. Por el contrario, no existe diferencia significativa respecto al *porcentaje de testores típicos* al haber encontrado el 100% en todos los casos; además, no existe diferencia significativa respecto al *tiempo* y el *número de iteraciones realizadas*.

		Variables de Salida			
Meta	**#Exp**	% Testores	%Testores Típicos	Iteraciones Realizadas	Tiempo (s)
AG	**253**	72.54	100	2.33	2.8939
EDA	**327**	78.76	100	2.33	2.7623

Tabla 42 *Resultados promedio de los mejores experimentos aplicados al contexto de hemofilia*

De acuerdo con lo anterior se aplicaron pruebas de U de Mann Whitney para cada una de las variables de salida, en las cuales se determinó que en el caso del *porcentaje de testores* son estadísticamente diferentes. Por su parte, ambas metaheurísticas resultan

estadísticamente iguales respecto al *porcentaje de testores típicos*, el *número de iteraciones* realizadas y el *tiempo* que utilizaron. En conclusión, la metaheurística más adecuada para realizar la búsqueda de testores típicos en hemofilia es el EDA al observar mayor porcentaje de *testores* y un menor *tiempo* de ejecución, como se muestra en la Tabla 42.

A continuación, se describen la interpretación de los conjuntos de testores típicos encontrados para cada uno de los escenarios de experimentación.

4.3 Resultados Médicos

Como se mención al inicio del apartado 4, los resultados obtenidos expuestos en este documento tienen dos puntos de vista importantes: el punto de vista computacional expuesto en el apartado 4.2, y el punto de vista médico el cual es expuesto a continuación. Desde este punto de vista se describe el significado de los conjuntos de testores típicos de cada uno de los escenarios de experimentación.

4.3.1 Resultados en Cáncer de Mama

Para el escenario de *cáncer de mama* se obtuvo un total de dos *testores típicos*, los cuales representan dos *subconjuntos irreducibles de características* con los que es posible catalogar un objeto nuevo de célula de cáncer mamario en su clase correspondiente (*benigna* o *maligna*). Como se recordará, este escenario contó con 10 características (ver apartado 3.3.1) a partir de los cuales se realizó una *búsqueda exhaustiva* y dos *búsquedas metaheurísticas*, por medio de un *AG* y un *EDA*, de *testores típicos*. Los *testores típicos* son interpretados por medio del cálculo del *peso informacional* que, de acuerdo con el apartado 2.4, funge como una puntuación para cada característica involucrada que determina su importancia en la clasificación de objetos.

CARACTERÍSTICA	PESO INFORMACIONAL
1. Radio	50%
2. Textura	100%
3. Perímetro	100%
4. Área	50%
5. Lisura	100%
6. Compacidad	100%
7. Concavidad	100%
8. Puntos cóncavos	100%
9. Simetría	100%
10. Dimensión fractal	100%

Tabla 43 *Peso informacional obtenido de la búsqueda de testores típicos para cáncer de mama*

En la tabla 43 se muestra el *peso informacional* obtenidos para el contexto del *cáncer de mama.* Como se puede observar, ocho de las características presentan una puntuación de 100%, lo cual significa que son características imprescindibles. En otras palabras, para clasificar un nuevo caso de célula de cáncer mamario es necesario conocer el valor en dichas características. Por su parte, las características radio y perímetro presentan un peso informacional de 50%, lo que significa que es posible clasificar una célula sin conocer el valor de una de las dos variables. Por ejemplo, se puede realizar una clasificación conociendo las características imprescindibles y el radio de la célula, sin conocer su perímetro. Por lo tanto, es posible omitir una de las dos características, pero no ambas.

En el caso de haber encontrado una o más características con un *peso informacional* de 0%, significaría que dicha característica no es necesaria. Por tanto, esta característica no aportaría información importante para realizar la clasificación de un objeto y así, se reduciría el problema observado. Recordando que este es un objetivo de la *selección de características*.

Los pesos informacionales fueron revisados por el médico patólogo Luis Muñoz Fernández del Centenario Hospital Miguel Hidalgo, Aguascalientes, Ags.

4.3.2 Resultados en Hemofilia

Para el caso de la hemofilia se identificaron trece *testores típicos* a partir de los cuales, se realizó el cálculo del *peso informacional* para cada una de las 11 características evaluadas, las cuales son descritas en el apartado 3.3.2.

CARACTERÍSTICA	PESO INFORMACIONAL
1. Edad (años)	23.08%
2. IMC	61.54%
3. Tipo de enfermedad	23.08%
4. Artropatía	7.69%
5. Articulaciones con daño	38.46%
6. VIH	23.08%
7. VHC	15.38%
8. VHB	15.38%
9. Inhibidores	100%
10. Numero de hemorragias al año	38.46%
11. Modalidad de tratamiento	53.85%

Tabla 44 *Peso informacional obtenido del conjunto de testores típicos para hemofilia*

En la Tabla 44 se enlistan las características involucradas en este escenario de experimentación con su *peso informacional* correspondiente. En este caso, los inhibidores son la característica que posee mayor impacto en la clasificación en *hemofilia* al obtener un 100% de *peso informacional* y, por lo tanto, la hace imprescindible. En otras palabras, en el 100% de los casos será necesario conocer esta característica para clasificar correctamente un caso de *hemofilia*.

Por su parte, la variable de *artropatía* puede ser descartada debido a su bajo *peso informacional* y a que su información está contenida dentro de la característica de *articulaciones con daño*, la cual tiene un peso informacional más alto.

Los pesos informacionales presentados, fueron analizados por la Dra. Cardiel de la Clínica N°1 del Instituto Mexicano del Seguro Social, así como por la Dra. Beatriz Morquecho como médico general privado.

CAPÍTULO 5

Conclusiones

Con la realización de este proyecto de tesis, se cuenta con un amplio panorama del alcance de las ciencias computacionales gracias a su influencia en el resto de las áreas del conocimiento provocando cambios importantes en sus respectivos paradigmas. Tal es el caso de la medicina, la cual, claramente se ha visto beneficiada con la interacción continua de áreas ingenieriles como la mecatrónica y la electrónica en la creación de aparatos de diagnóstico y tratamiento de enfermedades, así como la informática que apoya en la generación y administración de la información generada.

Gracias a esta interacción, en la actualidad se cuenta con grandes cantidades de información por lo que se requiere de mecanismos más aptos para su análisis. Por esta razón, la inteligencia artificial es el área indicada al proponer nuevas más eficientes formas de extraer información y generar el conocimiento adecuado para la toma de decisiones. Ejemplo

de ello se tiene la selección de características, utilizados en este proyecto, como parte del enfoque lógico-combinatorio del reconocimiento de patrones aplicando la teoría de testores.

La aplicación de dicha teoría resultó toda una experiencia al comprender su alcance al posibilitar la modelación de problemas lo más apegados a la realidad sin importar el área del conocimiento en la que se aplica. Tal fue el caso de este proyecto, en el que se analizaron dos patologías médicas de las cuales, se localizaron los testores típicos que representan la información mínima para clasificar objetos en sus clases correspondientes. Por otra parte, no hay que perder de vista que la búsqueda exhaustiva de testores típicos tienen una complejidad exponencial, por lo que se requirió del apoyo de más herramientas de la inteligencia artificial para permitir la creación de solución alternativa a dicha búsqueda.

El uso de metaheurísticas permitió crear una búsqueda alternativa de testores típicos, el cual, además, representó otro mundo de conocimiento para entender y desarrollar. Así pues, el proyecto fue creciendo con la coalición de áreas dentro de las ciencias computacionales. Así pues, se exploraron el algoritmo de estimación de la distribución y el algoritmo genéticos, los cuales resultan muy flexibles para ser hibridados de acuerdo con el problema que se busca solucionar. Gracias a esto, se logró experimentar el desarrollo de un nuevo operador, al que se le llamó operador de alteración, que es capaz de trabajar en conjunto con los operadores básicas de cada metaheurística y apoya positivamente en la exploración del espacio de soluciones al realizar pequeñas búsquedas en la vecindad de soluciones ya evaluadas con el objetivo de mejorar la población antes de ser procesada por el resto de los operadores de cada metaheurística.

Para el diseño e integración del operador, así como la implementación de las metaheurísticas se hizo uso de las herramientas de construcción propuestas por la ingeniería de software, las cuales facilitaron en gran medida la partición de los sistemas en componentes independientes permitiendo hacer cambios sin afectar el funcionamiento del resto.

Finalmente, se comprueba el alcance y el beneficio que puede traer la realización de proyectos que involucren la interacción de diferentes áreas, tanto para las mismas áreas como para el crecimiento personal al grado de requerir mayor esfuerzo para comprender el punto de vista de aquellas áreas ajenas al perfil de las personas involucradas.

5.1 Objetivos Cubiertos

En este apartado se describen los apartados de este trabajo de tesis en los que se cubren cada uno de los objetivos definidos en el apartado 1.2.

- **Objetivo general:**

  Cubierto propiamente por parte de la aplicación de la metodología descrita en el apartado 3, específicamente en el apartado 3.2.4 y 3.2.5 aplicando las metaheurísticas hibridadas. Por su parte, dichas metaheurísticas son descritas en 4.2.3 y 4.2.4.

- **Objetivo Particular 1:**

  Este objetivo se cubre como parte de la metodología, específicamente descrito en el apartado 3.2.2. Los datos obtenidos fueron base para la evaluación del desempeño de las metaheurísticas hibridadas.

- **Objetivo Particular 2:**

  La aplicación construida para la búsqueda exhaustiva que cubrió el objetivo particular 1 incluye un módulo especial que calcula el peso informacional de cada característica involucrada en el problema evaluados. Por su parte, el apartado 4.3 expone la interpretación de dichos resultados.

- **Objetivo Particular 3:**

  El funcionamiento de las metaheurísticas tipo AG y EDA se expone propiamente en el apartado 4.2.3 y 4.2.4 cuyos operadores que los hibridan se pueden consultar en el apartado 4.2.1 y 4.2.2. Por parte del ajuste de ambas metaheurísticas es cubierto en los apartados 4.2.4 y 4.2.5.

- **Objetivos Particular 4:**

  Este objetivo cubierto en la implementación de las aplicaciones encargadas de la búsqueda exhaustiva y metaheurística de los testores típicos. Los modelos adaptados se pueden consultar en el Anexo A.

- **Objetivo Particular 5:**

  El operador desarrollado se denominó operador de alteración, cuyo funcionamiento es descrito en el apartado 4.2.2 y, a su vez, es incorporado al AG y al EDA en los apartados 4.2.3 y 4.2.4 respectivamente.

- **Objetivo Particular 6:**

  La validación de los resultados se incluye en el apartado 4.3 en el cual, se muestran los resultados obtenidos desde la perspectiva médica.

5.2 Contribuciones

La principal contribución de este trabajo de tesis es el diseño, implementación e incorporación de un nuevo operador que hibridan a la metaheurísticas trabajadas en este proyecto: AG y EDA. Este operador se denominó operador de alteración, el cual es capaz de trabajar perfectamente con los operadores básicos de cada metaheurística. Este nuevo operador, como su nombre lo indica, altera cada individuo (solución) en una población evaluada de manera que el resto de los operadores recibe una población alterada. Esta población contiene mejores soluciones obtenidas a partir de las soluciones originales. Básicamente, el operador modifica un porcentaje del individuo en busca de mejores soluciones en su vecindad sustituyendo la solución original con la solución mejorada. Como se puede observar a lo lago de la de descripción de resultados en los apartados 4.2.4 y 4.2.5, la incorporación de este nuevo operador resultó ser determinante en la evaluación del desempeño de los algoritmos metaheurísticos.

Por otra parte, se encuentran el conjunto de testores típicos de dos patologías médicas: cáncer de mama y hemofilia. Cada conjunto permitió encontrar las características que con mayor incidencia para el contexto de cada enfermedad. En este caso, se buscó clasificar objetos por medio de los testores típicos. Para el caso de cáncer de mama se encontraron dos testores típicos, los cuales representan las características mínimas necesarias para clasificar células en benignas o malignas. Por parte de la hemofilia, se identificaron 13 testores típicos con los cuales se pueden clasificar casos en leves, moderados y graves.

5.2.1 Trabajo a Futuro

De acuerdo con la participación de los especialistas consultados para el escenario de experimentación de hemofilia, se propone la inclusión de nuevas variables como la existencia de accidentes, cirugías y la aplicación de transfusiones de sangre con el objetivo de observar el comportamiento de los testores típicos encontrados. Lo anterior, debido a la existencia de combinaciones de niveles de gravedad en hemofilia como, por ejemplo, pacientes con hemofilia leve con comportamiento grave debido a accidentes, extracciones molares o intervenciones quirúrgicas. Además, debido a que las medidas de control en la donación de sangre han evolucionado, pacientes con edad avanzada pudieron haber recibido sangre de baja calidad que pudiera influir en el comportamiento de la enfermedad.

Por parte de las metaheurísticas implementadas, se propone que sean aplicadas a otras patología y problemas de otras áreas del conocimiento cuyas bases de datos sean más robustas para así, evaluar su desempeño en estos problemas.

5.2.2 Productos Realizados

- Gallegos, D. Torres, F. Álvarez, A. Torres, " Feature Subset Selection and Typical Testors Applied to Breast Cancer Cells", Research in Computing Science: Advances in Human Computer Interaction and Artificial Intelligence, vol 121, pp.151-163. Anexo H.1.

- Gallegos, F. Álvarez, D. Torres, A. Torres, "Análisis de Células Cancerígenas Aplicando Teoría de Testores", Tecnología Educativa Revista CONAIC, Vol 3, Num 3 pp.70-77. Anexo H.2.

- Gallegos, D. Torres, F. Álvarez and A. Torres, " Identificación de características de células de cáncer de mama por medio de testores típicos ", Research in Computing Science: Advances in Computer Vision, Signal Processing and Virtual Environments, vol 140, pp. 43-54. Anexo H.3.

Bibliografía

[1] S. O. Lugo-Reyes, G. Maldonado-Colín, and C. Murata, "Inteligencia artificial para asistir el diagnóstico clínico en medicina," *Artificial Intelligence to Assist Clinical Diagnosis in Medicine.*, Article vol. 61, no. 2, pp. 110-120, 2014.

[2] A. B. Tucker, *Computer science handbook*, 2nd ed.. ed. Boca Raton, Fla.: Boca Raton, Fla. : Chapman & Hall/CRC, 2004.

[3] M. D. Torres Soto, A. Torres Soto, M. d. l. L. Torres Soto, L. Bermudez Rosales, and E. E. Ponce de León Sentí, "Factores Predisponentes en Relajación Residual Neuromuscular," *Research in Computing Science*, vol. 93, pp. 163–174Available: http://www.rcs.cic.ipn.mx/2015_93/

[4] M. D. Torres Soto, A. Torres Soto, and E. Ponce de León Sentí, "Algoritmo Genético y Testores Típicos en el Problema de Selección de Subconjuntos de Características," Available: http://www.iiisci.org/journal/CV$/risci/pdfs/C415DR.pdf

[5] V. R. Avila Cruz, "Medicina y Computación," ed: Recuperado el viernes de mayo de, 2015.

[6] M. d. C. Expósito Gallardo and R. Ávila Ávila, "Aplicaciones de la inteligencia artificial en la Medicina: perspectivas y problemas," *ACIMED*, vol. 17, pp. 0-0, 2008.

[7] S. Monestel Umaña and L. Samaha, "Uso de teorías de comunicación para disminuir el error diagnóstico," *Revista Médica de la Universidad de Costa Rica*, vol. 8, no. 2, pp. 35-43, 2014.
[8] H. Singh and S. Weingart, "Diagnostic errors in ambulatory care: dimensions and preventive strategies," *Advances in health sciences education: theory and practice*, pp. 57-61Available: https://link.springer.com/content/pdf/10.1007%2Fs10459-009-9177-z.pdf
[9] J. M. Ceriani Cernadas, "Errores de diagnóstico en la práctica médica," *Archivos argentinos de pediatría*, vol. 113, no. 3, pp. 194-195, 2015.
[10] B. N. Ramos Domínguez, "Calidad de la atención de salud: Error médico y seguridad del paciente," *Revista Cubana de Salud Pública*, vol. 31, pp. 0-0, 2005.
[11] M. Granados García, A. R. Martín, and J. Hinojosa Gómez, *Tratamiento del cáncer: oncología médica, quirúrgica y radioterapia*. Distrito Federal, UNKNOWN: Editorial El Manual Moderno, 2016.
[12] E. Garrido Fuente, "Neoplasia de mama," El Cid Editor, Córdoba, ARGENTINA2016, Available: http://ebookcentral.proquest.com/lib/univeraguascalientessp/detail.action?docID=4507593.
[13] E. Garrido Fuente, "Factores de riesgos ambientales y genéticos: influencia en el cáncer de mama," El Cid Editor, Córdoba, AR2015, Available: http://site.ebrary.com/lib/univeraguascalientessp/docDetail.action?docID=11087163.
[14] J. Cárdenas Sánchez, J. E. Bargalló Rocha, A. Erazo Valle, A. Poitevin Chachón, C. Vicente Valero, and V. Pérez Sánchez, "Consenso Mexicano sobre diagnóstico y tratamiento del cáncer mamario," ed. México: IMSS, 2017, pp. 1-147.
[15] INEGI, "Estadísticas a Propósito del... Día Mundial de la Lucha contra el Cáncer de Mama," in *Estadisticas Nacionales*, ed. México: Instituto Nacional de Estadística y Geografía, 2015.
[16] UNAM, "19 de Octubre: Día mundial de la lucha con el cáncer de mama," ed: Universidad Nacional Autónoma de México, 2017.
[17] A. B. Ortega, "CÁNCER DE MAMA-MÉXICO," ed. México: Universidad Nacional Autónoma de México, 2010, pp. 1-19.
[18] A. Bergés *et al.*, "Reporte sobre Hemofilia en México," ed. México: Federación de Hemofilia de la República Mexicana, 2016, pp. 1-24.
[19] FHRM. (2017). *Hemofilia*. Available: http://www.hemofilia.org.mx/web16/
[20] WFH, "World Federation of Hemophilia, Report on the Annual Global Survey 2016," World Federation of Hemophilia, Montréal, Québec, Canada2017, Available: http://www1.wfh.org/publications/files/pdf-1690.pdf.
[21] Diario de Yucatán, "Mayor calidad de vida," in *Diario de Yucatán*, ed. Yucatán, México: Diario de Yucatán, 2016.
[22] J. García-Chávez and A. Majluf-Cruz, "Hemofilia," *Gaceta Médica de México*, no. Hemofilia, p. 14Available: http://www.anmm.org.mx/GMM/2013/n3/GMM_149_2013_3_308-321.pdf
[23] J. Ruíz Shulcloper, "Acerca del surgimiento del Reconocimiento de Patrones en Cuba," *Revista Cubana de Ciencias Informáticas*, vol. 7, no. 2, 2013.

[24] J. A. Carrasco Ochoa and J. F. Martínez Trinidad, "Reconocimiento de Patrones," *Komputer Sapiens, Revista de Divulgación de la Sociedad Mexicana de Inteligencia Artificial*, vol. 3, pp. 5-10Available: http://smia.mx/komputersapiens/index.php?option=com_content&view=article&id=72&Itemid=84

[25] H. Vega Huerta, A. Cortez Vasquez, A. Maria Huayna, L. Alarcon Loayza, and P. Romero Naupari, "Reconocimiento de patrones mediante redes neuronales artificiales.(Report)," *Revista de investigacion de Sistemas e Informatica*, vol. 6, no. 2, p. 17, 2009.

[26] Y. Villuendas Rey, *Esquema para el pre-procesamiento de conjuntos de entrenamiento de clasificadores del vecino más cercano basado en extensiones a la teoría de los conjuntos aproximados*. Havana, CUBA: Editorial Universitaria, 2014.

[27] C. V. Sanz, "Razonamiento Evidencial Dinámico, Un Método de Clasificación Aplicado al Análisis de Imágenes Hiperespectrales," Doctorado, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina, 2002.

[28] F. A. Sánchez Garfias, J. L. Díaz de León Santiago, and C. Yáñez márquez, "Reconocimiento automático de patrones: conceptos básicos," vol. 10, ed. México: Research in Computer Science, 2003, pp. 91-102.

[29] C. Yánez, "Memorias Asociativas Basadas en Relaciones de Orden y Operaciones Binarias," vol. 6, ed. México, D.F.: Computación y Sistemas, 2002, pp. 300-311.

[30] A. Fernández, "Selección de Características - Reconocimiento de Patrones 2013," ed. Uruguay: Universidad de la República de Uruguay, 2013.

[31] A. G. d. T. A. e. Computacion, "Clasificación Supervisada y No Supervisada," ed. Loxa, Ecuador: advancedtech.wordpress.com, 2008.

[32] J. Ruíz Shucloper, E. Alba Cabrera, and M. Lazo Cortés, "Introducción a la Teoría de Testores," ed: Departamento de Ingeniería Electrica, CINVESTAV-IPN, 1995, p. 197.

[33] Y. Santiesteban Alganza and A. Pons Porrata, "LEX: UN NUEVO ALGORITMO PARA EL CALCULO DE LOS TESTORES TIPICOS," *Revista Ciencias Matematicas*, Article vol. 21, no. 1, pp. 85-95, 2003.

[34] A. Pereira González, "Selección de Características para el Reconocimiento de Patrones con Datos de Alta Dimensionalidad en Fusión Nuclear," ed. España: Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), 2015.

[35] M. D. Network, "Selección de Características (Minería de Datos)," ed: msdn.microsoft.com, 2016.

[36] S. Pal and P. Mitra, C. P. LLC, Ed. *Pattern Recognition Algorithms for Data Mining*. Calcutta, India: Chapman & Hall / CRC, 2004, p. 244.

[37] A. Lias-Rodríguez and A. Pons-Porrata, "Un nuevo Algoritmo de Escala Exterior para el Cálculo de los Testores Típicos," p. 10Available: http://www.cerpamid.co.cu/sitio/files/publicaciones/1034921953BR_RECPAT09.pdf

[38] J. Ochoa Somuano, "Técnicas de Selección de Atributos para la Categorización Automática de Escenas Visuales," Maestría, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos, 2005.

[39] M. D. Torres, E. Ponce de León, C. A. Ochoa, A. Torres, and E. Díaz, "Mecanismos de Aceleración en Selección de Características Basada en el Peso Informacional de las Variables para Aprendizaje no Supervisado," *Revista Iberoamericana de Sistemas, Cibernética e Informática*, vol. 6, pp. 29-34

[40] J. A. Santos, A. Carrasco, and J. F. Martínez, "Feature Selection using Typical Testors applied to Estimation of Stellar Parameters," *Computación y Sistemas*, vol. 8, pp. 15-23

[41] M. O. Cotilla, "Un Recorrido por la Sismología de Cuba," 1 ed. Cuba: Editorial Complutense, S. A., 2006.

[42] A. Duarte Muñoz, *Metaheurísticas*. España: Dykinson, 2008.

[43] S. H. Zanakis and J. R. Evans, "HEURISTIC "OPTIMIZATION": WHY, WHEN, AND HOW TO USE IT," *Interfaces*, vol. 11, no. 5, pp. 84-91, 1981.

[44] M. García Torres, "Aplicación de Técnicas Metaheurísticas en Minería de Datos," Doctorado, Universidad de la Laguna, España, 2009.

[45] S. Alonso, O. Cordón, I. Fernandez de Viana, and F. Herrara, "La Metaheurística de Optimización Basada en Colonia de Hormigas: Modelos y Nuevos Enfoques," in *Mejora de Metaheurísticas mediante Hibridación y sus Aplicaciones*, ed. Granada, España: Universidad de Granada, 2003, pp. 1-49.

[46] Z. Michalewicz and D. Fogel, "How to Solve It: Modern Heuristics," ed. Berlin, Alemania: Springer, 2004.

[47] Á. García Sánchez, "Técnicas Metaheurísticas," ed. Madrid, España: Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Industriales, 2013, pp. 1-47.

[48] P. Rodríguez-Piñero Tolmos, "Introducción a los Algoritmos Genéticos y sus Aplicaciones," U. d. Valencia, Ed., ed. Valencia, España: Asociación Española de Profesores Universitarios de Matemáticas para la Economía y la Empresa, 2002, pp. 1-9.

[49] M. Gestal, D. Rivero, J. R. Rabuñal, J. Dorado, and A. Pazos, U. d. Coruña, Ed. *Introducción a los Algoritmos Genéticos y la Programación Genética* (Monografías). Coruña: Digitalia, 2010, p. 76.

[50] J. J. Romero, C. Dafonte, Á. Gómez, and F. J. Penousal, *Inteligencia Artificial y Computación Avanzada* (Colección informática). Fundación Alfredo Brañas, 2007, p. 400.

[51] D. E. Goldberg, *Genetic Algorithms in Search, Optimization & Machine Learning*. Alabama, 1989.

[52] F. Carretero López, "Optimización Global con Algoritmos Genéticos," Licenciatura, Universidad Politécnica de Cataluña, Cataluña, España, 2010.

[53] R. Pérez Rodríguez and A. Hernández Aguirre, *Un Algoritmo de Estimación de la Distribuciones para el Problema de Secuenciamiento en Configuración Jobshop Flexibe* (Comunicaciones del CIMAT). Guanajuato, México: CIMAT, 2015, p. 90.

[54] A. Moujahid, I. Inza, and P. Larrañaga, "Algoritmos Genéticos," pp. 1-34Available: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t2geneticos.pdf

[55] C. Reeves, "Genetic Algorithms," in *Handbook of Metaheuristics*, F. Glover and G. A. Kochenberger, Eds. Boston, MA: Springer US, 2003, pp. 55-82.

[56] R. Caballero Fernández, J. Molina Luque, M. Luque Gallegps, A. Torrico González, and T. Gómez Núñez, "Algoritmos Genéticos para la Resolución de Problemas de Programación por Metas Entera. Aplicación a la Economía de la Educación," *Revista Virtual PRO Procesos Industriales*, vol. 2, Available: https://www.revistavirtualpro.com/biblioteca/algoritmos-geneticos-para-la-resolucion-de-problemas-de-programacion-por-metas-entera-aplicacion-a-la-economia-de-la-educacion

[57] K. A. D. Jong, "An analysis of the behavior of a class of genetic adaptive systems," University of Michigan, 1975.

[58] A. Mouhahid, I. Inza, and P. Larrañaga, "Tema 3: Algoritmos de Estimación de Distribuciones," D. d. C. d. l. C. e. I. Artificial, Ed., ed. País Vasco, España: Universidad del País Vasco, 2008, pp. 1-11.

[59] P. Bermejo, J. Gámez, A. Martínez, and J. Puerta, "Algoritmos de Estimación de Distribuciones para la Selección Simultánea de Instancias y Atributos," presented at the VIII Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados, Albacete, 2012. Available: http://www.congresomaeb2012.uclm.es/papers/paper_54.pdf

[60] P. Larrañaga, J. A. Lozano, and H. Mühlenbein, "Algoritmos de Estimación de Distribuciones en Problemas de Optimización Combinatoria," vol. 7, ed. Valencia, España: Revista Iberoamericana de Inteligencia Artificial, 2003.

[61] J. I. Rincón Miranda, M. D. Torres Soto, and A. Torres Soto, "Clusterización de Hongos en Base a su Información Proteómica por Medio de un EDA-UMDA," *Décima Quinta Conferencia Iberoamericana en Sistemas, Cibernética e Informática*, vol. 1, no. 1, pp. 1-7Available: http://www.iiis.org/CDs2016/CD2016Summer/books/CISCI-p.pdf

[62] P. Jalote, *A Concise Introduction to Software Engineering*. London: London : Springer London, 2008.

[63] P. Jalote, "An integrated approach to software engineering," SpringerLink, Ed., 3th Ed.. ed. Boston, MA: Boston, MA : Springer Science+Business Media, Inc., 2005.

[64] I. Sommerville, *Ingeniería del software*. Madrid: Madrid : Pearson Addison-Wesley, 2005.

[65] R. S. Pressman, *Ingeniería del software : un enfoque práctico*, 6a. ed.. ed. México, D.F.: México, D.F. : McGraw-Hill, 2005.

[66] E. S. d. I. I. ESEI, "Tema 4. Diseño Arquitectónico," ed. Vigo, España: Universidad de Vigo, 2011, p. 15.

[67] P. Kuchana, C. P. Company, Ed. *Software Arquitecture Design Patterns in Java*. USA: Auerbach Publications, 2004.

[68] E. Gamma, *Design patterns: elements of reusable object-oriented software*. Pearson Education India, 1995.

[69] J. Coplien, *Software Patterns*. USA: SIGS Books & Multimedia, 2005, p. 69.

[70] A. M. Guarachi. (2008) Inteligencia Artificial en Medicina. *RITS*. 33-37. Available: http://www.revistasbolivianas.org.bo/pdf/rits/n1/n1a08.pdf

[71] J. E. Fox. (1991) Sistemas Expertos y su aplicación en medicina. *IATREIA, Revista médica Universidad de Antioquia*. Available: https://aprendeenlinea.udea.edu.co/revistas/index.php/iatreia/article/view/3457/3219

[72] M. A. Pérez. (2008) Sistemas expertos para la asistencia médica. *Enterate en línea: Internet Cómputo y Telecomunicaciones*. Available: http://www.enterate.unam.mx/artic/2008/marzo/art5.html

[73] A. Conchas, "Como la inteligencia artificial cambió la medicia," *Machine Learning*, Available: https://www.inbest.me/comunidad/c%C3%B3mo-la-inteligencia-artificial-cambiar%C3%A1-la-medicina

[74] Á. Santiago, "Inteligencia Artificial el siguiente paso para el Internet de las Cosas," *Instituto Mexicano de Teleservicios Noticias*, Available: http://imt.com.mx/noticias/inteligencia-artificial-el-siguiente-paso-para-el-internet-de-las-cosas/

[75] R. Gestión, "Inteligencia artificial y salud, un mercado en pleno auge," *Gestión Tecnología*, Available: https://gestion.pe/tecnologia/inteligencia-artificial-salud-mercado-pleno-auge-137789

[76] Medline Plus. (2015). *Cáncer*. Available: https://medlineplus.gov/spanish/ency/article/001289.htmAvailable: https://medlineplus.gov/cancer.html

[77] NCI. (2015). *What is Cancer?* Available: http://www.cancer.gov/about-cancer/understanding/what-is-cancer

[78] Mountain View Hospital. (2015). *Crecimiento y Desarrollo Celular Normal*. Available: https://mountainview-hospital.com/hl/?/36702/Causas-de-C%C3%A1ncer~Crecimiento-y-Desarrollo-Celular-Normal/sp

[79] American Cancer Society. (2016). *¿Qué es el cáncer? Una guía para pacientes y sus familias*. Available: http://www.cancer.org/espanol/cancer/aspectosbasicossobreelcancer/que-es-el-cancer

[80] Societé Canadienne du Cancer. (2014). *Types of Tumours*. Available: http://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/types-of-tumours/?region=on

[81] Cancer Research UK. (2017). *How cancers grow*. Available: http://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancers-grow

[82] F. S. Ouchen, "TEMA 14: Neoplasias. Definiciones. Nomenclatura. Características," ed: Eusalud, 2008.

[83] T. Ocete Calvo. (2016). *Diferencia entre tumor y cáncer*. Available: http://www.bekiasalud.com/articulos/diferencias-tumor-cancer/

[84] E. Garrido Fuente, "Medios diagnósticos en la detección precoz del cáncer de mamas," 2015.

[85] breastcancer.org. (2016). *U.S. Breast Cancer Statistics*. Available: http://www.breastcancer.org/symptoms/understand_bc/statistics

[86] ACS. (2017). *Acerca del Cáncer de Seno*. Available: https://www.cancer.org/es/cancer/cancer-de-seno/acerca/que-es-el-cancer-de-seno.html

[87] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis," *International Symposium on Electronic Imaging: Science and Technology*, vol. 1905, pp. 861-870

[88] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology," *Human Pathology*, vol. 26, no. 7, pp. 792-796, 1995.
[89] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine- needle aspirates," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 130, no. 5, p. 511, 1995.
[90] A. Srivanstava *et al.*, Guías para el Tratamiento de la Hemofilia, F. M. d. Hemofilia, ed., 2 ed.: World Federation of Hemophilia, 2012. [Online]. Available: http://www1.wfh.org/publication/files/pdf-1513.pdf.
[91] The Hemophilia Alliance, "Manual sobre Hemofilia, Guía para las familias," C. C. s. Hematology, Ed., ed. USA: The Hemophilia Alliance, 2014, p. 48.
[92] Diario de Yucatán, "Llevan una vida normal: Los pacientes con hemofilia pueden practicar ejercicio," in *Diario de Yucatán*, ed. Yucatán, México: Diario de Yucatán, 2017.
[93] W. H. Wolberg, N. Street, and O. L. Mangasarian, "Wisconsin Diagnostic Breast Cancer (WDBC)," U. o. California, Ed., ed. USA, 1995.
[94] B. B. Mandelbrot, "The fractal geometry of nature," ed: New York: W.H. Freeman, 1982.
[95] I. Osuna-Ramírez, B. Hernández-Prado, J. C. Campuzano, and J. Salmerón, "Indice de masa corporal y percepción de la imagen corporal en una población adulta mexicana: la precisión del autorreporte," *Salud Pública de México*, vol. 48, pp. 94-103, 2006.
[96] G. M. Moreno, "Definición y clasificación de la obesidad," *Revista Médica Clínica Las Condes*, vol. 23, no. 2, pp. 124-128, 2012.
[97] M. para el Estudio and M. de Pediatría, "Avances en el tratamiento de la hemofilia," *Rev Med Inst Mex Seguro Soc*, vol. 43, no. Supl 1, pp. 135-138, 2005.
[98] Q. Moncerrad and L. Xavier, "Perfil epidemiológico del VIH SIDA en pacientes de el Hospital Leon Becerra Camacho de Milagro, periodo 2014-2015," Universidad de Guayaquil. Facultad de Ciencias Médicas. Escuela de Medicina, 2016.
[99] M. Barbeito Barros, "Prevalencia del VHC en pacientes con insuficiencia renal crónica y su impacto en el tratamiento antiviral," 2016.
[100] World Health Organization. (2017). *Hepatitis B*. Available: http://www.who.int/mediacentre/factsheets/fs204/es/
[101] C. K. Kasper. (2004) Diagnóstico y Tratamiento de Inhibidores de los Factores VIII y IX. *Tratamiento de la Hemofilia*. Available: http://www1.wfh.org/publication/files/pdf-1179.pdf
[102] IMSS, "Guía de Referencia Rápida Diagnóstico y Tratamiento de Hemofilia Pediátrica," ed. México: Consejo de Salubridad General, Gobierno Federal, 2012.
[103] IMSS, "Guía de Referencia Rápida Diagnóstico y Tratamiento de Hemofilia en Adultos," ed. México: Consejo de Salubridad General, Gobierno Federal, 2009, p. 9.
[104] G. Rudolph, "Convergence Analysis of Canonical Genetic Algorithms," vol. 5, ed. USA: IEEE, 1994, pp. 96-101.

ANEXOS

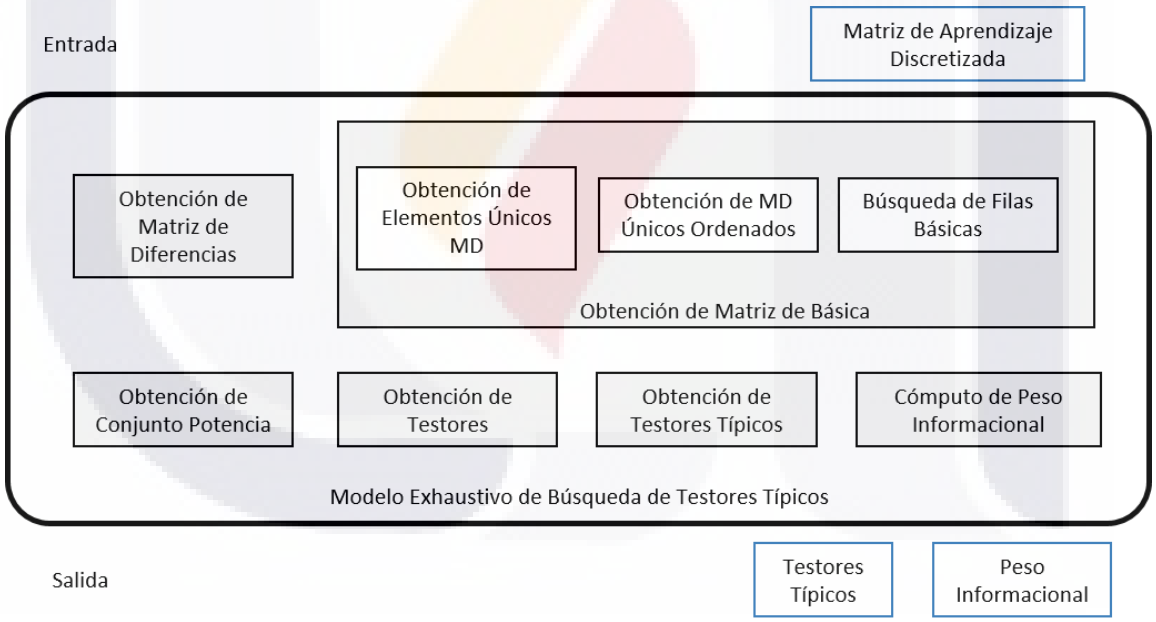
Anexos

A. Modelos de Diseño de Software

A.1 Arquitectura de Software: Búsqueda Exhaustiva de Testores Típicos

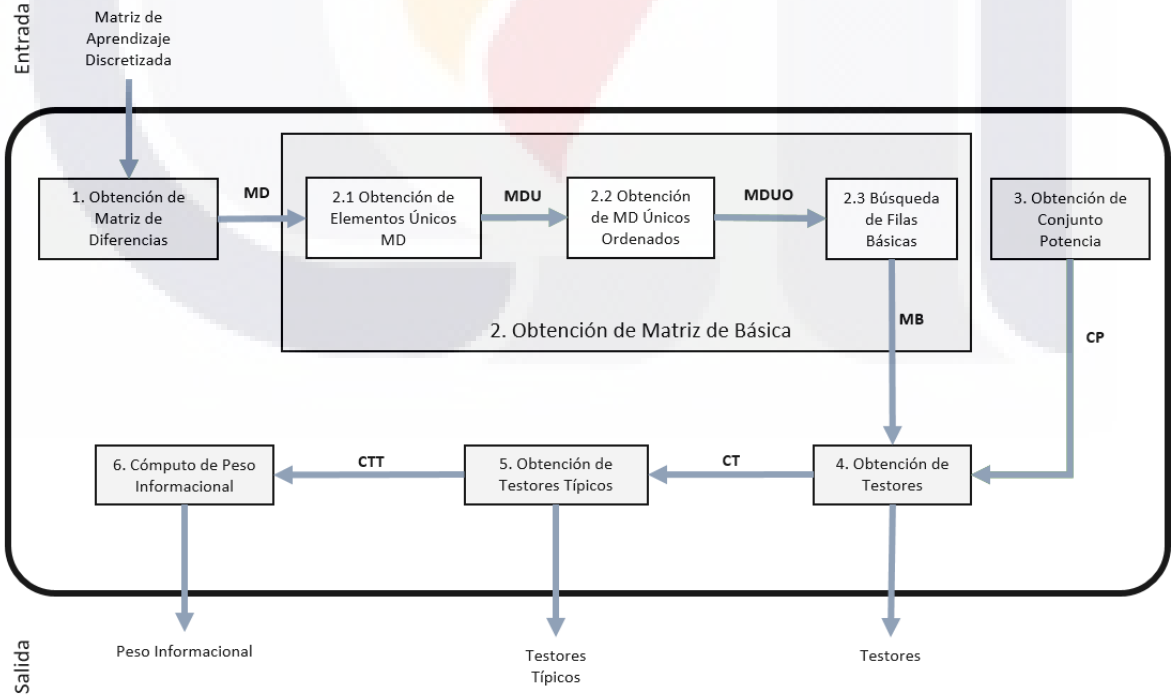
Modelo Estático

Este modelo identifica los componentes básicos que requiere la implementación de la *búsqueda exhaustiva de testores típicos*. Como se puede observar en la figura, cada uno de los componentes se encarga de un paso concreto del proceso descrito en el apartado 2.4. Cada uno de los componentes o módulos del sistema fueron identificados de manera que cada uno resulte una unidad de construcción de software. De esta manera, cada componente pudo ser gestionado de manera independiente permitiendo modificaciones sin alterar al resto.



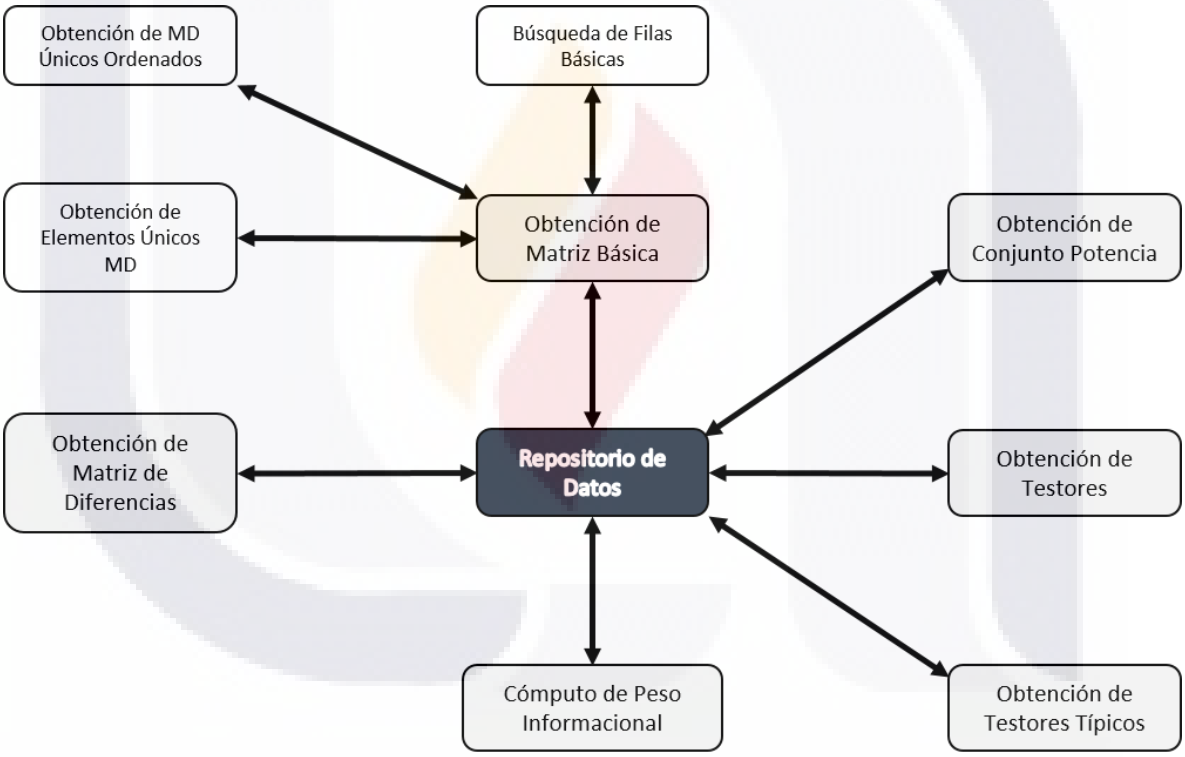
Modelo Dinámico

El *modelo dinámico* basa su estructura en el *modelo estático*, cuya diferencia radica en que en este nuevo modelo se observa la evolución de la información durante la ejecución del sistema. Como se observa en la figura, la aplicación necesita una *matriz de aprendizaje discretizada* que es recibida por el módulo 1. Ese módulo entrega como salida un *matriz de diferencias (MD)*, que a su vez es recibida por el módulo 2 por medio del submódulo 2.1, el cual obtiene una *MD con elementos únicos (MDU).* Las filas en la *MDU* son ordenadas de menor a mayor de acuerdo a la cantidad de bits 1 que posean, recordando que la representación de cada fila son cadenas binarias. Así pues, se obtiene una *MDU ordenada (MDUO)*. El submódulo 2.3 obtiene propiamente las filas básicas para formar la *matriz básica (MB)*. El módulo 3, obtiene el *conjunto potencia (CP)* con cadenas binarias que representan posibles *testores*. En el módulo 4 se extrae el *conjunto de testores (CT)* del *CP* por medio de la *MB*. Por su parte, el módulo 5 identifica el *conjunto de testores típicos (CTT)*, el cual es utilizado por el módulo 6 para calcular el *peso informacional* de cada característica involucrada en el análisis. Para más información, consultar el apartado 2.4.



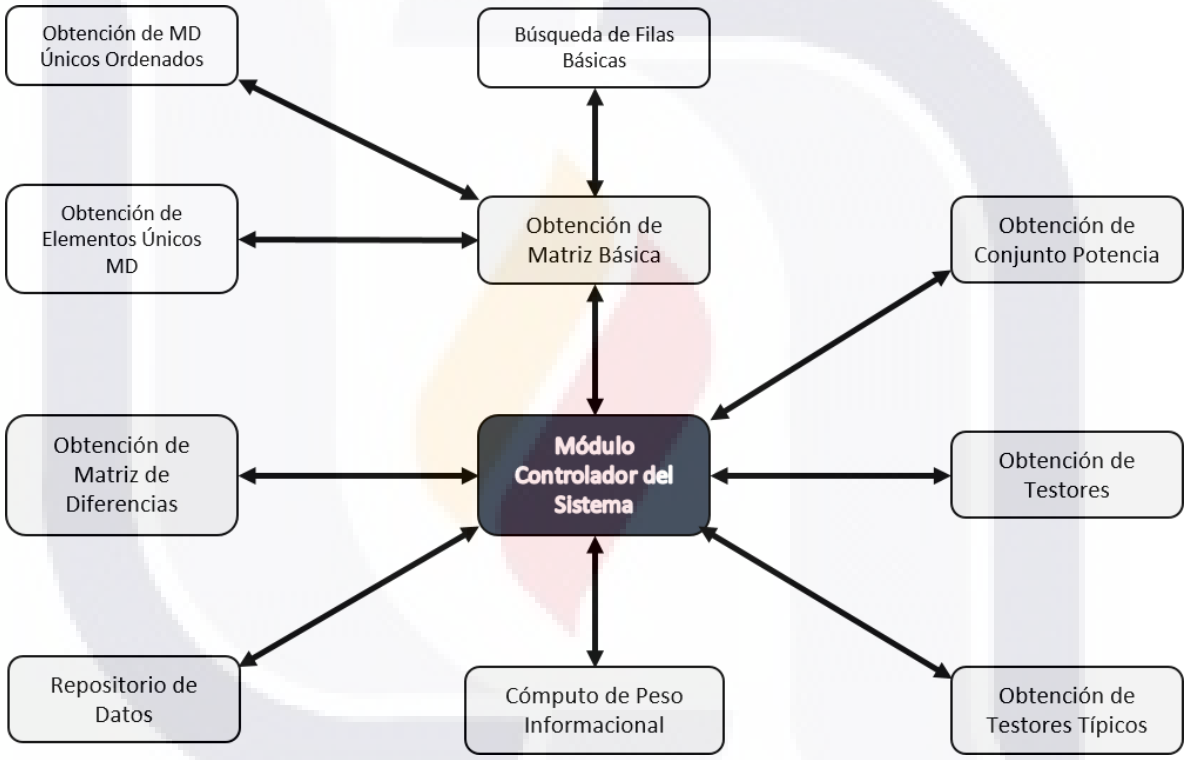
Modelo de Organización del Sistema

Cada uno de los módulos identificados almacenan sus salidas en archivos de texto, los cuales son leídos por un módulo determinado. Por esta razón, la *estrategia de organización* fue seleccionada fue el *modelo de repositorio*. Este modelo, como se observa en la figura, integra un nuevo, integra la representación del *repositorio central* en el que los datos pueden ser almacenados por un cada módulo y son leídos por otros. De esta manera, no es necesario que los módulos conozcan cómo es utilizada la información que generan por parte de otros módulos.



Modelo de Control

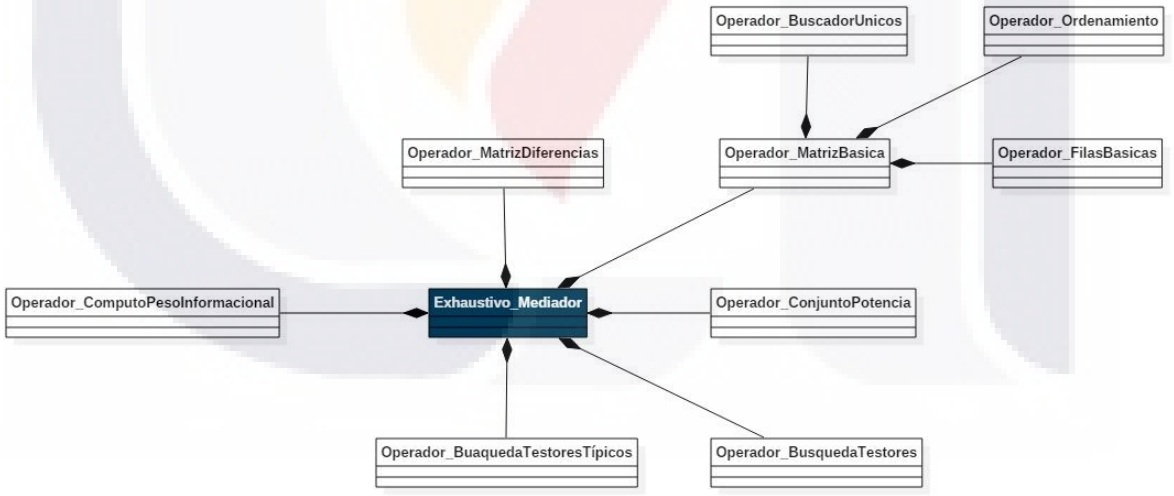
El modelo de control utilizado para mejorar el desempeño de la descomposición del sistema en submódulos fue el *modelo centralizado* que se observa en la figura. Este modelo integra un nuevo componente cuya tarea principal es gestionar la ejecución de cada uno de los submódulos, es decir, este nuevo componente es un mecanismo de control para que cada componente reciba y entregue la información correspondiente en el momento preciso.



A.2 Patrón de Diseño: Búsqueda Exhaustivo de Testores Típicos

En este anexo se expone el *patrón de diseño* adaptado para solucionar la implementación del sistema encargado de la búsqueda de la búsqueda exhaustiva de testores típicos. El *modelo de patrón de diseño* seleccionado fue el *modelo de mediador*, el cual se adapta al diseño arquitectónico definido en el Anexo A. Este modelo permitió encapsular el funcionamiento del sistema en un solo componente, es decir, es el componente que define manera en que el conjunto de componentes interactúa entre sí. De esta manera el algoritmo abstracto fue programado en el componente mediador, en el cual, se encarga de ejecutar cada uno de los componentes restantes en el momento adecuado, otorgando los recursos necesarios y recibiendo las salidas.

Finalmente, el patrón de diseño facilitó la implementación y pruebas del sistema permitiendo, además, realizar cambios a los componentes sin la necesidad de alterar toda la estructura del sistema.

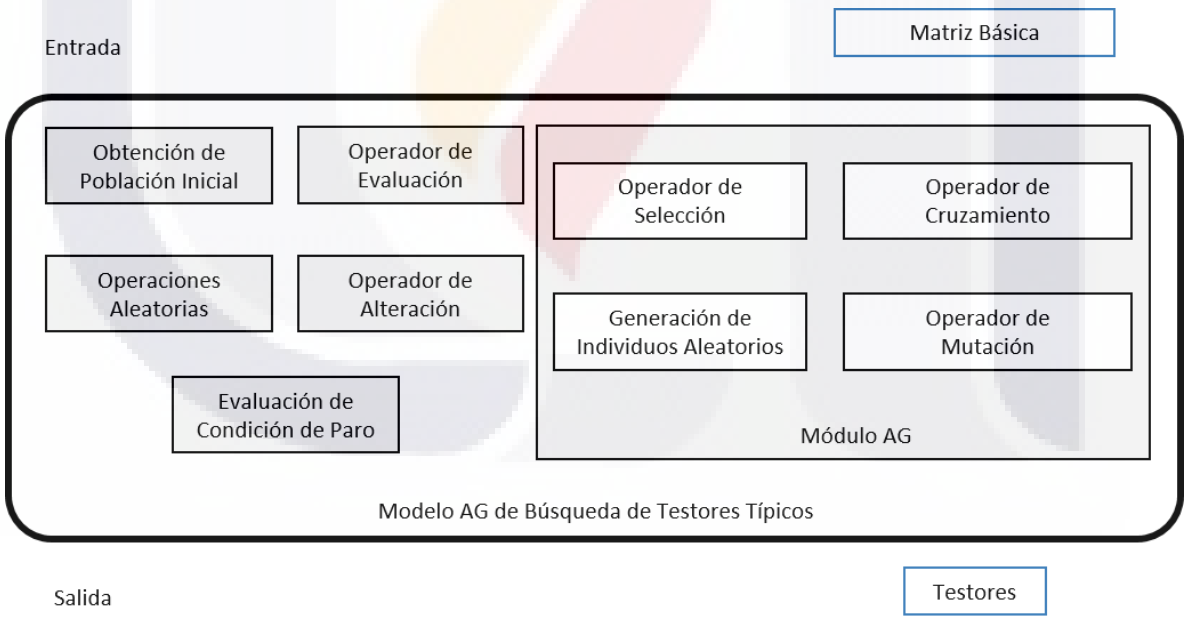


A.3 Arquitectura de Software: Búsqueda de Testores Típicos por AG

Modelo Estático

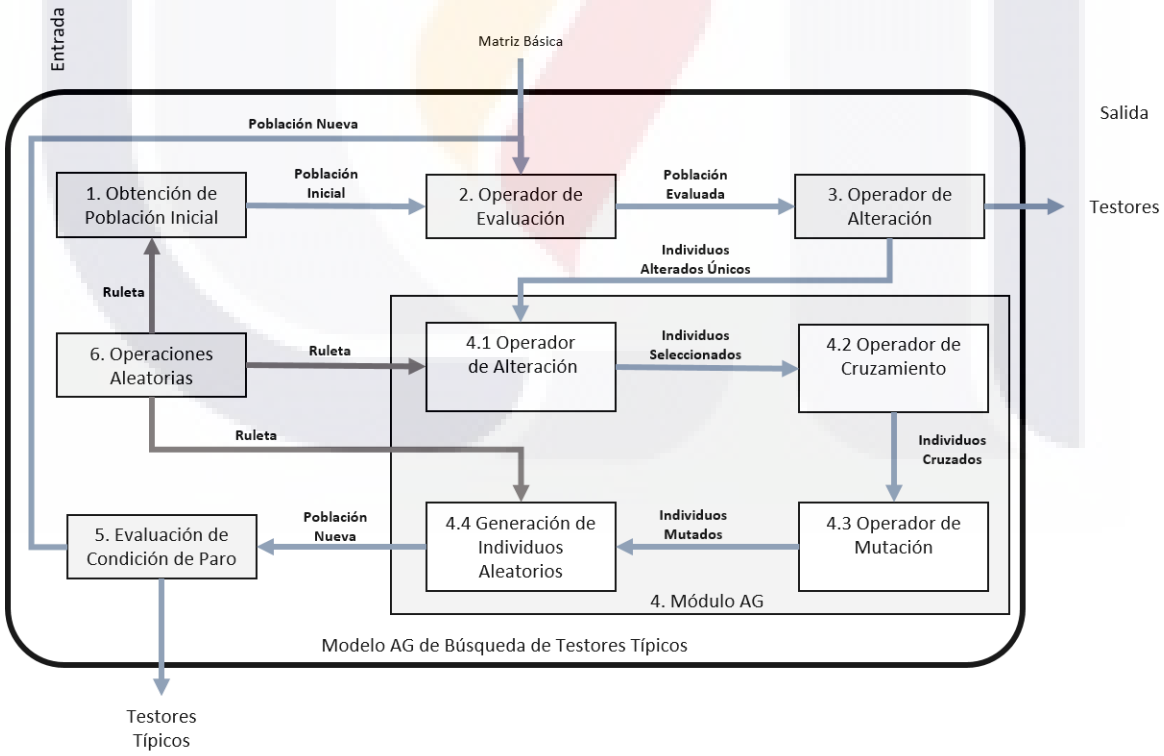
Para la búsqueda de testores típicos por medio de AG, se describe en la figura a continuación el modelo estático. Este modelo define los componentes en los que se divide el sistema como unidades de construcción.

Como se describió en la metodología del apartado 3, se trabajó con dos metaheurísticas que fueron hibridadas para mejorar el desempeño de la búsqueda del espacio de soluciones. Como se puede observar en la figura se incluyó un nuevo operador al proceso de la metaheurística AG, el operador de alteración, el cual fue descrito a detalle en el apartado 4.2.2. De manera general, el *AG híbrido* implementado requiere de la *matriz básica* del problema como entrada, la cual es utilizada para evaluar la población en cada iteración.



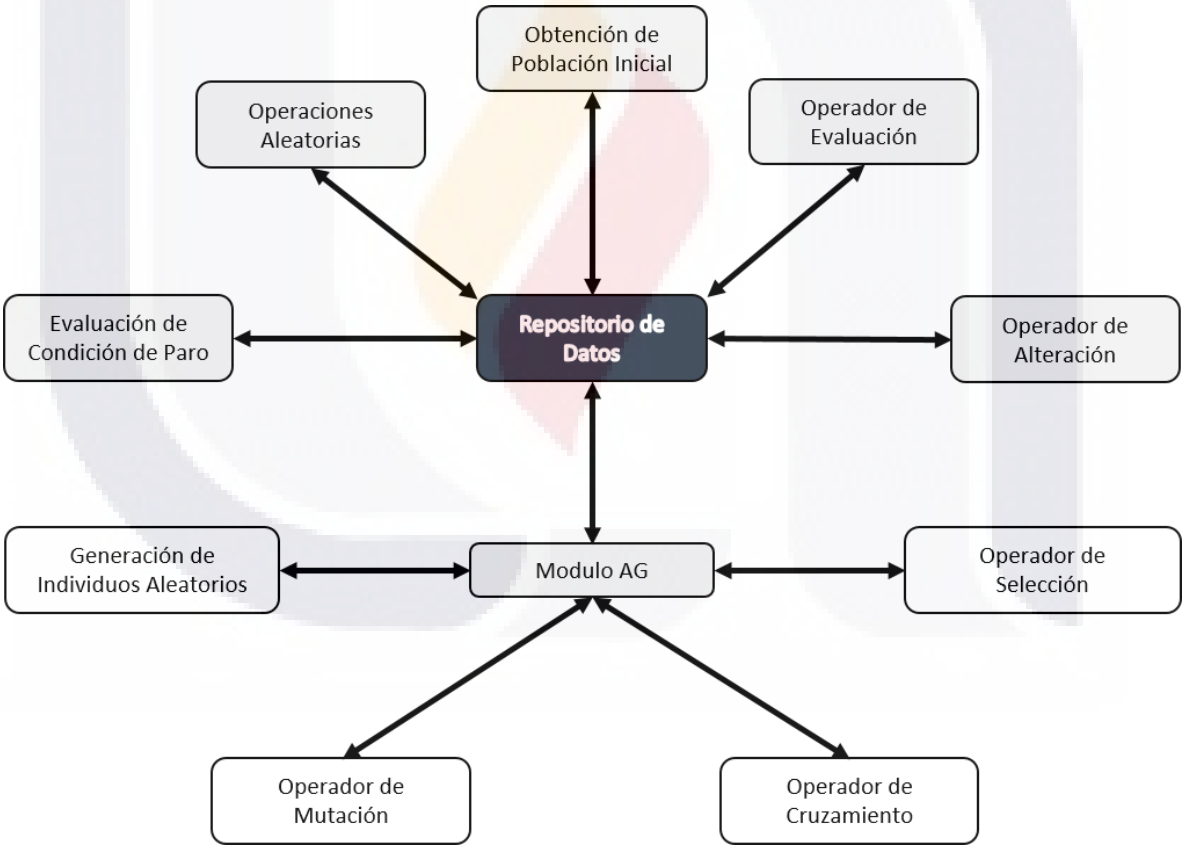
Modelo Dinámico

El *modelo dinámico* muestra el procesamiento del *algoritmo genético hibridado*, cuyo funcionamiento es descrito a detalle en el apartado 4.2.3. El procesamiento inicia con la generación de una población inicial, la cual se crea por medio del procedimiento descrito en el apartado 4.2.1. Como se mencionó en el modelo anterior, el recurso de entrada es la *matriz básica* utilizada por el operador de evaluación como se muestra en el diagrama a continuación. El siguiente paso fue procesar la población por el operador de alteración detallado en el apartado 4.2.2, con la intención de obtener más soluciones buenas a partir de las ya obtenidas. A partir de este punto se procesa la población por medio de los operadores comunes del AG simple del apartado 2.5.3. Para finalizar el procesamiento, se evalúa la condición de paro que, en caso de no ser cumplida, la nueva población generada por los operadores sustituye a la población anterior comenzando una nueva iteración.



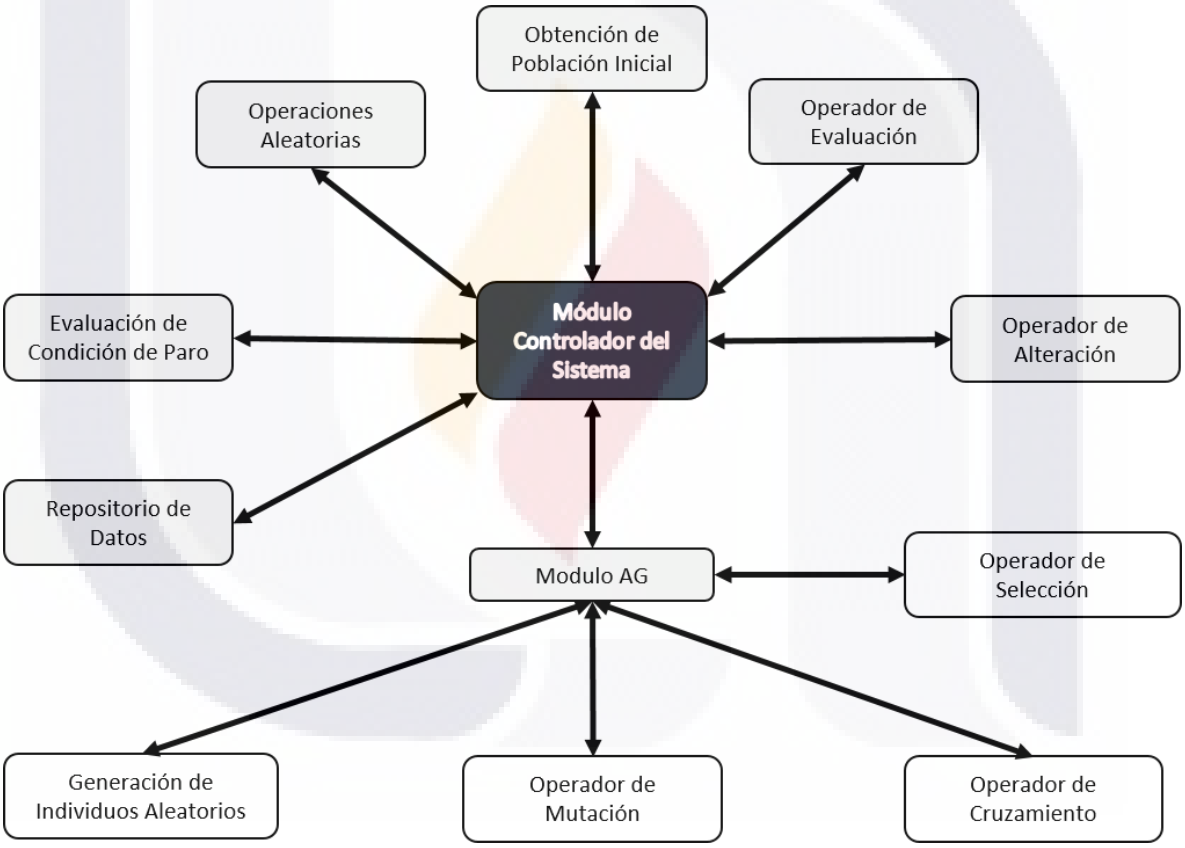
Modelo de Organización del Sistema

Debido a que el problema a tratar crece exponencialmente según la cantidad de variables o características involucradas, se determinó trabajar con archivos de texto. De esta manera, las poblaciones generadas son almacenadas en archivos y así, ser procesadas por cada operador del algoritmo. Por tanto, el modelo de organización seleccionado fue el *modelo de repositorio*, en el cual, como se observa en la figura a continuación, se integra *un módulo de repositorio* en el que cada componente obtiene recursos y almacena sus resultados. De esta manera, no es necesario que los módulos conozcan cómo es utilizada la información que generan por parte de otros módulos.



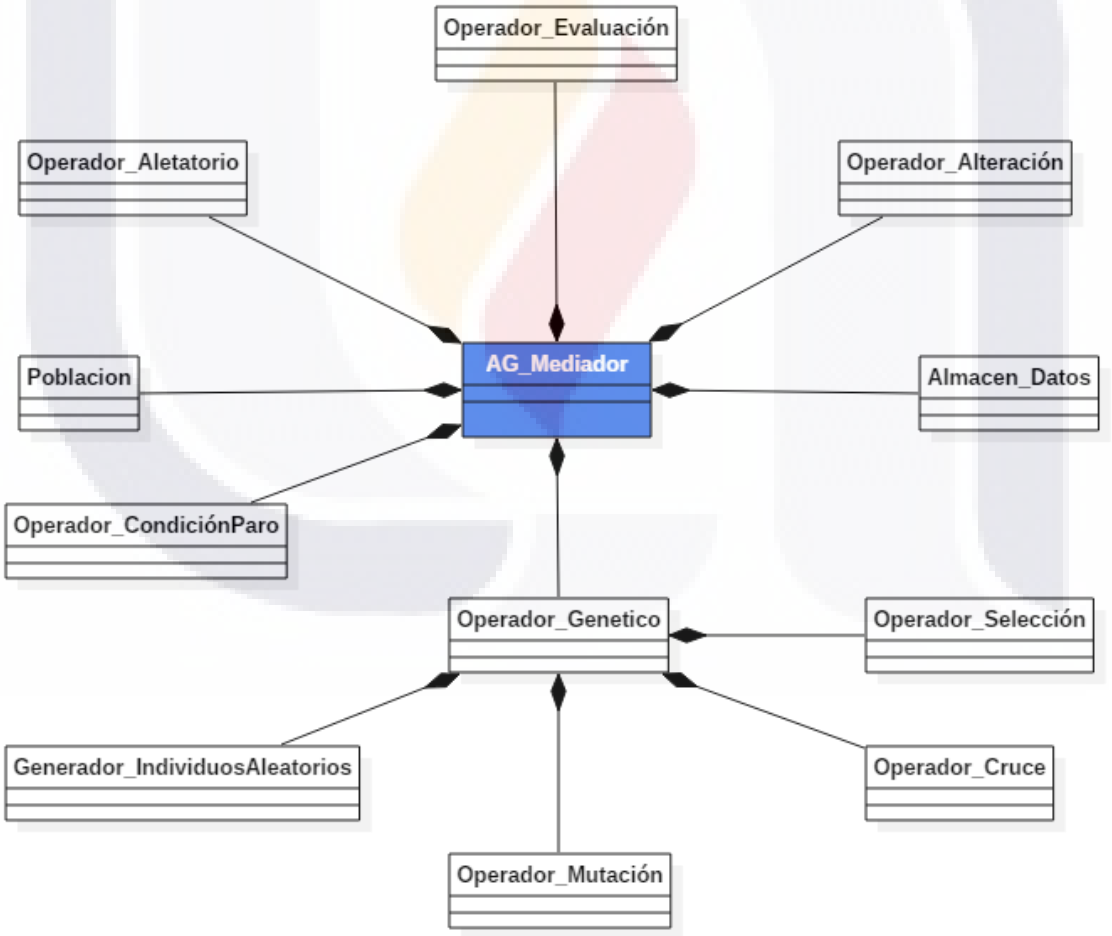
Modelo de Control

Para asegurar el correcto funcionamiento del algoritmo en tiempo de ejecución se empleó un modelo de control conocido como modelo centralizado. Este modelo requiere de la integración de un nuevo componente o módulo al sistema, cuya tarea principal es la gestión de cada uno de los componentes que integran el sistema. De esta manera se encarga de dar un orden de ejecución y de que cada uno de los componentes reciba los recursos necesarios, así como de recibir las salidas, que son recursos para el resto de los módulos.



A.4 Patrón de Diseño: Búsqueda de Testores Típicos por AG

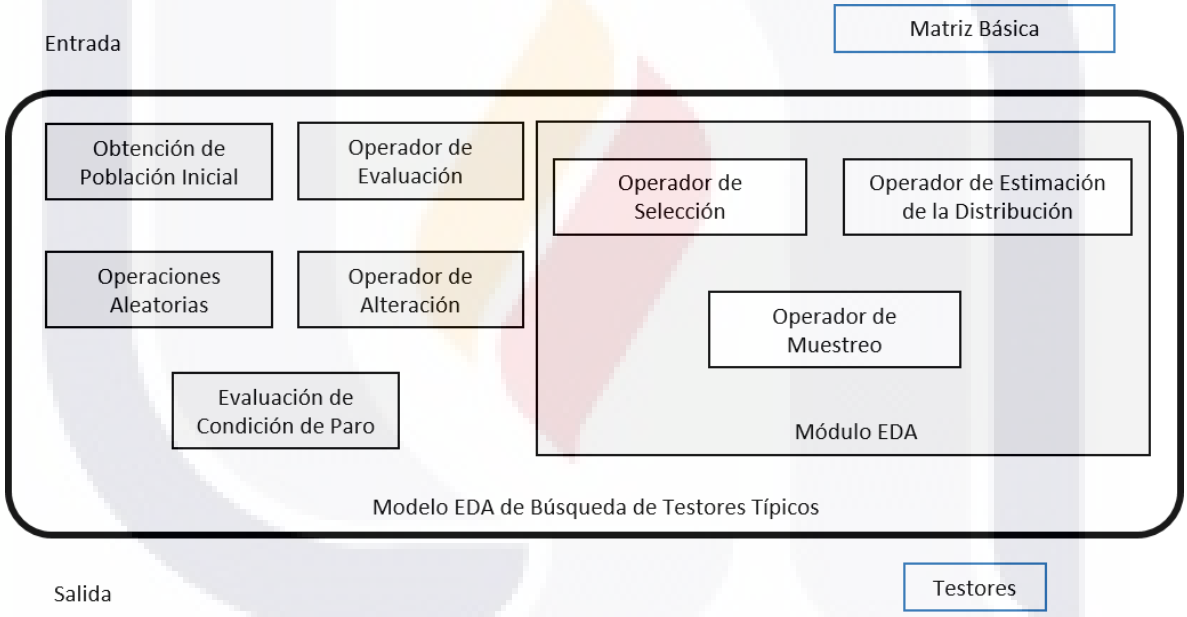
Para facilitar la implementación de un sistema encargado de la búsqueda de *testores típicos* por medio del *AG híbrido* descrito en el apartado 4.2.3 se hizo uso del *modelo mediador*, el cual forma parte de los *patrones de diseño* que propone la *ingeniería de software*. El modelo seleccionado permitió encapsular el funcionamiento del *algoritmo genético* con las hibridaciones diseñadas en nuevo componente de manera que, dicho componente define la manera en que interactúa el resto. Finalmente, el patrón de diseño facilitó la implementación y pruebas del sistema permitiendo, además, realizar cambios a los componentes sin la necesidad de alterar toda la estructura del sistema.



A.5 Arquitectura de Software: Búsqueda de Testores Típicos por EDA

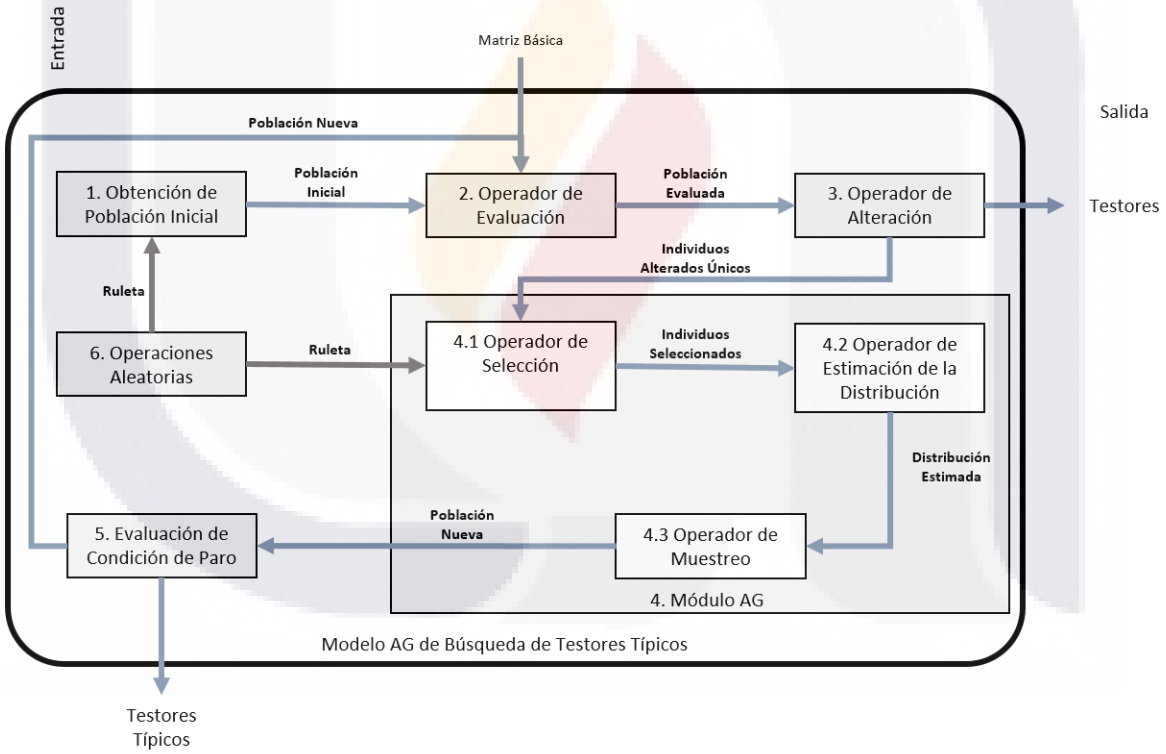
Modelo Estático

La figura a continuación describe el modelo estático para la bús*queda de testores típicos* por medio del *algoritmo de estimación de la distribución híbrido*, el cual es descrito en el apartado 4.3.4. En este modelo se identifican los componentes básicos del sistema, donde cada uno de ellos cumple un paso en el algoritmo. Al igual que en el *AG*, el *EDA* requiere de la *matriz básica* como herramienta de evaluación de las soluciones en los conjuntos generados en cada iteración hasta que la condición de paro sea satisfecha.



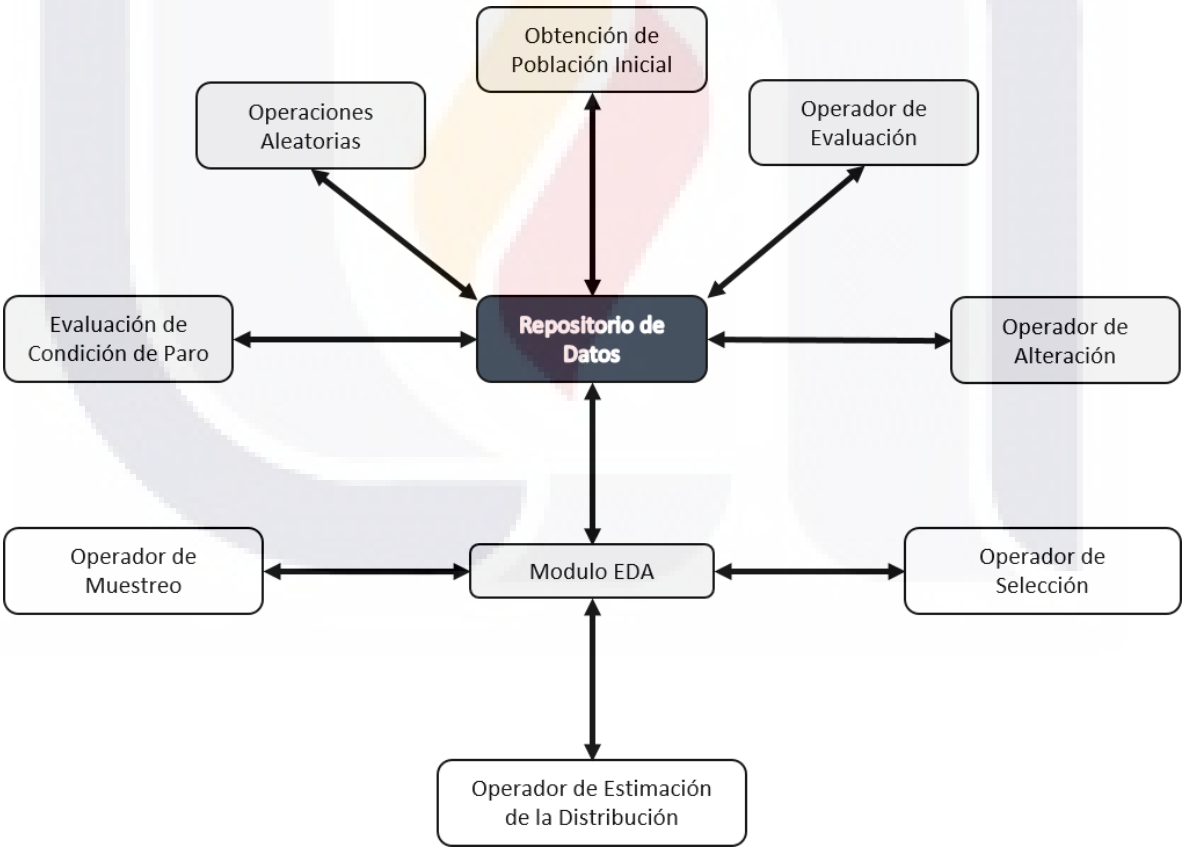
Modelo Dinámico

El diagrama a continuación muestra el procesamiento de la información durante la ejecución del *AG Híbrido*. Como se puede observar, el procesamiento comienza con la generación de una *población inicial* con ayuda del módulo de *operaciones aleatorias*. A continuación, dicha población es *evaluada* y *alterada* para después crear una *nueva población* por medio de los operadores *EDA* (*selección, estimación y muestreo*). Al final de la iteración se evalúa la *condición de paro* para determinar si el algoritmo continúa con otra iteración o finaliza el procesamiento. Para conocer más sobre el EDA general, el lector puede consultar el apartado 2.5.4.



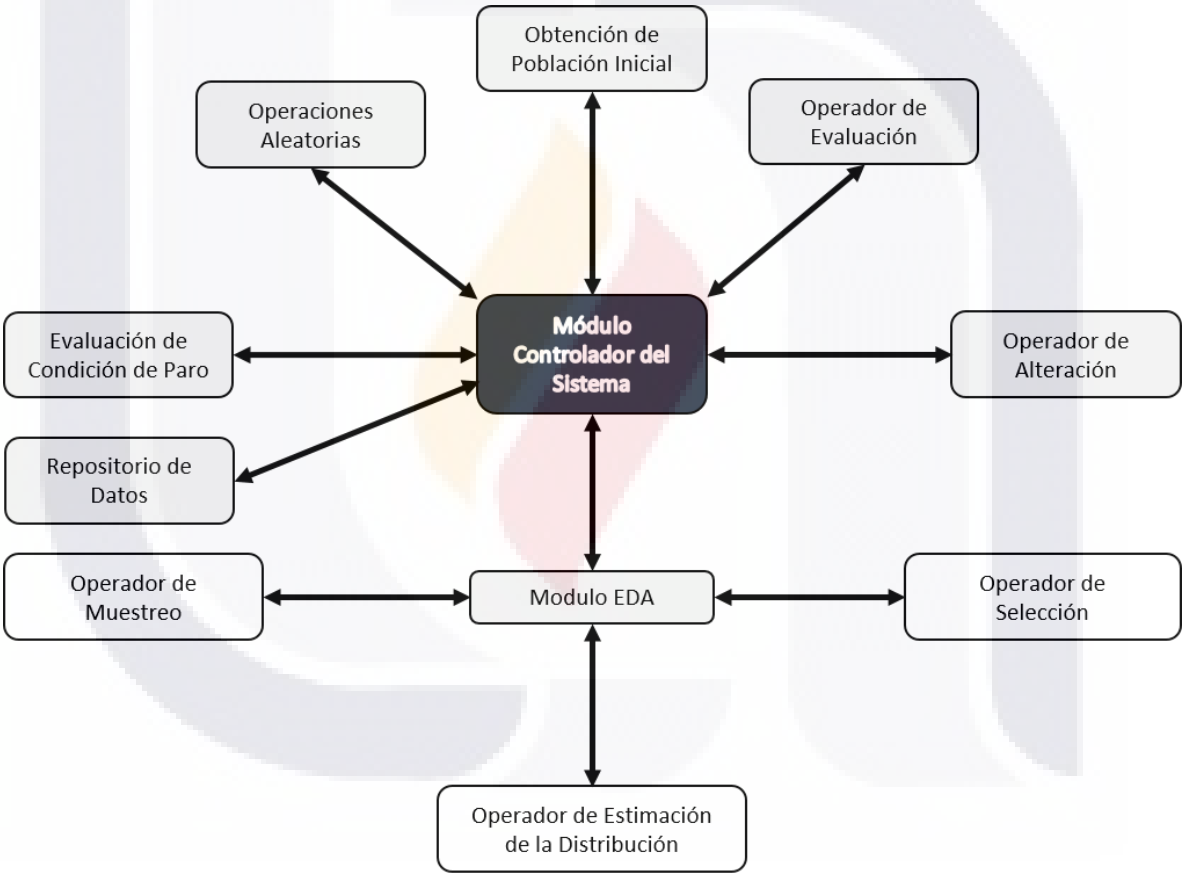
Modelo de Organización del Sistema

Como se ha mencionado, la búsqueda de testores típicos es un problema exponencial se acuerdo con el número de características con las que se trabaja. Por esta razón, se decidió trabajar con almacenamiento secundario por medio de archivos de texto, los cuales almacenan las poblaciones de soluciones, los testores y los testores típicos encontrados. Para facilitar el acceso al almacenamiento por parte de los componentes se seleccionó el modelo de organización de repositorio, el cual integra un nuevo módulo de repositorio en el que cada uno de los componentes obtiene recursos y almacena sus resultados. De esta manera, no es necesario que los módulos conozcan cómo es utilizada la información que generan por parte de otros módulos.



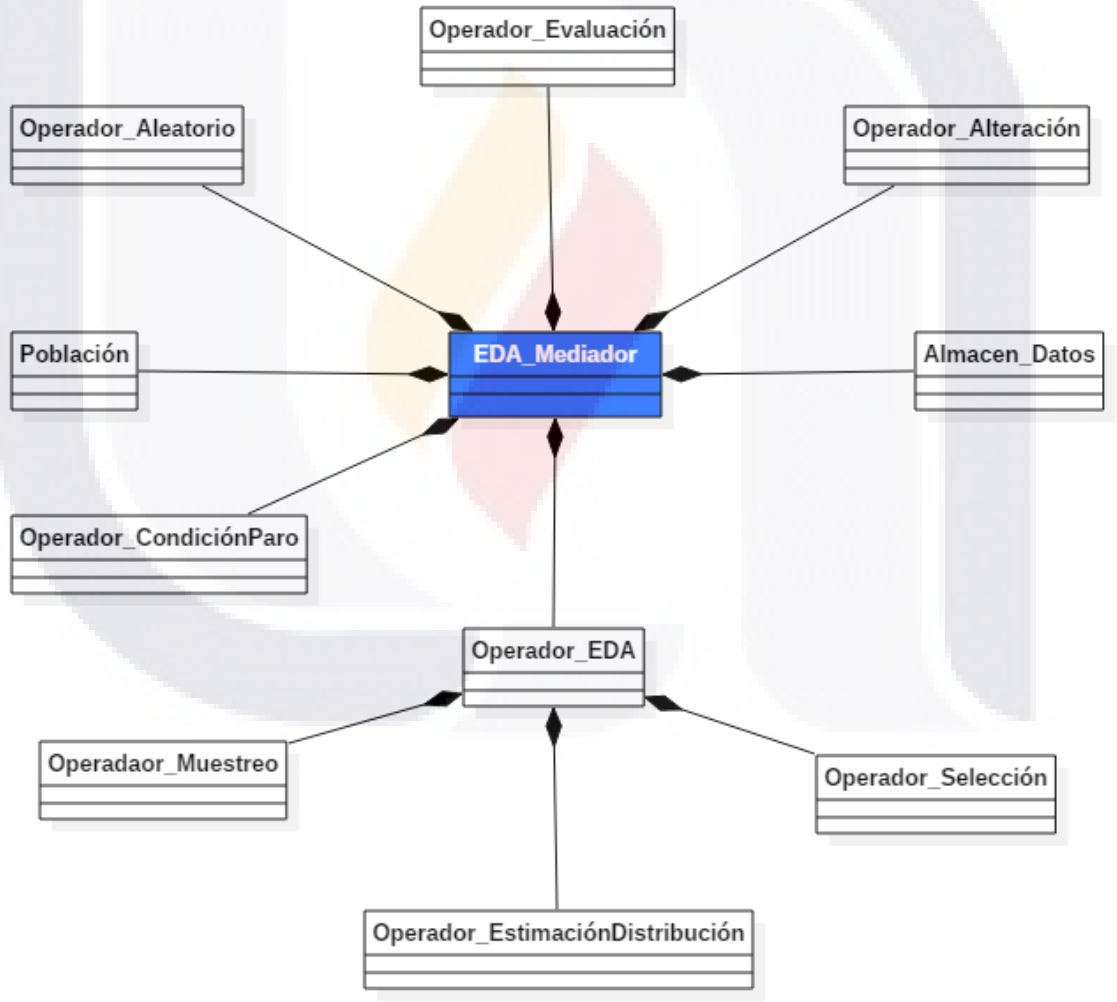
Modelo de Control

En este modelo de integra un nuevo *módulo al sistema*, el cual es conocido como *modelo centralizado*. El objetivo de este nuevo módulo es asegurar el correcto funcionamiento del algoritmo en tiempo de ejecución. En otras palabras, se encarga de dar un orden de ejecución y de que cada uno de los componentes reciba los recursos necesarios, así como de recibir las salidas, que son recursos para el resto de los módulos.



A.6 Patrón de Diseño: Búsqueda de Testores Típicos por EDA

El *patrón de diseño* seleccionado para dar solución a la implementación del *EDA híbrido para la búsqueda de testores típicos* (ver apartado 4.3.2) se seleccionó el *modelo mediador*, el cual permite encapsular el funcionamiento del algoritmo de manera que, define la forma en que interactúa el resto de los componentes. Finalmente, el *patrón de diseño* facilitó la implementación y pruebas del sistema permitiendo, además, realizar cambios a los componentes sin la necesidad de alterar toda la estructura del sistema.



B. Pruebas Estadísticas para Afinación de AG Aplicado a Cáncer de Mama

B.1 Prueba de Levene

ANOVA de un factor

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PORCENTAJE TESTORES ECONTRADOS	Inter-grupos	101996.465	399	255.630	1.298	.000
	Intra-grupos	2284069.527	11600	196.903		
	Total	2386065.992	11999			
PORCENTAJE TESTORES TIPCOS ENCONTRADOS	Inter-grupos	659.167	399	1.652	1.337	.000
	Intra-grupos	14333.333	11600	1.236		
	Total	14992.500	11999			
TIEMPO UTILIZADO (S)	Inter-grupos	15.098	399	.038	2.279	.000
	Intra-grupos	192.608	11600	.017		
	Total	207.706	11999			
ITERACIONES REALIZADAS	Inter-grupos	181896.458	399	455.881	15.329	.000
	Intra-grupos	344973.533	11600	29.739		
	Total	526869.992	11999			

➔ Se rechaza H_0, por tanto, los grupos no son homocedásticos existiendo entonces diferencias entre las varianzas.

B.2 Prueba de Kolomogorov Smirnov

Prueba de Kolmogorov-Smirnov para una muestra

		PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES TIPCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
N		12000	12000	12000	12000
Parámetros normales[a,b]	Media	92.238897	99.98	.166389	5.57
	Desviación típica	14.1016100	1.118	.1315684	6.626
Diferencias más extremas	Absoluta	.476	.508	.174	.248
	Positiva	.291	.491	.174	.248
	Negativa	-.476	-.508	-.144	-.245
Z de Kolmogorov-Smirnov		52.176	55.695	19.081	27.115
Sig. asintót. (bilateral)		.000	.000	.000	.000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

➔ La población de experimentos no sigue una distribución normal, por lo tanto, usar estadística no paramétrica.

B.3 Prueba de Kruskal Wallis

Estadísticos de contraste[a,b]

	PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES TIPCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
Chi-cuadrado	511.567	527.553	1525.145	5046.489
gl	399	399	399	399
Sig. asintót.	.000	.000	.000	.000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: N? EXPERIMENTO

➔ Existe diferencia significativa en los 400 grupos estudiados en las 4 variables continuas.

C. Pruebas Estadísticas para Afinación de EDA Aplicado a Cáncer de Mama

C.1 Prueba de Levene

ANOVA de un factor

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PORCENTAJE TESTORES ECONTRADOS	Inter-grupos	83492.810	399	209.255	1.392	.000
	Intra-grupos	1744144.817	11600	150.357		
	Total	1827637.628	11999			
PORCENTAJE TESTORES T?PCOS ENCONTRADOS	Inter-grupos	37770.000	399	94.662	3.351	.000
	Intra-grupos	327666.667	11600	28.247		
	Total	365436.667	11999			
TIEMPO UTILIZADO (S)	Inter-grupos	34.606	399	.087	1.396	.000
	Intra-grupos	720.636	11600	.062		
	Total	755.242	11999			
ULTIMA ITERACION	Inter-grupos	609066.637	399	1526.483	10.842	.000
	Intra-grupos	1633270.833	11600	140.799		
	Total	2242337.470	11999			

➔ Se rechaza H_0, por tanto, los grupos no son homocedásticos existiendo entonces diferencias entre las varianzas.

C.2 Prueba de Kolmogorov Smirnov

Prueba de Kolmogorov-Smirnov para una muestra

		PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ULTIMA ITERACION
N		12000	12000	12000	12000
Parámetros normales[a,b]	Media	94.650005	99.38	.206076	8.58
	Desviación típica	12.3416299	5.519	.2508827	13.670
Diferencias más extremas	Absoluta	.508	.532	.246	.289
	Positiva	.332	.456	.246	.265
	Negativa	-.508	-.532	-.224	-.289
Z de Kolmogorov-Smirnov		55.686	58.294	26.907	31.713
Sig. asintót. (bilateral)		.000	.000	.000	.000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

➔ La población de experimentos no sigue una distribución normal, por lo tanto, usar estadística no paramétrica.

C.3 Prueba de Kruskal Wallis

Estadísticos de contraste[a,b]

	PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
Chi-cuadrado	511.567	527.553	1525.145	5046.489
gl	399	399	399	399
Sig. asintót.	.000	.000	.000	.000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: N? EXPERIMENTO

➔ Existe diferencia significativa en los 400 grupos estudiados en las 4 variables continuas.

D. Pruebas Estadísticas para la Contrastación de Metaheurísticas Aplicadas a Cáncer de Mama

D.1 Prueba de Levene

ANOVA de un factor

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PORCENTAJE TESTORES ECONTRADOS	Inter-grupos	166.666	1	166.666	1.442	.235
	Intra-grupos	6703.690	58	115.581		
	Total	6870.357	59			
PORCENTAJE TESTORES TÍPCOS ENCONTRADOS	Inter-grupos	.000	1	.000	.	.
	Intra-grupos	.000	58	.000		
	Total	.000	59			
TIEMPO UTILIZADO (S)	Inter-grupos	8.728	1	8.728	12.203	.001
	Intra-grupos	41.484	58	.715		
	Total	50.212	59			
ITERACIONES REALIZADAS	Inter-grupos	1440.600	1	1440.600	15.252	.000
	Intra-grupos	5478.333	58	94.454		
	Total	6918.933	59			

- ➔ Los algoritmos AG y EDA aplicados al contexto del cáncer de mama son homocedásticos respecto al porcentaje de testores encontrados.
- ➔ En ambos algoritmos se alcanza el objetivo de encontrar los testores típicos.
- ➔ Los algoritmos no son homocedásticos respecto al tiempo utilizado y el número de iteraciones realizadas.

D.2 Prueba de Kolmogorov Smirnov

Prueba de Kolmogorov-Smirnov para una muestra

		PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES TÍPCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
N		60	60	60	60
Parámetros normales[a,b]	Media	96.111115	100.00	.609533	9.13
	Desviación típica	10.7910483	.000[c]	.9225251	10.829
Diferencias más extremas	Absoluta	.524		.276	.238
	Positiva	.359		.276	.201
	Negativa	-.524		-.271	-.238
Z de Kolmogorov-Smirnov		4.059		2.140	1.846
Sig. asintót. (bilateral)		.000		.000	.002

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

c. La distribución no tiene varianza para esta variable. No es posible realizar la prueba de Kolmogorov-Smirnov para una muestra.

➔ Los algoritmos no siguen una distribución normal por lo que se usaron pruebas estadísticas no paramétricas.

D.3 Prueba de Kruskal Wallis

Estadísticos de contraste[a,b]

	PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES TÍPCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
Chi-cuadrado	1.431	.000	26.875	23.542
gl	1	1	1	1
Sig. asintót.	.232	1.000	.000	.000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: N° EXPERIMENTO

- ➔ Existe diferencia significativa en los grupos respecto al tiempo utilizado y el número de iteraciones realizadas.
- ➔ No existe diferencia en los grupos respecto al porcentaje de testores típicos encontrados. Ambos grupos (AG y EDA) encuentran el 100% de los testores típicos en el 100% de las réplicas ejecutadas.
- ➔ No existe diferencia en los grupos respecto al porcentaje de testores encontrados.

E. Pruebas Estadísticas para Afinación de AG Aplicado a Hemofilia

E.1 Prueba de Levene

ANOVA de un factor

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PORCENTAJE TESTORES ECONTRADOS	Inter-grupos	977507.262	399	2449.893	75.254	.000
	Intra-grupos	377639.221	11600	32.555		
	Total	1355146.484	11999			
PORCENTAJE TESTORES T?PCOS ENCONTRADOS	Inter-grupos	.000	399	.000	.	.
	Intra-grupos	.000	11600	.000		
	Total	.000	11999			
TIEMPO UTILIZADO (S)	Inter-grupos	1717.321	399	4.304	12.330	.000
	Intra-grupos	4049.159	11600	.349		
	Total	5766.480	11999			
ITERACIONES REALIZADAS	Inter-grupos	1980.859	399	4.965	15.080	.000
	Intra-grupos	3818.800	11600	.329		
	Total	5799.659	11999			

- ➔ Las variables porcentaje de testores, tiempo e iteraciones realizadas no son homocedásticas.
- ➔ El porcentaje de testores típicos en hocedástico.

E.2 Prueba de Kolmogorov Smirnov

Prueba de Kolmogorov-Smirnov para una muestra

		PORCENTAJE TESTORES ECONTRADOS	TIEMPO UTILIZADO (S)	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	ITERACIONES REALIZADAS
N		12000	12000	12000	12000
Parámetros normales[a,b]	Media	56.800251	2.091947	100.00	1.99
	Desviación típica	10.6272426	.6932388	.000[c]	.695
Diferencias más extremas	Absoluta	.068	.106		.346
	Positiva	.068	.106		.346
	Negativa	-.055	-.074		-.310
Z de Kolmogorov-Smirnov		7.438	11.573		37.868
Sig. asintót. (bilateral)		.000	.000		.000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

c. La distribución no tiene varianza para esta variable. No es posible realizar la prueba de Kolmogorov-Smirnov para una muestra.

➔ Los variables porcentaje de testores encontrados, tiempo utilizado y las iteraciones realizadas no siguen una distribución normal. Por lo tanto, seguir estadística no paramétrica.

E.3 Prueba de Kruskal Wallis

Estadísticos de contraste[a,b]

	PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
Chi-cuadrado	9013.204	.000	3967.256	4356.170
gl	399	399	399	399
Sig. asintót.	.000	1.000	.000	.000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: N? EXPERIMENTO

- ➔ Existe diferencia significativa en los grupos en cuanto a porcentaje de testores encontrados, el tiempo y el número de iteraciones realizadas.
- ➔ No existe diferencia significativa desde el punto de vista del porcentaje de testores típicos encontrados.

F. Pruebas Estadísticas para Afinación de EDA Aplicado a Hemofilia

F.1 Prueba de Levene

ANOVA de un factor

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PORCENTAJE TESTORES ECONTRADOS	Inter-grupos	1405232.296	399	3521.885	69.078	.000
	Intra-grupos	591419.260	11600	50.984		
	Total	1996651.556	11999			
PORCENTAJE TESTORES T?PCOS ENCONTRADOS	Inter-grupos	.000	399	.000	.	.
	Intra-grupos	.000	11600	.000		
	Total	.000	11999			
TIEMPO UTILIZADO (S)	Inter-grupos	1306.050	399	3.273	11.917	.000
	Intra-grupos	3186.303	11600	.275		
	Total	4492.353	11999			
ULTIMA ITERACION	Inter-grupos	1650.167	399	4.136	15.628	.000
	Intra-grupos	3069.800	11600	.265		
	Total	4719.967	11999			

- ➔ Los grupos son homocedásticos respecto al porcentaje de testores típicos encontrados
- ➔ Los grupos no son homocedásticos respecto al porcentaje de testores, el tiempo utilizado y el número de iteraciones realizadas.

F.2 Prueba de Kolmogorov Smirnov

Prueba de Kolmogorov-Smirnov para una muestra

		PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ULTIMA ITERACION
N		12000	12000	12000	12000
Parámetros normales[a,b]	Media	60.772450	100.00	1.696558	1.95
	Desviación típica	12.8996704	.000[c]	.6118774	.627
Diferencias más extremas	Absoluta	.074		.138	.346
	Positiva	.072		.138	.346
	Negativa	-.074		-.097	-.337
Z de Kolmogorov-Smirnov		8.154		15.144	37.848
Sig. asintót. (bilateral)		.000		.000	.000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

c. La distribución no tiene varianza para esta variable. No es posible realizar la prueba de Kolmogorov-Smirnov para una muestra.

- ➔ Los grupos no siguen una distribución normal desde respecto a las variables porcentaje de testores, tiempo y las iteraciones realizadas.
- ➔ El porcentaje de testores típicos es constante en todos los grupos encontrando el 100% del conjunto esperado.
- ➔ Usar estadística no paramétrica.

F.3 Prueba de Kruskal Wallis

Estadísticos de contraste[a,b]

	PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ULTIMA ITERACION
Chi-cuadrado	9174.755	.000	4147.609	4296.235
gl	399	399	399	399
Sig. asintót.	.000	1.000	.000	.000

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: N? EXPERIMENTO

- ➔ No existe diferencia significativa respecto al porcentaje de testores típicos encontrados.
- ➔ Existe diferencia significativa respecto al porcentaje de testores, el tiempo utilizado y el número de iteraciones realizadas.

G. Pruebas Estadísticas para la Contrastación de Metaheurísticas Aplicadas a Hemofilia

G.1 Prueba de Levene

ANOVA de un factor

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PORCENTAJE TESTORES ECONTRADOS	Inter-grupos	759.996	1	759.996	29.106	.000
	Intra-grupos	1436.101	55	26.111		
	Total	2196.097	56			
PORCENTAJE TESTORES T?PCOS ENCONTRADOS	Inter-grupos	.000	1	.000	.	.
	Intra-grupos	.000	55	.000		
	Total	.000	56			
TIEMPO UTILIZADO (S)	Inter-grupos	.127	1	.127	.128	.721
	Intra-grupos	54.494	55	.991		
	Total	54.621	56			
ITERACIONES REALIZADAS	Inter-grupos	.078	1	.078	.240	.626
	Intra-grupos	17.852	55	.325		
	Total	17.930	56			

- ➔ Las metaheurísticas son homocedásticas respecto al tiempo utilizado y el número de iteraciones realizadas.
- ➔ Los algoritmos no son homcedásticos respecto al porcentaje de testores encontrados.
- ➔ En ambos casos, las réplicas de los experimentos logran encontrar el 100% de los testores típicos.

G.2 Prueba de Kolmogorov Smirnov

Prueba de Kolmogorov-Smirnov para una muestra

		PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
N		57	57	57	57
Parámetros normales[a,b]	Media	76.005823	100.00	2.848070	2.30
	Desviación típica	6.2622701	.000[c]	.9876111	.566
Diferencias más extremas	Absoluta	.144		.221	.455
	Positiva	.144		.221	.455
	Negativa	-.130		-.152	-.299
Z de Kolmogorov-Smirnov		1.087		1.671	3.438
Sig. asintót. (bilateral)		.188		.007	.000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

c. La distribución no tiene varianza para esta variable. No es posible realizar la prueba de Kolmogorov-Smirnov para una muestra.

- ➔ Las metaheurísticas siguen una distribución normal respecto al porcentaje de testores encontrados.
- ➔ No existe normalidad en cuanto al tiempo utilizado y el número de iteraciones realizadas.

G.3 Prueba de Kruskal Wallis

Estadísticos de contraste[a,b]

	PORCENTAJE TESTORES ECONTRADOS	PORCENTAJE TESTORES T?PCOS ENCONTRADOS	TIEMPO UTILIZADO (S)	ITERACIONES REALIZADAS
Chi-cuadrado	13.226	.000	3.035	.725
gl	1	1	1	1
Sig. asintót.	.000	1.000	.081	.394

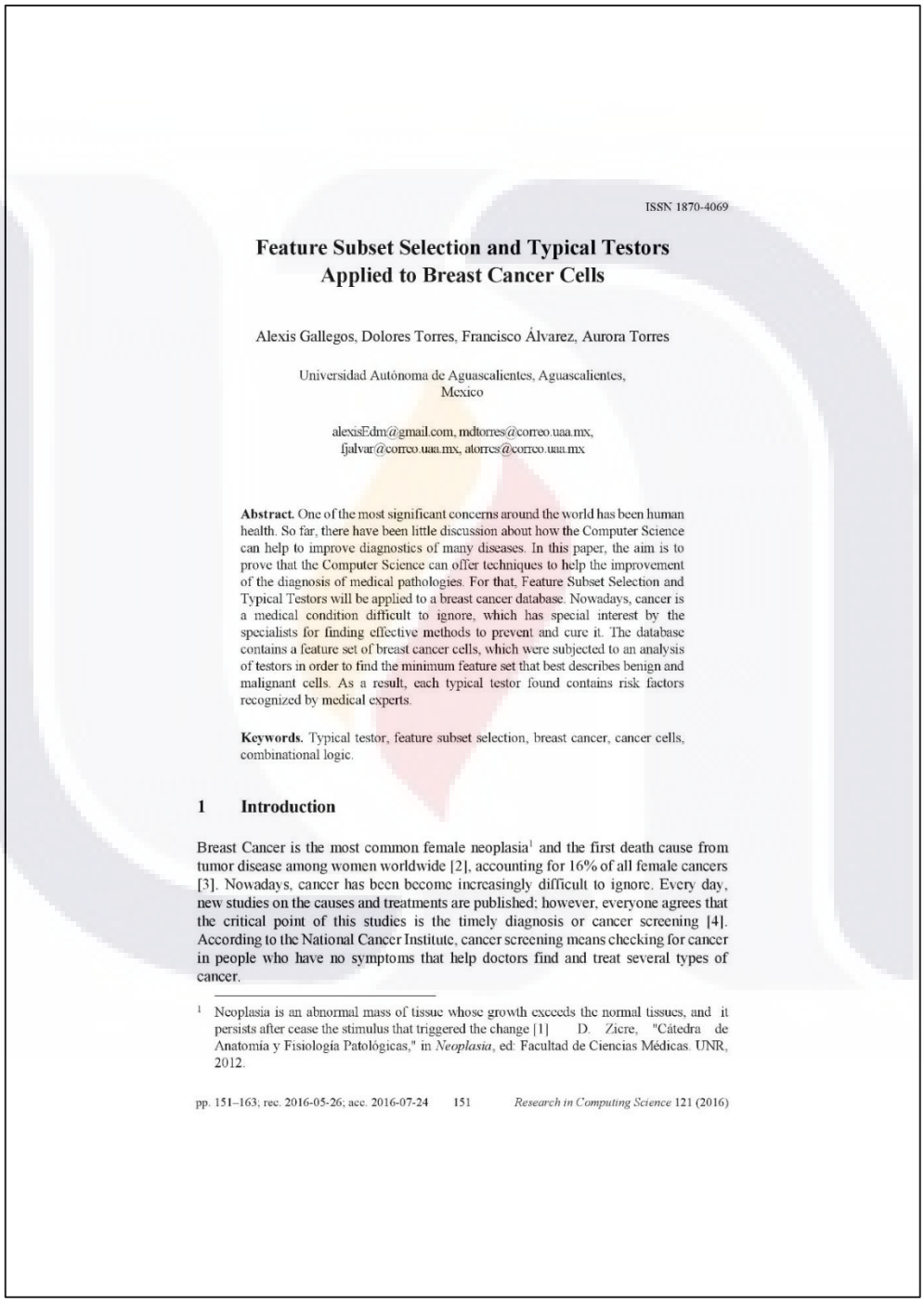
a. Prueba de Kruskal-Wallis

b. Variable de agrupación: N? EXPERIMENTO

- ➔ Existe diferencia significativa entre los experimentos respecto al porcentaje de testores encontrados.
- ➔ No existe diferencia significativa entre los experimentos de acuerdo al porcentaje de testores típicos encontrados, el tiempo utilizado y el número de iteraciones realizadas.

H. Productos

H.1 Feature Subset Selection and Typical Testors Applied to Breast Cancer Cells



ISSN 1870-4069

Feature Subset Selection and Typical Testors Applied to Breast Cancer Cells

Alexis Gallegos, Dolores Torres, Francisco Álvarez, Aurora Torres

Universidad Autónoma de Aguascalientes, Aguascalientes, Mexico

alexisEdm@gmail.com, mdtorres@correo.uaa.mx, fjalvar@correo.uaa.mx, atorres@correo.uaa.mx

Abstract. One of the most significant concerns around the world has been human health. So far, there have been little discussion about how the Computer Science can help to improve diagnostics of many diseases. In this paper, the aim is to prove that the Computer Science can offer techniques to help the improvement of the diagnosis of medical pathologies. For that, Feature Subset Selection and Typical Testors will be applied to a breast cancer database. Nowadays, cancer is a medical condition difficult to ignore, which has special interest by the specialists for finding effective methods to prevent and cure it. The database contains a feature set of breast cancer cells, which were subjected to an analysis of testors in order to find the minimum feature set that best describes benign and malignant cells. As a result, each typical testor found contains risk factors recognized by medical experts.

Keywords. Typical testor, feature subset selection, breast cancer, cancer cells, combinational logic.

1 Introduction

Breast Cancer is the most common female neoplasia[1] and the first death cause from tumor disease among women worldwide [2], accounting for 16% of all female cancers [3]. Nowadays, cancer has been become increasingly difficult to ignore. Every day, new studies on the causes and treatments are published; however, everyone agrees that the critical point of this studies is the timely diagnosis or cancer screening [4]. According to the National Cancer Institute, cancer screening means checking for cancer in people who have no symptoms that help doctors find and treat several types of cancer.

[1] Neoplasia is an abnormal mass of tissue whose growth exceeds the normal tissues, and it persists after cease the stimulus that triggered the change [1] D. Zicre, "Cátedra de Anatomía y Fisiología Patológicas," in *Neoplasia*, ed: Facultad de Ciencias Médicas. UNR, 2012.

Early detection is important because when abnormal tissue or cancer is found timely, it may be easier to treat. By the time symptoms appear, cancer has begun to spread and is harder to treat [5]. However, there are effective methods for screening only for some cancers.

Despite the usefulness of early detection, it can have some risks, as well as the methods used. For example, screening test can present a false-positive results; it means that the test indicates that the cancer is present when this is not true. Also, the test can have false-negative results, this indicates that cancer is not present even though it is.

Furthermore, over diagnosis is possible, this happens when screening test correctly shows that a person has cancer, but the cancer is slow growing and would not have harmed that person in his or her lifetime [5]. This justifies the need to improve the diagnosis of cancer.

The clinical diagnosis is a cognitive process that starts from the sensitive concrete thinking. It is linked to objective reality; it develops in abstract thinking and has the criterion of truth in practice [6]. It involves training, experience, pattern recognition and calculation of conditional probability, among other components, that human treatment has and therefore, is not free of errors that can cause sickness, damages, expenses and even death, especially in sensitive diseases such as cancer [7].

The errors represent an estimated 150 out of 1000 patients with misdiagnosis [8]. As such, the medical field is one of the areas that could be most benefit from close interaction with Computing Science and Mathematics to improve processes such as medical diagnosis [7]. Reason why it is decided to apply comprehensive mathematical methods to support diagnosis and prognosis of diseases such as cancer, in this case breast cancer.

This paper has been divided into three parts and organized the following way. The first part deals with important concepts in Typical Testors, Featured Subset Selection in computer science, Breast Cancer and its impact around the world. Next, the second section will examine the framework of this analysis, such as the previous research related to Testors Theory and the review of the methodology applied to breast cancer cells. Finally, the third section describes the results of the methodology and its review.

2 Important Concepts

2.1 Typical Testors

Testors Theory was formulated as an independent scientific direction of Mathematical Cybernetics in the 60's in the former Union of Soviet Socialist Republics (USSR), whose origin is linked to the use of mathematical logic methods for locating faults in electrical circuits that perform Boolean functions [9].

Later, testors were used to perform supervised classification and selection of variables in problems of geology [9, 10]. The use given in this article to the testors and typical testors is related to feature subset selection, whose precursor is Dmitriev, Zhuravlev and colleagues [10].

In this way, a testor is a subset of features that distinguishes objects from different classes [10]. According to Santiesteban and Pons [11], Shulcloper [9], and Torres [10], a typical testor is a testor that it is no longer possible to remove any feature without losing its status of testor. Otherwise, a typical testor is already formed by the minimum set of features needed to ensure the identification of the class to which a specific object belongs.

Typical testors determinne issues such as evaluation of informational weight of traits and selection of variables. They can reduce the dimension of the space of representation of objects [11] and the can be used as a set of support for classification algorithms [12]. So, consequently, the aim of this study is to prove that testor analysis can help to classify cells based on a real dataset; this will be explained in chapter 3.

2.2 Featured Subset Selection

Regularly, Featured Subset Selection (FSS) [13] is used to reduce dimensionality [14], which is used to efficiently reduce the number of variables, attributes or characteristics with which should describe the objects and to find their influence in a problem. This is an alternative method that starts by using the set of typical testors, taking out irrelevant or redundant features [11, 14].

FSS really has importance because reducing the number of features may help to decrease the cost of acquiring data and also make the classification models easier to understand [14, 15]. Also, the number of features could affect the accuracy of classification. Some authors have also studied the bias feature subset selection for classification learning [14].

The FSS problem has been studied by the statistics and machine learning communities for many years with high attention because of the enthusiastic research in data mining [16]. There are many potential benefits of variable and feature selection [17]:

- Facilitating data visualization and data understanding,
- Reducing the measurement and storage requirements,
- Reducing training and utilization times,
- Defying the curse of dimensionality to improve prediction performance.

Selecting the most relevant variables is usually suboptimal for building a predictor, particularly if the variables are redundant [17]. For example, the brute-force selection method evaluates exhaustively all possible combinations of the input features, and then finds the best subset [16]. Later, chapters 3 and 4 will describe a brute-force method applied to the already mentioned dataset related to breast cancer cell with the aim of evaluating if a cell is malign or benign and calculate the informational weight of every feature.

2.3 Informational Weight

The use of the informational weight for feature subset selection is an excellent tool that shows tangible results [14]. Informational weight of a feature is a score, in other words, is the measure of significance to predict whether an object belongs to a group or to another (Classification) [10, 18].

2.4 Breast Cancer

Firstly, cancer is a collection of related diseases. In all types of cancer, some body's cells begin to divide without stopping and spread into surrounding tissues. Cancer can start almost anywhere in the human body, which is made up of trillion of cells [19]. It is the result of mutations, or abnormal changes in the genes that regulate cell growth. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place [19, 20].

When cancer develops, this orderly process breaks down. Mutations can „turn on" certain genes and „turn off" others in a cell. The modified cell acquires the ability to divide without any control or order, which produces more identical cells and generates a tumor [19, 20].

Consequently, breast cancer is a malignant tumor that has been developed from the breast cells [21]. The breast is made up of glands called lobules that can make milk and thin tubes called ducts that carry milk from the lobules to the nipple, generally breast cancer originates in cells of those lobules [20, 21].

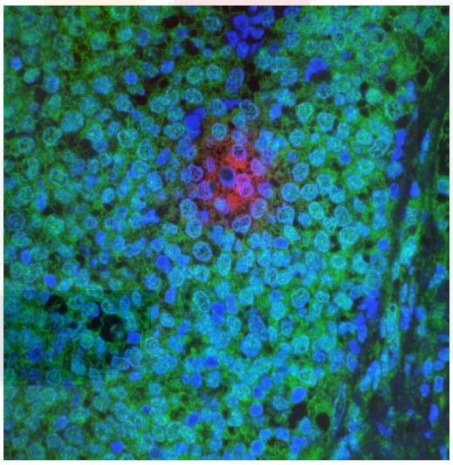


Fig. 1. Invasive breast cancer tumor [19].

Breast cancer has had a great impact worldwide, according to the Pan American Health Organization (PAHO) in the American continent breast cancer is the most common between the women with 29% of the cancer cases. PAHO estimates more than 596,000 new cases and more than 142,100 deaths in the region by 2030, mainly in the area of Latin America and the Caribbean [22]. Next figure shows the incidence of breast malignant tumors in women over age 20 years divided by age group, year 2014:

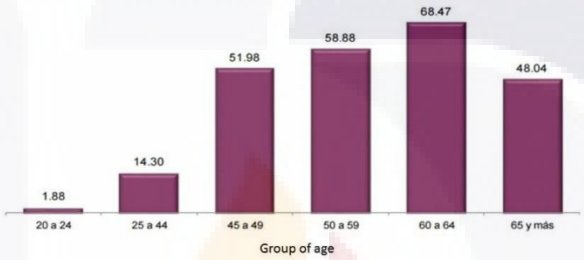


Fig. 2 Incidence of breast malignant tumors in women over age 20 years divided by age group. For 100 thousand women for each age group, INEGI [22].

In general, any type of cancer represents an important impact to the physical state of the person, his or her emotional sphere, a high cost of treatment and can even undermine the economy of the countries; so the prevention and a timely diagnosis are critical to address this problem [23]. Hence, this paper is focused in the application of Feature Subset Selection and Typical Testors to improve the diagnosis of cancer in body's cells. Next chapters will explain the study done.

3 Framework

The general methodology used for this paper is shown in figure 3 below:

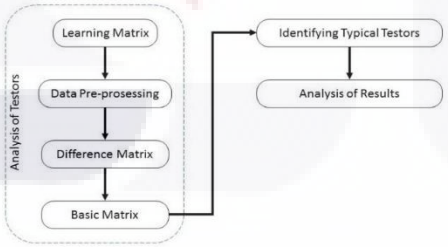


Fig. 3. General Methodology.

As seen if Figure 3 the methodology used needs a Learning Matrix (LM), which is the source of the information that contains descriptions of objects [9, 11]. For this paper, the LM comes from the University of California and their Machine Learning Repository. The repository is Wisconsin Diagnostic Breast Cancer [24].

The database contains the diagnosis and 10 features computed from a digitized image of a fine needle aspirate of a breast mass and describes characteristics of a cell nucleus present in the image [24]. The image below shows an example of this images.

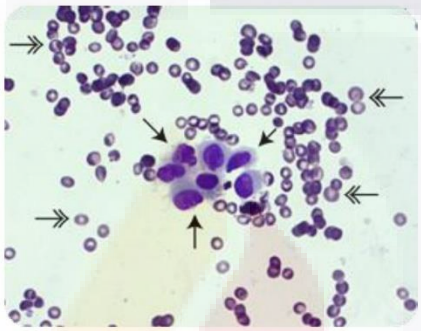


Fig. 4. Fine needle aspirate of a breast mass [25].

The real-valued features for each cell nucleus are [24, 26]:

1. Diagnosis (M=malignant, B=benign)
2. Radius,
3. Texture,
4. Perimeter,
5. Area,
6. Smoothness,
7. Compactness,
8. Concavity,
9. Concave points,
10. Symetry,
11. Fractal dimension

Diagnosis is the final result of evaluating the characteristics of the cell with a computer vision diagnostic system [26-28]. Each cell in the database has one of two possible diagnoses, it can be a malignant cell registered with an uppercase letter M or a benign cell registered with an uppercase letter B.

The radius of a cell was measured by averaging the length of radial lines segments defined by the centroid of the cell and the individual points in the boundary of the cell. The radial lines were defined by Street, Wolberg and Mangasarin in [26, 27] as can be seen in figure 5.

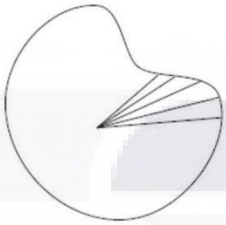


Fig. 5. Radial lines measured in a cell [26].

As mentioned earlier, each cell feature was extracted by a computer vision system, so, the texture was measured by finding variance of gray scale intensities in computer pixels [26, 27]. See Figure 6.

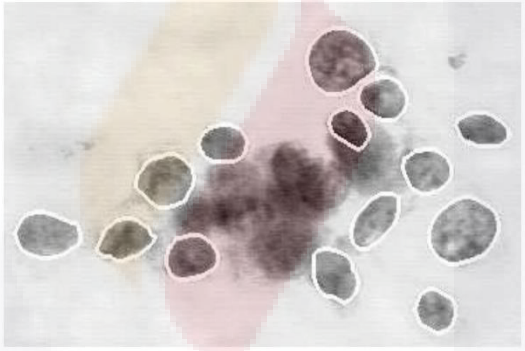


Fig. 6. Example of image taken by a computer vision system and boundary cell [26].

The perimeter is defined as the total distance between individual points named snake points in [26]. Those individual points comprise the white lines in the boundary of the cells (see Figure 6).

The area is measured by counting the number of pixels on the interior of the white line adding one-half of the pixels in the perimeter [26].

Meanwhile, the smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it [26], see Figure 5. Basically, the smoothness is a local variation in radius lengths [24].

Perimeter and area are combined to calculate a measure of compactness, which is a measure of shape [26, 27]. The compactness if given by the formula:

Compactness=perimeter^2/area.

This number is minimized by a circular disk and increases with the irregularity of the boundary and also increases for elongated cell nuclei, which can indicate an increased probability of malignancy [26].

Concavity analyzes the shape irregularities in a cell nucleus. Street, Wolberg and Mangasarian [26] measure the number and severity of concavities or indentations in a cell nucleus. They draw chords between non-adjacent white points and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord. See Figure 7.

Fig. 7. Chords used to compute Concavity [26].

Concave points uses similar measure than Concavity but this feature only measures the number, rather than the magnitude, of contour concavities [26].

Symmetry is found the longest chord through the center. Then, according to [26], the length difference between lines perpendicular to the longest chord to the cell boundary in both directions was measured. This is illustrated in Figure 8.

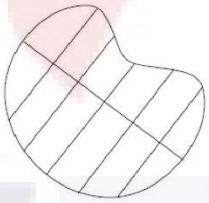


Fig. 8. Segments used in Symetry computation [26].

Finally, the Fractal Dimension is a shape feature [27], so, a higher value corresponds to a less contour and thus to a higher probability of malignancy [26]. The fractal dimension is aproximated using the coastline approximation by Madelbrot [26, 29]. The perimeter of the nucleus is measured using increasingly larger ‚rulers'. This is, as the ruler size increases, decreasing the precision of the measurement, the observed

perimeter decreases. Now, Plotting these to values on a log scale and measuring the downward slope gives the negative of an approximation of the fractal dimension [26]. This measurement is illustrated in Figure 9.

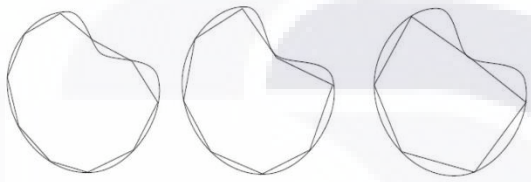


Fig. 9. Sequence of measurements for computing Fractal Dimension [26].

The database contains a total of 569 instances with 357 benign instances and 212 malignant instances. This dataset was preprocessed (Data Preprocessing, see Fig. 3), it means that is necessary a depth analysis of the database looking for duplicate instances in each class and contradictions (delete equal records but different diagnosis). Consequently, the database most contain unique instances.

At the end of Data Preprocessing, the dataset contains 130 benign instances and 146 malignant instances. Those 276 instances were analysed with the next steps of the methodology.

With the LM and C_1, ..., C_n comparison criteria specified [11] by a pathologist of each feature described earlier, a Difference Matrix (DM) is computed by comparing each instance from a class with each instance in the other classes following the comparison criteria of the corresponding feature. When a couple of feature are equal the matrix receives a 0, and 1 when the features are different [11]. A DM contains information that distinguishes objects from differents classes, which contains descriptions of objects [11, 30, 31].

Following this, the Basic Matrix (BM) is determined. The BM contains the basic differences from the DM without duplicates [11, 30]. According to Pons and Shulcloper [9, 32]: be i_q a row of the Difference Matrix, the row i_q is essential if there is not a row i_p that is subline of i_q, then the Basic Matrix contains the essential rows of the DM.

Next step is identifying typical testors. As mentioned earlier, typical testors are formed by the minimum set of features needed to ensure the identification of the class to which a specific object belongs.

The subset $\tau = \{X_{i_1}, \ldots, X_{i_s}\}$ of features from a LM is a Testor if each column from its BM is deleted, except those corresponding elements of τ, there is not a row composed by full of zeros. The subset τ is a Typical Testor whether any feature is deleted, the subset is no longer a testor [9, 11, 32].

Finally, the informational weight is calculated and it is possible begin an analysis of results. This information is explained below.

4 Results and Conclusions

At the end of the process, a validation is required which is done by a specialist in the area. For breast cancer and its cells, the specialists are the Oncologist and Pathologist. The final information must be attached to reality, for this reason a Computer Science specialist can not validate this information.

Finally, the tyical testors and their informational weight is shown below:

Table 1. Typical Testors.

Feature	Typical Testors (In column)	
	1	2
Radius	0	1
Texture	1	1
Perimeter	1	1
Area	1	0
Smoothness	1	1
Compactness	1	1
Concave points	1	1
Symetry	1	1
Fractal dimension	1	1

As can be seen, Table 1 shows the typical testors. The process found two typical testors and their resulting information that states that most of the features are critical to determine if a instance of the cell is malignant or benignant.

The Typical Testor 1 states a 0 in the radius feature and 1 in the other features. On the other hand, the Typical Testor 2 sets a 0 in the area and 1 in the other features. This means that the radius and the area ar interchangeable.

Table 2. Informational weight according to the typical testors.

Feature	Informational weight
Radius	50%
Texture	100%
Perimeter	100%
Area	50%
Smoothness	100%
Compactness	100%
Concave points	100%
Symetry	100%
Fractal dimension	100%

In other words, it is possible to classify a cell instance knowing at least one of the two features, the radius or the area. For example, an instance can be classified knowing its texture, perimeter, smoothness, compactness, concave points, symetry and fractal points but if radius is unknown, the area must be known. However, if the area is unknown, the radius must be known. Finally, in the best case, both values are known but is not possible to classify a cell if both features are unknown.

The informational weight is obtained by calculating a percentage factor that indicates the frequency of each variable in the set of typical testors [33]. Table 2 shows the informational weight of each feature.

The value of the informational weight represents the degree of importance of each feature analyzed in a classification process. A value of 100% indicates that the feature is critical and it can not be ignored in any case.

Also, it can be possible that one or more features have 0% of informational weight meaning that is not necessary, therefore the number of features is reduced and make the problem easier. Remember, this is one of the objective of the analysis.

5 Future Work

The analysis described in this paper is a exhaustive method to find the typical testors. Next step, is apply a Metaheuristics as alternative.

The main goal is to find typical testors through metaheutistc algorithms taking advantage of the already identified basic matrix. In this way, the alternative process will be an hybrid method.

Moreover, it is intended to apply both processes (hybrid and exhaustive) in more cases they represent the behavior of different pathologics.

Acknowledgements. We thank Dr. Luis Muñoz Fernandez for all his assistance and advices and Angela Paulina Pérez Díaz for her advices. We really appreciate your help.

References

1. Zicre, D.: Cátedra de Anatomía y Fisiología Patológicas. In: Neoplasia, ed: Facultad de Ciencias Médicas, UNR (2012)
2. Guerra Merino, I.: Factores pronostico del cáncer de mama en 108 mujeres menores de 36 años. Universidad Complutense de Madrid (2000)
3. CEAMEG: Cancer de Mama. Cancer de Mama, Vol. 1, pp. 1 (2014)
4. Canceronline. Detección Precoz de Cáncer. Available: http://www.canceronline.cl/index.php?option=com_content&view=article&id=48&Itemid=57
5. NIH.: Cancer Screening. Available: http://www.cancer.gov/about-cancer/screening (2015)
6. Pérez, N. M.: El diagnóstico médico: algunas consideraciones filosóficas. (2009)

7. Lugo-Reyes, S. O., Maldonado-Colín, G., Murata, C.: Inteligencia artificial para asistir el diagnóstico clínico en medicina. Artificial Intelligence to Assist Clinical Diagnosis in Medicine, Vol. 61, No. 2, pp. 110–120 (2014)
8. Reed, K.: HealthGrades Patient Safety in American Hospitals Study. Available: https://www.hospitals.healthgrades.com/
9. Ruíz, J., Alba, E., Lazo, M. : Introducción a la Teoría de Testores. Departamento de Ingeniería Electrica, CINVESTAV-IPN, pp. 197 (1995)
10. Torres, M. D., Torres, A., Torres, M. L., Bermudez, L., Ponce, E. E.: Factores Predisponentes en Relajación Residual Neuromuscular. Research in Computing Science, Vol. 93, pp. 163–174 (2015)
11. Santiesteban, Y., Pons, A.: LEX: A New Algorithm for the Calculus of all Typical Testors. Vol. 1, pp. 85–95
12. Lias-Rodríguez, A., Pons-Porrata, A.: Un nuevo Algoritmo de Escala Exterior para el Cálculo de los Testores Típicos. pp. 10, http://www.rcs.cic.ipn.mx/2015_93/Factores%20predisponentes%20en%20relajacion%20residual%20neuromuscular.pdf (2015)
13. Wang, G., Song, Q., Sun, H., Zhang, X.: A Feature Subset Selection Algorithm Automatic Recommendation Method. China: Cornell University Library, pp. 1–34 (2013)
14. Torres, D., Ponce de León, E., Torres, A., Ochoa, A., Díaz, E.: Hybridization of Evolutionary Mechanisms for Featured Subset Selection in Unsupervised Learning. MICAI 2009, Advances in Artificial Intelligence, pp. 610–621 (2009)
15. Pelikan, M., Sastry, K., Cantú-Paz, E.: Scalable Optimization vía Probabilistic Modeling: From Algorithms to Applications. Springer (2006)
16. Deng, K.: OMEGA: On-line Memory-Based General Purpose System Classifier. Doctor, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1998)
17. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, Vol. 3 (2003)
18. Cotilla, M. O.: Un Recorrido por la Sismología de Cuba. Cuba: Editorial Complutense, S. A. (2006)
19. N. C. Institute: What is Cancer? http://www.cancer.gov/about-cancer/understanding/what-is-cancer (2015)
20. Breastcancer.org.: ¿Qué es el Cáncer de mama? http://www.cancer.gov/about-cancer/understanding/what-is-cancer(2014).
21. NIH: Breast Cancer - Patient Version. National Cancer Institute.
22. INEGI: Estadísticas a Propósito del Día Mundial de la Lucha contra el Cáncer de Mama. In: Estadisticas Nacionales, ed. México, Instituto Nacional de Estadística y Geografía (2015)
23. INEGI: Estadísticas a Propósito del Día Mundial Contra el Cáncer. México, Instituto Nacional de Estadística y Geografía (2016)
24. Wolberg, W. H., Street, N., Mangasarian, O. L.: Wisconsin Diagnostic Breast Cancer (WDBC). California, Ed., USA (1995)
25. V. B. Imaging: Fine Needle Aspiration. http://www.breastimaging.vcu.edu/services/guided/fineneedle.html (2016)

26. Street, W. N., Wolberg, W. H., Mangasarian, O. L.: Nuclear Feature Extraction for Breast Tumor Diagnosis. In: International Symposium on Electronic Imaging: Science and Technology, Vol. 1905, pp. 861–870
27. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computerized breast cancer diagnosis and prognosis from fine- needle aspirates. Archives of surgery (Chicago, Ill.: 1960), Vol. 130, No. 5, pp. 511 (1995)
28. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, Vol. 26, No. 7, pp. 792–796 (1995)
29. Mandelbrot, B. B.: The fractal geometry of nature. New York, W.H. Freeman (1982)
30. Ochoa-Somuano, J.: Técnicas de Selección de Atributos para la Categorización Automática de Escenas Visuales. Maestría, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos (2005)
31. Martínez-Sánchez, N., García-Lorenzo, M. M., García-Valdivia, Z. Z.: Modelo para Diseñar Sistemas de Enseñanza-Aprendizaje Inteligentes Utilizando el Razonamiento Basado en Casos. Revista Avances en Sistemas e Informática, Colombia, Vol. 6 (2009)
32. Pons-Porrata, A.: Desarrollo de Algoritmos para la Estructuración Dinámica de Información y su Aplicación a la Detección de Sucesos. Doctorado, Departamento de Lenguajes y Sistemas Informáticos, Universidad Jaume 1, Castellón (2004)
33. Rodríguez de Léon, P.: Heurística lógico combinatoria para la selección de subconjuntos de características en diabetes mellitus. Tesis (maestría en informática y tecnologías computacionales), Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Aguascalientes, Ags., Méx. (2016)

H.2 Análisis de Células Cancerígenas Aplicando la Teoría de Testores Típicos

TECNOLOGÍA EDUCATIVA REVISTA CONAIC

Análisis de Células Cancerígenas Aplicando la Teoría de Testores Típicos
Cancer Cell Analysis Applying "Typical Testors" Theory

Alexis Gallegos [1], Dolores Torres [2], Francisco Álvarez [1], Aurora Torres [1]

[1] Universidad Autónoma de Aguascalientes- Campus Central, Departamento de Ciencias de la Computación, Aguascalientes, México
[2] Universidad Autónoma de Aguascalientes- Campus Central, Departamento de Sistemas de Información, Aguascalientes, México
alexisedm@gmail.com, mdtorres@correo.uaa.mx, fjalvar@correo.uaa.mx, atorres@correo.uaa.mx

Fecha de recepción: 18 de septiembre 2016
Fecha de aceptación: 10 de diciembre 2016

Resumen. El diagnóstico oportuno forma parte de una serie de recomendaciones que permiten detectar enfermedades en etapas tempranas y puedan ser atacadas antes de que conduzcan a problemas serios o incluso la muerte. Debido a su importancia, el diagnóstico médico es un proceso cognoscitivo que debe mantenerse en constante evolución con el objetivo de reducir la posibilidad de un diagnóstico erróneo. En casos de cáncer, el diagnóstico oportuno es determinante para conseguir un pronóstico positivo. Por tanto, se presenta un análisis sobre las características de las células de cáncer, específicamente en cáncer de mama. Dicha patología es de los tipos de cáncer más presente en mujeres alrededor del mundo sin importar el nivel de desarrollo de la región. El propósito del artículo es presentar el uso de testores típicos, es decir, como técnica de reducción de dimensiones para clasificar las células cancerígenas en malignas o benignas; así como el peso informacional de cada variable como índice de importancia.

Palabras Clave: Células de Cáncer, Cáncer de mama, Diagnóstico, Selección de Características.

Abstract. The timely diagnosis is part of a series of recommendations that can detect diseases in early stages to be attacked before they lead to serious problems or even death. Because of its importance, medical diagnosis is a cognitive process that must be kept in constant evolution in order to reduce the possibility of a misdiagnosis. In cases of cancer, timely diagnosis is crucial to achieve a positive prognosis. Thus, an analysis is presented on the features of cancer cells, specifically in breast cancer. This pathology is one of the most common types of cancer in women around the world, regardless of the level of development of the region. The purpose of the article is present the use of typical testors, a technique of reduction of dimensions to classify cancer cells into malignant or benign; as well as get the informational weight of each variable as an index of importance.

Keywords: Cancer cells, Breast Cancer, Diagnosis, Feature Subset Selection.

1 Introducción

La salud es una prioridad para la humanidad; más aún, con patologías cuyo diagnóstico oportuno y correcto puede ser la diferencia entre un pronóstico positivo y uno negativo. Razón por la cual, la medicina es uno de los campos que más podrían beneficiarse con la interacción cercana con la computación y las matemáticas para mejorar procesos como el diagnóstico médico[1].

El diagnóstico es una tarea fundamental para los médicos y forma la base para establecer un tratamiento adecuado[2]. Debido a que el proceso tiene tratamiento humano no se encuentra exento de posibles errores que pueden causar retrasos considerables en la atención médica oportuna del paciente[1].

El diagnóstico oportuno es tan importante en enfermedades comunes como un resfriado, así como para enfermedades de mayor impacto para la salud como lo es el cáncer. Así pues, debido a su importancia, el presente trabajo aplica la técnica de testores típicos a células de cáncer de mama para encontrar el peso informacional y así determinar la influencia de cada una de las variables presentes en esta patología.

El presente trabajo, consta de tres apartados más; de los cuales, el primero abordará el panorama sobre el cáncer y su impacto en la sociedad así como profundizar acerca de su diagnóstico oportuno.

Para el siguiente apartado se comentará un poco sobre la evolución de la metodología de testores típicos. Se tratará la metodología aplicada al análisis de las células de cáncer donde se observará la evolución del procesamiento de información y sus resultados.

Finalmente en el apartado de conclusiones, se analizará la información obtenida y una pequeña descripción de trabajos futuros con el propósito de aprovechar los resultados ya obtenidos.

2 Antecedentes

A diferencia de lo que se pueda pensar, el cáncer no es una enfermedad, sino muchas de ellas, en otras palabras, es un término usado para designar alrededor de 200 entidades distintas[3] constituyendo un serio problema para la humanidad debido a las altas tasas de incidencia y mortalidad presentes en el mundo[4], así como problemas de orden psicológico familiar, laboral y económico entre otros[5].

El cáncer es un trastorno caracterizado por un desequilibrio entre la proliferación celular y los mecanismos normales de muerte celular[3]. Las células sanas se multiplican cuando el cuerpo las necesita y mueren cuando se dañan o el cuerpo ya no las necesita, de manera que, cuando el material genético de la célula cambia provoca que crezcan y se dividan descontroladamente y no mueren de manera normal[6].

Existen muchos tipos de cáncer, debido a que puede aparecer en cualquier órgano o tejido. Según Cáncer.org y la American Cancer Society [7], la diferencia entre ellos es su velocidad de crecimiento y propagación; su respuesta al tratamiento. La causa de la mayoría de los tipos de cáncer sigue siendo desconocida, pero se han detectado múltiples factores de riesgo como el alcoholismo, problemas genéticos, obesidad, exposición a la radiación, etc.[6]

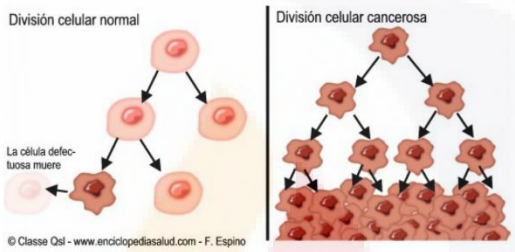


Fig. 1. División Celular normal y cancerosa[8].

La mayoría de los diferentes tipos de cáncer forman una masa anormal de tejido corporal sin ninguna función fisiológica conocida como tumor, masa o neoplasia[7, 9]. Pero, al hablar de un tumor no siempre implica que se tenga desarrollado un cáncer[10]. Existen dos tipos de tumor, el tumor benigno y el maligno. El primero de ellos no implica cáncer[10, 11], aparece en una determinada parte del cuerpo con tendencia de crecimiento lento permaneciendo en un mismo lugar. Rara vez causa problemas graves[12], una vez extirpados, los tumores benignos no suelen reaparecer.

De acuerdo a la Canadian Cancer Society[12], algunas células desarrollan cambios leves que desaparecen sin tratamiento. Otras células desarrollan anomalías genéticas y las nuevas células se vuelven cada vez más anormales hasta convertirse en cáncer. Este proceso puede tomar mucho tiempo hasta que suceda.

Los cambios premalignos pueden variar de acuerdo a su grado de anormalidad[12, 13]:

1. Hiperplasia, se trata del aumento anormal en el número de células. En la mayoría de los casos no significan cambios precancerosos.
2. Atipia, las células lucen significativamente atípicas bajo el microscopio. Algunas veces los cambios son causados por la cicatrización y la inflamación, en lugar de cambios precancerosos.
3. Metaplasia, las células se ven normales, no son del tipo normal como las células del tejido en que se encuentra.
4. Displasia, las células se desarrollan anormalmente, su apariencia y organización ya no son comunes. En la mayoría de los casos, se refiere a condiciones precancerosas.

Por otra parte, el tumor maligno puede dar paso a cualquier tipo de cáncer[10]. Las células son completamente deformes y desorganizadas provocando que el tumor crezca descontroladamente invadiendo tejidos cercanos, vasos sanguíneos o vasos linfáticos. Algunos de ellos pueden interferir en las funciones del cuerpo y causar la muerte[12]. Al igual que los tumores benignos, los malignos pueden desaparecer o retirarse pero éstos pueden reaparecer en algún momento.

Para el presente trabajo se analizan células de cáncer de mama, uno de los cánceres tumorales conocidos desde antiguas épocas (Egipto, 1600 a.C. aproximadamente)[14]. Consiste en un tumor maligno desarrollado a partir de células mamarias, generalmente aparecen en las células de los lobulillos, las glándulas productoras de leche[15].

Tal como es reportado por breastcancer.org[15], si no es tratado a tiempo, las células pueden invadir el tejido mamario sano circundante y llegar a los ganglios linfáticos de las axilas, encargados de eliminar sustancias extrañas del cuerpo. Si las células alcanzan los ganglios linfáticos, tendrán acceso al resto del cuerpo.

Hoy en día el cáncer constituye un problema importante de salud debido a su alta incidencia como causa de mortalidad prematura en la población, y a los problemas que genera en orden psicológico familiar, laboral y económico asociado especialmente a los cambios de estilo de vida[5, 14, 16].

En el mundo, el cáncer de mama es una de las principales causas de muerte, con alrededor de 500 mil decesos al año, de los cuales el 70% ocurren en países en desarrollo[16]. Solo en Canadá, conforma la tercera causa de muerte aumentando su tasa de incidencia alrededor de 15% en los últimos 20 años[4]. En España supone el 29% de todos los cánceres diagnosticados con 25, 215 casos para el año 2012[17].

Para el caso de Estados Unidos, una de cada 8 mujeres desarrollará cáncer de mama a lo largo de su vida, representando así el 12% de la población femenina. Además, para el 2016 se esperan 246,660 diagnósticos de cáncer invasivo[18].

En 1986, en Cuba, el cáncer de mama alcanza el primer lugar de tumores malignos con una tasa de 35.1 por cada 100 mil mujeres, situando a Cuba como el segundo lugar entre las 10 localizaciones más frecuentes de cáncer en ambos sexos[4].

Volviendo a México, según el Consenso Mexicano sobre el Diagnóstico y Tratamiento de Cáncer Mamario[16], se sitúa en el primer lugar de mortalidad por tumor maligno en mujeres mayores a 25 años desde 2007.

Nota: Se utilizó la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud (CIE-10), código C50. Excluye casos con edad no especificada

Fuente: Para 2007 a 2009: SSA, CENAVECE (2014). *Anuarios de Morbilidad 1984-2014*; y CONAPO (2008). *Proyecciones de la Población de México 2005-2050*. Procesó **INEGI**.
Para 2010 a 2014: SSA, CENAVECE (2014). *Anuarios de Morbilidad 1984-2014*; y CONAPO (2014). *Proyecciones de la Población 2010-2050*. Procesó **INEGI**.

Fig. 2. Incidencia de tumor maligno de mama en población de 20 años o más de 2007 a 2014 por cada 100,000 habitantes de cada sexo. INEGI [19].

Como se puede observar en la figura 2, se hace un análisis de la incidencia de cáncer de mama entre 2007 y 2014 mostrando que en los hombres la incidencia es muy baja y relativamente estable, mientras que en las mujeres manteniendo su tendencia a la alza alcanzando para 2014 28.75 casos por cada 100 mil mujeres mayores a 20 años. Mientras que en la figura 3 se analiza la incidencia en mujeres por entidad federativa en 2013, de los cuales Campeche (117.15 casos), Colima (94.24), Aguascalientes (63.33) y Veracruz (62.36) superan la media nacional de 28.90 casos por cada 100 mil mujeres[19].

El promedio de edad con diagnóstico de cáncer de mama es de 53 años en México, representando casi un década menos comparado con Estados Unidos de América, Canadá y algunos países de Europa, donde el promedio está alrededor de los 60 años. Hasta el 11% de las mujeres con diagnóstico de éste tipo de cáncer es de 40 años, siendo mayores que en el caso de países desarrollados[16].

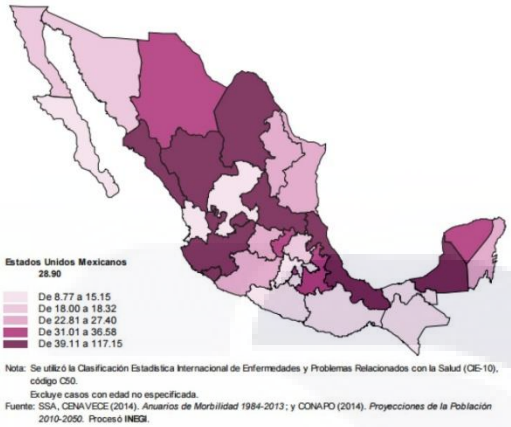


Fig. 3. Incidencia de tumor maligno de mama por entidad federativa de 2007 a 2014, por cada 100,000 mujeres mayores a 20 años. INEGI[19]

Ya con un panorama sobre el impacto que tiene el cáncer en el mundo y, más específicamente, en México se dará paso a tratar la metodología de testores típicos utilizada para el análisis de células benignas y malignas de cáncer de mamá.

3 Metodología

3.1 Conceptos Básicos

El enfoque Lógico Combinatoria del Reconocimiento de Patrones se realiza bajo dos perspectivas: la selección de subconjuntos de características y la selección de objetos[20]. Para el fin de este artículo se expondrá el cálculo de testores típicos correspondientes a la selección de subconjuntos de características.

El problema de selección de subconjuntos de características consiste en seleccionar un subconjunto de variables de un conjunto mayor, con el objetivo de que dicho subconjunto sea suficiente para clasificar y reducir el espacio de características[21]. Su importancia radica en los siguientes puntos[21, 22]:

- Eficiencia y precisión en la clasificación y descripción de fenómenos en los que interviene un número considerable de variables,
- Reducción en requerimientos de medición y almacenamiento,
- Reducción en tiempo de entrenamiento y utilización.

El término de testor fue introducido a los problemas de Reconocimiento de Patrones por Zhuravlev en la ex Unión Soviética[23]. Como idea básica se tiene que un testor es un subconjunto de características, el cual no puede confundir cualquier par de sub-descripciones de diferentes clases. Es otras palabras, es el conjunto características que permite distinguir entre dos clases de objetos[20, 24].

Por tanto, un testor puede definirse como: "un conjunto de características T es testor si y solo si cuando todas las características eliminadas, excepto aquellas en T, no hay algún par de sub-descripciones similares en dos clases diferentes"[23, 24]. A pesar de que Zhuravlev solo hace mención de dos clases en su definición, claramente puede extenderse a n clases sin salir del concepto[25].

Dentro del conjunto de los testores se encuentran los testores típicos, los cuales según Shucloper[25], se trata de testores que cada características es esencial, esto es, si al eliminar una característica el conjunto no resulta un testor[24, 25]. Formalmente es definido como: "Un conjunto de características T es un testor típico si y solo si T es un testor y no existe otro testor T^0, el cual $T^0 \subset T$"[23, 24].

Al encontrar un testor típico, se consigue una irreducible combinación de características, en el cual cada característica es determinante para mantener las diferencias entre clases[23].

Finalmente, el peso informacional es una herramienta que proporciona resultados tangibles[26], presentando matemáticamente la importancia de cada característica para que un objeto pertenezca a una clase o a otra[27, 28].

En apartado 3.1 se expone la descripción del cálculo de testores típicos y el peso informacional de cada característica sobre datos de células de cáncer de mama.

3.2 Framework

Para aplicar los conceptos anteriores, se tiene una base de datos de la Universidad de California proveniente de su Machine Learning Repository. La fuente de datos es Diagnóstico de Cáncer de Mama de Winsconsin (Winsconsin Diagnostic Breast Cancer)[29].

Dicha fuente de datos contiene características calculadas a partir de una imagen digitalizada de la célula perteneciente al tejido de mama. Cabe destacar que el análisis no se realizó con las imágenes de la célula, sino con la información que se obtuvo a partir del procesamiento de imágenes realizado por Nick Street, William Wolberg y O. L. Mangasarian en "Nuclear Feature Extraction for Breast Tumor Diagnosis"[30].

La base de datos descrita por la Universidad de California define 11 características, de las cuales, la primera, representa el diagnóstico de la célula y 10 son características del núcleo de las células obtenidas a partir del análisis de imágenes digitalizadas de una masa fina de tejido mama. Las características son explicadas a continuación[30, 31]:

1. Diagnóstico (Maligno o benigno),
2. Radio. Se refiere a la medida promedio de la longitud de líneas radiales definidas desde el centro a un punto de límite de la célula.
3. Textura. Es medida por medio de la varianza de la intensidad en escala de grises en los pixeles de la imagen digitalizada.
4. Perímetro. La distancia total comprendida en el borde de la célula localizada por puntos por medio de un sistema de visión por computadora.
5. Área. Se define en este caso como el número de pixeles que se encuentran en el interior del límite celular.
6. Suavidad, variación local en las longitudes del radio.
7. Compacidad. Combinación del perímetro y área de la célula en la fórmula perímetro^2/área. Es una característica de forma que se minimiza en una célula con límite regular e incrementa con células irregulares.
8. Concavidad. Analiza las irregularidades en la forma del núcleo de la célula. Street, Wolberg y Mangasarian miden el número y la severidad de las concavidades en el núcleo.

Fig. 4. Puntos cóncavos en el núcleo celular[30].

9. Puntos cóncavos, número de puntos cóncavos en el contorno.
10. Simetría. Es calculada localizando la línea más larga que pase por el centro de la célula y se trazan líneas perpendiculares a dicha línea. Se mide la diferencia de longitudes en las dos direcciones de la línea central.

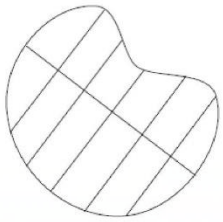


Fig. 5. Segmentos usados para el cálculo de la simetría[30].

11. Dimensión Fractal. Se calcula a partir de la "aproximación costa" definida por Madelbrot. Se trata de una característica de forma, a mayor valor corresponde a mayor probabilidad de malignidad.

La fuente de datos consta de 569 instancias, de los cuales 357 son registros de células benignas y 212 de células malignas. Dicha fuente conforma la matriz de aprendizaje con la descripción de los objetos a analizar[32].

Para realizar un pre-procesamiento de la información, es necesario un análisis profundo en búsqueda de registros duplicados y contradicciones (misma información pero diagnóstico diferente). Al finalizar el procesamiento se cuenta con 141 instancias benignas únicas y 146 instancias únicas malignas.

El siguiente paso es encontrar la Matriz de Diferencias (MD) basado en criterios de comparación obtenidos a partir de literatura especializa o de la experiencia de un especialista, en este caso un patólogo. Una MD se obtiene a partir de la comparación de cada instancia de una clase con las instancias del resto de las clases asignando en la MD un 0 cuando son iguales y un 1 cuando se consideran diferentes[25, 32].

Ahora, se determina la Matriz Básica (MB), la cual consta de las filas básicas de la matriz de diferencias[21, 32]. Una fila iq es básica si no existe una fila ip que sea subfila de iq, es decir; siendo ip e iq filas de la MD, se dice que ip es subfila de iq si para todo elemento de iq=0 se cumple que ip=0 y además existe al menos un elemento iq=1 en el que ip=0.[21, 25]

Lo siguiente es la identificación de Testores Típicos, los cuales son representados por una serie de registros que representan el conjunto de características mínimas necesarias para la identificación de clases de acuerdo con la descripción del objeto. Un subconjunto $\tau = \{X_{i_1}, \dots, X_{i_s}\}$ es un Testor Típico si al eliminar alguna característica deja de ser testor[25].

A continuación, en el apartado de Conclusiones se describirán los resultados obtenidos al proceso descrito.

4 Conclusiones

Como se pudo observar a lo largo del artículo, el cálculo de testores típicos puede ser aplicado a cualquier área del conocimiento. Por tanto es necesaria la presencia de especialistas en el tema abordado. En este caso fue necesaria el apoyo de un experto en Patología.

Los testores típicos y su peso informacional obtenidos se presentan la tabla 1 a continuación:

Características	Testores Típicos		Peso Informacional
	1	2	
Radio	0	1	50%
Textura	1	1	100%
Perímetro	1	1	100%
Área	1	0	50%
Suavidad	1	1	100%
Compacidad	1	1	100%
Puntos Cóncavos	1	1	100%
Simetría	1	1	100%
Dimensión Fractal	1	1	100%

Tabla 2. Testores Típicos y Peso Informacional.

Se observa que ocho de las características observadas son críticas para determinar si una célula es benigna o maligna, éstas son las marcadas con 100% de peso informacional. En cambio el Radio y el Área obtuvieron el 50%, esto significa que para clasificar una célula se puede prescindir de una u otra variable pero no ambas, es decir, se debe conocer al menos una de ellas para tomar una decisión precisa.

Es importante mencionar que es posible que una característica obtenga un peso informacional de 0%, lo cual significa que no influye en la clasificación y puede ser eliminada del conjunto de características. Así, es posible reducir el número de características y facilitar la compresión del problema y su comportamiento.

Finalmente como trabajo a futuro se tiene la creación de una Metaheurística con el propósito de contrastar sus resultados con los del cálculo de Testores Típicos; por tanto, se busca entrenar la metaheurística y clasificar nuevas instancias de células de acuerdo con la matriz básica.

Referencias

[1] S. O. Lugo-Reyes, G. Maldonado-Colín, and C. Murata, "Inteligencia artificial para asistir el diagnóstico clínico en medicina," *Artificial Intelligence to Assist Clinical Diagnosis in Medicine*., Article vol. 61, no. 2, pp. 110-120, 2014.

[2] J. Díaz Novás, B. Gallego Machado, and A. León González, "El diagnóstico médico: bases y procedimientos," vol. 22, ed. Cuba: Biblioteca Virtual en Salud de Cuba, 2006, p. 11.

[3] *Tratamiento del cáncer: oncología médica, quirúrgica y radioterapia*. Distrito Federal, MÉXICO: Editorial El Manual Moderno, 2016.

[4] E. Garrido Fuente, "Neoplasia de mama," El Cid Editor, Córdoba, AR2016, Available: http://site.ebrary.com/lib/univeraguascalientessp/docDetail.action?docID=11203437.

[5] E. Garrido Fuente, "Factores de riesgos ambientales y genéticos: influencia en el cáncer de mama," El Cid Editor, Córdoba, AR2015, Available: http://site.ebrary.com/lib/univeraguascalientessp/docDetail.action?docID=11087163.

[6] M. Plus. (2015). *Cáncer*. Available: https://medlineplus.gov/spanish/ency/article/001289.htmAvailable: https://medlineplus.gov/cancer.html

[7] S. A. c. e. Cáncer. (2016). *¿Qué es el cáncer? Una guía para pacientes y sus familias*. Available: http://www.cancer.org/espanol/cancer/aspectosbasicossobrelcancer/que-es-el-cancer

[8] (2007). *Oncología y Cáncer*. Available: http://www.enciclopediasalud.com/categorias/oncologia-y-cancer/articulos/cancer

[9] "Tumor," ed. U.S.A.: University of Meryland, Medical Center, 2012.

[10] T. Ocete Calvo. (2016). *Diferencia entre tumor y cáncer*. Available: http://www.bekiasalud.com/articulos/diferencias-tumor-cancer/

[11] (2014). *Tumor*. Available: https://medlineplus.gov/spanish/ency/article/001310.htm

[12] *Types of Tumours*. Available: http://www.cancer.ca/en/cancer-information/cancer-101/what-is-cancer/types-of-tumours/?region=on

[13] (2014). *Cómo entender su informe de patología: Hiperplasia atípica*. Available: http://www.cancer.org/espanol/servicios/comocomprendersudiagnostico/entiendasuinformedepatologia/patologiadelseno/patologia-hiperplasia-atipica

[14] E. Garrido Fuente, "Medios diagnósticos en la detección precoz del cáncer de mamas," 2015.

[15] breastcancer.org. (2014). *¿Qué es el Cáncer de mama?* Available: http://www.cancer.gov/about-cancer/understanding/what-is-cancer

[16] J. Cárdenas Sánchez, J. E. Bargalló Rocha, A. Erazo Valle, A. Poitevin Chachón, C. Vicente Valero, and V. Pérez Sánchez, "Consenso Mexicano sobre diagnóstico y tratamiento de cáncer mamario," vol. 6, ed. México: Masson Doyman México S.A., 2015, p. 149.

[17] A. Santaballa Bertrán. (2015). *Cáncer de Mama*. Available: http://www.seom.org/en/informacion-sobre-el-cancer/info-tipos-cancer/cancer-de-mama-raiz/cancer-de-mama?format=pdf

[18] breastcancer.org. (2016). *U.S. Breast Cancer Statistics*. Available: http://www.breastcancer.org/symptoms/understand_bc/statistics

[19] INEGI, "Estadísticas a Propósito del... Día Mundial de la Lucha contra el Cáncer de Mama," in *Estadisticas Nacionales*, ed. México: Instituto Nacional de Estadística y Geografía, 2015.

[20] Y. Villuendas Rey, *Esquema para el pre-procesamiento de conjuntos de entrenamiento de clasificadores del vecino más cercano basado en extensiones a la teoría de los conjuntos aproximados*. Havana, CUBA: Editorial Universitaria, 2014.

[21] M. D. Torres Soto, A. Torres Soto, and E. Ponce de León Sentí, "Algoritmo Genético y Testores Típicos en el Problema de Selección de Subconjuntos de Características," Available: http://www.iiisci.org/journal/CVS/risci/pdfs/C415DR.pdf

[22] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," vol. 3, ed: Journal of Machine Learning Research, 2003.

[23] J. A. Carrasco-Ochoa, J. Ruiz-Shulcloper, and L. A. De la Vega Doría, "Feature Selecction Using Typical E: Testors, Working on Dynamical Data," *Progress in Pattern Recognition, Image Analysis and Applications,* Available: https://books.google.com.mx/books?id=_mzefGf2hG4C&pg=PA685&lpg=PA685&dq=zhuravlev+testores&source=bl&ots=JgQLDgGDmq&sig=aGnuAA2gFHhhV1c-seEUWzrONtA&hl=es-419&sa=X#v=onepage&q=zhuravlev%20testores&f=false

[24] J. A. Santos, A. Carrasco, and J. F. Martínez, "Feature Selection using Typical Testors applied to Estimation of Stellar Parameters," *Computación y Sistemas*, vol. 8, pp. 15-23

[25] J. Ruíz Shucloper, E. Alba Cabrera, and M. Lazo Cortés, "Introducción a la Teoría de Testores," ed: Departamento de Ingeniería Electrica, CINVESTAV-IPN, 1995, p. 197.

[26] D. Torres, E. Ponce de León, A. Torres, A. Ochoa, and E. Díaz, "Hybridization of Evolutionary Mechanisms for Featured Subset Selection in Unsupervised Learning," *MICAI 2009: Advances in Artificial Intelligence,* pp. 610-621Available: https://books.google.com.mx/books?id=atqYTbU8gO0C&printsec=frontcover#v=onepage&q&f=false

[27] M. D. Torres Soto, A. Torres Soto, M. d. l. L. Torres Soto, L. Bermudez Rosales, and E. E. Ponce de León Sentí, "Factores Predisponentes en Relajación Residual Neuromuscular," *Research in Computing Science*, vol. 93, pp. 163–174Available: http://www.rcs.cic.ipn.mx/2015_93/

[28] M. O. Cotilla, "Un Recorrido por la Sismología de Cuba," 1 ed. Cuba: Editorial Complutense, S. A., 2006.

[29] W. H. Wolberg, N. Street, and O. L. Mangasarian, "Wisconsin Diagnostic Breast Cancer (WDBC)," U. o. California, Ed., ed. USA, 1995.

[30] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis," *International Symposium on Electronic Imaging: Science and Technology*, vol. 1905, pp. 861-870

[31] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology," *Human Pathology*, vol. 26, no. 7, pp. 792-796, 1995.

[32] Y. Santiesteban Algaza and A. Pons Porrata, "LEX: A New Algorithm for the Calculus of all Typical Testors," vol. 1, p. 85–95

H.3 Identificación de Características de Células de Cáncer de Mama por Medio de Testores Típicos

ISSN 1870-4069

Identificación de características de células de cáncer de mama por medio de testores típicos

Alexis Gallegos, Dolores Torres, Francisco Álvarez, Aurora Torres

Universidad Autónoma de Aguascalientes, Departamento de Ciencias de la Computación, Aguascalientes, Aguascalientes, México

alexisEdm@gmail.com, mdtorres@correo.uaa.mx, fjalvar@correo.uaa.mx, atorres@correo.uaa.mx

Resumen. Una de las preocupaciones más importantes del mundo ha sido la salud humana, especialmente en enfermedades como el cáncer. Por esta razón, este artículo se enfoca en la aplicación de las Ciencias Computacionales, específicamente, la Selección de Subconjuntos de Características y Testores típicos para mejorar el diagnóstico de cáncer. En este caso, se procesó una base de datos de cáncer de mama. Ésta base de datos fue publicada por la Universidad de California para aprendizaje máquina. Los datos describen características del núcleo de las células obtenido de una imagen digitalizada de un aspirado con aguja fina de masa mamaria clasificando cada célula como maligna o benigna. Finalmente, el método proveerá el peso informacional de cada característica. Ésta información permitirá saber si una característica realmente describe una célula y así, clasificar nuevas instancias con la información correcta.

Palabras clave: peso informacional, testor típico, selección de subconjuntos, cáncer de mama, lógica combinacional.

Identification of Breast Cancer Cell Features by Means of Typical Testors

Abstract. One of the most significant concerns around the world has been human health, especially in diseases such as cancer. For this reason, this paper is focused on the application of Computer Science, specifically, Feature Subset Selection and Typical Testors to improve the diagnosis of cancer. In this case, a breast cancer cell database was processed. This database was published by the University of California for machine learning. The data describes features of the cell nuclei obtained from a digitized image of a fine needle aspirate of a breast mass classifying every cell as malignant or benign. Finally, the method will provide the informational weight of each feature. This information will let know if a feature actually describes a cell and then, classify new instances with the right data.

Keywords: informational weight, typical testor, subset selection, breast cancer, combinational logic.

pp. 43–54; rec. 2017-03-06; acc. 2017-05-06

1. Introducción

El cáncer de mama es el cáncer más común y la principal causa de muerte por enfermedad tumoral en mujeres alrededor en el mundo [1], lo que representa el 16% de los cánceres en mujeres [2]. Hoy en día, el cáncer se ha vuelto cada vez más difícil de ignorar. Cada día, nuevos estudios sobre las causas y tratamientos son publicados; sin embargo, todo coinciden que el punto crítico de estos estudios es la detección temprana [3].

De acuerdo con el Instituto Nacional de Cáncer [4], la detección de cáncer significa la comprobación de cáncer o de condiciones que pueden convertirse en cáncer en personas que no presentan síntomas.

La detección temprana es importante debido a que cuando un tejido anormal o cáncer es encontrado a tiempo, puede ser más fácil de tratar. Al momento que los síntomas aparecen, el cáncer ha comenzado a extenderse y es más difícil de tratar [4].

A pesar de la utilidad de la detección temprana, pueden existir algunos riesgos, así como los métodos utilizados. Por ejemplo, una prueba de detección puede presentar resultados falsos positivos; significa que la prueba indica la presencia de cáncer cuando no es verdad. Por otro lado, la prueba puede tener resultados falsos negativos indicando que el cáncer no está presente, aunque si lo este.

Por otra parte, el sobrediagnóstico es posible, el cual sucede cuando la prueba de detección muestra que una persona tiene cáncer, pero el cáncer es de crecimiento lento y no habría perjudicado a la persona en toda su vida [4]. Lo anterior justifica la necesidad de mejorar el diagnóstico de cáncer.

El diagnóstico clínico es un proceso cognitivo que parte del pensamiento concreto sensible. Está relacionado con la realidad objetiva; se desarrolla en el pensamiento abstracto y tiene el criterio de verdad en la práctica [5]. Involucra práctica, experiencia, reconocimiento de patrones y cálculo de probabilidad condicional, entre otros componentes. Sin embargo, el diagnóstico tiene tratamiento humano, por lo tanto, no está libre de errores que pueden causar enfermedad, daños, gastos extra e incluso la muerte, especialmente en enfermedades sensibles como el cáncer [6].

Los errores representan un estimado de 150 de cada 1000 pacientes con diagnóstico erróneo [7]. Por esta razón, el campo de la medicina es una de las áreas que pueden beneficiarse mejor de una interacción cercana con las Ciencias Computacionales y las Matemáticas para mejorar procesos como lo es el diagnóstico médico [6]. Siendo así, la razón por la que se decide aplicar métodos matemáticos integrales para apoyar el diagnóstico de enfermedades como el cáncer, en este caso, cáncer de mama.

Este artículo está dividido en tres secciones y organizado de la siguiente manera. La primera sección trata conceptos importantes en Selección de Subconjuntos de Características y testores típicos en Ciencias Computaciones, el cáncer de mama y su impacto en el mundo. La sección siguiente examina marco de trabajo del análisis, siendo una revisión de la metodología aplicada a las células de cáncer de mama. Finalmente, la tercera sección describe los resultados de la metodología y su revisión.

2. Conceptos importantes

2.1. Selección de subconjuntos de características

Normalmente, la Selección de Subconjuntos de Características (FSS, por sus siglas en inglés: Feature Subset Selection) [8] es usado para reducir la dimensionalidad [9], lo que significa que reduce el número de variables, atributos o características con las cuales se describen los objetos y encontrar su influencia en un problema. Este un método alternativo que inicia usado el conjunto de testores típicos, descartando características irrelevantes o redundantes [9, 10].

La importancia de la FSS recae en la reducción del número de características, el cual puede ayudar a disminuir el costo de adquisición de información y hacer que los modelos de clasificación sean más fáciles de entender [9, 11]. Además, el número de características podría afectar la precisión de la clasificación. Algunos autores también han estudiado la Selección de Subconjuntos de Características para el aprendizaje de clasificación [9].

Los problemas de FSS han sido estudiados con gran atención por estadísticos y comunidades de aprendizaje máquina durante años debido a la investigación entusiasta de la minería de datos [12]. Existen muchos beneficios potencias de la selección de características, como lo son [13]:

- Facilita la visualización de información y su entendimiento,
- Reduce requerimientos de medición y almacenamiento,
- Reduce tiempos de capacitación y utilización,
- Reduce la dimensionalidad para mejorar el rendimiento de la predicción.

La selección de las variables más relevantes suele ser subóptima para construir un predictor, sobre todo si las variables son redundantes [13]. Por ejemplo, el método de selección por fuerza bruta evalúa exhaustivamente todas las posibles combinaciones de las características de entrada y así encontrar el mejor subconjunto [12]. Más adelante, en las secciones 3 y 4 describirán el método de fuerza bruta aplicado a la base de datos ya mencionada con el objetivo de evaluar si una célula es maligna o benigna y calcular el peso informacional de cada característica.

2.2. Testores típicos

La teoría de testores fue formulada como una dirección científica independiente de Cibernética Matemática en los años 60 en la formada Unión de Repúblicas Socialistas Soviéticas (USSR), cuyo origen está vinculado con el uso de lógica matemática para localizar fallas en circuitos electrónicos que realizan funciones booleanas [14].

Más tarde, los testores fueron utilizados para realizar clasificación supervisada y selección de variables en problemas de geología [14, 15]. El uso dato a los testores y testores típicos para éste artículo está relacionado con la Selección de Subconjuntos de Características, cuyos precursores son Dmitriev, Zhuravlev, y colegas [15].

De este modo, un testor es un subconjunto de características que distingue objetos de diferentes clases [15]. De acuerdo con Santiesteban y Pons [10], Shulcloper [14] y

Torres [15], un testor típico es un testor al que no es posible eliminar alguna característica sin perder su estado de testor. En otras palabras, un testor típico ya está formando por el conjunto mínimo de características necesarias para asegurar la identificación de la clase a la que pertenece un objeto específico.

Los testores típicos determinan cuestiones como la evaluación del peso informacional de los rasgos y la selección de variables. Pueden reducir el espacio de representación de los objetos [10] y pueden ser usados como un conjunto de soporte para la algoritmos de clasificación [16]. En consecuencia, el objetivo de este estudio es probar que el análisis de testores puede ayudar a clasificar las células basadas en un conjunto de datos real; esto se explicara en la sección 3.

2.3. Peso informacional

El uso del peso informacional para Selección de Subconjuntos de Características es una excelente herramienta que muestra resultados tangibles [9]. El peso informacional es una puntuación, es decir, es una medida de significancia para predecir si un objeto pertenece a un grupo o a otro (clasificación) [15, 17]. Más información en la sección 4.

2.4. Cáncer de mama

El cáncer es una colección de enfermedades relacionadas que causa que algunas células del cuerpo comiencen a dividirse sin detenerse y se extiendan a tejidos cercanos. El cáncer puede generarse en casi cualquier parte del cuerpo, el cual está hecho de millones de células [18]. Se trata del resultado de mutaciones o cambios anormales en los genes que regulan el crecimiento de la célula. Normalmente, una célula crece y divide para formar nuevas células según como el cuerpo lo necesite. Cuando las células envejecen y resultan dañadas, mueren, y nuevas células las reemplazan [18, 19].

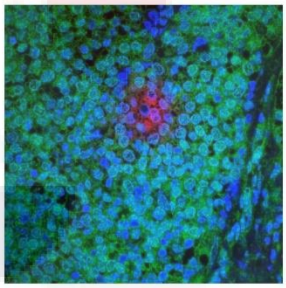


Fig. 1. Tumor invasivo de cáncer de mama [18].

Cuando el cáncer se desarrolla, el proceso celular se descompone. Las mutaciones pueden "activar" ciertos genes y "desactivar" otras en la célula. La célula modificada

adquiere la habilidad de dividirse sin ningún control u orden, lo que produce células idénticas y generan un tumor [18, 19].

En consecuencia, el cáncer de mama es un tumor maligno que se ha desarrollado de células de mama [20]. La mama está hecha de glándulas llamadas lóbulos que pueden producir leche y tubos delgados llamados ductos que llevan leche de los lóbulos al pezón, generalmente, el cáncer de mama se origina en las células de éstos lóbulos [19, 20].

El cáncer de mama tiene gran impacto en el mundo. Según la Organización Panamericana de Salud (PAHO), en América el cáncer de mama es el más común en mujeres con el 29% de los casos de cáncer. PAHO estima más de 596,000 casos nuevos y más de 142,100 muertes en la región para 2030, principalmente en Latinoamérica y el Caribe [21]. La siguiente figura muestra la incidencia de tumores malignos de mama en mujeres mayores a 20 años divididos por grupo de edad, en el año 2014:

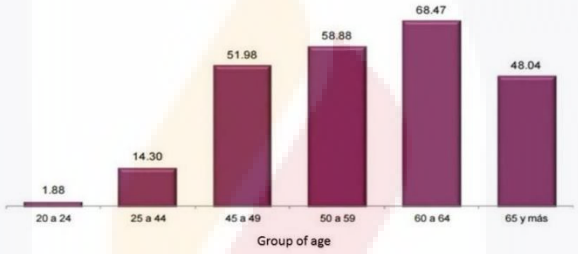


Fig. 2. Incidencia de tumores malignos de mama en mujeres mayores de 20 años dividido por grupo de edad. Por 100 mil mujeres por grupo de edad. INEGI [21].

En general, cualquier tipo de cáncer representa un impacto importante en el estado físico de la persona, su esfera emocional, un alto costo de tratamiento y puede incluso, socavar la economía de los países; así que la prevención y el diagnóstico temprano son críticos para abordar el problema [22]. Por lo tanto, este trabajo se concentra en la aplicación del Selección de Subconjuntos de Características y Testores Típicos para mejorar el diagnóstico de cáncer en células de cáncer. Las secciones siguientes se explicará el estudio realizado.

3. Marco de trabajo

La metodología general utilizada (ver Fig. 3) inicia con una Matriz de Aprendizaje (MA). La MA es la fuente de información que contiene la descripción de los objetos [14, 23]. Para éste trabajo, la MA viene de la Universidad de California y su Repositorio de Aprendizaje Máquina. La base de datos seleccionada es Diagnóstico de Cáncer de Mama de Wisconsin [24].

Fig. 3. Metodología general.

La base de datos contiene el diagnóstico y 10 características obtenidas de una imagen digitalizada de una aspiración de tejido de mama con aguja fina y describe las características del núcleo de la célula presentada en la imagen [25]. La imagen a continuación muestra un ejemplo de las imágenes descritas.

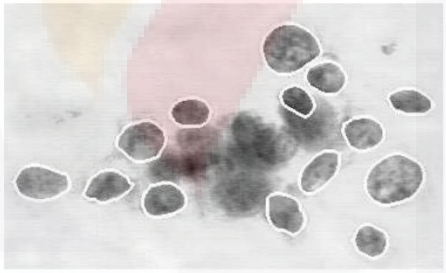


Fig. 4. Ejemplo de una imagen tomada por un sistema de visión por computadora y el contorno de la célula [26].

Las características evaluadas para cada núcleo de célula son [25, 26]:

1. Diagnóstico (M=maligno, B=benigno),
2. Radio,
3. Textura,
4. Perímetro,

5. Área,
6. Suavidad,
7. Compacidad,
8. Concavidad,
9. Puntos cóncavos,
10. Simetría,
11. Dimensión fractal.

El diagnóstico es el resultado final de la evaluación de las características de la célula con un sistema de diagnóstico de visión por computadora [26-28]. Cada célula en la base de datos tiene uno de dos posibles diagnósticos, puede ser célula maligna registrada con la letra M o benigna registrado con la letra B.

El radio de la célula fue medido promediando la longitud de los segmentos de líneas radiales definidos por el centroide de la célula y los puntos individuales en el límite de la célula. Las líneas radiales fueron definidas por Street, Wolberg y Magasarian en [26, 27] como se puede observar en la Fig. 5.

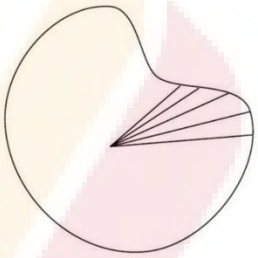


Fig. 5. Líneas radiales medidas en una célula [26].

Como se mencionó anteriormente, cada característica de la célula fue extraída por sistema de visión por computadora, por tanto, la textura fue medida encontrando la varianza en intensidades de escala de grises en los pixeles de la computadora [26, 27]. (Ver Fig. 4).

El perímetro es definido como la distancia total entre puntos individuales llamados puntos serpiente en [26]. Estos puntos individuales comprenden las líneas blancas en el perímetro de las células (ver Fig. 4).

El área es obtiene contando el número de pixeles en el interior de la línea blanca añadiendo la mitad de los pixeles en el perímetro [26].

Mientras tanto, la suavidad del núcleo de la célula se calcula midiendo la diferencia entre la longitud de una línea radial y la longitud principal que la rodea [26]. Básicamente, la suavidad es la variación local en las longitudes de radio [25].

El perímetro y el área son combinados para calcular la medida de compacidad; la cual es una medida de forma [26, 27]. La compacidad está dada por la fórmula:

$$\text{Compacidad} = \text{perímetro}^2/\text{área}.$$

Este número es minimizado por un disco circular e incrementa con la irregularidad del perímetro y aumenta también para núcleos celulares alargados, lo que puede indicar mayor probabilidad de malignidad [26].

La concavidad analiza las irregularidades de forma en el núcleo de la célula. Street, Wolberg y Mangasarian miden el número y la severidad de las concavidades y hendiduras en el núcleo de la célula. Ellos dibujan cuerdas entre cada punto blanco no adyacente y miden hasta qué punto el límite real del núcleo se encuentra en el interior de cada cuerda (ver Fig. 6).

Fig. 6. Cuerdas usadas para calcular la concavidad [27].

Los puntos cóncavos usan una medida similar a la concavidad, pero ésta característica solo mide el número, más que la magnitud, de las concavidades del contorno [26].

La simetría se obtiene encontrando la línea más larga que pase por el centro. Entonces, de acuerdo con [26], se trazan líneas perpendiculares a dicha línea para medir la diferencia de longitudes en las dos direcciones de la lineal central (ver Fig. 7).

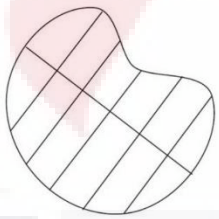


Fig. 7. Segmentos usados en el cálculo de la simetría [26].

Finalmente, la dimensión fractal es una característica de forma [27], es decir, a mayor valor corresponde a un menor contorno y por tanto a una mayor probabilidad malignidad [26]. La dimensión fractal se aproxima usando la aproximación de costa de Mandelbrot [26, 29]. El perímetro del núcleo es medido usando "reglas" cada vez más grande. Esto es, a medida que aumenta el tamaño de la regla, decrece la precisión de la medición, el perímetro observado disminuye. Ahora, trazando estos valores a una escala

logarítmica y medir la pendiente descendente da el negativo de una aproximación de la dimensión fractal [26] (ver Fig. 8).

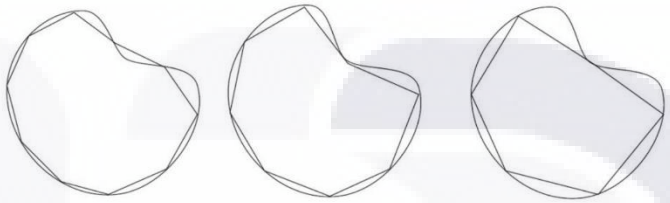


Fig. 8. Secuencia de medidas para calcular dimensión fractal [26].

La base de datos contiene un total de 569 instancias, 357 de ellas son instancias benignas y 212 instancias malignas. Esta base de datos fue pre-procesada (ver Fig. 3), esto significa que es necesario un análisis profundo de la base de datos buscando instancias duplicadas en cada clase y contradicciones (eliminar registros iguales pero diagnósticos diferentes). Consecuentemente, la base de datos debe contener instancias únicas.

El siguiente paso de la metodología requiere una matriz de trabajo, la cual se obtiene de matriz de aprendizaje pre-procesada. La matriz de trabajo se compone de datos discretizados. Cada característica es discretizada de acuerdo a la literatura del problema y el consejo de un experto, quién confirma los criterios de comparación [10].

Finalmente, el elemento principal de la metodología es el peso informacional. Para calcularlo es necesario aplicar la teoría de testores típicos mencionada en la sección 2.2. Como resumen, los testores típicos están formados por el conjunto mínimo necesario para asegurar la identificación de clases en la que un objeto específico pertenece. Para más información ver las referencias [10, 14].

4. Resultados y conclusiones

Al final del proceso, el peso informacional es calculado de acuerdo con los testores típicos encontrados. Como se puede observar en la Tabla 1, el radio y el área del núcleo de la célula tienen 50% de peso informacional. Esto significa que es posible clasificar una instancia de célula conociendo a menos una de las dos características, el radio o el área, pero el resto de las características debe conocerse debido a que obtuvieron un 100%. Por ejemplo, una instancia puede ser clasificada conociendo su textura, perímetro, suavidad, compacidad, puntos cóncavos, simetría y la dimensión fractal, pero si el radio es desconocido, el área debe conocerse. Por otro lado, si se desconoce el área, el radio debe conocerse. Finalmente, en el mejor caso se da cuando ambos valores se conocen mientras que no es posible clasificar una célula si ambos datos son desconocidos.

Table 1. Peso informacional de acuerdo con lo testores típicos.

Feature	Informational weight
Radius	50%
Texture	100%
Perimeter	100%
Area	50%
Smoothness	100%
Compactness	100%
Concave points	100%
Symmetry	100%
Fractal dimension	100%

El peso informacional se obtiene calculando un factor de procentaje que indica la frecuencia de cada variable en el conjunto de testores típicos [30]. El valor del peso informacional representa el grado de importancia de cada característica analizada en un proceso de clasificación. Un valor de 100% indica que la característica es crítica no puede ser ignorada en ningún caso.

Además, es posible que una o más características obtengan 0% de peso informacional, lo que significa que no es necesaria. Por lo tanto, el número de características se reduce y hace el problema más sencillo. Recuerde que éste es uno de los objetivos del anális.

El peso informacional puede ser validado por la teoría del problema o un especialista, de manera que la información final se apegue a la realidad. Para este experimento, un patólogo valido el peso informacional y el comportamiento de los datos.

Referencias

1. Guerra-Merino, I.: Factores pronostico del cáncer de mama en 108 mujeres menores de 36 años. Universidad Complutense de Madrid (2000)
2. CEAMEG: Cancer de Mama. Vol. 1, No. Cancer de Mama, pp. 1 (2014)
3. Canceronline: Detección Precoz de Cáncer. Available: http://www.canceronline.cl/index.php?option=com_content&view=article&id=48&Itemid=57
4. NIH: Cancer Screening. Available: http://www.cancer.gov/about-cancer/screening (2015)
5. Pérez-Guirado, N. M.: El diagnóstico médico: algunas consideraciones filosóficas (2009)
6. Lugo-Reyes, S. O., Maldonado-Colín, G., Murata, C.: Inteligencia artificial para asistir el diagnóstico clínico en medicina. Artificial Intelligence to Assist Clinical Diagnosis in Medicine, Vol. 61, No. 2, pp. 110–120 (2014)
7. Reed, K.: HealthGrades Patient Safety in American Hospitals Study. Disponible en: https://www.hospitals.healthgrades.com/
8. Wang, G., Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A Feature Subset Selection Algorithm Automatic Recommendation Method. China Journal of Artificial Inteligence Research (2013)

9. Torres, D., Ponce de León, E., Torres, A., Ochoa, A., Díaz, E.: Hybridization of Evolutionary Mechanisms for Featured Subset Selection in Unsupervised Learning. In: MICAI 2009, Advances in Artificial Intelligence, pp. 610–621
10. Santiesteban-Algaza, Y., Pons-Porrata, A.: LEX: A New Algorithm for the Calculus of all Typical Testors. Vol. 1, pp. 85–95.
11. Pelikan, M., Sastry, K., Cantú-Paz, E.: Scalable Optimization vía Probabilistic Modeling: From Algorithms to Applications. Springer (2006)
12. Deng, K.: OMEGA: On-line Memory-Based General Purpose System Classifier. Doctor Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1998)
13. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, Vol. 3 (2003)
14. Ruíz-Shucloper, J., Alba-Cabrera, E., Lazo-Cortés, M.: Introducción a la Teoría de Testores. Departamento de Ingeniería Electrica, CINVESTAV-IPN, pp. 197 (1995)
15. Torres-Soto, M. D., Torres-Soto, A., Torres-Soto, L., Bermudez-Rosales, L., Ponce de León-Sentí, E. E.: Factores Predisponentes en Relajación Residual Neuromuscular. Research in Computing Science, Vol. 93, pp. 163–174 (2015)
16. Lias-Rodríguez, A., Pons-Porrata, A.: Un nuevo Algoritmo de Escala Exterior para el Cálculo de los Testores Típicos. Research in Computing Science, Vol. 93 (2015)
17. Cotilla, M. O.: Un Recorrido por la Sismología de Cuba. 1 ed., Cuba, Editorial Complutense, S. A. (2006)
18. National Cancer Institute: What is Cancer? Disponible en: http://www.cancer.gov/ about-cancer/understanding/what-is-cancer (2015)
19. Breastcancer.org: ¿Qué es el Cáncer de mama? Disponible en: http://www.cancer.gov/about-cancer/understanding/what-is-cancer (2014)
20. NIH: Breast Cancer - Patient Version. National Cancer Institute
21. INEGI: Estadísticas a Propósito del... Día Mundial de la Lucha contra el Cáncer de Mama. Estadisticas Nacionales, México: Instituto Nacional de Estadística y Geografía (2015)
22. INEGI: Estadísticas a Propósito del Día Mundial Contra el Cáncer. México, Instituto Nacional de Estadística y Geografía (2016)
23. Santiesteban, Y., Pons, A.: LEX: un nuevo algoritmo para el calculo de los testores tipicos. Revista Ciencias Matematicas, Vol. 21, No. 1, pp. 85–95 (2003)
24. Mangasarian, O. L., Street, W. N.: Breast cancer diagnosis and prognosis via linear programming. Operations Research, Vol. 43, No. 4, pp. 570 (1995)
25. Wolberg, W. H., Street, N., Mangasarian, O. L.: Wisconsin Diagnostic Breast Cancer (WDBC). California, Ed., USA (1995)
26. Street, W. N., Wolberg, W. H., Mangasarian, O. L.: Nuclear Feature Extraction for Breast Tumor Diagnosis. In: International Symposium on Electronic Imaging, Science and Technology, Vol. 1905, pp. 861–870
27. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. Archives of surgery (Chicago, Ill.: 1960), Vol. 130, No. 5, pp. 511 (1995)
28. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L.: Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, Vol. 26, No. 7, pp. 792–796 (1995)
29. Mandelbrot, B. B.: The fractal geometry of nature. New York, W.H. Freeman (1982)

30. Rodríguez-de Léon, P.: Heurística lógico combinatoria para la selección de subconjuntos de características en diabetes mellitus. Tesis (maestría en informática y tecnologías computacionales), Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Aguascalientes, Ags., México (2016)