UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

MASTER IN INNOVATION AND RESEARCH
IN INFORMATICS

Metodología para la selección de características usando algoritmos genéticos

T E S I S

que presenta

JOSÉ ANTONIO ESTRADA PAVÍA

Director de tesis: Dr. José Ramón Herrero Zaragoza
Codirector de tesis: Dr. Oscar Camacho Nieto
Codirector de tesis: Dr. Mario Aldape Pérez

México, D.F. Enero de 2016

Índice

Índice de figuras

Índice de cuadros

Agradecimientos

A CONACYT por el apoyo prestado en la realización de este trabajo.

Al Instituto Pólitecnico Nacional, por darme la oportunidad de estudiar en tan grandiosa institución.

Al CIDETEC por una escuela modelo en donde tuve la fortuna de ser parte.

A la Universitat Politecnica de Catalunya y a la Facultad de Informatica de Barcelona que me dieron la oportunidad de seguir desarrollando mi investigación mas allá de las fronteras del país.

Y en especial: gracias a todos ustedes que hicieron posible esto.

1. Introducción

Este trabajo de tesis se enfoca en el estudio en un área importante dentro del reconocimiento de patrones, la selección de características. Los objetivos primordiales de esta área son: disminuir la complejidad dimensional de los patrones de un conjunto fundamental y al mismo tiempo, elevar la precisión predictiva de algún algoritmo de clasificación.

En el presente trabajo se propone una metodología para la selección de características usando un algoritmo genético. Esta metodología puede ser aplicada a diversos problemas de clasificación, sobre todo aquellos en los que los patrones tienen dimensiones elevadas y las soluciones triviales como la búsqueda de características relevantes mediante algoritmos de fuerza bruta no son una opción computacionalmente viable.

En la sección 1.1 se muestra el objetivo de la tesis, en la sección 1.2 se incluye la motivación que dio lugar a este trabajo de tesis y en la sección 1.3 se presenta el planteamiento del problema. Después, en la sección 1.4 se enuncian las contribuciones originales; por último, en la sección 1.5 se describe la organización del presente trabajo.

1.1. Objetivo

El objetivo del presente trabajo de tesis es: una metodología para la selección de características usando algoritmos genéticos. Mostrar la utilidad de esta metodología a través de pruebas con conjuntos de datos públicos y realizar un estudio experimental que demuestre la eficiencia de la nueva metodología.

1.2. Motivación

El reconocer es una actividad humana que se realiza en la vida cotidiana, esta actividad tan común para nosotros, se ha llevado al campo de la computación a través del reconocimiento de patrones. Para un ser humano es sencillo saber qué es cada objeto, ya que de forma automática el ser humano abstrae las características principales de un conjunto de características combinadas. Por ejemplo, la representación de una silla puede definirse como una superficie donde alguien puede sentarse, generalmente con cuatro patas y un respaldo, puede tener un respaldo grande o chico, puede tener un asiento inclinado o liso y sigue siendo una silla. En este ejemplo el cerebro ya ha seleccionado las características más importantes de la silla y así la logra reconocer aunque esta tenga modificaciones.

En el campo de la computación, se ha tratado de imitar esta misma abstracción de las características principales, seleccionando características representativas de un conjunto de datos que logren describir de forma precisa el objeto o el evento que representan esos datos. Esto ha dado lugar al desarrollo de lo que hoy conocemos como selección de características.

Formalmente la selección de características es la aplicación de un algoritmo o conjunto de estos que busca encontrar el conjunto de características dimensionalmente menor y que además aporte la mayor cantidad de información. Esto

con el objetivo de mejorar la eficiencia y el procesamiento de patrones por algún clasificador [1]. Recientemente se ha reportado la aplicación de selección de características, utilizando computación paralela y memorias asociativas [2], cabe señalar que esta técnica pierde su viabilidad cuando los conjuntos de datos tienen un elevado número de características, es en esta situación cuando se debe idear un método más eficiente para la selección de características en conjuntos de datos altamente dimensionales.

Existen problemas que debido a lo complejo de su naturaleza generan gran cantidad de características, en estos problemas es deseable la reducción de las mismas para facilitar el procesamiento de sus patrones y reducir la información redundante o de poca importancia. La selección de características se puede ver como un problema combinacional, donde se tiene un conjunto finito de combinaciones que representa la posibilidad de usar o no una característica para fines de clasificación.

Los conjuntos de entrenamiento con grandes cantidades de características son complejos de analizar con técnicas de búsqueda exhaustiva, como la usada en [2]. En el presente trabajo se propone el uso de los algoritmos genéticos como método alternativo para realizar la selección de características. Los algoritmos genéticos son técnicas de búsqueda basadas en mecanismos de selección natural y operaciones genéticas, dentro de la naturaleza sirven para encontrar soluciones a problemas combinacionales [3]. Sabiendo esto, se puede concluir que un algoritmo genético es una alternativa computacionalmente viable para tratar procesos de selección de características en conjuntos de datos altamente dimensionales.

1.3. Planteamiento del problema

Se estima que cada 20 meses el volumen de datos en el mundo se duplica [4]. Todos estos datos deben ser tratados, analizados y procesados para generar información útil para el consumo de las empresas o los particulares. Tomando como premisa esta situación, surge la necesidad de simplificar el análisis y el tratamiento de estos datos para convertirlos en información. Las ciencias de la computación, a través del reconocimiento de patrones y diversos algoritmos clasificadores, proporciona una alternativa para resolver esta problemática.

El desempeño de los algoritmos de clasificación depende directamente de los datos con que se les entrene, estos datos son proporcionados a los clasificadores en forma de patrones.

Los conjuntos de entrenamiento son abstracciones de datos relacionados, que tienen una interpretación lógica para el usuario. El objetivo final del clasificador es asignar una etiqueta de clase a una instancia desconocida por el mismo.

Los patrones sor valores agrupados en filas y columnas. Cada fila describe al objeto o evento de interés y cada columna representa una característica del mismo. Las filas en los conjuntos de datos se conocen como instancias, cada instancia del conjunto de entrenamiento tiene asociada una etiqueta, la cual puede ser un valor numérico o una cadena de caracteres y se le llama clase.

Con la creciente cantidad de características en los patrones, surgen problemas como "The curse of dimensionality"[5]. Esta problemática se refiere a que entre más características tenga un patrón, menor va a ser el espacio para diferenciar entre las diversas instancias proporcionadas. Este es un problema para los métodos de clasificación ya que ante esta situación tienden a reducir su capacidad de discriminación.

Otro problema asociado a los patrones altamente dimensionales es el tiempo de procesamiento. El tiempo que un algoritmo clasificador necesita para procesar un patrón aumenta de acuerdo con la dimensión del patrón usado.

1.4. Contribuciones

- Un algoritmo genético original propuesto en este trabajo de tesis.
- Una metodología para la selección de características utilizando el algorimto propuesto.

1.5. Organización de la tesis

Las secciones 1.1, 1.2, 1.3 y 1.4, describen el objetivo, la motivación, el planteamiento del problema y las contribuciones. A continuación se explica la organización del resto del documento de tesis.

El Capítulo 2 proporciona las bases en que se fundamenta la presente tesis y consta de dos secciones. La primera sección está dedicada a la selección de características, proporcionando conceptos necesarios para la correcta comprensión del trabajo, en la segunda sección se presenta una introducción a los algoritmos genéticos abarcando su definición y sus operaciones.

Dentro del capítulo 3, se hace una recopilación de los trabajos más recientes relacionados con el tema de esta tesis.

El capítulo 4 incluye los algoritmos que se utilizan dentro del trabajo de tesis. Las operaciones del algoritmo genético son descritas a detalle en este capítulo.

El capítulo 5 es la parte medular del trabajo de tesis, en este capitulo se describe el proceso que se llevo acabo para obtener la metodología propuesta.

En el capítulo 6 se presentan una serie de experimentos que muestran funcionamiento de la metodología propuesta. Esta es probada con diferentes conjuntos de datos públicos.

El capítulo 7 consta de dos secciones, una dedicada a las conclusiones y otra donde se describe el trabajo futuro. En la sección de trabajo a futuro, se enumeran los diferentes aspectos no cubiertos por el actual trabajo de tesis y que podrian ser de interes para el lector.

Para concluir, se presentan las referencias bibliográficas.

2. Antecedentes

Este capítulo consta de dos secciones. En la sección 2.1 se describen los conceptos básicos sobre selección de características, en la sección 2.2, por otro

lado, se describen los algoritmos genéticos, sus motivaciones, su funcionamiento, sus operaciones y se incluye un ejemplo de uso.

2.1. Selección de características

La tarea de la selección de características consiste en obtener un subconjunto de características que proporcione la mayor cantidad de información útil [4], por lo tanto después de que se realizó la selección de características el conjunto de datos obtenido debe contener los datos más relevantes, de hecho las buenas técnicas de selección de características deben ser capaces de detectar e ignorar características irrelevantes. El resultado de esto es que la calidad de el conjunto de datos se incrementa después de la selección de características. Este es un fenómeno deseable, ya que al incrementar la calidad del conjunto de datos, los clasificadores incrementarán su eficiencia de clasificación, con esto se realizan mayor número de clasificaciones de forma correcta.

Por sentido común se espera que al incrementar el número de características, se incrementará la eficiencia de clasificación ya que se incluye más información para distinguir una clase de la otra, desafortunadamente esto no es verdad si el tamaño del conjunto de entrenamiento no se incrementa en igual medida que se agregan características. Un conjunto de datos altamente dimensional aumenta la probabilidad de que un algoritmo encuentre resultados falsos que no son válidos en general [4].

Una de las razones por las que al tener un conjunto de datos altamente dimensional, aumente la probabilidad de que un algoritmo encuentre resultados falsos es la llamada *The Curse of Dimensionality* [5]. *The Curse of Dimensionality* plantea que 2 puntos cercanos en un espacio de 2 dimensiones pueden estar a distancias exactamente iguales en un espacio de 100 dimensiones. Esto dificulta a los algoritmos clasificadores, el hacer predicciones sobre instancias de datos no clasificados que no fueron enseñados en el entrenamiento.

Existen 2 cualidades que se deben ser consideradas por los métodos de selección de características: relevancia y redundancia. Una característica es relevante si provee información valiosa para el proceso de clasificación, en caso contrario es irrelevante [4]. Una característica se considera redundante si está altamente correlacionada con otras características, esto es: posee en los diferentes ejemplos proporcionados, rasgos parecidos o iguales. Los subconjuntos de características deseables a encontrar en la selección de características deben poseer propiedades de relevancia y no redundancia.

El número de características también es clave para determinar el número total de hipótesis que pueden ser deducidas de los datos proporcionados. Una hipótesis es el resultado de la función que predice a que clase pertenece una instancia en específico, basándose en los datos proporcionados. Entre más características se tengan, es más grande el espacio de hipótesis, pero si se disminuye el espacio de hipótesis, es más fácil encontrar la hipótesis correcta. El punto anterior se debe tener en consideración ya que también el incremento lineal en el número de características, deriva en un incremento exponencial del espacio de hipótesis [5]. La selección de características puede, de forma eficiente, reducir

el espacio de hipótesis, quitando características irrelevantes y redundantes. Dado un conjunto de datos de tamaño fijo, la reducción de dimensionalidad de los patrones también reduce el tiempo de entrenamiento en los algoritmos de clasificación [5].

Para realizar la selección de características se han estudiado diferentes propuestas: modelos de selección, estrategias de búsqueda, medidas de la calidad de las características y evaluación de las características con un clasificador. Los 3 modelos típicos de selección de características son *Filter*, *Wrapper* y *Embedded* [5]. Un modelo *Embedded* de selección de características integra la selección de características en la construcción del modelo, un ejemplo de este modelo es el algoritmo de árbol de decisión por inducción [5]. En un modelo *Wrapper* se emplea un algoritmo de aprendizaje y se usa el rendimiento de este algoritmo para determinar la calidad de las características seleccionadas. En un modelo *Filter* se utilizan algoritmos independientes al clasificador, estos algoritmos filtran los datos proporcionados para evaluar que tan bueno es incluir una característica en la clasificación [1] estos modelos están basados en métricas de evaluación aplicadas directamente sobre los datos.

En algunos algoritmos se usan estrategias secuenciales para seleccionar las características, un ejemplo de estos es el *Sequential Forward Selection* (SFS) [5], en este algoritmo se selecciona una característica a la vez, hasta que al añadir otra característica no mejore la calidad del subconjunto seleccionado. En este algoritmo existe la condición que la característica seleccionada no salga del conjunto de características. De forma similar el algoritmo *Sequential Backward Selection* (SBS) [5] elimina una característica a la vez, después de que esta característica fue eliminada, no se vuelve a tomar en cuenta. Estas 2 estrategias son heurísticas y no garantizan encontrar el subconjunto óptimo de características. Existen otras alternativas a estos métodos como los algoritmos aleatorios de selección de características.

La evaluación de las características seleccionadas comúnmente conlleva dos tareas, una es comparar el antes y después de la selección de características [5]. El objetivo de esta tarea es observar si la selección de características logra sus objetivos. Los aspectos a evaluar pueden ser el número de características seleccionadas, el tiempo empleado, la escalabilidad y la eficiencia del clasificador usado. La segunda tarea es comparar dos algoritmos de selección de características para ver si uno es mejor que otro para una tarea en específico.

Finalmente la selección de características aporta los siguientes beneficios [4]:

- Facilita la visualización de los datos: con esto las tendencias en los datos se pueden reconocer más fácilmente.
- Reduce los requerimientos para las mediciones y almacenamiento: existen datos que son difíciles de medir y ocupan gran cantidad de espacio cuando son almacenados, al reducir las características se reducen las medidas a realizar del experimento analizado.
- Reduce los tiempos de entrenamiento y clasificación: con conjuntos de datos más pequeños el tiempo de los algoritmos de clasificación puede ser

reducido, sobre todo si el algoritmo es computacionalmente complejo.

- Mejora el rendimiento de clasificación: el rendimiento se aumenta ya que la selección de características remueve características redundantes o irrelevantes para los algoritmos de clasificación.

2.2. Algoritmos genéticos

En la naturaleza, un proceso de evolución ocurre cuando se cumplen las siguientes condiciones [6]:

- Una entidad tiene la habilidad de reproducirse
- Existe una población de entidades con capacidades reproductivas
- Existe alguna variación entre las entidades
- Algunas diferencias en la habilidad de sobrevivir al ambiente son asociadas con la variabilidad de la entidad

En un entorno natural, las variaciones de la entidad están codificadas en los cromosomas, estas variaciones impactan en los individuos modificando su comportamiento y sus estructuras físicas. Dentro de un entorno competitivo, con recursos limitados, las entidades con mayor adaptabilidad al medio, son las que sobreviven, crecen y se reproducen pasando esas variaciones favorables a su descendencia, al contrario de aquellas entidades cuyas variaciones no les fueron útiles para sobrevivir al ambiente. Estos conceptos de supervivencia del más apto y selección natural son descritos en [7].

Después de varias generaciones y debido a la diversidad en la población, los miembros mejor adaptados, crecen y se reproducen, estos producen descendientes con las características que les sirvieron a los padres para sobrevivir así que estos descendientes están mejor adaptados al medio. Cuando las diferencias de la población son visibles, medibles y su adaptabilidad es mayor, se dice que la población ha evolucionado [6]. En este proceso los individuos con mayor adaptabilidad son los que prevalecen.

En el libro de Holland *Adaptation in Natural and Artificial Systems* [8] se proporciona un marco general para analizar los sistemas adaptativos, naturales o artificiales y se muestra como el proceso evolutivo se puede aplicar a los sistemas artificiales. Cualquier problema de adaptación en general puede ser modelado en términos genéticos. Después de que es modelado en estos términos, el problema puede ser resuelto por los llamados Algoritmos Genéticos (AG) [6].

Los algoritmos genéticos de forma general son procedimientos de búsqueda basados en los principios de la selección natural. Pueden ser definidos como algoritmos matemáticos altamente paralelos que transforman una población de objetos matemáticos individuales, en una nueva población mejor adaptada al medio, utilizando operaciones modeladas con los principios darwinianos de la

reproducción y la supervivencia del más apto, logran encontrar o crear a los mejores objetos matemáticos, que representan la mejor solución del problema [6].

El primer AG fue desarrollado por John H. Holland en 1960 este permitía a las computadoras evolucionar encontrando soluciones para problemas combinacionales y de búsqueda, particularmente difíciles [3].

La mayoría de los algoritmos genéticos tiene cuatro elementos en común:

Una población inicial de cromosomas (soluciones probables), un proceso de selección de acuerdo con su adaptabilidad, capacidad de reproducción para generar nuevos individuos y un rango de mutación de los nuevos individuos. Los cromosomas comúnmente son representados como arreglos binarios de n elementos, también llamados alelos, cada uno de estos alelos, puede tomar el valor de 0 o 1. Cada cromosoma puede ser visto como un punto de búsqueda en un espacio de posibles soluciones [9].

La forma más simple de un algoritmo genético involucra tres operaciones: selección, reproducción y mutación. Cada una es explicada a continuación.

- Selección: La función de esta operación es elegir de forma elitista padres entre la población para crear descendencia. Los valores de aptitud proporcionados por la función fitness, son tomados como criterio para decidir qué individuo de la población será un posible candidato para el proceso de reproducción [10]. El propósito de la selección es acentuar la supervivencia de los individuos más aptos, esperando que estos individuos se reproduzcan y generen descendencia que posea mayores valores de aptitud [9].

- Reproducción: Esta operación selecciona de manera aleatoria un punto en el cromosoma y combina las subsecuencias antes y después del punto seleccionado para crear descendencia. Por ejemplo los cromosomas 1110|0001 y 0000|1111 pueden ser reproducidos por el punto medio de cada uno, obteniendo los descendientes 11101111 y 00000001. El objetivo de esta operación es recombinar cromosomas, también llamados esquemas y generar diferentes soluciones, cada vez más cercanas a la solución óptima del problema [9].

- Mutación: De manera aleatoria, esta operación cambia algunos bits en los cromosomas. Por ejemplo si el cromosoma 00010101 se muta en la primera y segunda posición, el cromosoma resultante será 11010101. La mutación se puede producir en cada alelo del cromosoma, con una probabilidad usualmente baja. La mutación asegura a la población contra una convergencia a un resultado local [9].

A continuación se presenta un AG estándar de De Jong [11]:

Generar una población aleatoria de m padres

Repetir:

Evaluar y guardar el valor aptitud f(i) para cada individuo i en la población de padres

Definir la probabilidad de seleccionar algún padre de la población utilizando un método de selección

Generar m descendientes seleccionando aleatoriamente a los padres

Seleccionar la descendencia que sobrevivirá

Termina repetir

3. Estado del arte

La selección de características es un problema de optimización importante dentro del ámbito de la clasificación de patrones. A través de la selección de características se pretende maximizar la eficiencia de clasificación y minimizar el número de características usadas. Desde 1998 se ha probado que los algoritmos genéticos son muy útiles resolviendo problemas complejos de optimización **[12]**, en este caso, la selección del mejor subconjunto de características en un problema de clasificación es presentado como un caso particular de un problema de optimización.

En 1999 se usaron los algoritmos genéticos para la selección de características utilizando redes neuronales en un problema de detección para tumores en la piel, como indicador de la calidad proporcionada por los subconjuntos de características (función de aptitud o función Fitness) se usó el rango de clasificación proporcionado por *Nearest Neighbors Classifier*. Este cálculo se realizó usando la metodología *Leaving One Out*. [13].

En el año 2006 debido a la gran cantidad de datos obtenidos a través de chips de ADN, se buscaron nuevas formas de analizar grandes cantidades de información genómica. Esta información se debe depurar y de ella extraer las características relevantes, para esto se usó un algoritmo genético con especiación y un cromosoma modificado para poder representar gran cantidad de características en una representación reducida, añadido a esto como función de aptitud se uso una red neuronal **[14]**.

En el campo de la medicina se han usado diversas combinaciones de algoritmos genéticos y selección de características, como ejemplo de esto en el 2013 se realizó un trabajo para proporcionar diagnósticos diferenciales de la enfermedad erythemato-squamous, este proyecto usó Algoritmos genéticos, una red bayesiana y selección de características [15].

Pero no solo en el campo de la medicina, también en el ambiente empresarial han sido ampliamente utilizados. En 2014 se utilizó un algoritmo genético híbrido en combinación con redes neuronales, para identificar el subconjunto óptimo de características dentro de un conjunto patrones de riesgos crediticios, este algoritmo se usó para evaluar una base de datos de créditos, obtenida de un banco Croata **[16]**.

También en 2014 pero en el campo de reconocimiento de escritura a mano alzada, se propuso un algoritmo genético del tipo generacional con una función de aptitud tipo filtro llamado *Separability Index*. El *Separability Index*, es una extensión del método *Fisher Linear Discriminant*, esta aproximación dio buenos

resultados analizando 4 de las principales bases de datos escritura a mano alzada [17].

4. Algoritmos

En esta sección se describen los algoritmos que se utilizan dentro del trabajo de tesis. Diversas operaciones de los algoritmos genéticos son descritas a detalle con ejemplos.

4.1. Operadores genéticos

Los AG dependen de tres operaciones básicas para su correcto funcionamiento, estas operaciones son: selección, reproducción (crossover) y mutación. conforme los AG se han desarrollado, los expertos en la materia han propuesto diversas formas de realizar estas operaciones, considerando algunas deficiencias de las operaciones originales y tratando de mejorarlas, a continuación se muestra una pequeña recopilación de diversas formas de realizar estas operaciones encontradas en la literatura.

4.1.1. Selección

La operación selección da preferencia para reproducir a los individuos que mejor estén evaluados por la función de adaptación (Fitness). Esto se logra planteando métodos que favorezcan el paso de los genes de buenas soluciones e impidan el paso de genes provenientes de peores soluciones. Esto con el propósito de que la operación de reproducción (crossover) trabaje con los mejores cromosomas disponibles.

A grandes rasgos las operaciones de selección están divididas en: selección proporcional, selección basada en ranking, selección por torneos, selección por rango y la selección basada en género. A continuación se presenta una descripción general de cada uno y ejemplos.

Métodos de selección proporcional La selección por proporción describe un conjunto de modelos de selección que elige a los individuos de acuerdo con su valor de aptitud f. En estos modelos la probabilidad de selección p de un individuo de la i-ésima clase en la t-ésima generación está dada por la siguiente expresión [18]:

$$p_{i,t} = \frac{f_i}{\sum_{j=1}^{n} m_{j,t} f_j}$$

Donde n es el número total de individuos y m es la sumatoria del valor de todos los individuos.

Se han sugerido diversos métodos para muestrear esta distribución de probabilidad, incluyendo *Monte Carlo* o *Roulette Wheel Selection*, *Stochastic Remainder Selection* y *Stochastic Universal Selection* [18]. A continuación se explicarán los métodos *Roulette Wheel Selection* y *Stochastic Remainder Sampling.*

Roulette Wheel Selection El método *Roulette Wheel Selection* es conceptualmente equivalente a darle a cada elemento de la población una porción de una ruleta, en donde el área de la ruleta asignada al elemento sea igual al Fitness del mismo. Al girar la ruleta, el marcador de selección de individuo se detiene en un punto específico, este marcador, tendrá mayor probabilidad de detenerse en un elemento cuya área sea mas grande. El primer paso es calcular la sumatoria del Fitness de todos los individuos, después dividir el Fitness particular del individuo entre la sumatoria del Fitness total y así obtener su probabilidad de selección, esto se expresa en la siguiente fórmula [9].

$$p_i = \frac{f_i}{\sum_{j=1}^{n} f_j}$$

Se crea un arreglo con el valor Fitness de cada elemento, después se recorre la lista utilizando la probabilidad del individuo para decidir si se selecciona o no, esto se repite hasta tener todas las parejas seleccionadas.

Ejemplo:

Se tiene una población de 6 elementos con los Fitness que a continuación se presentan:

No.	Fitness
1	12
2	10
3	11
4	30
5	5
6	14

- Se calcula la probabilidad de selección de cada individuo

P(1) = 12/82= 0.14
P(2) = 10/82= 0.12
P(3) = 11/82= 0.13
P(4) = 30/82= 0.36
P(5) = 5/82= 0.06
P(6) = 14/82= 0.17

- Se obtiene un número aleatorio entre 0 y 1, si el número obtenido es menor o igual a la probabilidad de selección de cada individuo, se seleccionará el elemento, de lo contrario, no se hará.

Número aleatorio obtenido para P(1): .61, .61>P(1) entonces,
no se selecciona.
Número aleatorio obtenido para P(2): .45, .45>P(2) entonces,
no se selecciona.
Número aleatorio obtenido para P(3): .33, .33>P(3) entonces,
no se selecciona.
Número aleatorio obtenido para P(4): .22, .22>P(4) entonces,

se selecciona.
Número aleatorio obtenido para P(5): .64, .64>P(5) entonces,
no se selecciona.
Número aleatorio obtenido para P(6): .77, .77>P(6) entonces,
no se selecciona.
Número aleatorio obtenido para P(1): .34, .34>P(1) entonces,
no se selecciona.
Número aleatorio obtenido para P(2): .15, .15>P(2) entonces,
no se selecciona.
Número aleatorio obtenido para P(3): .07, .07<P(3) entonces,
se selecciona.
Número aleatorio obtenido para P(4): .29, .29<P(4) entonces,
se selecciona.
Número aleatorio obtenido para P(5): .64, .64>P(5) entonces,
no se selecciona.
Número aleatorio obtenido para P(6): .87, .87>P(6) entonces,
no se selecciona.
Número aleatorio obtenido para P(1): .25, .25>P(1) entonces,
no se selecciona.
Número aleatorio obtenido para P(2): .01, .01<P(2) entonces,
se selecciona.
Número aleatorio obtenido para P(3): .34, .34>P(3) entonces,
no se selecciona.
Número aleatorio obtenido para P(4): .96, .96>P(4) entonces,
no se selecciona.
Número aleatorio obtenido para P(5): .80, .80>P(5) entonces,
no se selecciona.
Número aleatorio obtenido para P(6): .17, .17<P(6) entonces,
se selecciona.
Número aleatorio obtenido para P(1): .56, .56>P(1) entonces,
no se selecciona.
Número aleatorio obtenido para P(2): .90, .90>P(2) entonces,
no se selecciona.
Número aleatorio obtenido para P(3): .70, .70>P(3) entonces,
no se selecciona.
Número aleatorio obtenido para P(4): .23, .23<P(4) entonces,
se selecciona.
Los individuos seleccionados son 4,3,4,2,6 y 4.

Stochastic Remainder Sampling Este método asigna padres de forma determínistica, tomando como referencia un valor *mi* determinado por el Fitness y posteriormente usa el metodo *Roulette Wheel* para seleccionar los elementos faltantes de forma estocástica [19].

El número de copias por individuo es calculado como:

$$mi = \frac{if}{f} \tag{1}$$

Donde if es el Fitness del individuo y f es el Fitness promedio.

Solo la parte entera de mi es usada para seleccionar individuos en la siguiente generación de la población.

El algoritmo a utilizar es el siguiente:

- Se calcula el fitness de cada individuo asi como el promedio de todos los individuos y se calcula el valor mi para cada individuo.
- Se utiliza la parte entera del valor mi para realizar copias del individuo, con esto se genera un pool de individuos.
- Se seleccionan los individos mas aptos del pool de individuos a través del metodo *Roulette Wheel*.

Ejemplo:

Se tiene una población de 6 elementos con los Fitness que a continuación se presentan:

No.	Fitness
1	12
2	10
3	11
4	30
5	5
6	14

- Se calcula el Fitness promedio.

$if = 13{,}67$

- Se calcula el valor mi dado por la ecuación 1 para cada individuo

$mi(1) = 13.67/12 = 1.14$
$mi(2) = 13.67/10 = 1.367$
$mi(3) = 13.67/11 = 1.24$
$mi(4) = 13.67/30 = 0.455$
$mi(5) = 13.67/5 = 2.73$
$mi(6) = 13.67/14 = .976$

- Se relizan las copias del individuo.

El pool de individuos se muestra a continuación.
1,2,3,5,5

- Se seleccionan los individos mas aptos del pool de individuos a traves del metodo *Roulette Wheel* presentado en la sección anterior.

Los elementos seleccionados son 5,5,2,2,1,3

Selección basada en ranking Baker en 1985 introdujo la noción de la selección basada en ranking a la práctica de los algoritmos genéticos [18]. En este tipo de selección se ordena la población del mejor al peor, después se le asignan las copias que cada individuo debe recibir de acuerdo con una función constante de asignación y se realiza una selección proporcionada de acuerdo con la asignación realizada. A continuación se presenta el método Linear Ranking Selection

Linear Ranking Selection En el modelo linear ranking selection propuesto por Baker en 1985 los individuos son acomodados en un ranking, de acuerdo con su valor de Fitness. Aquellos con elevados valores de Fitness están en primeros lugares y aquellos con bajos valores de Fitness están en un ranking bajo. Los individuos son seleccionados con una probabilidad linealmente proporcional al rank de los individuos de la población [19].

Ejemplo:

Se tiene una población de 6 individuos con los Fitness que a continuación se presentan:

No.	Fitness
1	12
2	10
3	11
4	30
5	5
6	14

- Ordenamos la población del mayor al menor (del mejor al peor)

No.	Fitness
4	30
6	14
1	12
3	11
2	10
5	5

- Asignamos copias de los elementos a un nuevo conjunto de individuos padres probables para ser seleccionados, utilizando una proporción basada en su ranking.

Esto es al valor que esta primero en el ranking se le asignarán mas copias, al segundo valor menos y al tercer valor menos, hasta llegar al último valor del ranking, que contendrá el menor número de copias. Esto se debe hacer de forma proporcional.

Se determina de manera trivial un valor de 6, el primer elemento del ranking obtendrá 6 copias y los elementos subsecuentes obtendrán una copia menos que el individuo inmediatamente superior en el ranking.

No.	Fitness	Copias
4	30	6
6	14	5
1	12	4
3	11	3
2	10	2
5	5	1

- Se crea un pool de individuos con los datos anteriores.

4,4,4,4,4,4,6,6,6,6,6,1,1,1,1,3,3,3,2,2,5

- Se selecciona al azar 6 elementos.

Se obtienen de forma aleatoria valores entre 1 y 21, los valores aleatorios obtenidos son los siguientes: 11,9,20,18,8,2 estos valores indican la posición del valor a seleccionar.

Tomamos los individuos de las pocisiones obtenidas de forma aleatoria.

Los elementos seleccionados son 6,6,2,3,6 y 4.

Selección por torneos Se selecciona un número aleatorio de individuos de la población (con o sin remplazo) para que participen en un torneo, después se selecciona al mejor o los mejores individuos del torneo, para su posterior utilización, y se repite el proceso las veces que se desee (usualmente hasta que el pool de individuos a reproducir este lleno). Los torneos generalmente son entre dos individuos, también se pueden hacer torneos largos, con más de dos individuos [18]. A continuacion se explicará en detalle y con un ejemplo los métodos *Binary Tournament Selection, Boltzmann Tournament Selection y Correlative Tournament Selection.*

Binary Tournament Selection En esta variante de la selección por torneos, dos individuos son elegidos de manera aleatoria, el mejor de estos dos individuos es seleccionado con una probabilidad fija p, donde $.5 < p$ y aproximadamente igual a uno.

Ejemplo:

- Se generó la siguiente población inicial.

No.	Fitness
1	12
2	10
3	11
4	30
5	5
6	14

Se realizará un torneo binario (entre dos elementos) en el cual a cada individuo se le asignará una pareja de manera aleatoria, los dos individuos compiten y el que tenga el valor más alto, se mantendrá y el del valor bajo se eliminará.

- La asignación de parejas se realizó de la siguiente forma

No.	Pareja
1	1
2	2
3	1
4	3
5	3
6	2

Las dos parejas con el número 1 (los individuos 1 y 3), se seleccionan para el torneo, la competencia en el torneo usa sus fitness, para determinar quien se duplicará y quien se eliminará. El individuo No. 1 se selecciona al ser el mejor Fitness de los dos, lo mismo se hace con todas las parejas.

Los individuos seleccionados son 1,1,6,6,4 y 4.

Boltzmann Tournament Selection *Boltzmann Tournament Selection* (BTS) es un método de selección inspirado en el recocido simulado (simulated annealing) [19]. BTS mantiene la diversidad en la población mediante el muestreo, de tal manera que las muestras con el tiempo se vuelvan un distribución Boltzmann.

Sin embargo, el factor de la deriva genética trabaja en contra de BTS mediante la limitación de la cantidad de diversidad que puede mantener.

Una población contiene n elementos. En cada generación, un torneo de tres individuos es llevado a cabo en cada espacio de la población. El primer individuo es seleccionado de forma uniformemente aleatoriamente, el segundo individuo debe ser seleccionado con un valor de Fitness diferente del primero por un rango x. La mitad de las veces se seleccionará un tercer individuo con un valor de Fitness diferente del primero y del segundo por el rango x (este proceso es llamado selección estricta), y la otra mitad de las veces, se selecciona un tercer individuo con un valor Fitness diferente del valor Fitness del primer elemento (esto es llamado selección relajada). Si el algoritmo es exitoso, creará una porción estable de individuos de acuerdo con la distribución de Boltzmann. Si se selecciona uniformemente aleatorio un elemento de la población este estará sesgado hacia los mejores individuos.

Ejemplo:

Se presenta el siguiente espacio de la población.

No.	Fitness
1	12
2	10
3	11
4	30
5	5
6	14

Se lleva acabo un torneo de tres individuos.

- Se selecciona un individuo de forma uniformemente aleatoria, este individuo es el número 6.
- Se elige un rango x de 3, con este rango se elige otro individuo diferente por este rango, este individuo es el número 2.

Se realiza una selección estricta.

- Se selecciona un tercer individuo con un valor de Fitness diferente del primero y del segundo por el rango x, este es el individuo número 5.
- Se selecciona un individuo de forma uniformemente aleatoria, este individuo es el número 3.
- Se elige un rango x de 3, con este rango se elige otro individuo diferente por este rango, este individuo es el número 5.

Se realiza una selección relajada.

- Se selecciona un tercer individuo con un valor Fitness diferente del valor Fitness del primer elemento, este es el individuo número 2.

Los elementos seleccionados son 6,2,5,3,5 y 2.

Correlative Tournament Selection Este método es una extensión de la selección ordinaria por torneos [19]. En lugar de seleccionar dos padres de manera aleatoria para reproducirlos, se selecciona un par de padres que estén altamente correlacionados por naturaleza, esto con el objetivo de realizar una reproducción efectiva. La distancia de Hamming es usada como función para decidir si los individuos tienen cierta correlación entre sí.

Ejemplo:

Se generó la siguiente población inicial.

No.	Fitness	Binario
1	12	01100
2	10	01010
3	11	01011
4	30	11110
5	05	00101
6	14	01110

Se calcula la distancia de Hamming de los individuos.
Hamming(1,2) = 2
Hamming(1,3) = 3
Hamming(1,4) = 2
Hamming(1,5) = 2
Hamming(1,6) = 1
Hamming(2,3) = 1

Hamming(2,4) = 2
Hamming(2,5) = 4
Hamming(2,6) = 1
Hamming(3,4) = 3
Hamming(3,5) = 3
Hamming(3,6) = 2
Hamming(4,5) = 4
Hamming(4,6) =1
Hamming(5,6) =3

Los individuos pueden ser seleccionados de los individuos cuya distancia de Hamming es 1.

Los pares de padres seleccionados son 1-6, 2-3, 2-6, 4-6

Selección por rango A cada elemento de la población se le asigna un ranking numérico basado en su Fitness, la selección es basada en ese ranking. Se selecciona un subconjunto de los mejores individuos. La ventaja de este método es que puede prevenir que los elementos con alto Fitness sean dominantes al principio, a expensas de los individuos con menos aptitud, que reduciría la diversidad genética [19].

Ejemplo:

Se generó la siguiente población inicial.

No.	Fitness
1	12
2	10
3	11
4	30
5	5
6	14

Se ordenan los elementos del mayor al menor basado en su Fitness.

No.	Fitness
4	30
6	14
1	12
3	11
2	10
5	5

- Un rango de selección es establecido.

Se estableció un rango del primer elemento al segundo.

Los elementos seleccionados son 4,4,4,6,6,6.

Selección basada en género El concepto base tras de esta técnica proviene de la selección sexual natural, donde el concepto de seleccionar una pareja es diferente entre individuos femeninos y masculinos. Los individuos de las especies generalmente pertenecen a la clase macho o hembra. El que dos individuos diferentes que están especializados en sus propios campos cooperen de forma común en la naturaleza indica una ventaja, ya que esta forma de cooperación prevalece hasta nuestros días. En un ambiente natural esta cooperación y especialización hacen posible la generación de descendencia sana y mejor adaptada al medio [19].

Restricted mating A los individuos sólo se les está permitido reproducirse con individuos que tengan una etiqueta en común. Esta etiqueta representa la similitud en sus genes. La distancia de Hamming es usada como función para decidir si los individuos presentan similitud en sus genes.

Ejemplo:

Se generó la siguiente población inicial.

No.	Fitness	Binario
1	12	01100
2	10	01010
3	11	01011
4	30	11110
5	05	00101
6	14	01110

Se calcula la distancia de Hamming de los individuos.

Hamming(1,2) = 2
Hamming(1,3) = 3
Hamming(1,4) = 2
Hamming(1,5) = 2
Hamming(1,6) = 1
Hamming(2,3) = 1
Hamming(2,4) = 2
Hamming(2,5) = 4
Hamming(2,6) = 1
Hamming(3,4) = 3
Hamming(3,5) = 3
Hamming(3,6) = 2
Hamming(4,5) = 4
Hamming(4,6) =1
Hamming(5,6) =3

Los individuos se etiquetarán, de acuerdo a su distancia de Hamming, unos con otros.

No.	Fitness	Etiqueta
1	12	1
2	10	2
3	11	2,3
4	30	4
5	5	0
6	14	1,3,4

De acuerdo con el método, solo a los individuos 1-6, 2-3, 3-6, 4-6 se les esta permitido reproducirse.

Correlative Family- based selection En este método un individuo llamado arbitrariamente xi, que representa el individuo con más alto Fitness de la familia, es seleccionado. La familia está conformada por los dos padres y sus descendientes. Se calcula la distancia de Hamming entre xi y todos los individuos. El individuo con mayor distancia de Hamming es seleccionado para sobrevivir a la siguiente generación junto con xi. Esto se hace para mantener la diversidad de la población así como mantener al individuo con el mayor Fitness.

Ejemplo:

Se presenta la siguiente familia.

No.	Fitness	Binario
1	12	01100
2	10	01010
3	11	01011
4	30	11110

- Se selecciona el individuo número 4 como xi
- Se calcula la distancia entre xi y todos los individuos.

Hamming(xi,1) = 2
Hamming(xi,2) = 2
Hamming(xi,3) = 3

- Se selecciona el individuo con mayor distancia de Hamming. Se selecciona el individuo 3, junto con xi.

Los elementos seleccionados son 3 y 4.

Con estos individuos se generará una nueva familia para repetir el procedimiento.

4.1.2. Crossover

Una de las características que definen a los algoritmos genéticos es la operación crossover (reproducción) esta operación se basa en el concepto natural de reproducción. Los elementos más aptos (con mayor Fitness) se seleccionan

de la población y se reproducen. La razón para usar este operador es que posee la habilidad de recombinar soluciones que han sido bien calificadas con base en su Fitness, esto se hace esperando que al recombinarlas una mejor solución sea encontrada.

Existen diversas formas de realizar la operación crossover, a continuación se explicarán algunas de ellas.

El crossover de un solo punto, es la forma más sencilla de realizar una reproducción. Se elige un punto aleatorio en los dos individuos, después de seleccionar este punto, los bits antes del punto seleccionado y después de este se combinan y generan la nueva descendencia.

Ejemplo:

padre: 1101011
madre: 0011011

- Se selecciona aleatoriamente un punto p, desde 1 hasta $n-1$ donde n es la longitud de los individuos

Consideremos que $p = 3$
El corte quedará de la siguiente forma

padre: 1101|011
madre: 1110|001

- Ahora se realiza el crossover, concatenando los bits del padre antes del punto p con los bits de la madre después del punto p y viceversa

padre + madre: 1101001
madre +padre: 1110001

El objetivo de realizar esta operación es combinar buenas soluciones para intentar obtener una mejor solución. El método crossover de un solo punto tiene desventajas, con este método no es posible encontrar todas las posibles soluciones [9]. Por ejemplo, no se pueden combinar instancias de 11xxxxx1 con xxxx11xx para formar instancias de 11xx11x1, donde los valores x representan cualquier valor de 0 o 1.

El método de crossover de un punto trata algunos lugares de las soluciones de forma preferencial, los segmentos intercambiados de ambos padres siempre contiene los últimos bits de la cadena. Para reducir este sesgo posicional varios profesionales de los AG usan dos puntos en el crossover, en donde las dos posiciones son seleccionadas de forma aleatoria y sus bits son intercambiados. Esta forma de realizar la operación de crossover tiende a disminuir la formación de soluciones parecidas a los padres con individuos de gran tamaño, además de que puede combinar de más formas a los individuos que el crossover de un solo punto. Aunado a esto, los segmentos intercambiados no necesariamente tendrán los últimos bits de la cadena [9].

Aunque el crossover de dos puntos proporciona mayor cantidad de combinaciones para la solución, existen combinaciones que este método no puede

realizar, los expertos de los AG han probado con diferentes tipos de crossover, por ejemplo existe un método de crossover que utiliza para los puntos de corte, números de una distribución de poisson con base en la longitud del individuo.

Existen en la literatura muchos otros métodos para realizar la operación de crossover, pero el éxito o el fracaso de un método de crossover en particular depende de diversos factores, como la función Fitness usada, la forma en que se codifican las soluciones, y otros detalles del AG en particular [9].

4.1.3. Mutación

La operación reproducción es la más innovadora, creando variación en las poblaciones dentro de los algoritmos genéticos [9], pero esta operación en su esquema de construcción de nuevas solucionespuede llevar a los individuos a encontrar un máximo o un mínimo local. Para contrarrestar esto, se desarrolló la operación de mutación, que proporciona un seguro contra el estancamiento del algoritmo en un mínimo o máximo local.

Esta operación cambia algunos bits del individuo, de manera aleatoria.

Por ejemplo el individuo: 00101010

Puede mutarse a:

10101010 Cambiando el bit 0

00111010 Cambiando el bit 3

Algunos estudios comparativos han sido realizados para demostrarla eficacia de la mutación contra el crossover. No obstante, en 1993, Spears formalmente verificó la idea intuitiva de que: Aun cuando la mutación y el crossover tienen la misma habilidad para proponer soluciones, la operación crossover es una forma más robusta de construir soluciones [9].

No se puede elegir solo una operación, ya sea reproducción o mutación, se debe balancear la reproducción, la selección y la mutación para obtener buenos resultados. El lograr este balance depende también de los detalles en la codificación y la función Fitness utilizada. Incluso los parámetros de mutación y reproducción pueden variar mientras se ejecuta el algoritmo para mejorar dinámicamente la eficiencia de este [9].

4.2. Algoritmos genéticos propuestos

Se compararon cuatro algoritmos genéticos simples, con diferentes combinaciones de operaciones genéticas, para estos algoritmos el porcentaje de mutación fue sintonizado. Los algoritmos genéticos simples siguen el diagrama de flujo de la Figura 1.

Además de los 4 algoritmos genéticos simples se utilizó un Algoritmo Genético Dominante (AGD), este algoritmo se detalla en la Figura 2.

Los algoritmos genéticos a comparar fueron:

- Algoritmo Genético Simple con método de selección Binary Tournament y método de crossover Singlepoint crossover nombrado AGS1.

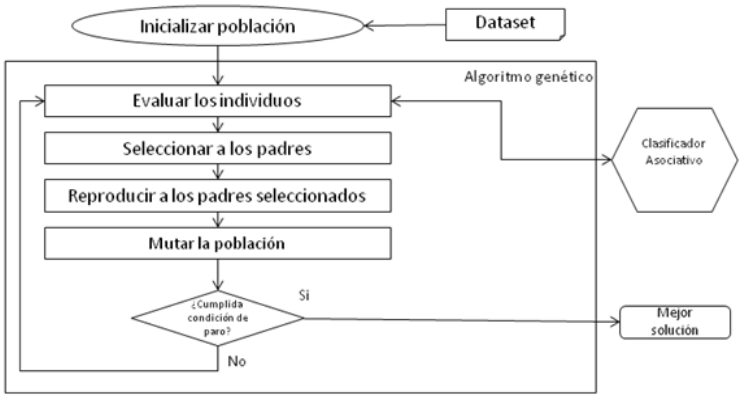

Figura 1: AG Simple

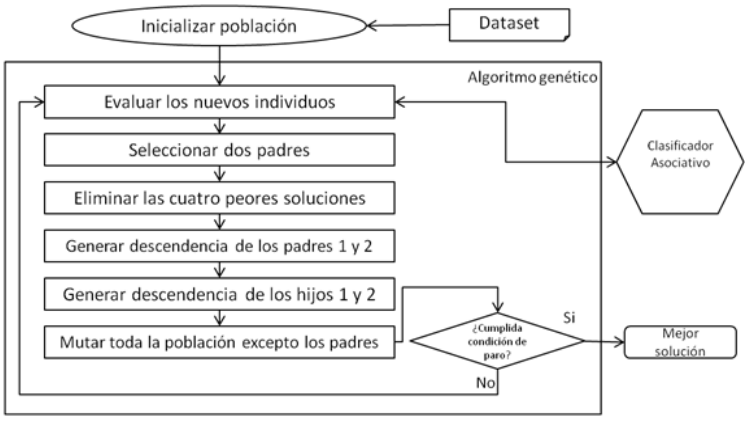

Figura 2: AG Dominante

- Algoritmo Genético Simple con método de selección Stochastic Remainder Sampling y método de crossover Singlepoint crossover nombrado AGS2.
- Algoritmo Genético Simple con método de selección Binary Tournament y método de crossover Doublepoint crossover nombrado AGS3.
- Algoritmo Genético Simple con método de selección Stochastic Remainder Sampling y método de crossover Doublepoint crossover nombrado AGS4.
- Algoritmo Genético Dominante nombrado AGD.

El algoritmo Genético Dominante se describe a continuación.

El algoritmo AGD nació de dos ideas principales, una estricta conservación de las mejores soluciones y una vasta exploración de diferentes puntos de solución.

La población inicial se define con una cantidad de 6 elementos, estos elementos se generan de manera aleatoria, el número de elementos en la población se mantiene constante en todo momento, esto se hace para restringir en cierta medida el espacio de soluciones.

El primer paso de este algoritmo es evaluar a los nuevos individuos, esto se realiza calculando el Fitness de cada elemento.

La estricta conservación de las mejores soluciones se hace al conservar solo las 2 mejores soluciones en todas las generaciones, esto se logra con el proceso de selección, que ordena las soluciones y elimina a las 4 peores soluciones.

Se realiza el proceso de selección y finalmente se mutan los hijos, para obtener mayor diversidad y lograr explorar puntos diferentes en el espacio de soluciones.

4.2.1. Descripción del Algoritmo Genético Dominante (AGD)

A continuación se presentan detalles del AGD, el algoritmo propuesto en este trabajo de tesis.

En la Figura 2 se muestra un diagrama del AGD cuyos pasos se listarán a continuación:

- Paso 1: Inicializar la población: La población como en cualquier otro AG se inicializa de manera aleatoria, restringiéndola en todo momento a 6 elementos.
- Paso 2: Evaluar los nuevos individuos: Los individuos se evalúan según la función Fitness que se presenta en 5.1 solo si no habían sido evaluados. En la primera generación se evalúan todos los individuos.
- Paso 3: Seleccionar a los padres: Para la selección, se utiliza un método tipo ranking; se ordenan las soluciones en una lista de mayor a menor seleccionando las 2 primeras soluciones.
- Paso 4: Eliminar las 4 peores soluciones: Se eliminan los últimos cuatro elementos de la lista anterior.

- Paso 5: Generar descendencia de los padres 1 y 2. Esto nos asegura que se pueda buscar un mejor resultado basado en la información relevante de los padres.
- Paso 6: Generar descendencia de los hijos 1 y 2. Este paso es por lo que el algoritmo se llama dominante, se reproducen los hijos, lo que da como resultado una copia de los hijos obtenidos en el paso anterior, esto nos deja con la población con 2 mejores soluciones y 2 copias de los hijos de estas soluciones.
- Paso 7: Mutar toda la población excepto los padres. En este paso se asegura que las dos copias de los hijos no sean iguales aplicando una mutación a cada solución. El mantener a los padres sin modificaciones asegura que siempre que se encuentre una buena solución esta se mantenga hasta que sea remplazada por una mejor.
- Paso 8: Si se cumplió la condición de paro, entregar el mejor resultado, de lo contrario ir al paso 2

4.3. Ejemplos de aplicación

A continucación se muestra un ejemplo del uso de los principales algoritmos usados en este trabajo de tesis, así como una corrida detallada del AGD para una mejor comprensión del mismo.

4.3.1. Ejemplo de uso: algoritmo genético.

A continuación se presenta el ejemplo de un algoritmo genético sencillo, para un caso de optimización trivial. En este ejemplo, solo se realizan las operaciones de selección y crossover. Más detalles de estas operaciones pueden ser revisados en la sección 4.2**.**

Se requiere maximizar la función $f(x) = x^2$

Donde $64 > x > 0 \quad x \in \mathbb{Z}^+$

Primero se deben de codificar los valores, para este ejemplo se utilizará una codificación binaria de 6 bits.

Una vez elegida la codificación, se requiere generar individuos al azar, este conjunto de individuos se le llama población.

Se genera una población inicial de 6 individuos, estos deben ser creados de forma aleatoria, cada bit del individuo o cromosoma tomará un valor de 0 o 1.

Se generó la siguiente población inicial.

Individuo	Valor Bin	x	f(x)
1	010101	21	441
2	010011	19	361
3	011001	25	625
4	110001	49	2401
5	010100	20	400
6	011001	25	625

$f(x)$ promedio es de 808.83

Se realizará un torneo binario en el cual a cada individuo se le asignará una pareja de manera aleatoria, los dos individuos compiten y el que tenga el valor más alto, se mantendrá y el del valor bajo se eliminará.

La asignación de parejas se realizó de la siguiente forma:

Individuo	Pareja
1	1
2	2
3	1
4	3
5	3
6	2

La columna Pareja indica a los participantes del torneo, en donde solo se selecciona un individuo de los dos posibles candidatos.

Se realizan 3 torneos que a continuación se detallarán:

Torneo 1:

El individuo 1 y el individuo 3 competirán en un torneo, donde ganará quien tenga el valor fitness mas alto. En este caso el ganador fue el individuo 3, por lo tanto el individuo 3 con el valor binario "011001"se selecciona dos veces.

Torneo 2:

El individuo 2 y el individuo 6 competirán en un torneo, donde ganará quien tenga el valor fitness mas alto. En este caso el ganador fue el individuo 6, por lo tanto el individuo 6 con el valor binario "011001"se selecciona dos veces.

Torneo 3:

El individuo 4 y el individuo 5 competirán en un torneo, donde ganará quien tenga el valor fitness mas alto. En este caso el ganador fue el individuo 4, por lo tanto el individuo 4 con el valor binario "110001"se selecciona dos veces.

La población después de la realización de los torneos se muestra a continuación:

Individuo	Valor Bin
1	011001
2	011001
3	011001
4	011001
5	110001
6	110001

Ahora se realizará la operación de reproducción o crossover.

Se seleccionan al azar tres parejas y se les aplica el operador de reproducción.

Para este ejemplo se usará la reproducción de un punto. Se selecciona al azar un número c desde 1 hasta $n-1$, donde n es la longitud de los individuos, en este número c, se hará un corte de los individuos y se seleccionarán los c primeros bits de la madre y los bits restantes del padre, con esto se genera el primer individuo. Después se seleccionan los c primeros bits del padre y los bits restantes de la madre, para generar un segundo individuo.

Se usarán las siguientes parejas, seleccionadas al azar

Individuo	Pareja
1	2
2	3
3	1
4	2
5	1
6	3

Consideremos los siguientes números generados al azar para los valores de c:
$c1 = 1, c2 = 3, c3 = 4$
Para c1 con la pareja 1 se cruzarán los individuos 3 y 5 cuyos valores son:
3) 011001
5) 010100

3|5 0|10100
5|3 0|11001

Los hijos resultantes son los siguientes:

010100
011001

Para c2 con la pareja 2 se cruzarán los individuos 1 y 4 cuyos valores son:
1) 010101
4) 110001

1|4 01|0001
4|1 11|0101

Los hijos resultantes son los siguientes:

010001
110101

Para c3 con la pareja 3 se cruzarán los individuos 2 y 6 cuyos valores son:
2) 010011
6) 011001

2|6 0100|01
6|2 0110|11

Los hijos resultantes son los siguientes:

010001
011011

La nueva población es:

No.	Valor Bin	x	f(x)
1	010100	20	400
2	011001	25	625
3	010001	17	289
4	110101	53	2809
5	010001	17	289
6	011011	27	729

f(x) promedio es de 856.833

El incremento del f(x) promedio, indica que la nueva generación es mejor que la generación anterior. Si se sigue el algoritmo de forma iterativa, cada vez el resultado de la función se aproximará al máximo óptimo.

4.3.2. Corrida detallada del AGD:

- Paso 1: Se genera la población incial.

A continuación se muestra la población inicial, esta población se generó de forma aleatoria para un conjunto de datos con diecinueve características.

Individuos
0000110101011110011
0001110001110000101
0110110111100111100
0000100010011111110
0010101011011001010
1000100110100000010

- Paso2: Se calcula el Fitness de la población con la función Fitness mostrada en la sección 5.1.

Al ser la primera generación, se calcula el Fitness a todos los elementos.

Individuos	Fitness
0000110101011110011	56.711052631578944
0001110001110000101	57.02684210526316
0110110111100111100	58.33526315789474
0000100010011111110	58.80894736842105
0010101011011001010	63.65894736842105
1000100110100000010	71.89263157894737

- Paso 3: Se realiza la operación se selección.

Individuos	Fitness
1000100110100000010	71.89263157894737
0010101011011001010	63.65894736842105
0000100010011111110	58.80894736842105
0110110111100111100	58.33526315789474
0001110001110000101	57.02684210526316
0000110101011110011	56.711052631578944

- Paso 4: Eliminar las 4 peores soluciones

Individuo	Fitness
1000100110100000010	71.89263157894737
0010101011011001010	63.65894736842105

- Paso 5: Generar descendencia de los padres 1 y 2.

Se realiza la primera reproducción con los padres seleccionados, los individuos que no han sido evaluados se marcan con -1.

Individuos	Fitness
1000100110100000010	71.89263157894737
0010101011011001010	63.65894736842105
1000100111011001010	-1.0
0010101010100000010	-1.0

- Paso 6: Generar descendencia de los hijos 1 y 2.

Se realiza una segunda reproducción, en esta operción, se utilizan los individuos generados en la reproducción anterior como padres, para generar los nuevos individuos.

Individuos	Fitness
1000100110100000010	71.89263157894737
0010101011011001010	63.65894736842105
1000100111011001010	-1.0
0010101010100000010	-1.0
1000100110100000010	-1.0
0010101011011001010	-1.0

- Paso 7: Mutar toda la población excepto los padres.

Se muta la población, la operación de mutación solamente se realiza sobre los individuos que no han sido evaluados.

Individuos	Fitness	Genes mutados
1000100110100000010	71.89263157894737	N/A
0010101011011001010	63.65894736842105	N/A
1000100111011001000	-1.0	17
0010101010100000010	-1.0	N/A
0000100110100000010	-1.0	0
0000001111111001010	-1.0	2,4,7,10

- Paso 8: Si se cumplió la condición de paro, entregar el mejor resultado, de lo contrario ir al paso 2

Al no cumplirse la condicion de paro, se pasa al paso 2

- Paso 2: Evaluar los nuevos individuos:

Se calcula el Fitness de la nueva población, en este punto, se puede observar que una mejor solución ha surgido, la solución con 69.14 %, esta nueva solución remplazará a uno de los padres.

Individuos	Fitness
1000100110100000010	71.89263157894737
0010101011011001010	63.65894736842105
1000100111011001000	63.816842105263156
0010101010100000010	67.04263157894736
0000100110100000010	69.14052631578947
0000001111111001010	63.65894736842105

- Paso 3: Se realiza la operación se selección.

Al seleccionar la población, la solución con 69.14 %se mantiene dominante sobre las demás.

Se puede observar que existen dos elementos repetidos, debido a que se generaron 2 pares de hijos iguales, pero estas soluciones al no ser mejores que otras soluciones, desaparecerán en el paso siguiente.

Individuos	Fitness
1000100110100000010	71.89263157894737
0000100110100000010	69.14052631578947
0010101010100000010	67.04263157894736
1000100111011001000	63.816842105263156
0010101011011001010	63.65894736842105
0000001111111001010	63.65894736842105

- Paso 4: Eliminar las 4 peores soluciones

Individuos	Fitness
1000100110100000010	71.89263157894737
0000100110100000010	69.14052631578947

- Paso 5: Generar descendencia de los padres 1 y 2.

Se realiza la primera reproducción con los padres seleccionados, los individuos que no han sido evaluados se marcan con -1.

Individuos	Fitness
1000100110100000010	71.89263157894737
0000100110100000010	69.14052631578947
1000100110100000010	-1.0
0000100110100000010	-1.0

- Paso 6: Generar descendencia de los hijos 1 y 2.

Se realiza una segunda reproducción, en esta operación, se utilizan los individuos generados en la reproducción anterior como padres, para generar los nuevos individuos.

Individuos	Fitness
1000100110100000010	71.89263157894737
0000100110100000010	69.14052631578947
1000100110100000010	-1.0
0000100110100000010	-1.0
1000100110100000010	-1.0
0000100110100000010	-1.0

- Paso 7: Mutar toda la población excepto los padres.

Se muta la población, la operación de mutación solamente se realiza sobre los individuos que no han sido evaluados.

Individuos	Fitness	Genes mutados
1000100110100000010	71.89263157894737	N/A
0000100110100000010	69.14052631578947	N/A
0000100010100000010	-1.0	0,7
1000100111000010011	-1.0	0,9,10,14,18
1000101100110000110	-1.0	6,8,11,16
0000000110100000010	-1.0	4

- Paso 8: Si se cumplió la condición de paro, entregar el mejor resultado, de lo contrario ir al paso 2

Se supone que se cumplió la condicion de paro. El resultado es el siguiente:

Solución	Fitness
1000100110100000010	71.89263157894737

5. Desarrollo

A continuación se presenta el desarrollo de la metodología propuesta de forma detallada. Primero se muestra la función fitness que guia al algoritmo genetico, después se describe la experimentación realizada para, finalmente presentar en forma la metodología.

5.1. Función Fitness

El objetivo de la función Fitness es obtener un valor que sea más alto, si se tiene una mayor eficiencia de clasificación, pero también que esta función obtenga como resultado un valor mayor mientras menos características se le especifiquen. A continuación se muestra la función usada para lograr este objetivo.

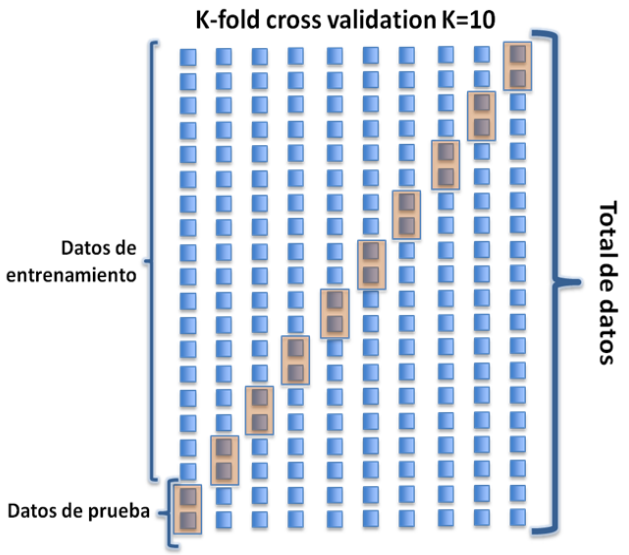

Figura 3: Método K-fold

$$Fitness(x, y, z) = y * ,97 + (100 - ((\frac{100}{z}) * x)) * ,03$$

Donde:
x: Son las características seleccionadas
y: Es la eficiencia de clasificación
z: Es el número de características disponibles

Para calcular la eficiencia de clasificación se utilizó un método de validación cruzada *k-fold cross validation* **[20]**, con un k=10, el valor entregado por este método se usa como parte del resultado final en un 97 %, el otro 3 % se evalúa con el número de características que ofrece la solución, se realizó de esta forma debido a que existen diversas combinaciones de características que ofrecen la misma eficiencia de clasificación y para poder decidir si una combinación de características es mejor que otra, se necesitaba otro parámetro de diferenciación.

El método de validación cruzada k-fold cross validation prueba la eficiencia de clasificación de cada conjunto, este método consiste en dividir el conjunto total de casos en K subconjuntos disjuntos. De estos K subconjuntos uno se reserva como datos de validación para probar el modelo y los restantes se utilizan como datos de entrenamiento. El proceso se repite K veces (con cada uno de los K usado sólo una vez como datos de validación). El método se presenta de forma gráfica en la Figura 3.

5.2. Variación de parámetros

Una de las decisiones importantes para implementar un algoritmo genético son los valores que se eligen para los diversos parámetros. Algunos de los parámetros son; el tamaño de la población, el rango de reproducción y el rango de mutación. Generalmente estos parámetros interactúan uno con otro de forma no lineal, por lo tanto no pueden ser optimizados uno a la vez. Existe una gran discusión sobre la selección de parámetros. El consenso científico se centra en utilizar aquellos parámetros que han reportado buen desempeño en la literatura vigente. [9].

De Jong en 1975 realizó un estudio sobre como diferentes parámetros afectan a los algoritmos genéticos, usando un pequeño conjunto de funciones de prueba. Los experimentos de De Jong, indicaron que el mejor tamaño de población fue de 50-100 individuos, el mejor rango para realizar la operación crossover fue de ~0.6 para cada pareja de padres y el mejor rango de mutación fue de 0.001 por cada bit. Estos parámetros, son altamente usados en la comunidad de los Algoritmos Genéticos, aunque no se sabe como estos parámetros funcionarán con otros problemas fuera de el conjunto de prueba usado por De Jong [9].

Para el Algoritmo Genético Simple se varió el porcentaje de mutación para mejorar su eficiencia. Para el tamaño de la población se utilizó un tamaño de 6, esto fue para poder comparar el porcentaje de exploración del AGS con el AGD.

Dado que no existe un antecedente de los parámetros a usar en el Algoritmo Genético Dominante, se realizó una variación del porcentaje de mutación y se utilizó una población de 6 elementos, la población se seleccionó de esta forma para reducir en la medida de lo posible el espacio de búsqueda explorado por el algoritmo.

Para esta experimentación se emplearon 3 conjuntos de datos medianos: Breast Cancer, Heart, Hepatitis. 2 conjuntos de datos grandes Lung Cancer y Arritmia. Todos los conjuntos de datos fueron obtenidos del UCI machine learning repository [21]. En la Tabla 1 se exponen las características de cada conjunto de datos.

Tabla 1: Características de los conjuntos de datos.

Conjunto de datos	**No. de características**	**Espacio de soluciones**	**Instancias**
Breast Cancer	9	512	683
Heart	13	8,192	270
Hepatitis	19	524,288	80
Lung Cancer	56	1.44E+17	27
Arritmia	279	9.71E+83	420

La metodología utilizada en la variación de parámetros se explicará a continuación.

Se realizaron 20 corridas del algoritmo, de estas 20 corridas se calculó la varianza del Fitness, el promedio Fitness y el promedio del número de características seleccionadas por el algoritmo. Como criterio principal de evaluación de la calidad de la mutación usada, se tomó el valor de promedio Fitness.

Para cada prueba se utilizó aproximadamente un 5 % de exploración del espacio de soluciones en conjuntos de datos medianos y un total de 600 individuos evaluados para conjuntos de datos grandes.

Para explorar la influencia de la mutación en los algoritmos genéticos, se propuso la realización de los siguientes experimentos.

- Experimento 1: Se varió el porcentaje de mutación desde 100 % hasta 0 % tomando solo valores enteros en intervalos de 20 unidades de separación los valores de mutación fueron 100 %, 80 %, 60 %, 40 %, 20 % y 0 %.
- Experimento 2: Se realizó una exploración más fina, con tres valores cercanos a los mejores valores encontrados en el experimento anterior.
- Experimento 3: Se Exploraron valores de mutación dos unidades antes y dos unidades después de los mejores valores encontrados en el experimento dos.

Los experimentos se realizaron con los cinco conjuntos de datos propuestos en la Tabla 1.

En las tablas de resultados se incluye la varianza. Al tener una varianza baja, se asegura que los resultados son uniformes en todas las corridas del algoritmo.

5.2.1. Variación de parámetros AGD

Experimento 1:

En las Tablas 2, 3, 4, 5 y 6 se muestran los resultados del primer experimento en los cinco conjuntos de datos descritos en la Tabla 1.

Tabla 2: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	1.54	0.34	0.55	0.78	0.56	2.09
Promedio Fitness	51.19	53.87	54.88	55.53	57.69	51.91
Promedio No. Caract.	140.95	135.75	136.45	129.65	127.95	137.00

Tabla 3: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	0.11	0.09	0.05	0.05	0.09	0.11
Promedio Fitness	94.99	94.99	95.14	95.14	95.10	95.02
Promedio No. Caract.	4.55	4.85	4.80	4.70	5.10	5.20

Tabla 4: Resultados de mutación del conjunto de datos Heart.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	6.04	1.68	2.61	1.04	2.39	13.64
Promedio Fitness	69.41	76.08	77.34	79.72	80.28	71.26
Promedio No. Caract.	6.65	5.10	5.30	5.30	3.90	4.95

Tabla 5: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	9.49	0.89	0.41	8.18	14.97	22.98
Promedio Fitness	67.76	76.20	78.22	80.99	83.13	69.08
Promedio No. Caract.	8.55	7.60	7.10	5.50	5.20	7.85

Tabla 6: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	12.41	2.18	1.24	3.16	7.40	18.69
Promedio Fitness	73.70	77.79	78.75	79.87	81.40	74.26
Promedio No. Caract.	27.35	27.15	23.65	20.95	16.80	27.80

Experimento 2:

Se puede apreciar que la mutación con mayor promedio Fitness se encontraba en general entre los valores 40 % y 20 %, los valores utilizados para el experimento 2 fueron 35 %, 30 % y 25 %.

Los resultados se muestran en las Tablas 7, 8, 9, 10 y 11. En estas se considera el mayor valor promedio Fitness obtenido en el experimento anterior para efectos de comparación.

Tabla 7: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	0.59	0.53	0.70	0.56
Promedio Fitness	55.94	56.45	57.00	57.69
Promedio No. Caract.	132.15	130.45	128.35	127.95

Tabla 8: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	40 %	35 %	30 %	25 %
Varianza Fitness	0.05	0.02	0.04	0.03
Promedio Fitness	95.14	95.47	95.47	95.23
Promedio No. Caract.	4.70	5.25	5.40	4.55

Tabla 9: Resultados de mutación del conjunto de datos Heart.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	0.06	0.04	0.06	2.39
Promedio Fitness	80.59	80.62	80.58	80.28
Promedio No. Caract.	4.25	4.10	4.30	3.90

Tabla 10: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	17.04	18.72	17.14	14.97
Promedio Fitness	83.39	84.71	85.63	83.13
Promedio No. Caract.	5.65	4.10	3.75	5.20

Tabla 11: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	2.89	2.64	3.22	7.40
Promedio Fitness	79.86	79.77	80.73	81.40
Promedio No. Caract.	11.09	10.25	13.53	16.80

Experimento 3:

Después del experimento anterior se seleccionaron los valores que reportaran el mayor promedio Fitness y se propusieron valores 2 anteriores y 2 posteriores, en las tablas 12, 13, 14, 15 y 16 se muestran los resultados.

Tabla 12: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	22 %	20 %	18 %
Varianza Fitness	0.36	0.56	0.81
Promedio Fitness	57.35	57.69	57.81
Promedio No. Caract.	127.1	127.95	129.65

Tabla 13: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	32 %	30 %	28 %
Varianza Fitness	0.03	0.04	0.02
Promedio Fitness	95.44	95.47	95.50
Promedio No. Caract.	5.35	5.40	5.15

Tabla 14: Resultados de mutación del conjunto de datos Heart.

Porcentaje	32 %	30 %	28 %
Varianza Fitness	0.04	0.04	0.04
Promedio Fitness	80.62	80.62	80.62
Promedio No. Caract.	4.1	4.10	4.1

Tabla 15: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	27 %	25 %	23 %
Varianza Fitness	15.98	17.14	15.45
Promedio Fitness	82.93	85.63	83.08
Promedio No. Caract.	5.50	3.75	5.15

Tabla 16: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	27 %	25 %	23 %
Varianza Fitness	4.21	3.22	3.94
Promedio Fitness	80.48	80.73	80.19
Promedio No. Caract.	19.50	13.53	17.65

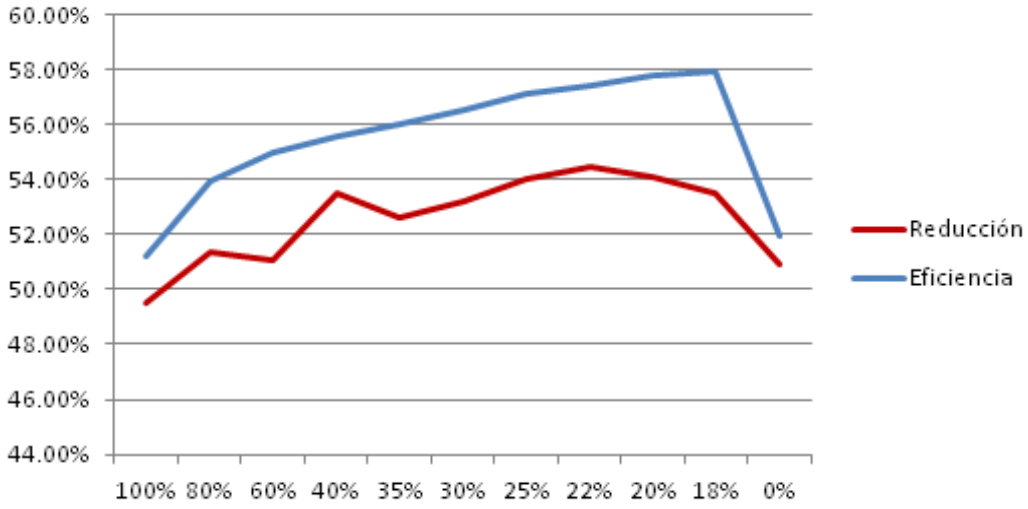

Figura 4: *Variación de porcentajes de mutación usando AGD en el conjunto de datos Arritmia.*

En las Figuras 4 ,5 ,6 ,7 y 8 se observan las gráficas que muestran la eficiencia de clasificación y la reducción de características variando el porcentaje de mutación en cada conjunto de datos. Para estas gráficas se consideraron todos los valores de mutación utilizados en la experimentación.

En la Tabla 17 se muestran los mejores resultados obtenidos en la experimentación variando el porcentaje de mutación.

Tabla 17: Resumen de la experimentación con mutación.

Dataset	Porcentaje de mutación	Eficiencia de clasificación	Reducción de características
Arritmia	18 %	57.95 %	53.53 %
Breast Cancer	30 %	97.00 %	40 %
Heart	25 %	81.1 %	66.9 %
Hepatitis	25 %	85.80 %	80.26 %
Lung Cancer	20 %	81.75 %	70 %

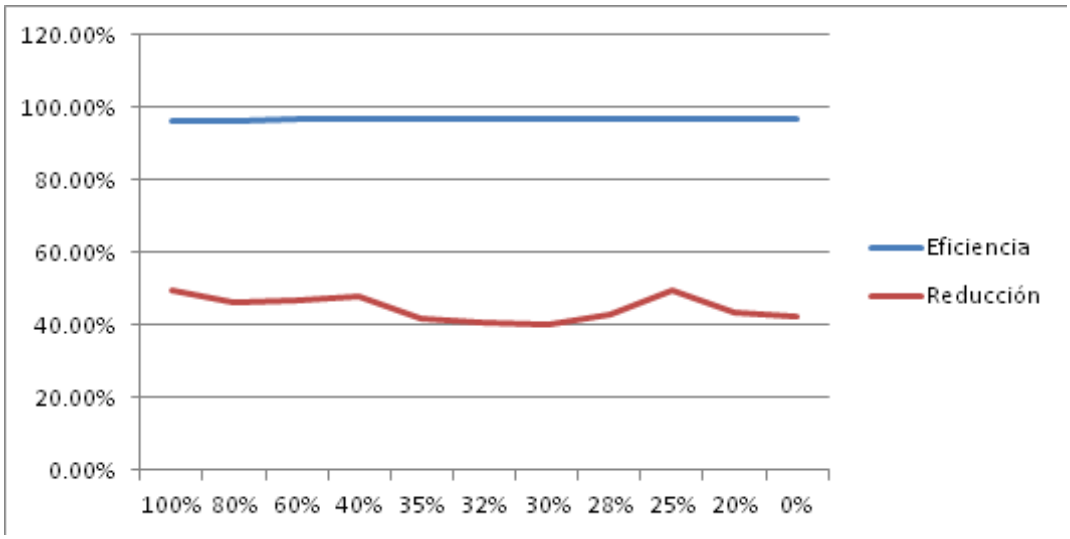

Figura 5: *Variación de porcentajes de mutación usando AGD en el conjunto de datos Breast Cancer.*

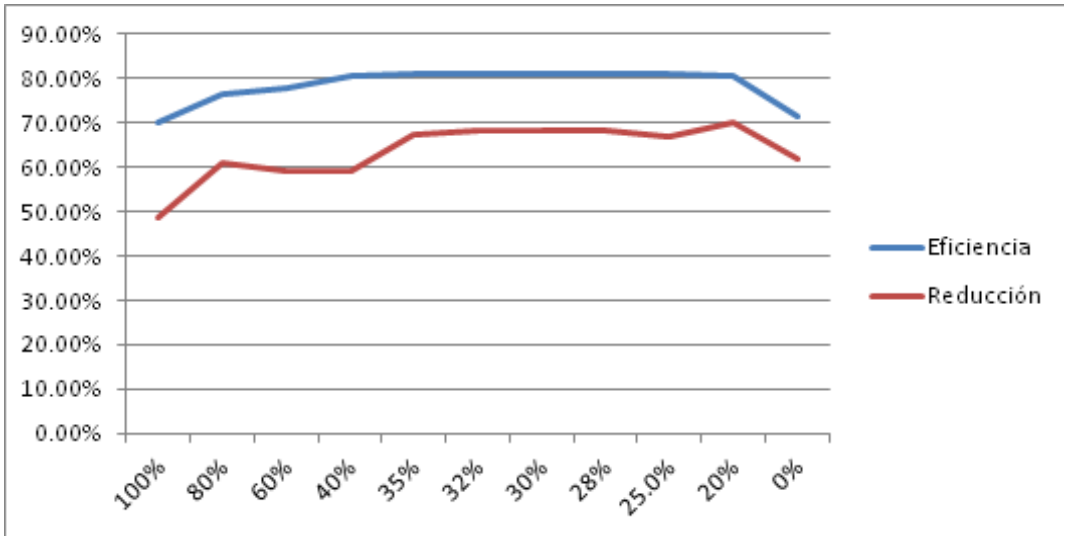

Figura 6: *Variación de porcentajes de mutación usando AGD en el conjunto de datos Heart.*

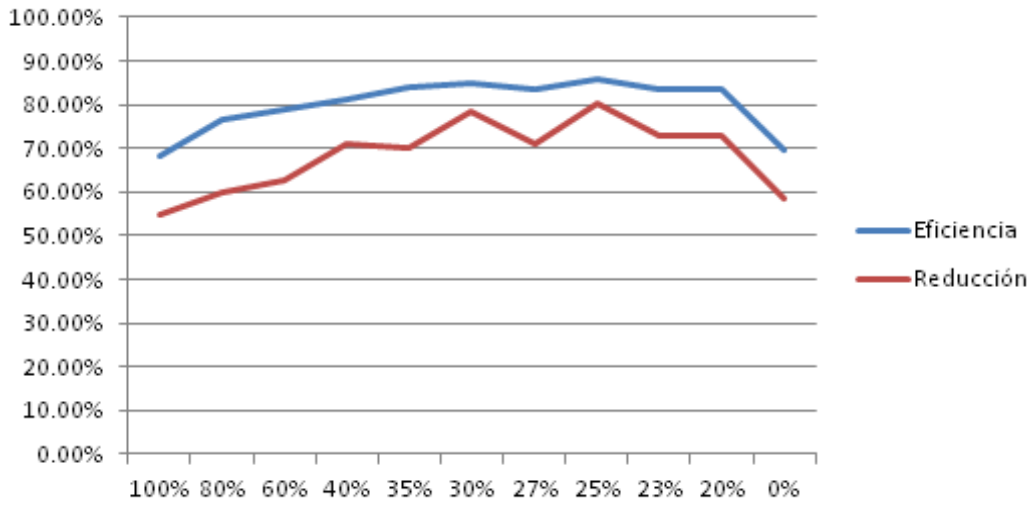

Figura 7: *Variación de porcentajes de mutación usando AGD en el conjunto de datos Hepatitis.*

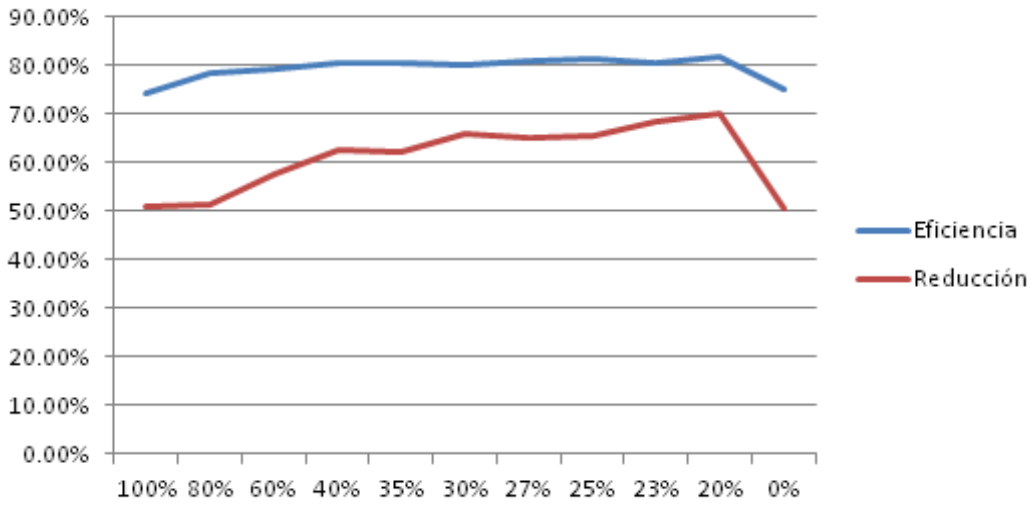

Figura 8: *Variación de porcentajes de mutación usando AGD en el conjunto de datos Lung Cancer.*

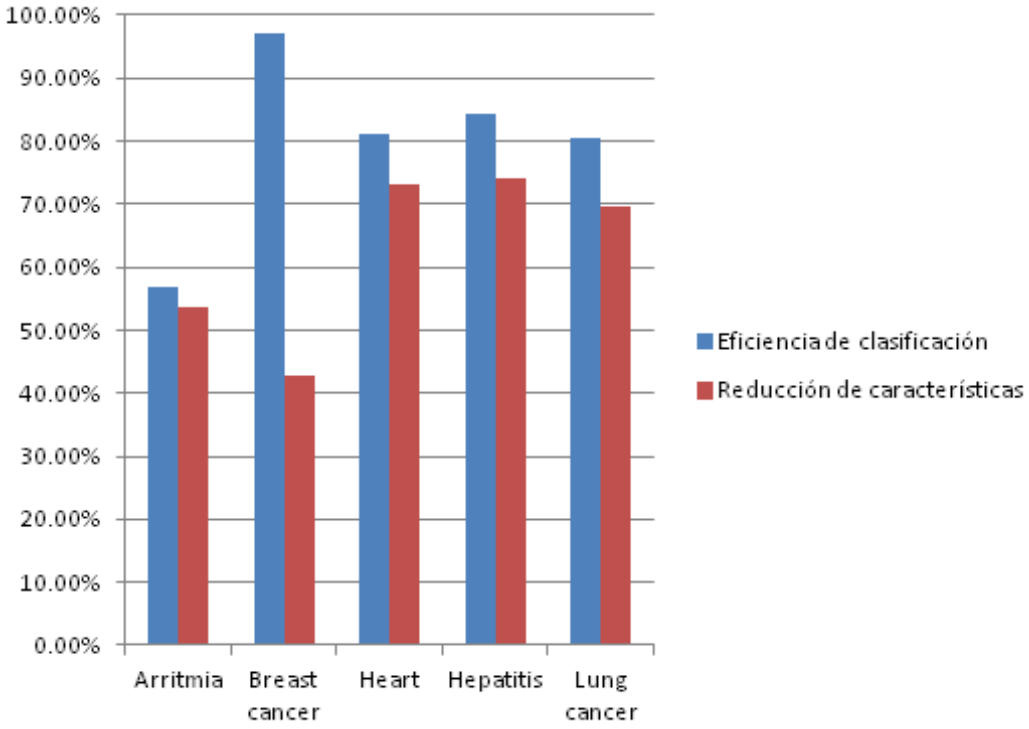

Figura 9: *Desempeño del AGD con 24 % de mutación.*

Considerando los mejores porcentajes de mutación el promedio es de 24 % de mutación.

En la Figura 9 se muestra el desempeño del AGD en los 5 conjuntos de datos utilizando una mutación del 24 %.

5.2.2. Variación de parámetros Algoritmos Genéticos Simples

La literatura recomienda un bajo porcentaje de mutación en los Algoritmos Genéticos, pero debido a que el problema de selección de características no es un problema numérico con valores cada vez más cercanos al óptimo, sino un problema de búsqueda espacial n dimensional, se pensó que el porcentaje de mutación, podría ser la clave, para elevar la eficiencia de los algoritmos genéticos en la búsqueda de una solución.

Variación de parámetros AGS1 (Algoritmo Genético Simple con método de selección *Binary Tournament* y método de crossover *Singlepoint Crossover*)

Experimento 1:

En las Tablas 18, 19, 20, 21 y 22 se muestran los resultados del primer experimento en los cinco conjuntos de datos descritos en la Tabla 1.

Tabla 18: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	1.75	3.04	2.63	1.88	2.04	2.07
Promedio Fitness	49.14	50.39	51.12	51.95	52.50	53.07
Promedio No. Caract.	142.30	138.90	134.45	137.70	136.25	137.20

Tabla 19: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	312.91	71.67	73.60	0.20	0.96	74.58
Promedio Fitness	66.86	92.58	92.72	94.75	94.77	93.06
Promedio No. Caract.	4.10	4.85	4.70	4.90	4.50	4.85

Tabla 20: Resultados de mutación del conjunto de datos Heart.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	24.90	16.76	16.53	16.00	9.33	9.94
Promedio Fitness	63.17	68.37	71.49	70.60	70.35	70.86
Promedio No. Caract.	6.80	5.70	5.45	5.10	6.15	5.85

Tabla 21: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	37.34	16.17	26.74	6.00	11.60	19.89
Promedio Fitness	68.76	68.22	67.12	67.75	69.21	67.92
Promedio No. Caract.	8.95	8.65	8.60	8.30	7.60	7.20

Tabla 22: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	20.20	20.15	19.19	11.35	11.99	3.88
Promedio Fitness	66.60	70.85	71.94	73.34	75.59	76.89
Promedio No. Caract.	28.60	28.00	25.90	27.85	27.50	26.75

Experimento 2:

Se puede apreciar que la mutación con mayor promedio Fitness se encontraba en general entre los valores 20 % y 0 %, así que para el experimento se propusieron los valores 15 %, 10 % y 5 %. Los resultados son mostrados en las Tablas 23, 24, 25, 26 y 27. En estas se considera el mayor valor promedio Fitness obtenido en el experimento anterior para efectos de comparación.

Tabla 23: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	15 %	10 %	5 %	0 %
Varianza Fitness	0.87	1.83	1.16	2.07
Promedio Fitness	52.98	53.74	55.21	53.07
Promedio No. Caract.	137.45	134.20	133.05	137.20

Tabla 24: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	20 %	15 %	10 %	5 %
Varianza Fitness	0.96	0.12	0.11	0.17
Promedio Fitness	94.77	95.00	95.12	95.27
Promedio No. Caract.	4.50	5.20	5.10	4.95

Tabla 25: Resultados de mutación del conjunto de datos Heart.

Porcentaje	60 %	15 %	10 %	5 %
Varianza Fitness	16.53	12.34	13.68	19.47
Promedio Fitness	71.49	72.20	73.06	74.48
Promedio No. Caract.	5.45	4.45	5.15	3.60

Tabla 26: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	20 %	15 %	10 %	5 %
Varianza Fitness	11.60	15.05	10.30	3.85
Promedio Fitness	69.21	71.00	72.26	76.86
Promedio No. Caract.	7.60	7.35	7.35	6.80

Tabla 27: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	15 %	10 %	5 %	0 %
Varianza Fitness	3.61	2.84	1.94	3.88
Promedio Fitness	76.88	76.87	78.77	76.89
Promedio No. Caract.	26.85	24.35	20.50	26.75

Experimento 3:

Después del experimento anterior se seleccionaron los valores que reportaran el mayor promedio Fitness y se propusieron valores 2 anteriores y 2 posteriores, en las tablas 28, 29, 30, 31 y 32 se muestran los resultados.

Tabla 28: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	0.45	1.16	1.71
Promedio Fitness	54.86	55.21	57.41
Promedio No. Caract.	133.60	133.05	130.70

Tabla 29: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	0.18	0.17	0.17
Promedio Fitness	94.92	95.27	95.10
Promedio No. Caract.	4.80	4.95	4.70

Tabla 30: Resultados de mutación del conjunto de datos Heart.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	12.24	19.47	17.79
Promedio Fitness	74.81	74.48	74.29
Promedio No. Caract.	4.30	3.60	2.55

Tabla 31: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	6.40	3.85	0.83
Promedio Fitness	75.06	76.86	78.19
Promedio No. Caract.	6.80	6.80	6.05

Tabla 32: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	1.33	1.94	4.55
Promedio Fitness	77.70	78.77	79.82
Promedio No. Caract.	25.20	20.50	16.45

En las Figuras 10, 11, 12, 13 y 14 se observan las gráficas que muestran la eficiencia de clasificación y la reducción de características variando el porcentaje de mutación en cada conjunto de datos. Para estas gráficas se consideraron todos los valores de mutación utilizados en la experimentación.

En la Tabla 33 se muestra los mejores resultados obtenidos en la experimentación variando el porcentaje de mutación.

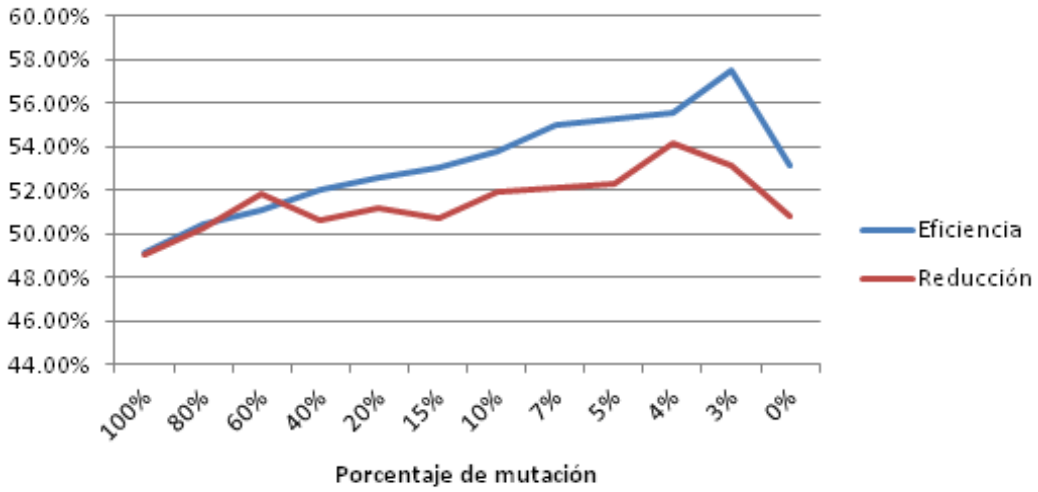

Figura 10: *Variación de porcentajes de mutación usando AGS1 en el conjunto de datos Arritmia.*

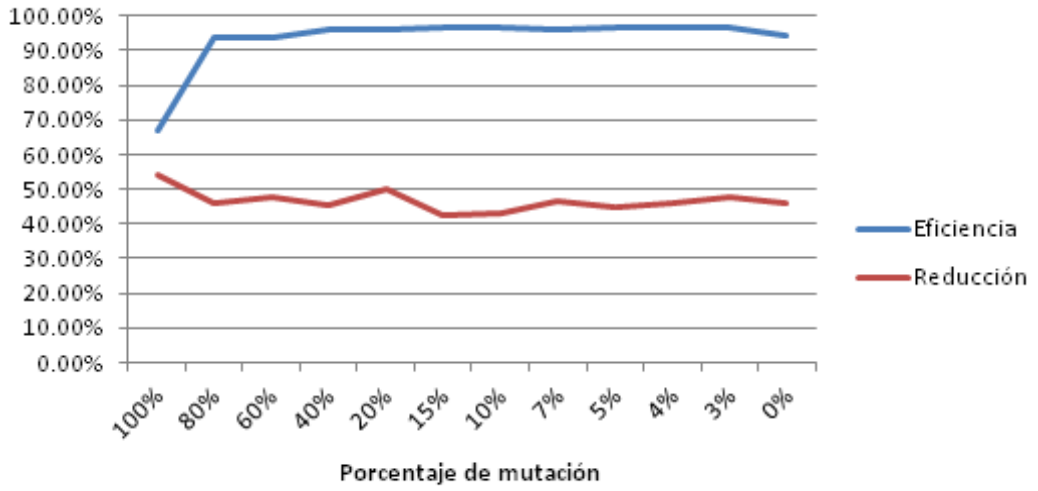

Figura 11: *Variación de porcentajes de mutación usando AGS1 en el conjunto de datos Breast Cancer.*

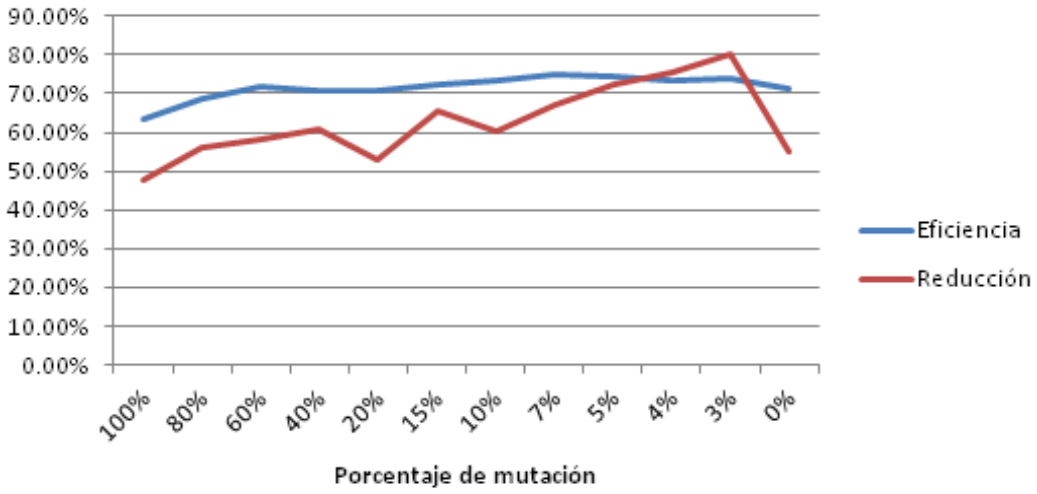

Figura 12: *Variación de porcentajes de mutación usando AGS1 en el conjunto de datos Heart.*

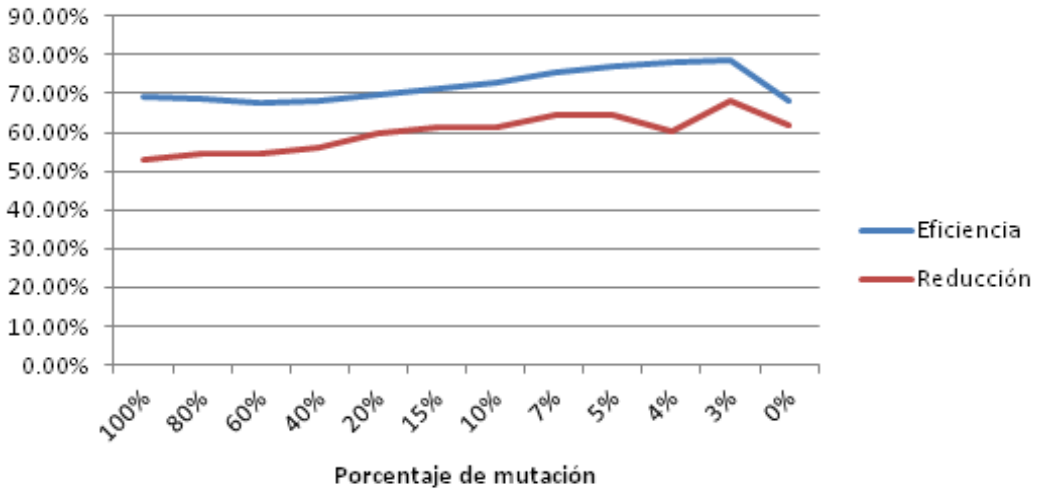

Figura 13: *Variación de porcentajes de mutación usando AGS1 en el conjunto de datos Hepatitis.*

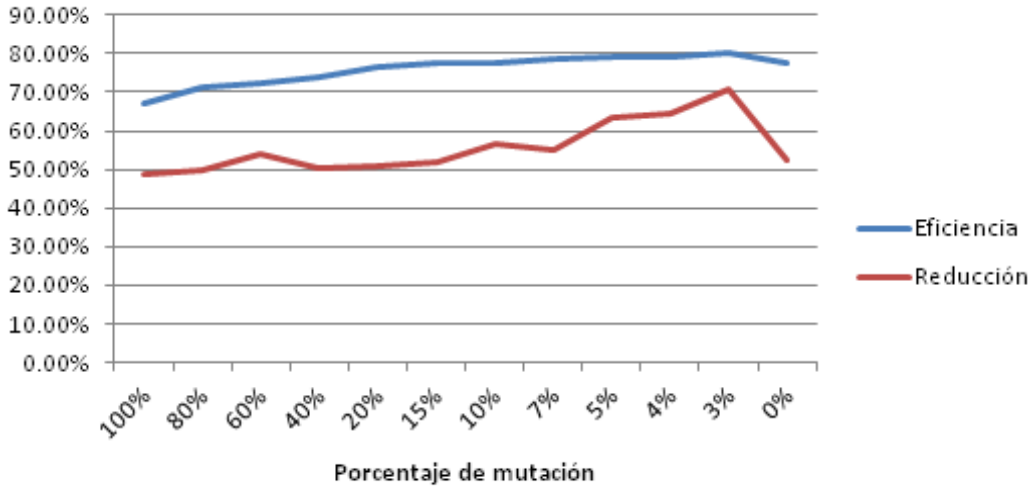

Figura 14: *Variación de porcentajes de mutación usando AGS1 en el conjunto de datos Lung Cancer.*

Tabla 33: Resumen de la experimentación con mutación.

Dataset	Porcentaje de mutación	Eficiencia de clasificación	Reducción de características
Arritmia	3 %	57.55 %	53.15 %
Breast Cancer	5 %	96.65 %	45.00 %
Heart	7 %	75.05 %	66.92 %
Hepatitis	3 %	78.50 %	68.16 %
Lung Cancer	3 %	80.10 %	70.63 %

Considerando los mejores porcentajes de mutación el promedio es de 4 % de mutación.

En la Figura 15 se muestra el desempeño del AGS1 en los 5 conjuntos de datos utilizando una mutación del 4 %.

Variación de parámetros AGS2 (Algoritmo Genético Simple con método de selección *Stochastic Remainder Sampling* y método de crossover *Singlepoint Crossover*)

Experimento 1:

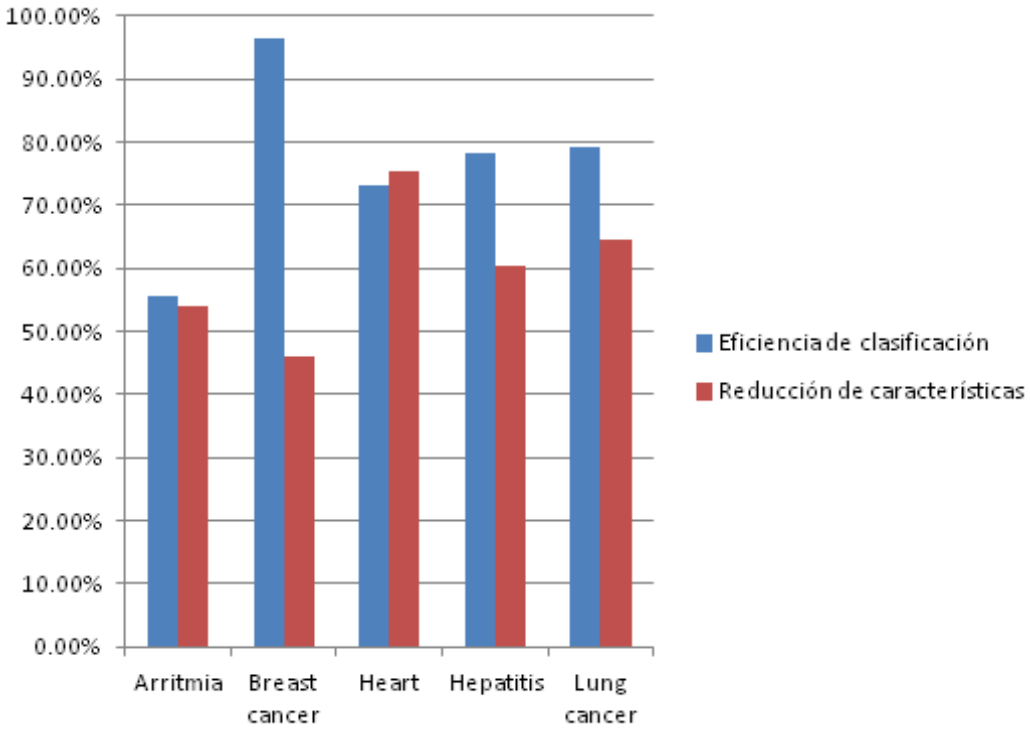

Figura 15: *Desempeño del AGS1 con 4 % de mutación.*

En las Tablas 34, 35, 36, 37 y 38 se muestran los resultados del primer experimento en los cinco conjuntos de datos descritos en la Tabla 1.

Tabla 34: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	6.02	2.25	1.66	2.67	3.18	9.06
Promedio Fitness	49.78	51.64	51.56	51.54	51.99	49.41
Promedio No. Caract.	137.05	139.80	137.95	140.50	139.30	139.65

Tabla 35: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	382.26	4.78	1.46	0.47	0.31	324.35
Promedio Fitness	70.56	94.07	94.51	94.74	94.70	82.42
Promedio No. Caract.	5.50	4.90	4.55	5.10	5.55	4.75

Tabla 36: Resultados de mutación del conjunto de datos Heart.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	26.62	23.07	15.25	16.27	11.02	14.26
Promedio Fitness	63.16	70.74	70.08	70.27	70.40	62.28
Promedio No. Caract.	6.45	5.75	6.50	5.45	5.95	6.90

Tabla 37: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	40.59	6.38	31.48	12.83	8.35	13.07
Promedio Fitness	62.18	66.02	66.95	68.60	68.48	61.08
Promedio No. Caract.	9.15	9.40	8.40	8.45	8.85	9.35

Tabla 38: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	51.67	19.42	12.07	14.63	11.78	39.12
Promedio Fitness	66.95	71.66	71.23	72.09	73.33	67.48
Promedio No. Caract.	28.50	27.45	27.30	26.75	27.10	28.50

Experimento 2:

Se puede apreciar que la mutación con mayor promedio Fitness se encontraba en general entre los valores 40 % y 20 %, así que los valores elegidos para la experimentación fueron 35 %, 30 % y 25 %. Se obtuvieron los resultados mostrados en las tablas 39, 40, 41, 42 y 43, en las tablas se considera el mayor valor promedio Fitness obtenido en el experimento anterior para efectos de comparación.

Tabla 39: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	1.35	1.79	2.91	3.18
Promedio Fitness	51.76	51.06	51.41	51.99
Promedio No. Caract.	138.00	139.55	138.70	139.30

Tabla 40: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	40 %	35 %	30 %	25 %
Varianza Fitness	0.47	0.17	138.84	0.12
Promedio Fitness	94.74	94.80	90.92	94.96
Promedio No. Caract.	5.10	4.90	5.05	5.15

Tabla 41: Resultados de mutación del conjunto de datos Heart.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	12.71	8.70	17.68	11.02
Promedio Fitness	70.70	70.66	71.59	70.40
Promedio No. Caract.	5.30	6.50	5.20	5.95

Tabla 42: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	40 %	35 %	30 %	25 %
Varianza Fitness	12.83	11.13	14.53	16.41
Promedio Fitness	68.60	66.47	68.51	67.52
Promedio No. Caract.	8.45	9.30	8.70	9.40

Tabla 43: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	15.93	18.93	8.17	11.78
Promedio Fitness	72.21	72.62	71.16	73.33
Promedio No. Caract.	27.15	25.90	26.00	27.10

Experimento 3:

Después del experimento anterior se seleccionaron los valores que reportaran el mayor promedio Fitness y se propusieron valores 2 anteriores y 2 posteriores, en las tablas 44, 45, 46, 47 y 48 se muestran los resultados.

Tabla 44: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	22 %	20 %	18 %
Varianza Fitness	3.51	3.18	1.77
Promedio Fitness	51.84	51.99	51.62
Promedio No. Caract.	139.10	139.30	137.20

Tabla 45: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	27 %	25 %	23 %
Varianza Fitness	0.26	0.12	74.65
Promedio Fitness	94.65	94.96	92.74
Promedio No. Caract.	4.40	5.15	5.60

Tabla 46: Resultados de mutación del conjunto de datos Heart.

Porcentaje	27 %	25 %	23 %
Varianza Fitness	17.14	17.68	20.30
Promedio Fitness	69.22	71.59	69.68
Promedio No. Caract.	5.80	5.20	5.90

Tabla 47: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	42 %	40 %	38 %
Varianza Fitness	8.79	12.83	13.73
Promedio Fitness	67.58	68.60	68.26
Promedio No. Caract.	9.65	8.45	8.40

Tabla 48: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	22 %	20 %	18 %
Varianza Fitness	12.16	11.78	14.73
Promedio Fitness	75.06	73.33	71.30
Promedio No. Caract.	26.50	27.10	27.00

En las Figuras 16, 17, 18, 19 y 20 se observan las gráficas que muestran la eficiencia de clasificación y la reducción de características variando el porcentaje de mutación en cada conjunto de datos. Para estas gráficas se consideraron todos los valores de mutación utilizados en la experimentación.

En la Tabla 49 se muestra los mejores resultados de la experimentación variando el porcentaje de mutación.

Tabla 49: Resumen de la experimentación con mutación.

Dataset	Porcentaje de mutación	Eficiencia de clasificación	Reducción de características
Arritmia	20 %	52.05 %	50.07 %
Breast Cancer	25 %	96.40 %	42.78 %
Heart	25 %	71.95 %	60 %
Hepatitis	40 %	69 %	55.53 %
Lung Cancer	22 %	75.75 %	52.68 %

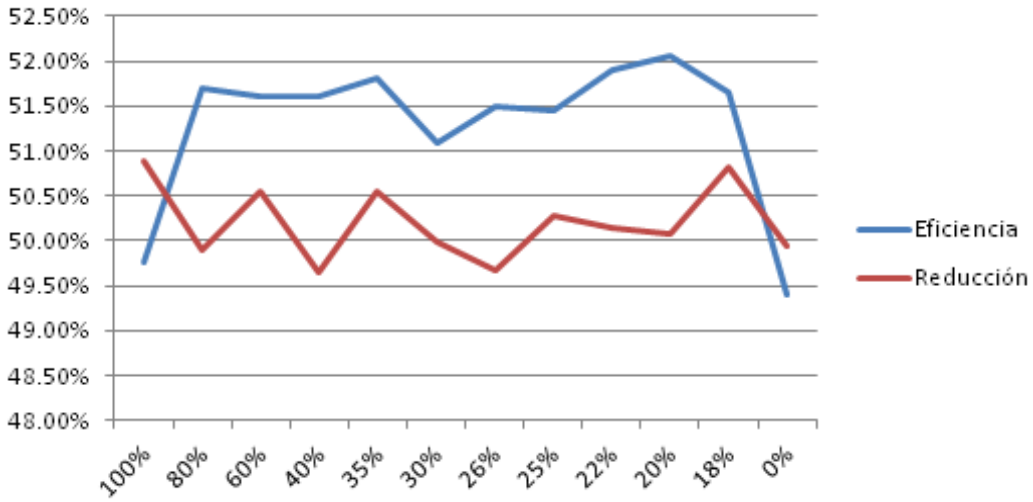

Figura 16: *Variación de porcentajes de mutación usando AGS2 en el conjunto de datos Arritmia.*

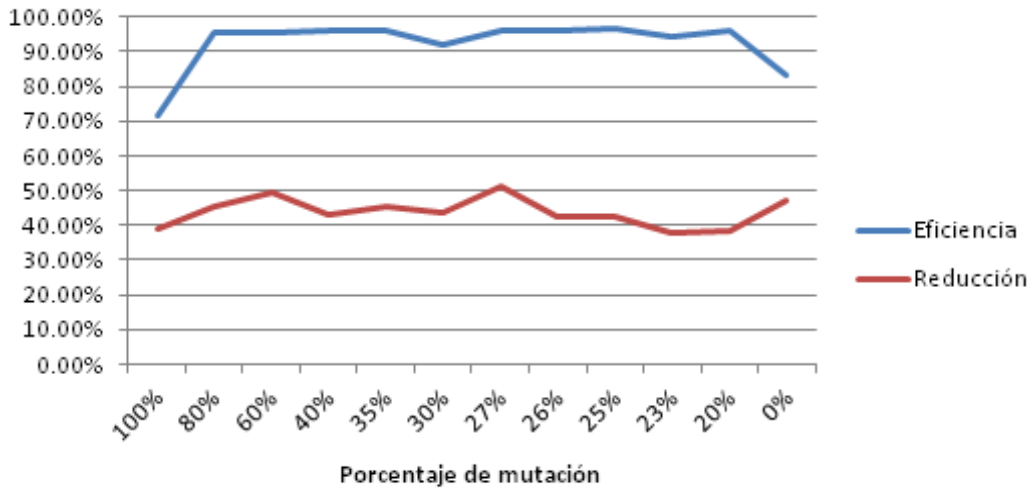

Figura 17: *Variación de porcentajes de mutación usando AGS2 en el conjunto de datos Breast Cancer.*

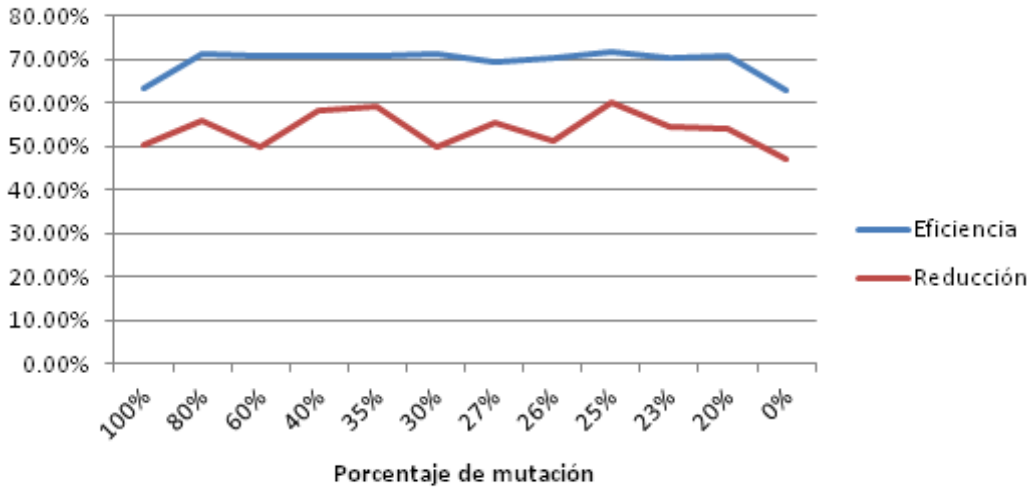

Figura 18: *Variación de porcentajes de mutación usando AGS2 en el conjunto de datos Heart.*

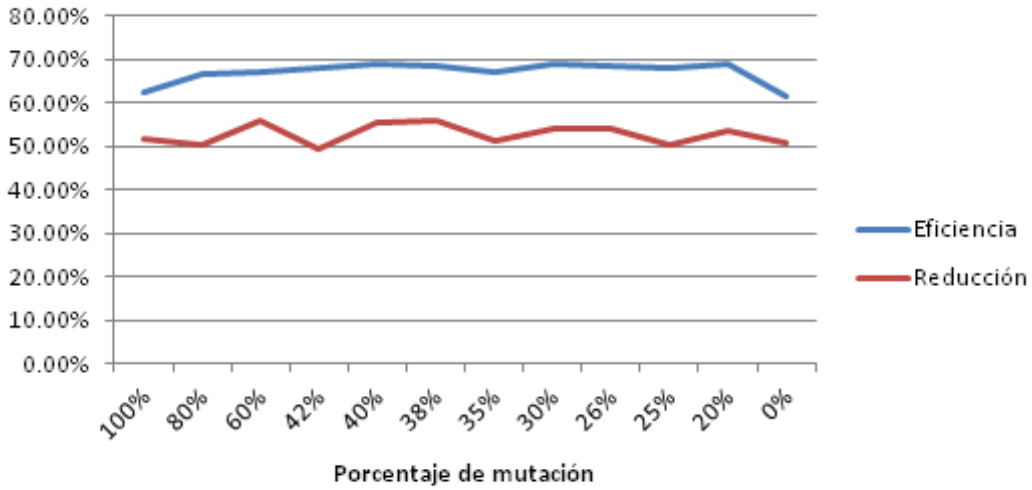

Figura 19: *Variación de porcentajes de mutación usando AGS2 en el conjunto de datos Hepatitis.*

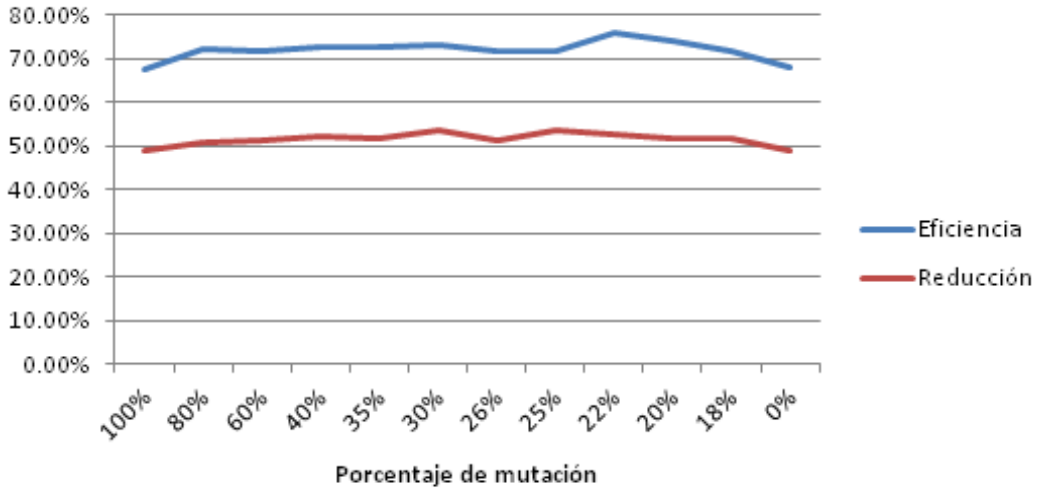

Figura 20: *Variación de porcentajes de mutación usando AGS2 en el conjunto de datos Lung Cancer.*

Considerando los mejores porcentajes de mutación el promedio es de 26 % de mutación.

En la Figura 21 se muestra el desempeño del AGS2 en los 5 conjuntos de datos utilizando una mutación del 26 %.

Variación de parámetros AGS3 (Algoritmo Genético Simple con método de selección *Binary Tournament* y método de crossover *Doublepoint Crossover*)

Experimento 1:

En las Tablas 50, 51, 52, 53 y 54 se muestran los resultados del primer experimento en los cinco conjuntos de datos descritos en la Tabla 1.

Tabla 50: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	3.10	1.30	1.89	1.55	1.60	1.36
Promedio Fitness	48.85	51.96	51.34	51.90	52.04	53.35
Promedio No. Caract.	137.95	141.20	135.90	143.05	138.75	138.25

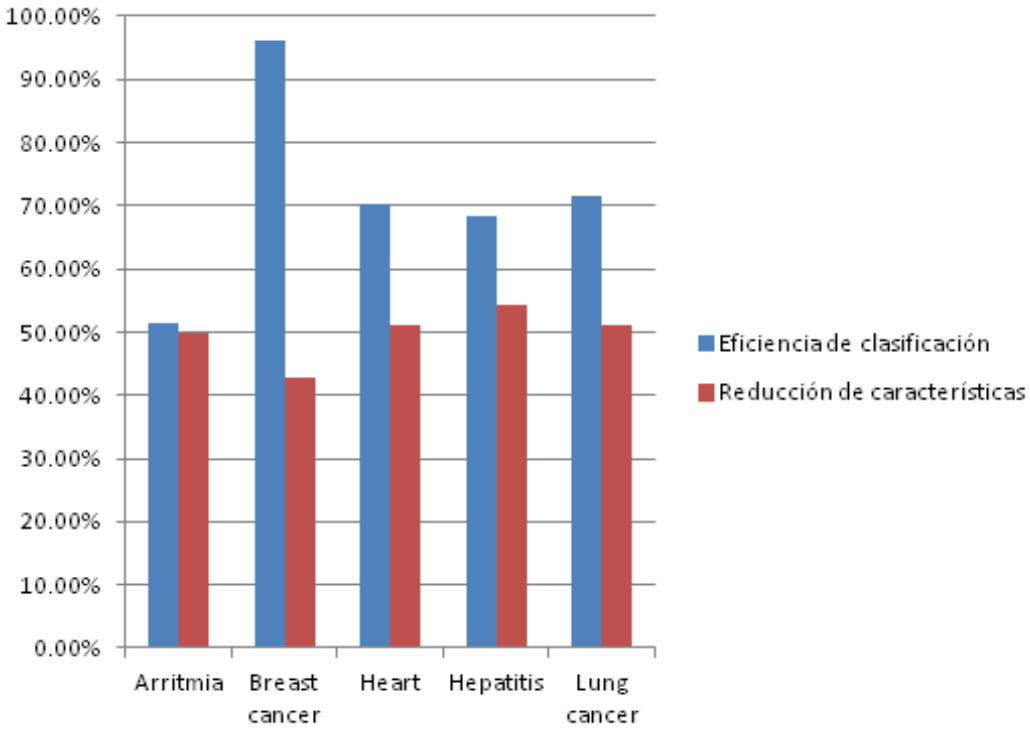

Figura 21: *Desempeño del AGS2 con 26 % de mutación.*

Tabla 51: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	367.87	73.61	73.70	0.29	0.18	0.15
Promedio Fitness	70.98	92.79	92.85	94.71	94.99	94.98
Promedio No. Caract.	4.10	5.60	4.60	4.85	5.40	5.10

Tabla 52: Resultados de mutación del conjunto de datos Heart.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	52.78	33.30	12.87	12.18	9.19	8.77
Promedio Fitness	63.85	70.07	69.34	68.68	70.37	69.39
Promedio No. Caract.	6.80	5.90	6.15	6.45	5.65	5.50

Tabla 53: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	34.80	18.13	10.83	15.72	9.30	15.77
Promedio Fitness	66.50	67.09	68.67	67.91	70.05	68.63
Promedio No. Caract.	9.15	9.10	8.60	8.80	6.90	6.70

Tabla 54: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	17.05	26.73	23.06	12.55	10.92	7.64
Promedio Fitness	69.18	69.90	72.03	73.62	74.87	75.74
Promedio No. Caract.	28.55	27.75	25.90	27.10	26.30	25.60

Experimento 2:

Se puede apreciar que la mutación con mayor promedio Fitness se encontraba en general entre los valores 20 % y 0 %, así que se utilizaron los valores 15 %, 10 % y 5 %.

Se obtuvieron los resultados mostrados en las Tablas 55, 56, 57, 58 y 59, en las tablas se considera el mayor valor promedio Fitness obtenido en el experimento anterior para efectos de comparación.

Tabla 55: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	15 %	10 %	5 %	0 %
Varianza Fitness	2.22	1.16	1.04	1.36
Promedio Fitness	53.01	53.66	55.48	53.35
Promedio No. Caract.	134.55	137.05	130.20	138.25

Tabla 56: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	20 %	15 %	10 %	5 %
Varianza Fitness	0.18	0.08	0.14	0.13
Promedio Fitness	94.99	95.12	95.14	95.13
Promedio No. Caract.	5.40	5.10	5.20	5.25

Tabla 57: Resultados de mutación del conjunto de datos Heart.

Porcentaje	20 %	15 %	10 %	5 %
Varianza Fitness	9.19	12.74	14.79	19.68
Promedio Fitness	70.37	72.64	73.68	74.91
Promedio No. Caract.	5.65	4.85	5.60	3.45

Tabla 58: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	20 %	15 %	10 %	5 %
Varianza Fitness	9.30	13.45	10.25	7.28
Promedio Fitness	70.05	71.10	73.75	76.36
Promedio No. Caract.	6.90	7.95	6.20	6.90

Tabla 59: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	15 %	10 %	5 %	0 %
Varianza Fitness	9.59	2.70	0.03	7.64
Promedio Fitness	75.77	77.39	78.56	75.74
Promedio No. Caract.	26.75	25.45	20.00	25.60

Experimento 3:

Después del experimento anterior se seleccionaron los valores que reportaran el mayor promedio Fitness y se propusieron valores 2 anteriores y 2 posteriores, en las tablas 60, 61, 62, 63 y 64 se muestran los resultados.

Tabla 60: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	1.01	1.04	1.69
Promedio Fitness	54.53	55.48	56.77
Promedio No. Caract.	136.90	130.20	127.05

Tabla 61: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	12 %	10 %	8 %
Varianza Fitness	0.14	0.14	0.12
Promedio Fitness	95.20	95.14	95.18
Promedio No. Caract.	5.50	5.20	5.25

Tabla 62: Resultados de mutación del conjunto de datos Heart.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	16.92	19.68	20.03
Promedio Fitness	75.69	74.91	74.52
Promedio No. Caract.	4.45	3.45	2.40

Tabla 63: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	7.19	7.28	0.68
Promedio Fitness	75.53	76.36	78.26
Promedio No. Caract.	7.25	6.90	6.25

Tabla 64: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	7 %	5 %	3 %
Varianza Fitness	0.84	0.03	0.67
Promedio Fitness	78.09	78.56	78.95
Promedio No. Caract.	23.40	20.00	16.30

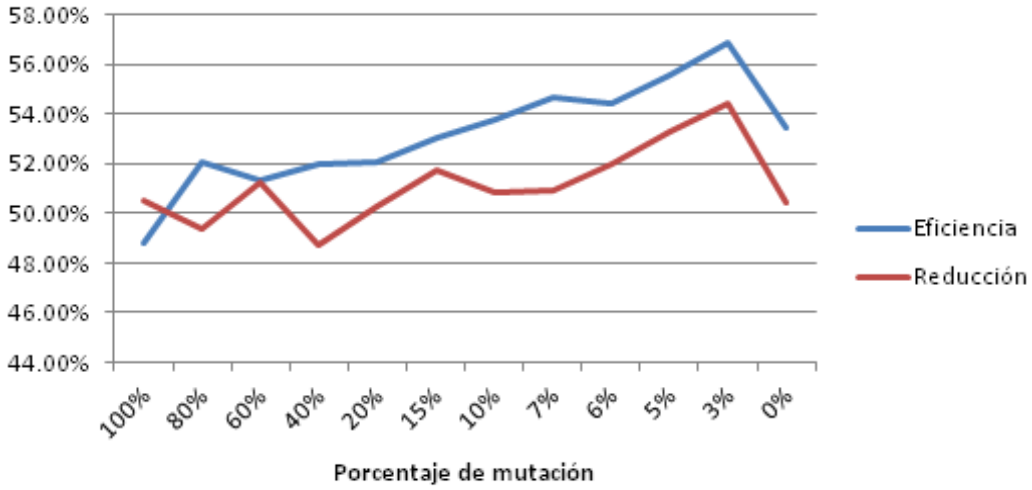

Figura 22: *Variación de porcentajes de mutación usando AGS3 en el conjunto de datos Arritmia.*

En las Figuras 22, 23, 24, 25 y 26, se observan las gráficas que muestran la eficiencia de clasificación y la reducción de características variando el porcentaje de mutación en cada conjunto de datos. Para estas gráficas se consideraron todos los valores de mutación utilizados en la experimentación.

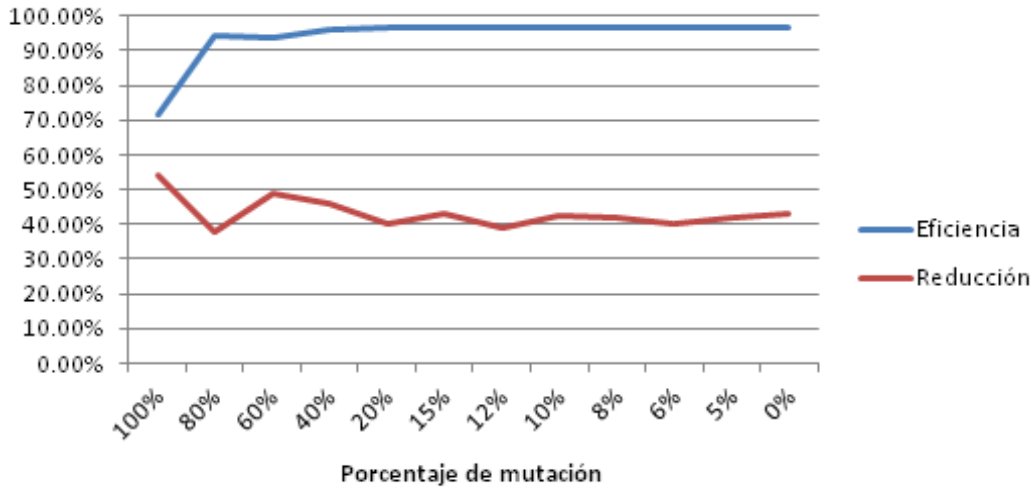

Figura 23: *Variación de porcentajes de mutación usando AGS3 en el conjunto de datos Breast Cancer.*

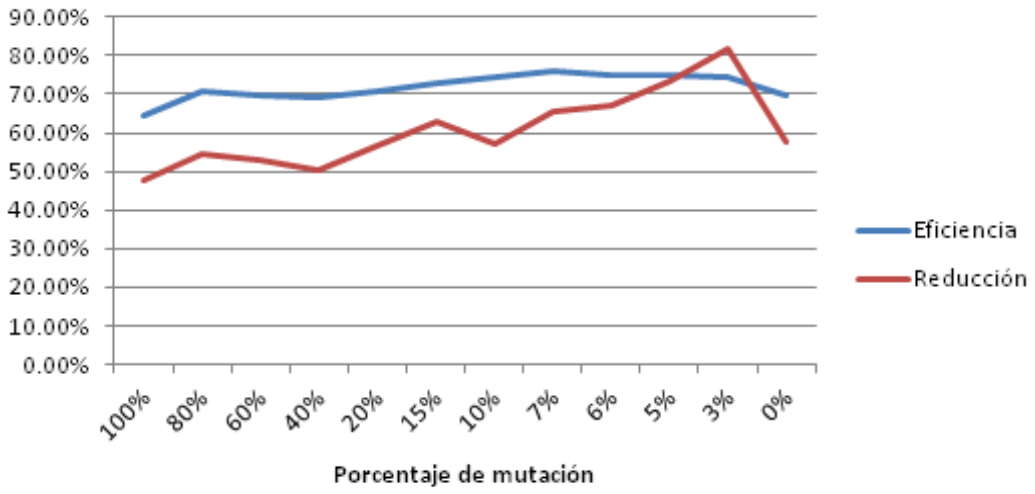

Figura 24: *Variación de porcentajes de mutación usando AGS3 en el conjunto de datos Heart.*

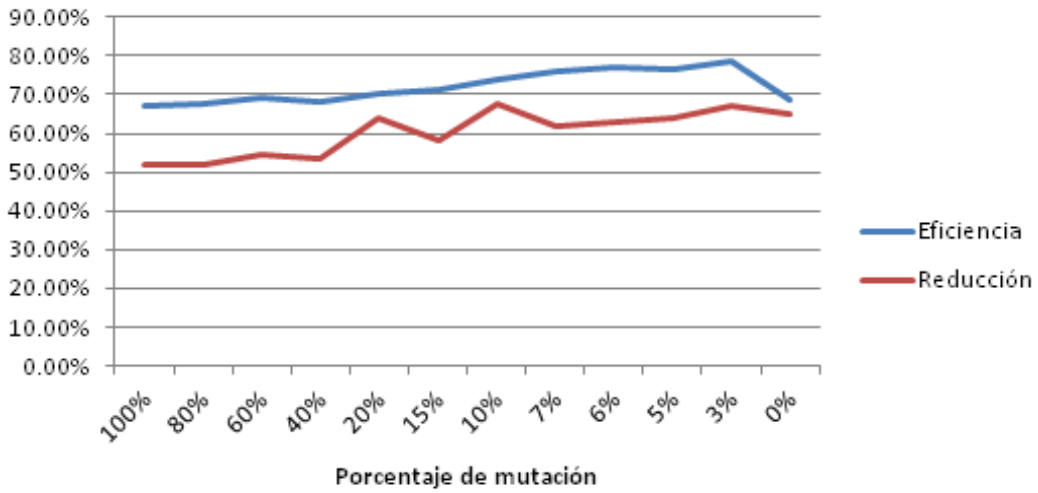

Figura 25: *Variación de porcentajes de mutación usando AGS3 en el conjunto de datos Hepatitis.*

Figura 26: *Variación de porcentajes de mutación usando AGS3 en el conjunto de datos Lung Cancer.*

En la Tabla 65 se muestra los mejores resultados de la experimentación variando el porcentaje de mutación.

Tabla 65: Resumen de la experimentación con mutación.

Dataset	Porcentaje de mutación	Eficiencia de clasificación	Reducción de características
Arritmia	3.00 %	56.85 %	54.46 %
Breast Cancer	12.00 %	96.75 %	38.89 %
Heart	7.00 %	76.00 %	65.77 %
Hepatitis	3.00 %	78.60 %	67.11 %
Lung Cancer	3.00 %	78.10 %	54.55 %

Considerando los mejores porcentajes de mutación el promedio es de 6 % de mutación.

En la Figura 27 se muestra el desempeño del AGS3 en los 5 conjuntos de datos utilizando una mutación del 6 %.

Variación de parámetros AGS4 (Algoritmo Genético Simple con método de selección *Stochastic Remainder Sampling* y método de crossover *Doublepoint Crossover*)

Experimento 1:

En las tablas 66, 67, 68, 69 y 70 se muestran los resultados del primer experimento en los cinco conjuntos de datos descritos en la Tabla 1.

Tabla 66: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	5.19	2.99	1.51	1.96	2.15	5.00
Promedio Fitness	48.49	51.36	51.23	51.65	51.12	50.48
Promedio No. Caract.	139.95	138.55	137.75	138.50	138.40	139.25

Tabla 67: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	377.20	144.36	0.74	0.63	0.34	249.12
Promedio Fitness	72.26	90.07	94.24	94.67	94.60	86.38
Promedio No. Caract.	5.50	4.50	5.30	5.00	4.90	5.15

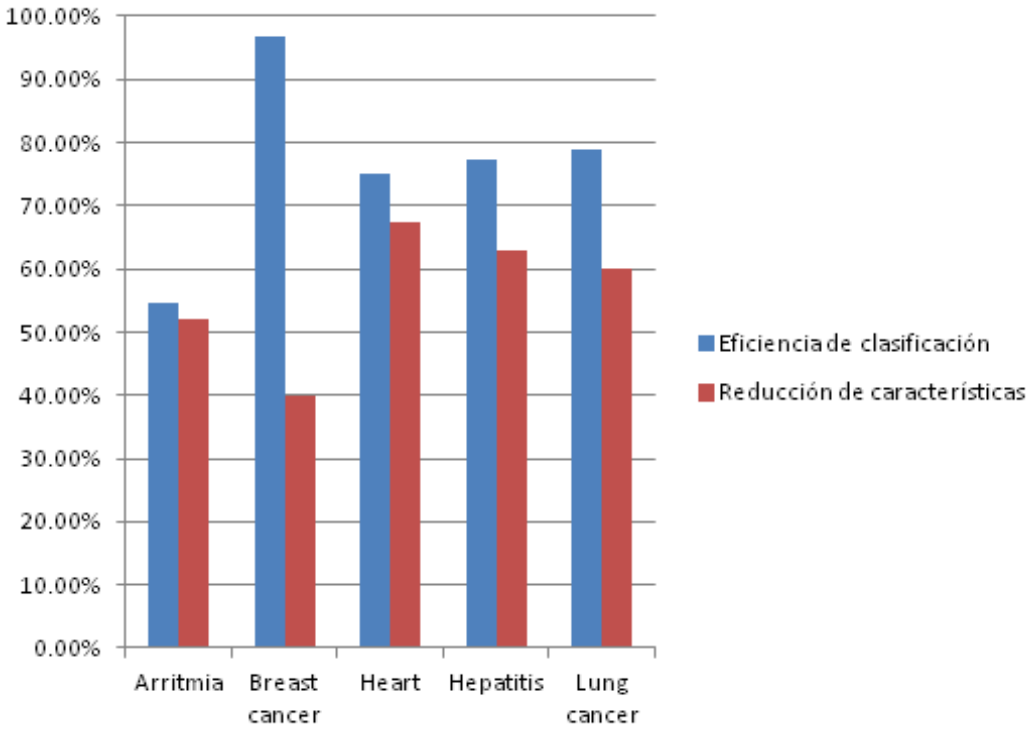

Figura 27: *Desempeño del AGS3 con 6 % de mutación.*

Tabla 68: Resultados de mutación del conjunto de datos Heart.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	17.17	18.57	21.85	11.09	16.88	40.30
Promedio Fitness	62.37	69.34	70.01	69.34	69.56	64.64
Promedio No. Caract.	7.55	6.15	5.55	5.50	5.80	6.30

Tabla 69: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	12.87	13.03	11.85	13.30	3.78	39.85
Promedio Fitness	59.20	66.04	67.06	67.97	68.79	62.11
Promedio No. Caract.	9.90	8.35	9.25	9.05	8.45	9.30

Tabla 70: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	100 %	80 %	60 %	40 %	20 %	0 %
Varianza Fitness	29.39	24.09	9.74	21.94	11.86	32.32
Promedio Fitness	64.65	71.94	71.57	73.20	74.41	66.60
Promedio No. Caract.	28.80	26.75	27.40	25.90	27.65	27.70

Experimento 2:

Se puede apreciar que la mutación con mayor promedio Fitness se encontraba en general entre los valores 40 % y 20 %, así que se utilizaron los valores 35 %,30 % y 25 %.

Se obtuvieron los resultados mostrados en las tablas 71, 72, 73, 74 y 75.

Tabla 71: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	40 %	35 %	30 %	25 %
Varianza Fitness	1.96	2.67	1.58	2.42
Promedio Fitness	51.65	51.45	51.14	51.36
Promedio No. Caract.	138.50	139.60	137.15	139.05

Tabla 72: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	40 %	35 %	30 %	25 %
Varianza Fitness	0.63	138.90	0.16	0.39
Promedio Fitness	94.67	90.91	94.78	94.78
Promedio No. Caract.	5.00	4.75	5.10	5.60

Tabla 73: Resultados de mutación del conjunto de datos Heart.

Porcentaje	60 %	35 %	30 %	25 %
Varianza Fitness	21.85	22.44	18.09	13.78
Promedio Fitness	70.01	70.52	70.60	70.62
Promedio No. Caract.	5.55	5.85	5.30	5.85

Tabla 74: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	10.52	11.18	10.92	3.78
Promedio Fitness	68.50	68.45	67.63	68.79
Promedio No. Caract.	7.55	7.85	9.05	8.45

Tabla 75: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	35 %	30 %	25 %	20 %
Varianza Fitness	18.93	15.61	16.01	11.86
Promedio Fitness	71.47	71.74	72.58	74.41
Promedio No. Caract.	27.30	27.80	28.30	27.65

Experimento 3:

Después del experimento anterior se seleccionaron los valores que reportaran el mayor promedio Fitness y se propusieron valores 2 anteriores y 2 posteriores, en las tablas 76, 77, 78, 79 y 80 se muestran los resultados.

Tabla 76: Resultados de mutación del conjunto de datos Arritmia.

Porcentaje	42 %	40 %	38 %
Varianza Fitness	1.37	1.96	1.75
Promedio Fitness	51.40	51.65	51.59
Promedio No. Caract.	139.25	138.50	139.75

Tabla 77: Resultados de mutación del conjunto de datos Breast Cancer.

Porcentaje	32 %	30 %	28 %
Varianza Fitness	0.17	0.16	0.31
Promedio Fitness	94.97	94.78	94.47
Promedio No. Caract.	4.80	5.10	4.70

Tabla 78: Resultados de mutación del conjunto de datos Heart.

Porcentaje	27 %	25 %	23 %
Varianza Fitness	14.25	13.78	11.04
Promedio Fitness	70.38	70.62	70.21
Promedio No. Caract.	6.05	5.85	5.95

Tabla 79: Resultados de mutación del conjunto de datos Hepatitis.

Porcentaje	22 %	20 %	18 %
Varianza Fitness	12.02	3.78	16.67
Promedio Fitness	68.52	68.79	67.48
Promedio No. Caract.	8.95	8.45	8.15

Tabla 80: Resultados de mutación del conjunto de datos Lung Cancer.

Porcentaje	22 %	20 %	18 %
Varianza Fitness	14.14	11.86	19.16
Promedio Fitness	72.67	74.41	72.13
Promedio No. Caract.	26.65	27.65	25.90

En las Figuras 28, 29, 30, 31 y 32 se observan las gráficas que muestran la eficiencia de clasificación y la reducción de características variando el porcentaje de mutación en cada conjunto de datos. Para estas gráficas se consideraron todos los valores de mutación utilizados en la experimentación.

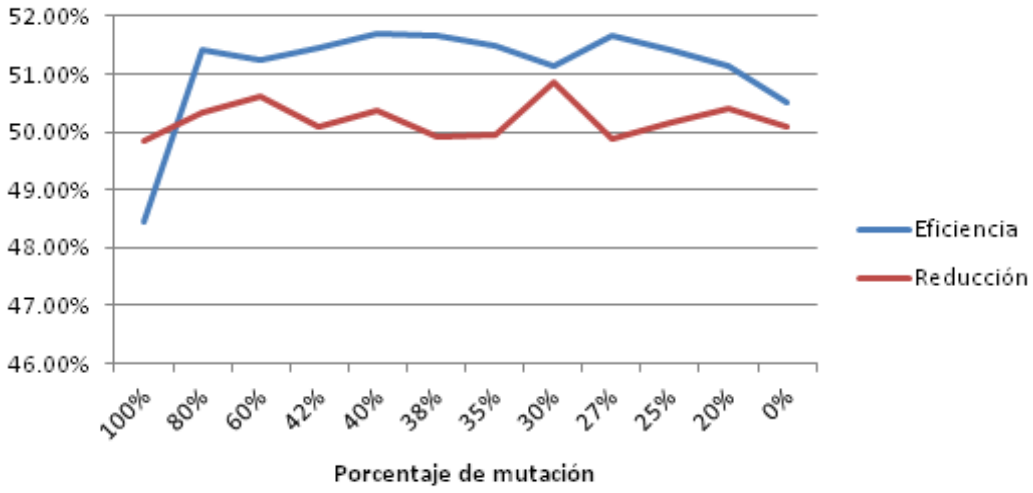

Figura 28: *Variación de porcentajes de mutación usando AGS4 en el conjunto de datos Arritmia.*

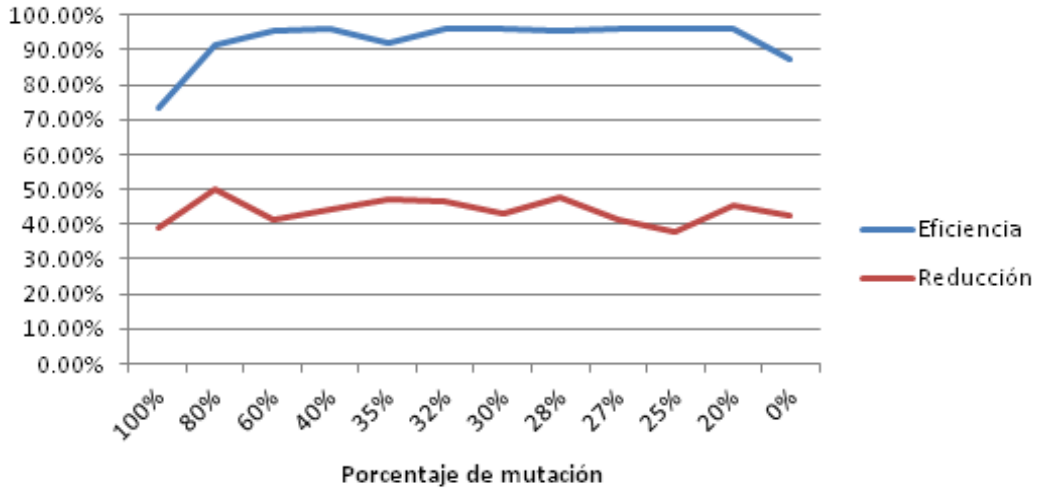

Figura 29: *Variación de porcentajes de mutación usando AGS4 en el conjunto de datos Breast Cancer.*

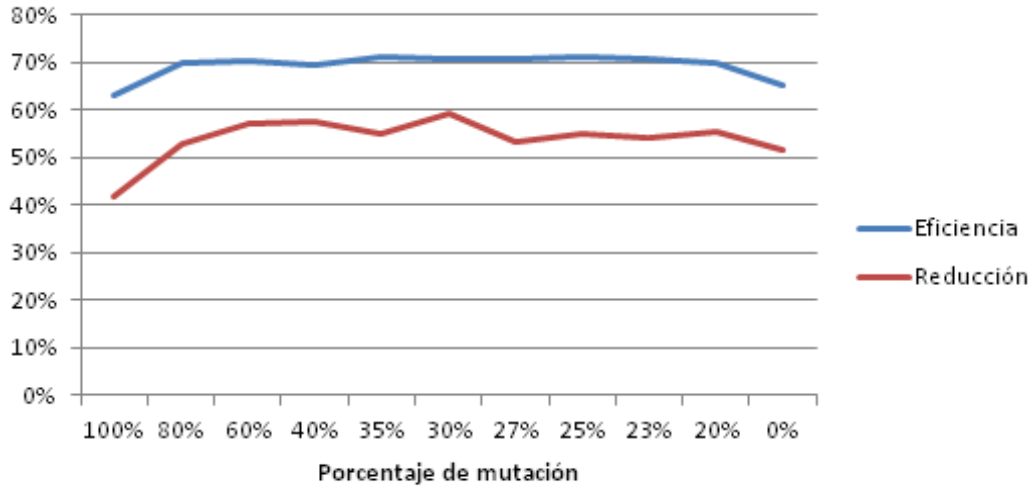

Figura 30: *Variación de porcentajes de mutación usando AGS4 en el conjunto de datos Heart.*

En la Tabla 81 se muestra los mejores resultados de la experimentación variando el porcentaje de mutación.

Tabla 81: Resumen de la experimentación con mutación.

Dataset	Porcentaje de mutación	Eficiencia de clasificación	Reducción de características
Arritmia	40 %	51.70 %	50.36 %
Breast Cancer	32 %	96.30 %	46.67 %
Heart	25 %	71.10 %	55 %
Hepatitis	20 %	69.20 %	55.53 %
Lung Cancer	20 %	75.15 %	50.63 %

Considerando los mejores porcentajes de mutación el promedio es de 27 % de mutación.

En la Figura 33 se muestra el desempeño del AGS4 en los 5 conjuntos de datos utilizando una mutación del 27 %.

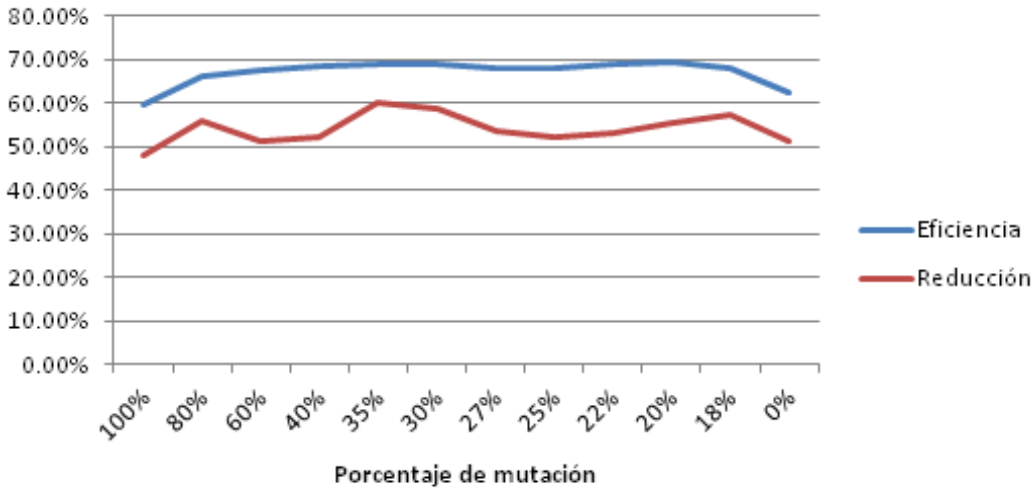

Figura 31: *Variación de porcentajes de mutación usando AGS4 en el conjunto de datos Hepatitis.*

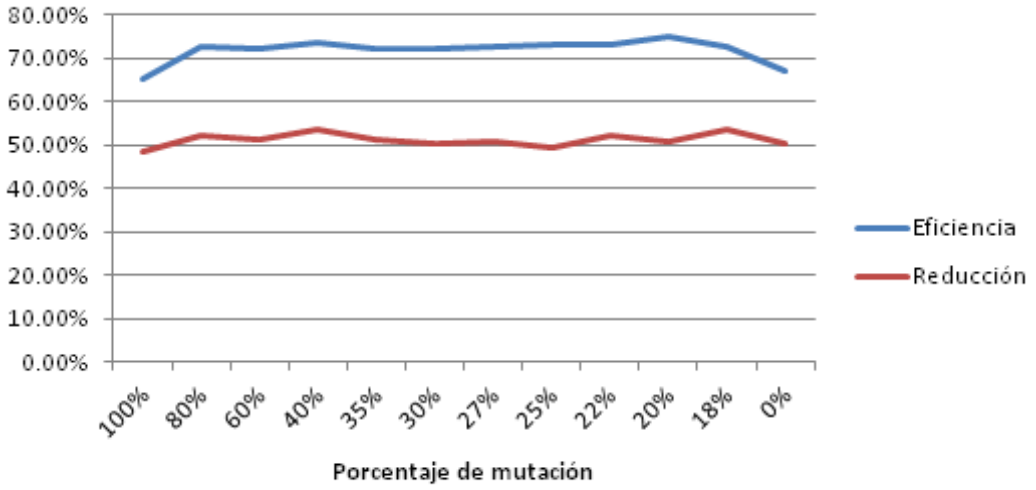

Figura 32: *Variación de porcentajes de mutación usando AGS4 en el conjunto de datos Lung Cancer.*

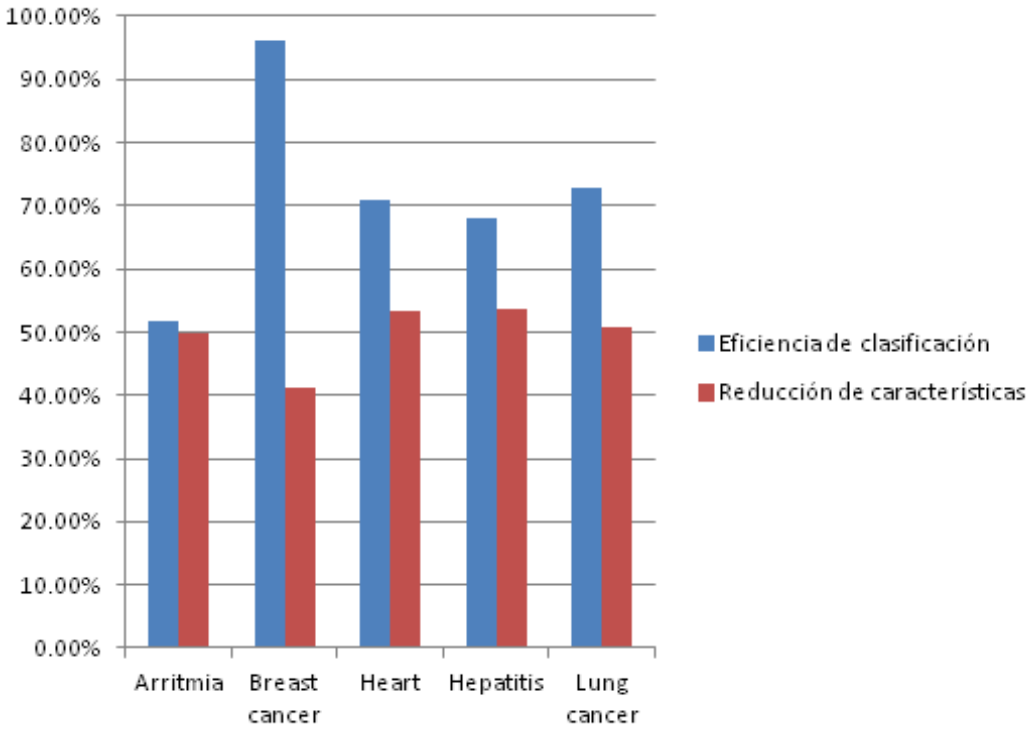

Figura 33: *Desempeño del AGS4 con 27 % de mutación.*

Conclusiones:

Se probaron diferentes porcentajes de mutación, se seleccionaron los mejores de cada algoritmo. En la Tabla 82 se resumen los resultados:

Tabla 82: Mejores porcentajes de mutación por algoritmo.

Algoritmo	Porcentaje de mutación
AGD	24 %
AGS1	4 %
AGS2	26 %
AGS3	6 %
AGS4	27 %

Finalmente se presenta la Tabla 83 donde son resumidos los valores finales obtenidos en la variación de mutación.

Tabla 83: Resumen de eficiencia de clasificación y reducción de características.

	Conjunto de Datos	Arritmia	Breast Cancer	Heart	Hepatitis	Lung Cancer
AGD	Eficiencia de clasificación	56.95 %	97.00 %	81 %	84.20 %	80.40 %
	Reducción de características	53.53 %	42.78 %	73.08 %	73.95 %	69.64 %
AGS1	Eficiencia de clasificación	55.60 %	96.45 %	73.15 %	78.15 %	79.05 %
	Reducción de características	54.16 %	46.11 %	75.38 %	60.53 %	64.38 %
AGS2	Eficiencia de clasificación	51.50 %	96.20 %	70.40 %	68.30 %	71.60 %
	Reducción de características	49.68 %	42.78 %	51.15 %	54.21 %	51.07 %
AGS3	Eficiencia de clasificación	54.45 %	96.75 %	75.15 %	77.25 %	78.90 %
	Reducción de características	52.03 %	40 %	67.31 %	62.89 %	59.91 %
AGS4	Eficiencia de clasificación	51.65 %	96.20 %	70.90 %	68.05 %	72.85 %
	Reducción de características	49.87 %	41.11 %	53.46 %	53.68 %	50.80 %

5.2.3. Experimentación AGD

Una vez obtenido el porcentaje de mutación indicado a usar en los diferentes algoritmos, podemos comparar el AGD y los AGS, a continuación se muestran

los objetivos de la experimentación así como la metodología usada en la experimentación de los diferentes algoritmos.

Los objetivos de la experimentación fueron:

- Observar el comportamiento del AGD y los AGS sobre diversos porcentajes de exploración del espacio de soluciones.
- Comparar las soluciones ofrecidas por el AGD y los AGS con las soluciones óptimas en los conjuntos de datos Breast Cancer, Heart y Hepatitis.
- Observar el aumento de la eficiencia de clasificación y la reducción de características en los conjuntos de datos Lung Cancer y Arritmia explorando pequeñas porciones del espacio de soluciones.

Para el cálculo de los óptimos, se realizó una evaluación por fuerza bruta de todas las combinaciones posibles de las características y para cada una se obtuvo su eficiencia de clasificación. Los óptimos solo se obtuvieron en los conjuntos se datos medianos, ya que la vasta cantidad de características de los conjuntos de datos grandes, imposibilita obtener los valores óptimos.

La experimentación se realizó de la siguiente forma:

Se realizaron 20 corridas del algoritmo, de estas 20 corridas se calculó la reducción de características y la eficiencia, el resultado se promedió y es presentado en forma de porcentaje.

Estas 20 corridas se realizaron utilizando diversos valores del espacio de búsqueda, estos valores fueron aproximadamente: 1 %, 3 %, 5 %, 10 %, 15 %, 20 % y 30 % para los conjuntos Heart, Hepatitis y Breast Cancer. Para los conjuntos de datos Lung Cancer y Arritmia se evaluaron generaciones del algoritmo, en cada generación 6 posibles soluciones fueron evaluadas. Para cada conjunto de datos se propuso un espacio de búsqueda de 10, 25, 50, 75, 100, 200, 300 y 350 generaciones.

Para cada AGS se realizaron 5 experimentos, con el porcentaje encontrado en la fase de variación de parámetros. Para el AGD se realizaron 5 experimentos con el porcentaje encontrado en la fase de variación de parámetros.

En los resultados de los conjuntos de datos Breast Cancer, Heart y Hepatitis se incluye una métrica llamada proximidad que representa la cercanía del Fitness de la solución encontrada por el algoritmo con respecto al Fitness de la solución optima encontrada por fuerza bruta.

De acuerdo con la expresión:

$$\text{proximidad} = \text{ABS}(\text{fitness de fuerza bruta - fitness obtenido})$$

Donde ABS es el valor absoluto.

Un valor de proximidad de CERO implica que la solución encontrada por el algoritmo es igual a la solución óptima encontrada por fuerza bruta.

A continuación se presentan los resultados obtenidos.
AGD mutación al 24 %

Tabla 84: Resultados del espacio de exploración del conjunto de datos Arritmia.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	54.35 %	55.40 %	56.45 %	56.95 %	57 %	57.85 %	58.40 %	58.35 %
Reducción	51.68 %	51.83 %	52.08 %	53.39 %	53.55 %	53.35 %	54.35 %	53.26 %

Tabla 85: Resultados del espacio de exploración del conjunto de datos Breast Cancer.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	96.65 %	96.85 %	97 %	97 %	97 %	97 %	97 %
Reducción	41.67 %	41.11 %	41.67 %	43.89 %	43.89 %	44.44 %	44.44 %
Proximidad	0.25	0.07	0.09	0.15	0.15	0.17	0.17

Tabla 86: Resultados del espacio de exploración del conjunto de datos Heart.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	78.85 %	81 %	81 %	81 %	81 %	81 %	81 %
Reducción	64.23 %	67.69 %	72.69 %	73.85 %	75 %	75.38 %	76.15 %
Proximidad	2.47	0.28	0.13	0.09	0.06	0.05	0.02

Tabla 87: Resultados del espacio de exploración del conjunto de datos Hepatitis.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	82.70 %	84.10 %	84.60 %	86.20 %	86.60 %	87.40 %	87.40 %
Reducción	73.16 %	74.47 %	75.79 %	80 %	82.63 %	85 %	85 %
Proximidad	6.60	5.20	4.68	3.00	2.53	1.69	1.69

Tabla 88: Resultados del espacio de exploración del conjunto de datos Lung Cancer.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	79.40 %	79.20 %	80.20 %	80.55 %	80.20 %	82 %	82.90 %	83.55 %
Reducción	56.52 %	64.20 %	64.55 %	67.68 %	69.02 %	70.09 %	68.13 %	65.45 %

AGS1 mutación al 4 %

Tabla 89: Resultados del espacio de exploración del conjunto de datos Arritmia.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	54.90 %	55.55 %	56.30 %	56.10 %	56.50 %	55.80 %	56.30 %	56.60 %
Reducción	51.11 %	51.94 %	53.32 %	53.58 %	53.12 %	53.41 %	52.63 %	53.66 %

Tabla 90: Resultados del espacio de exploración del conjunto de datos Breast Cancer.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	95.90 %	96.60 %	96.60 %	96.85 %	96.95 %	96.85 %	96.95 %
Reducción	45.56 %	43.89 %	45.56 %	45 %	43.89 %	45.56 %	42.78 %
Proximidad	0.87	0.24	0.19	0.04	0.10	0.05	0.07

Tabla 91: Resultados del espacio de exploración del conjunto de datos Heart.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	73.50 %	73.96 %	76.36 %	76.65 %	77.86 %	78.16 %	78.95 %
Reducción	68.46 %	74.23 %	72.69 %	69.62 %	65.77 %	69.62 %	66.54 %
Proximidad	7.53	6.92	4.64	4.44	3.39	2.98	2.30

Tabla 92: Resultados del espacio de exploración del conjunto de datos Hepatitis.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	78.20 %	77.90 %	78.20 %	78 %	78.05 %	77.80 %	78.35 %
Reducción	66.05 %	64.74 %	62.37 %	65.26 %	62.89 %	66.84 %	60.79 %
Proximidad	11.18	11.51	11.29	11.40	11.42	11.54	11.19

Tabla 93: Resultados del espacio de exploración del conjunto de datos Lung Cancer.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	78.25 %	79 %	79.65 %	79 %	79 %	79.20 %	79 %	78.85 %
Reducción	59.11 %	62.77 %	65.98 %	66.70 %	67.59 %	64.38 %	65.54 %	66.07 %

AGS2 mutación al 26 %

Tabla 94: Resultados del espacio de exploración del conjunto de datos Arritmia.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	51.55 %	51.05 %	52.55 %	51.75 %	51.85 %	50.85 %	51.75 %	51.20 %
Reducción	50.86 %	50.73 %	49.62 %	51.68 %	50.48 %	51.79 %	50.70 %	49.28 %

Tabla 95: Resultados del espacio de exploración del conjunto de datos Breast Cancer.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	95.90 %	96.06 %	96 %	96.20 %	96.16 %	96.25 %	96.05 %
Reducción	48.89 %	41.67 %	42.78 %	42.78 %	44.44 %	40 %	40 %
Proximidad	0.78	0.83	0.85	0.65	0.66	0.68	0.87

Tabla 96: Resultados del espacio de exploración del conjunto de datos Heart.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	69.05 %	69.26 %	71.75 %	70.80 %	70.85 %	71.55 %	71.30 %
Reducción	51.92 %	55 %	57.69 %	58.85 %	55 %	55.38 %	61.54 %
Proximidad	12.34	12.06	9.55	10.44	10.55	9.81	9.87

Tabla 97: Resultados del espacio de exploración del conjunto de datos Hepatitis.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	69.10 %	69.95 %	68.25 %	67 %	68.10 %	68.30 %	68.70 %
Reducción	54.47 %	58.42 %	56.32 %	51.58 %	55.53 %	54.74 %	47.63 %
Proximidad	20.35	19.41	21.12	22.48	21.29	21.12	20.95

Tabla 98: Resultados del espacio de exploración del conjunto de datos Lung Cancer.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	73.25 %	73.55 %	72.25 %	73.95 %	72.55 %	73.85 %	74.20 %	72.25 %
Reducción	51.43 %	51.25 %	52.05 %	51.16 %	52.23 %	51.70 %	53.04 %	48.39 %

AGS3 mutación al 6 %

Tabla 99: Resultados del espacio de exploración del conjunto de datos Arritmia.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	53.45 %	55.05 %	55.20 %	54.75 %	55.50 %	54.60 %	55 %	55.70 %
Reducción	52.17 %	50.88 %	51.47 %	52.31 %	51.97 %	52.35 %	51.33 %	51.25 %

Tabla 100: Resultados del espacio de exploración del conjunto de datos Breast Cancer.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	93.90 %	96.55 %	96.80 %	96.80 %	96.90 %	96.85 %	97 %
Reducción	49.44 %	43.89 %	40.56 %	43.33 %	43.33 %	41.67 %	41.67 %
Proximidad	2.71	0.28	0.13	0.06	0.04	0.05	0.09

Tabla 101: Resultados del espacio de exploración del conjunto de datos Heart.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	72.05 %	75.60 %	75.75 %	75.50 %	76.75 %	76.65 %	75.70 %
Reducción	65.77 %	68.08 %	67.31 %	65 %	66.15 %	65 %	68.46 %
Proximidad	9.02	5.50	5.38	5.69	4.45	4.58	5.39

Tabla 102: Resultados del espacio de exploración del conjunto de datos Hepatitis.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	76.50 %	76.85 %	77.25 %	76 %	77.75 %	77.86 %	76.80 %
Reducción	63.42 %	58.68 %	58.95 %	63.16 %	63.68 %	63.42 %	64.47 %
Proximidad	12.91	12.71	12.31	13.40	11.69	11.60	12.58

Tabla 103: Resultados del espacio de exploración del conjunto de datos Lung Cancer.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	78.10 %	79.05 %	78.85 %	78.70 %	78.70 %	78.55 %	79.05 %	78.25 %
Reducción	56.79 %	59.82 %	60.54 %	61.61 %	57.68 %	60.63 %	61.07 %	60.45 %

AGS4 mutación al 27 %

Tabla 104: Resultados del espacio de exploración del conjunto de datos Arritmia.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	50.90 %	51.10 %	51.60 %	51.55 %	51.95 %	52.20 %	51.75 %	51.65 %
Reducción	49.37 %	51.15 %	49.89 %	49.35 %	50.63 %	51.09 %	49.50 %	51.27 %

Tabla 105: Resultados del espacio de exploración del conjunto de datos Breast Cancer.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	96.05 %	93.95 %	96 %	93.90 %	96.05 %	94 %	94.20 %
Reducción	43.33 %	46.67 %	44.44 %	47.22 %	42.78 %	43.33 %	44.44 %
Proximidad	0.78	2.73	0.80	2.77	0.80	2.77	2.55

Tabla 106: Resultados del espacio de exploración del conjunto de datos Heart.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	71.50 %	70.55 %	69.55 %	70.20 %	69.85 %	70.15 %	69.50 %
Reducción	56.15 %	48.08 %	49.23 %	51.15 %	48.46 %	53.46 %	57.69 %
Proximidad	9.84	11.00	11.94	11.25	11.67	11.23	11.73

Tabla 107: Resultados del espacio de exploración del conjunto de datos Hepatitis.

Espacio de búsqueda	1 %	3 %	5 %	10 %	15 %	20 %	30 %
Eficiencia	67.40 %	68.50 %	68.75 %	69.65 %	67.15 %	67.50 %	67.85 %
Reducción	55.53 %	53.16 %	56.58 %	54.47 %	54.74 %	56.84 %	56.05 %
Proximidad	21.97	20.97	20.63	19.82	22.24	21.83	21.52

Tabla 108: Resultados del espacio de exploración del conjunto de datos Lung Cancer.

Generaciones	10	25	50	75	100	200	300	350
Eficiencia	72.15 %	74.55 %	72.55 %	73.50 %	72.50 %	73.40 %	73.90 %	71.55 %
Reducción	52.32 %	50.18 %	51.43 %	48.93 %	50.98 %	51.79 %	48.75 %	51.34 %

Como resumen de la experimentación se muestran las Tablas 109, 110, 111, 112 y 113 en ellas se presentan los mejores valores de cada algoritmo. Para los conjuntos de datos Breast Cancer, Hepatitis y Heart se incluye un valor llamado óptimo este valor representa la cercania de la solución entregada por el algoritmo con la solución óptima encontrada por fuerza bruta.

Tabla 109: Mejores valores de reducción y eficiencia del conjunto de datos Arritmia.

Algoritmo	AGD 24 %	AGS1 4 %	AGS2 26 %	AGS3 6 %	AGS4 27 %
Generaciones	300	350	75	350	200
Eficiencia	58.40 %	56.60 %	51.75 %	55.70 %	52.20 %
Reducción	54.35 %	53.66 %	51.68 %	51.25 %	51.09 %

Tabla 110: Mejores valores de reducción y eficiencia del conjunto de datos Breast Cancer.

Algoritmo	AGD 24 %	AGS1 4 %	AGS2 26 %	AGS3 6 %	AGS4 27 %
Espacio de búsqueda	20 %	15 %	20 %	30 %	1 %
Eficiencia	97 %	96.95 %	96.25 %	97 %	96.05 %
Reducción	44.44 %	43.89 %	40 %	41.67 %	43.33 %
Proximidad	0.17	0.10	0.68	0.09	0.78

Tabla 111: Mejores valores de reducción y eficiencia del conjunto de datos Heart.

Algoritmo	AGD 24 %	AGS1 4 %	AGS2 26 %	AGS3 6 %	AGS4 27 %
Espacio de búsqueda	30 %	30 %	5 %	15 %	1 %
Eficiencia	81 %	78.95 %	71.75 %	76.75 %	71.50 %
Reducción	76.15 %	66.54 %	57.69 %	66.15 %	56.15 %
Proximidad	0.02	2.30	9.55	4.45	9.84

Tabla 112: Mejores valores de reducción y eficiencia del conjunto de datos Hepatitis.

Algoritmo	AGD 24 %	AGS1 4 %	AGS2 26 %	AGS3 6 %	AGS4 27 %
Espacio de búsqueda	30 %	30 %	3 %	20 %	10 %
Eficiencia	87.40 %	78.35 %	69.95 %	77.86 %	69.65 %
Reducción	85 %	60.79 %	58.42 %	63.42 %	54.47 %
Proximidad	1.69	11.19	19.41	11.60	19.82

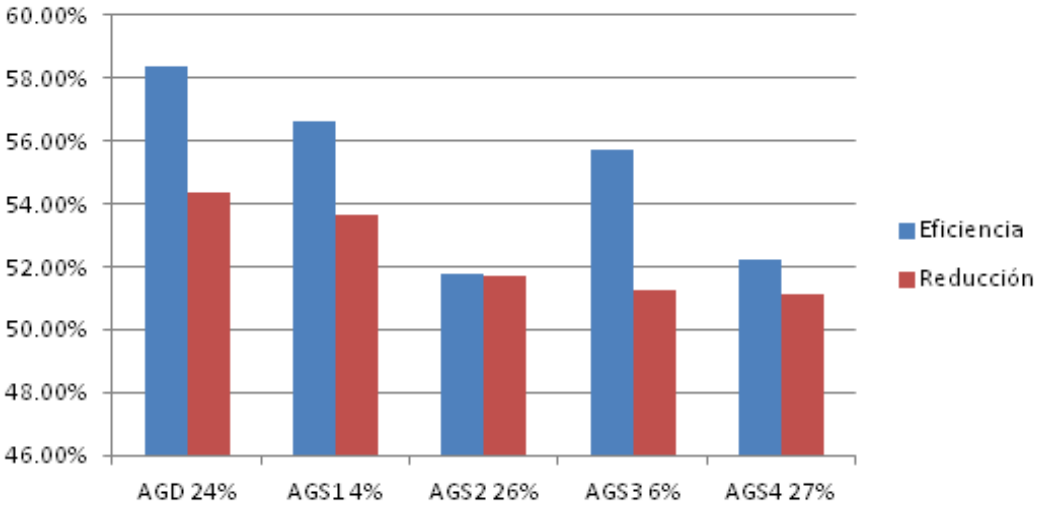

Figura 34: Mejores resultados por algoritmo para el conjunto de datos Arritmia

Tabla 113: Mejores valores de reducción y eficiencia del conjunto de datos Lung Cancer.

Algoritmo	AGD 24 %	AGS1 4 %	AGS2 26 %	AGS3 6 %	AGS4 27 %
Generaciones	350	50	75	300	25
Eficiencia	83.55 %	79.65 %	73.95 %	79.05 %	74.55 %
Reducción	65.45 %	65.98 %	51.16 %	61.07 %	50.18 %

En las Figuras 34, 35, 36, 37 y 38 se muestran los mejores valores de eficiencia y reducción para cada cada algoritmo.

En las Figuras 39, 40 y 41 se muestran los mejores valores de la métrica proximidad en los diferentes algoritmos.

En la Figura 42 se muestra una gráfica donde se calcularon los promedios de eficiencia y reducción en todos los algoritmos.

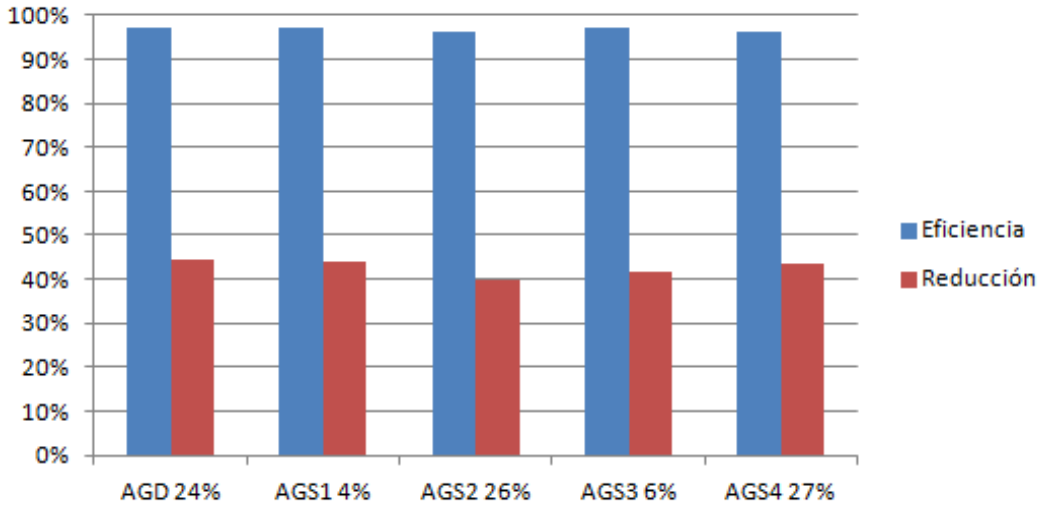

Figura 35: Mejores resultados por algoritmo para el conjunto de datos Breast Cancer

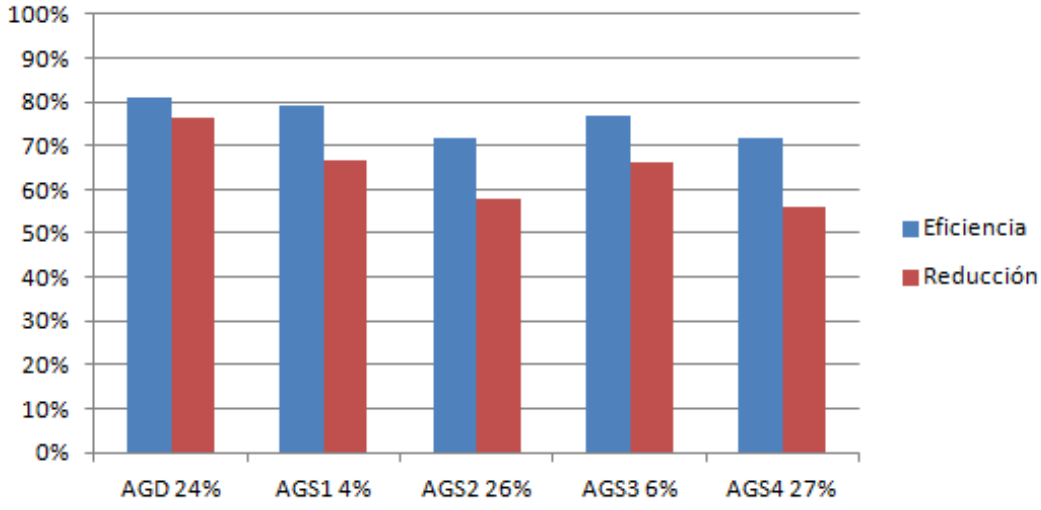

Figura 36: Mejores resultados por algoritmo para el conjunto de datos Heart

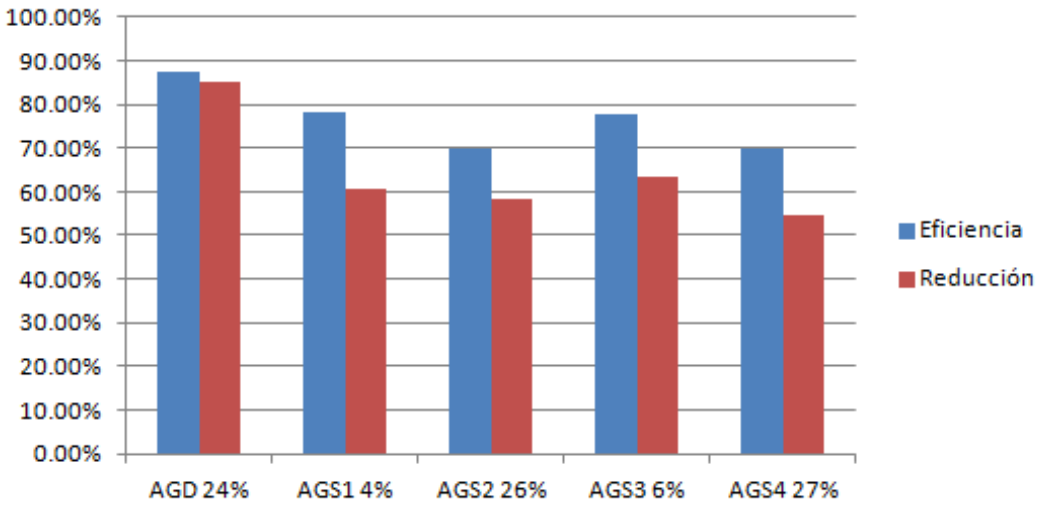

Figura 37: Mejores resultados por algoritmo para el conjunto de datos Hepatitis

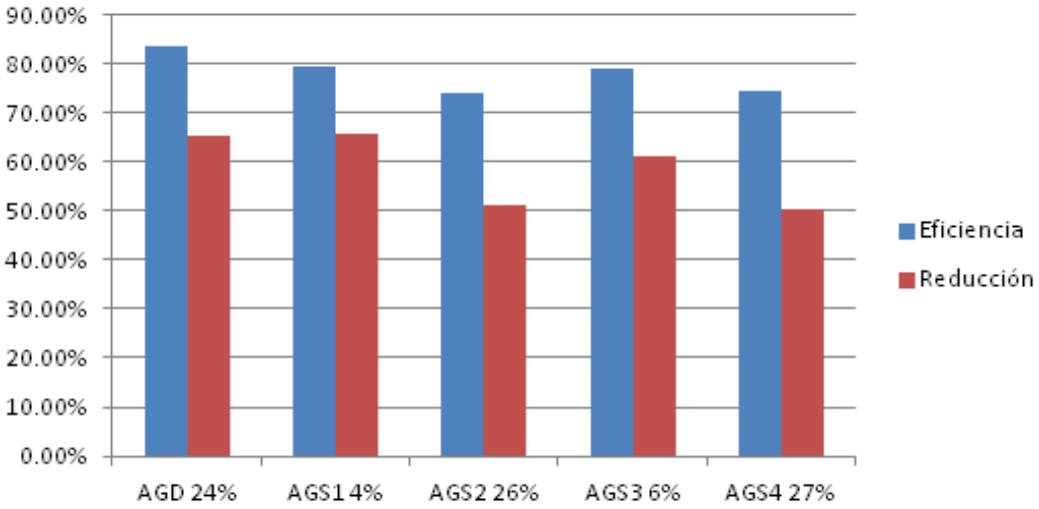

Figura 38: Mejores sesultados por algoritmo para el conjunto de datos Lung Cancer

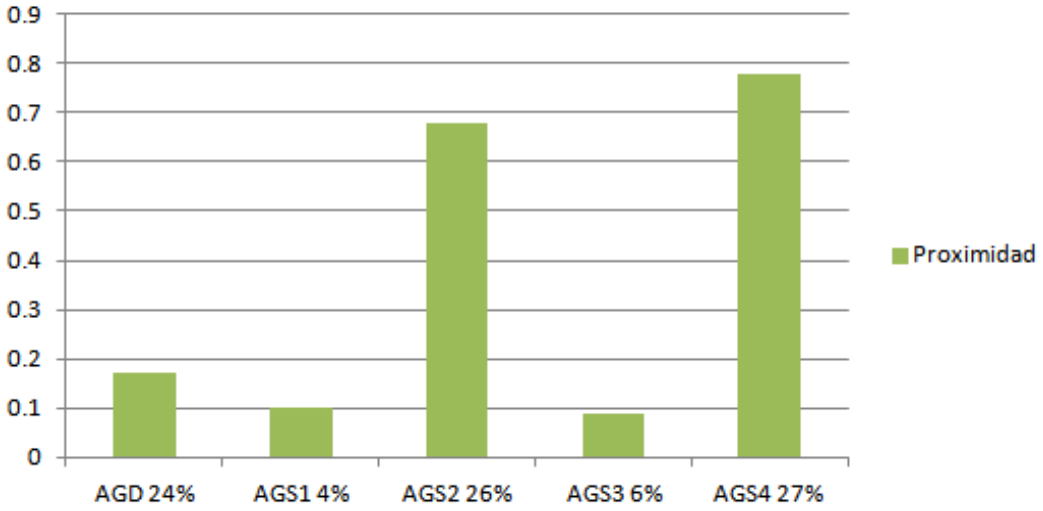

Figura 39: Mejores resultados de proximidad por algorimo para el conjunto de datos Breast Cancer

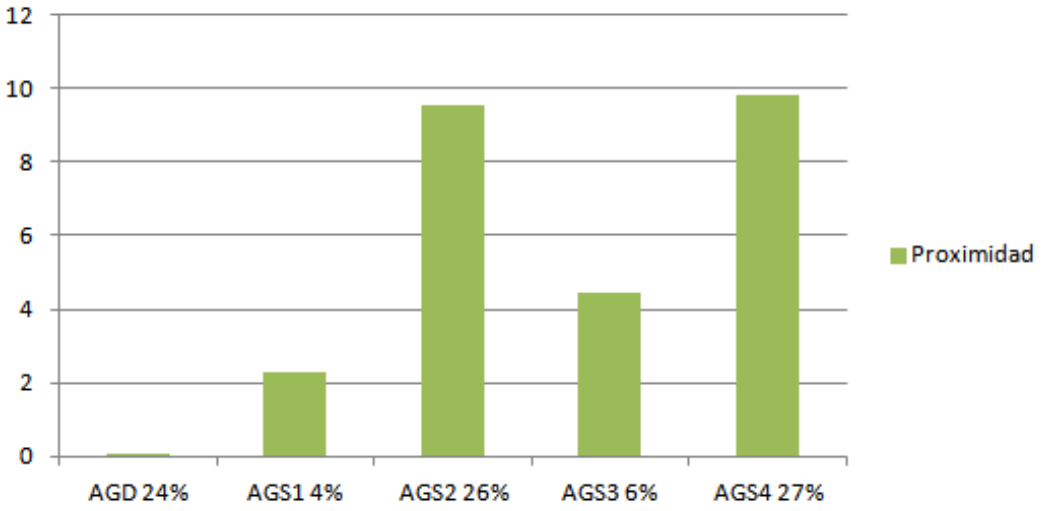

Figura 40: Mejores resultados de proximidad por algorimo para el conjunto de datos Heart

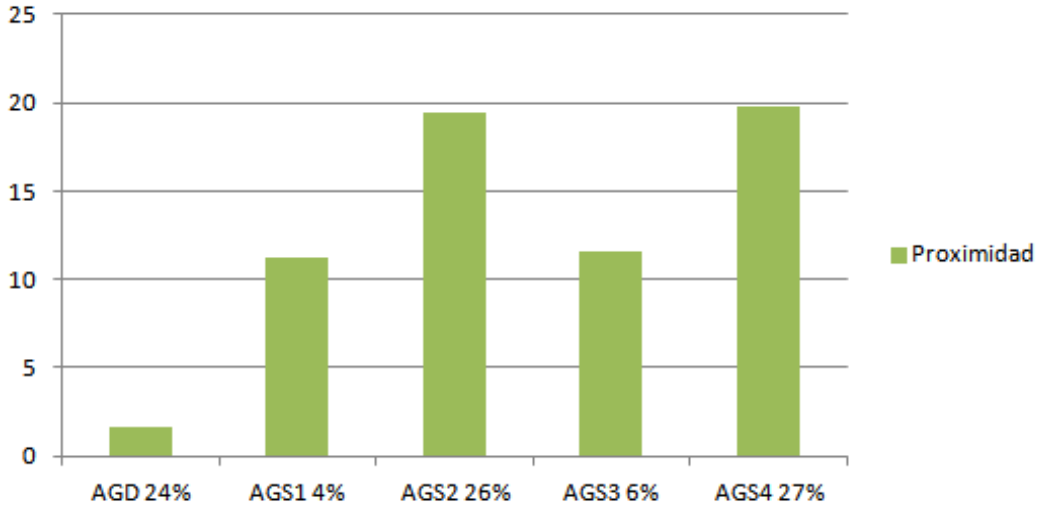

Figura 41: Mejores resultados de proximidad por algorimo para el conjunto de datos Hepatitis

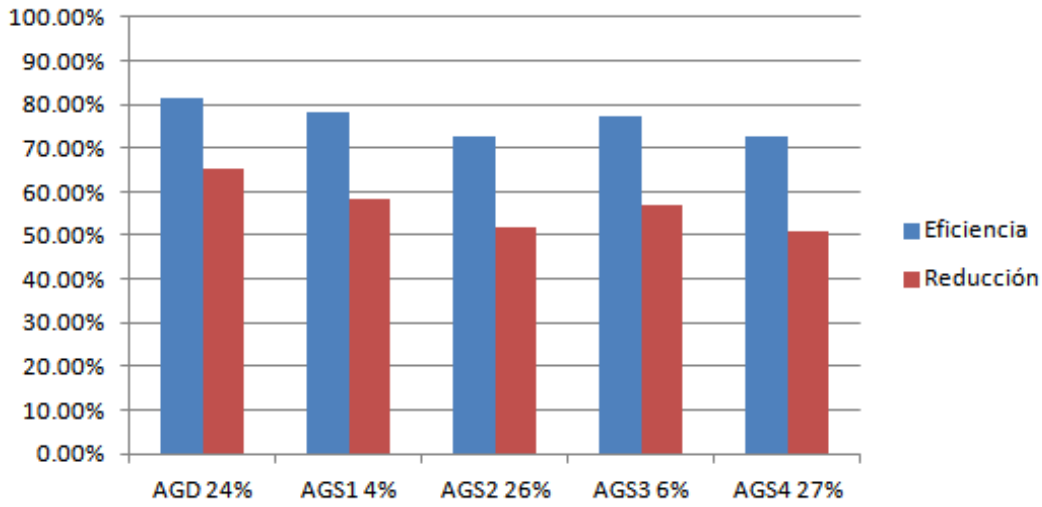

Figura 42: Mejores resultados de eficiencia y reducción promediados por algoritmo

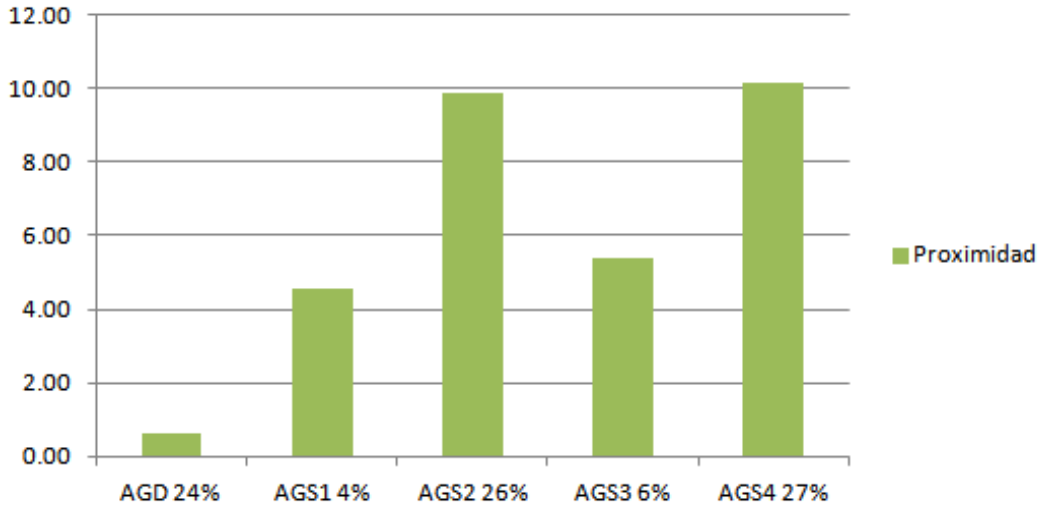

Figura 43: Mejores resultados de proximidad promediados por algoritmo

Finalmente en la Figura 43 se muestra una grafica donde se calculó el valor promedio de proximidad para cada algoritmo.

Se puede apreciar que el AGD sobresale al encontrar soluciones más eficientes y alcanzar una mayor proximidad con respecto a la solución óptima.

5.3. Construcción de la metodología DAGA

La metodología DAGA se creó para ser usada en conjuntos de datos con 20 características o más, en donde su alta dimensionalidad imposibilita el usar otras técnicas óptimas de selección de características como es la búsqueda por fuerza bruta.

La metodología DAGA (Dominant Adaptative Genetic Algorithm) debe su nombre al algoritmo genético que utiliza como parte central de la misma, hablamos del algoritmo AGD, que en el Capítulo 5.2.3 demostró tener una buena aproximación a las soluciones óptimas en conjuntos de datos pequeños.

Para un problema de clasificación dado, existen algoritmos clasificadores que proporcionan un mejor desempeño que otros Con esta idea en mente se propone una metodología que utiliza diferentes algoritmos de clasificación, para lograr la adaptabilidad del algoritmo AGD a diferentes problemas, mejorando así la eficiencia de clasificación.

Con el concepto de adaptabilidad incorporado en el AGD se logra una selección de características más eficiente, pero también se ofrece la posibilidad de seleccionar, de forma automática entre diferentes clasificadores, a aquel que ofrezca una mejor eficiencia de clasificación.

En la sección 4.2.1 se describe el AGD, este algoritmo se analizó variando su parámetro porcentaje de mutación. Como una primera aproximación en 5.2.1 se analizó, cuál era el porcentaje de mutación que permitía al algoritmo desempeñarse de forma eficiente en un grupo de conjuntos de datos.

Se decidió ir más allá con el desarrollo de este algoritmo. Se experimentó con variaciones de los porcentajes de mutación y cómo a través de estas variaciones evolucionaba la solución encontrada.

A continuación se describirán los conjuntos de datos usados, después se presentarán las 2 variaciones del AGD propuestas posteriormente se expondrá la experimentación realizada con los 2 algoritmos y se terminará con la definición de la metodología DAGA.

Para probar los algoritmos se usaron 5 conjuntos de datos públicos obtenidos del UCI machine learning repository [21]. En la tabla 114 se describen las características de estos conjuntos de datos.

Tabla 114: Conjuntos de fatos para la prueba de los algoritmos AGD1 y AGD2.

Conjunto de datos	**No. de características**	**Espacio de soluciones**	**Instancias**
Arritmia	278	4.86E+83	420
Promoters	57	1.44E+17	106
Dermatology	34	1.72E+10	358
Ionosphere	34	1.72E+10	351
Sick euthyroid	24	1.68E+07	82

En los siguientes párrafos se presentan 2 versiones del AGD, la llamada AGD1 y AGD2, estas nuevas versiones del AGD original tienen la particularidad de que en cada generación pueden variar su porcentaje de mutación en base a cierta condición dada, a continuación se detallarán estas condiciones para cada algoritmo.

El AGD1 utilizó las condiciones y porcentajes de mutación mostrados en la figura 44.

El AGD2 utilizó un porcentaje de mutación inicial de 50 %, las condiciones de cambio de porcentajes de mutación se describen en el párrafo siguiente.

Siempre que existiera un cambio en la mejor solución encontrada y el porcentaje de mutación fuera menor del 50 % se aumenta el porcentaje de mutación en 1 %. Cada 25 generaciones, se verifica si en 100 generaciones o más no ha habido cambio en la mejor solución, si es así, al porcentaje de mutación actual se le resta un 5 %. El algoritmo termina cuando su porcentaje de mutación es menor o igual a un 0 %.

La experimentación realizada con estos dos algoritmos consistió en 20 corridas, en cada una de las iteraciones del algoritmo se reporta el mejor fitness obtenido hasta el momento y el promedio de fitness por generación, con esto se pretende mostrar la evolución de las soluciones dentro del algoritmo, así como la rapidez con la que el algoritmo alcanza mejores soluciones.

En las figuras 45, 46, 47, 48 y 49 se presentan los resultados de la experimentación del AGD1 y AGD2 en los 5 conjuntos de datos. Se puede apreciar en

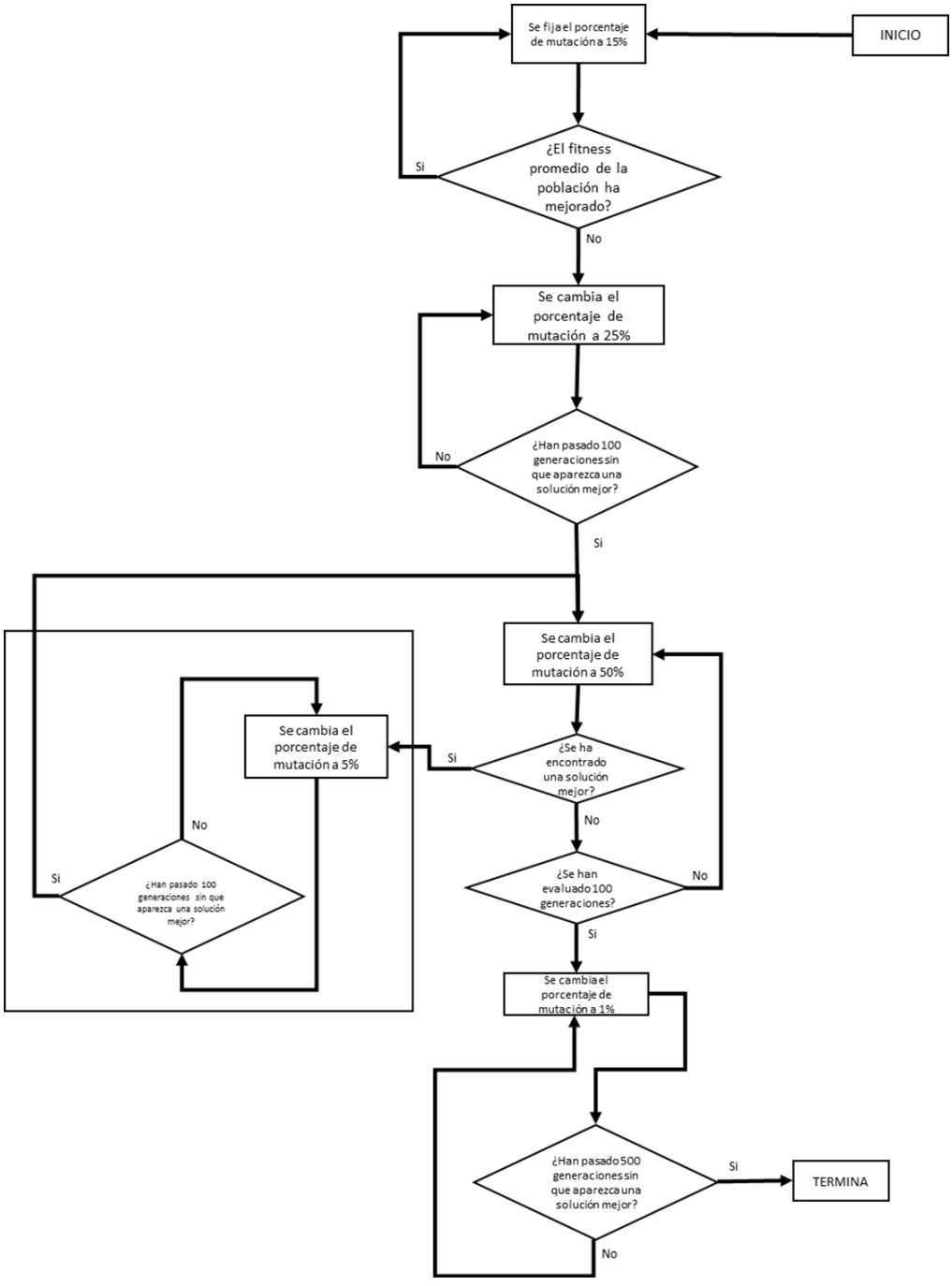

Figura 44: Condiciones y porcentajes de mutación utilizados por el AGD1

las figuras que el algoritmo AGD1 en promedio en la mayoría de los conjuntos obtiene mejores soluciones en un menor tiempo. Por otro lado, el AGD2 obtuvo un mejor resultado en las generaciones finales de la ejecución.

La metodología DAGA toma ventaja de los algoritmos AGD1 y AGD2, en particular de la rápida convergencia del AGD1 a una solución que eleva la eficiencia de clasificación y el resultado final del AGD2 que logra un mejor resultado en eficiencia de clasificación que el AGD1.

Para optimizar la solución presentada por la metodología DAGA, se incluye en esta un algoritmo tipo filtro llamado Information Gain [1]. La incorporación de este algoritmo a la metodología DAGA, logra reducir en algunos casos el número de características antes de realizar la tarea de selección de características llevada a cabo por el AGD.

Antes de presentar la metodología, cabe resaltar que los clasificadores usados así como el algoritmo Information Gain, fueron implementados directamente de la herramienta weka [22], permitiendo así, el poder replicar la experimentación en su totalidad, así como tener detalles específicos y documentados sobre la implementación de los algoritmos usados.

5.4. Metodología DAGA

Para implementar la metodología DAGA, lo primero que se debe hacer es seleccionar un algoritmo tipo filtro, en este trabajo se decidió seleccionar el algoritmo Information Gain implementado en weka, se realizó una pequeña experimentación con este y se obtuvieron buenos resultados.

Después de que se tiene seleccionado el algoritmo tipo filtro, se procede con la selección de los algoritmos clasificadores a utilizar. Para este trabajo se decidió utilizar los algoritmos SOM, Random Forest y J48, esta decisión se tomó en base a una lectura de la biografía especializada y el análisis de las opciones más usadas dentro del ámbito científico.

Una vez determinado el algoritmo tipo filtro y los clasificadores a usar, se procede con la correcta implementación de los de los algoritmos AGD1 y AGD2.

Se utiliza el AGD1 con un número de generaciones pequeño (en este trabajo 10 generaciones) para seleccionar el algoritmo clasificador que ofrezca una mayor eficiencia de clasificación para el conjunto de datos analizado. El uso del AGD1 con un pequeño número de generaciones es posible ya que este algoritmo es capaz de encontrar soluciones buenas de forma rápida.

Una vez que se halla elegido un clasificador haciendo uso del AGD1, se procede a el uso del AGD2. El AGD2 retoma el resultado parcial encontrado por el AGD1, sigue explorando el espacio de soluciones hasta cumplir su condición de paro y retornar el mejor subconjunto de características encontrado.

En la Figura 50 se muestra graficamente la metodología DAGA.

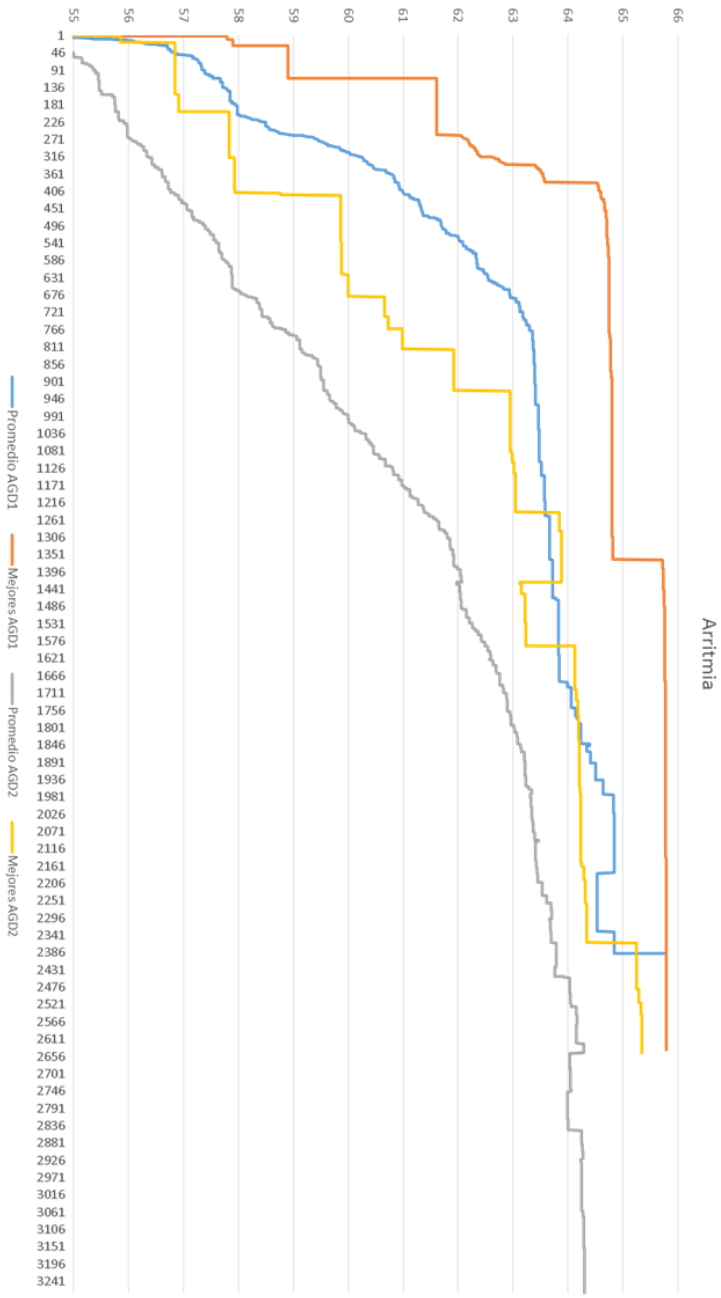

Figura 45: Resultados del algoritmo AGD1 y AGD2 en el conjunto de datos Arritmia

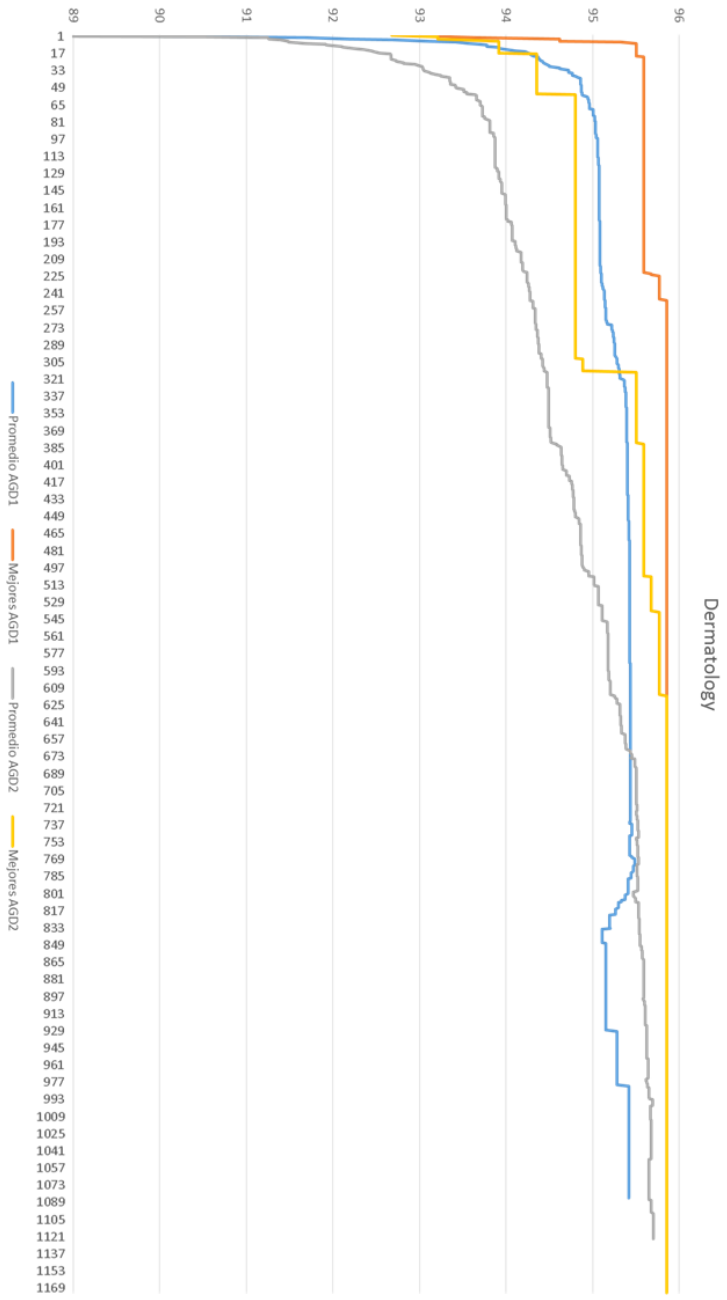

Figura 46: Resultados del algoritmo AGD1 y AGD2 en el conjunto de datos Arritmia

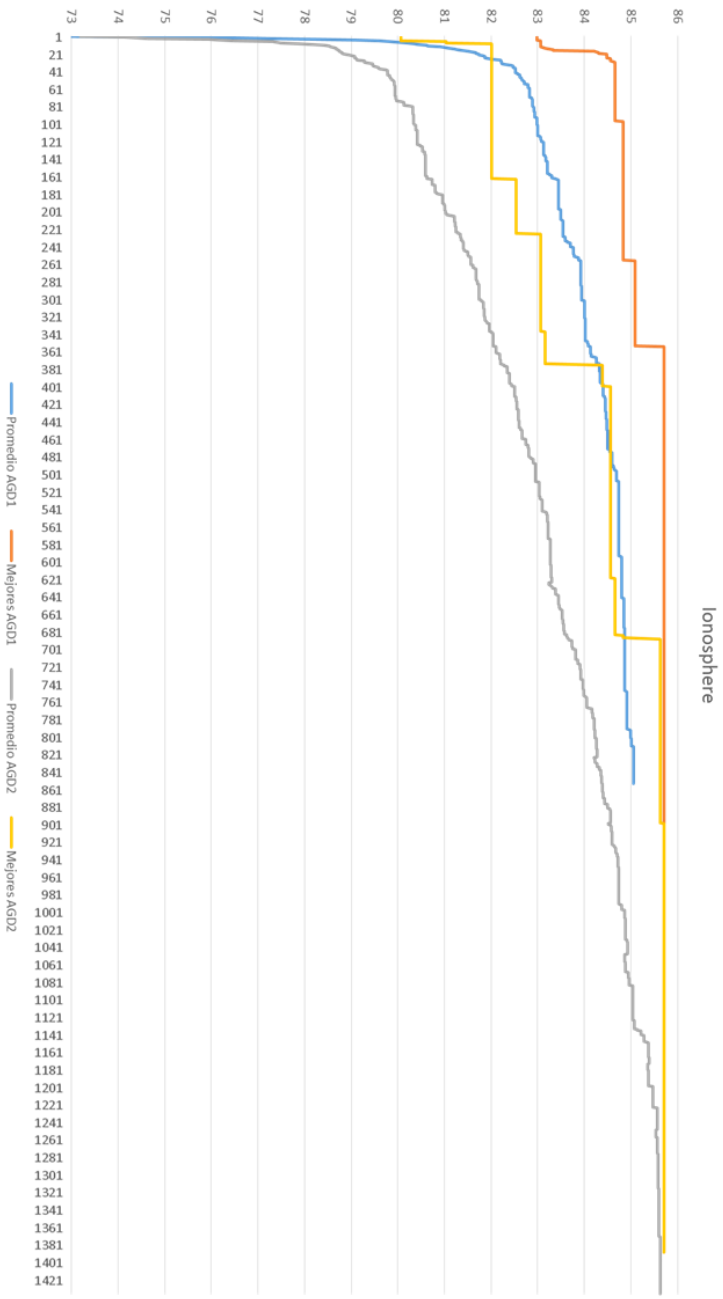

Figura 47: Resultados del algoritmo AGD1 y AGD2 en el conjunto de datos Ionosphere

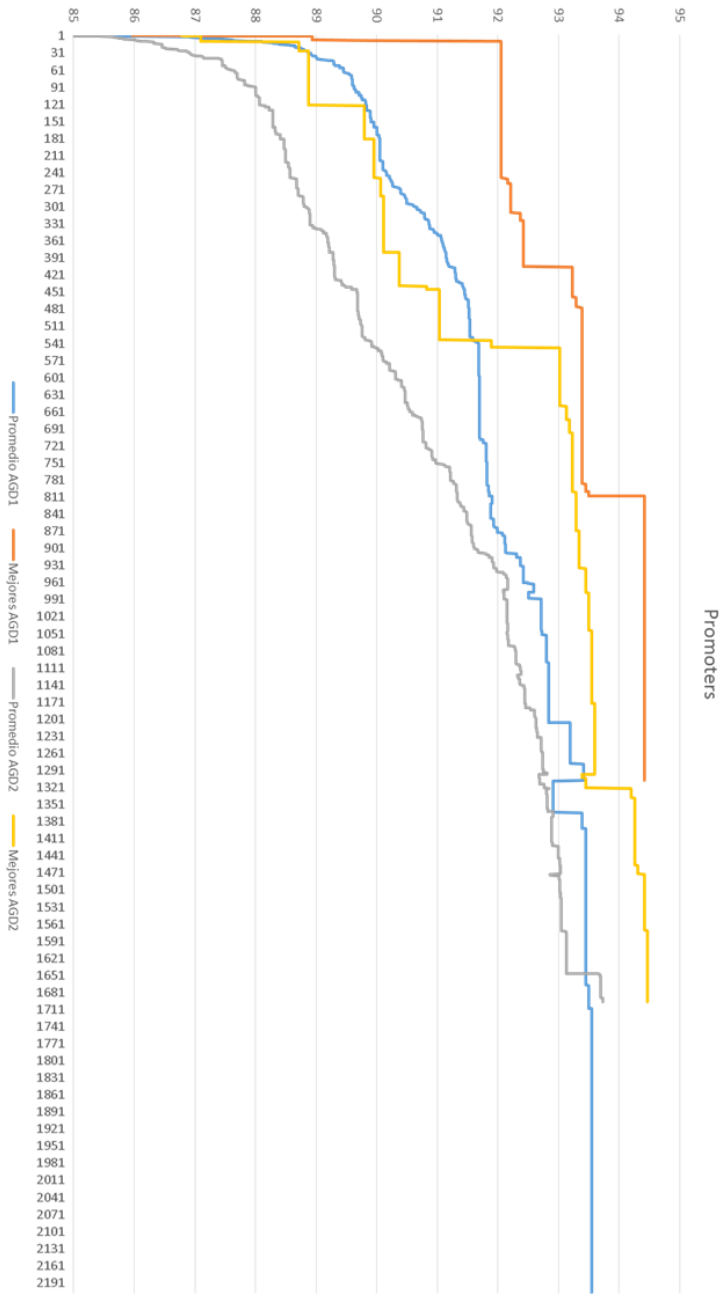

Figura 48: Resultados del algoritmo AGD1 y AGD2 en el conjunto de datos Promoters

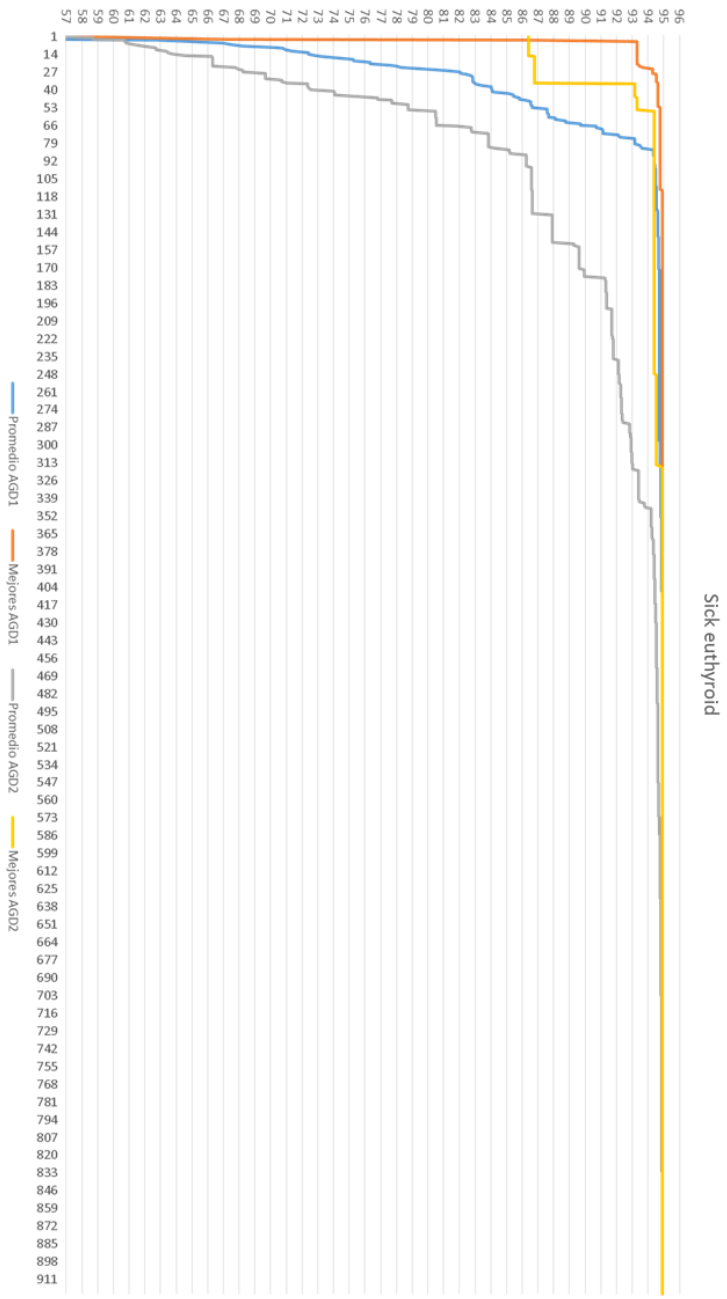

Figura 49: Resultados del algoritmo AGD1 y AGD2 en el conjunto de datos Sick euthyroid

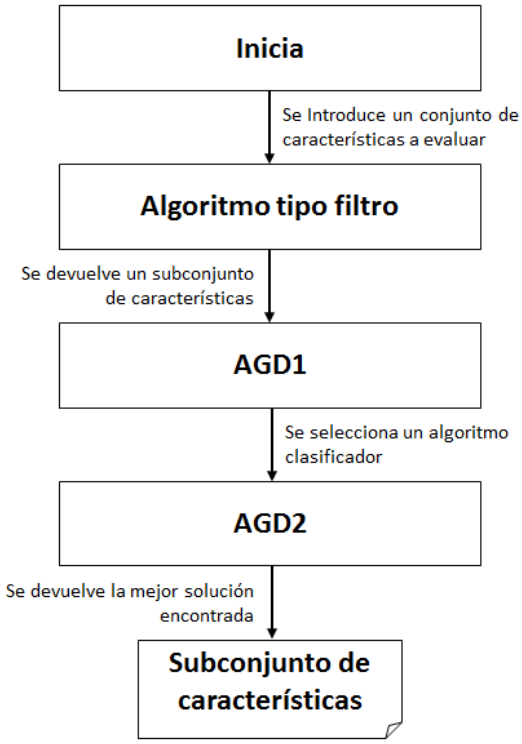

Figura 50: Diagrama de la metodología DAGA

6. Experimentación

Para la experimentación con la metodología DAGA, se utilizaron 13 conjuntos de datos de más de 20 características obtenidos del UCI machine learning repository [21], estos conjuntos de datos se describen en la tabla 115. Se realizaron 5 corridas utilizando la metodología DAGA, en la Tabla 116 se presentan ros resultados de la experimentación realizada. A continuación se describe el significado de cada columna de la tabla 116.

La columna Algoritmo muestra el algoritmo que selecciono la metodología DAGA, con la siguiente nomenclatura:

- RF: Implementación en WEKA del algoritmo Random Forest
- SVM: Implementación en WEKA del algoritmo Support Vector Machine
- J48: Implementación en WEKA del algoritmo C4.5

La columna Eficiencia muestra la eficiencia de clasificación alcanzado por el subconjunto de características encontrado por DAGA. Reducción muestra cuanto se redujo en porcentaje el número de características respecto al conjunto original. La columna Desv. Esta. muestra la desviación estándar de la eficiencia de clasificación en las 5 corridas de la metodología DAGA.

Tabla 115: Conjuntos de datos para la experimentación con la metodología DAGA

Conjunto de datos	**No. de características**	**Espacio de soluciones**	**Instancias**
LSVT_voice	310	2.09E+93	126
Arritmia	278	4.86E+83	420
Semeion	256	1.16E+77	1593
Promoters	57	1.44E+17	106
Lung_cancer	56	7.21E+16	27
Dermatology	34	1.72E+10	358
Ionosphere	34	1.72E+10	351
Sick euthyroid	24	1.68E+07	82
Parkinsons	22	4.19E+06	195
Spect	22	4.19E+06	267
Spect_f	44	1.76E+13	267
Movement_libras	20	1.05E+06	360
Soybean	35	3.44E+10	47

Tabla 116: Resultados de la experimentación con la metodología DAGA

Conjunto de datos	**Algoritmo**	**Eficiencia**	**Reducción**	**Desv. Esta.**
LSVT_voice	SVM	96.03 %	92.90 %	0.00
Arritmia	RF	81.19 %	86.69 %	0.35
Semeion	SVM	95.17 %	48.83 %	0.11
Promoters	RF	96.23 %	89.47 %	0.00
Lung_cancer	SVM	96.30 %	92.86 %	0.00
Dermatology	RF	98.88 %	58.82 %	0.15
Ionosphere	RF	95.73 %	52.94 %	0.24
Sick euthyroid	RF	97.35 %	87.50 %	0.00
Parkinsons	RF	95.38 %	72.73 %	0.28
Spect	RF	87.27 %	68.18 %	0.37
Spect_f	RF	88.39 %	86.36 %	1.62
Movement_libras	RF	86.11 %	73.33 %	0.58
Soybean	J48	100.00 %	94.29 %	0.00
Promedio		93.39 %	77.30 %	

Como conclusión de la experimentación, se presenta la Tabla 117 en donde se pueden comparar los resultados obtenidos por DAGA con respecto los resultados orignales obtenidos por el uso de algoritmos clasificadores. La columna Efic, muestra la eficiencia de clasificación alcanzada por el algoritmo clasificador, la columna Efic. DAGA muestra la eficiencia de clasificación obtenida a través de la metodología DAGA, Caract. muestra el número de características orignales mientras que en la columna Caract. DAGA se muestran las características seleccionadas después del uso de la metodología DAGA.

Tabla 117: Resultados de la experimentación con la metodología DAGA

Conjunto de datos	Algoritmo	Efic.	Efic. DAGA	Caract.	Caract. DAGA
LSVT_voice	SVM	86.51 %	96.03 %	310	22
Arritmia	RF	70.95 %	81.19 %	278	37
Semeion	SVM	93.85 %	95.17 %	256	131
Promoters	RF	88.68 %	96.23 %	57	6
Lung_cancer	SVM	81.48 %	96.30 %	56	4
Dermatology	RF	97.49 %	98.88 %	34	14
Ionosphere	RF	93.16 %	95.73 %	34	16
Sick euthyroid	RF	97.25 %	97.35 %	24	3
Parkinsons	RF	91.79 %	95.38 %	22	6
Spect	RF	82.02 %	87.27 %	22	7
Spect_f	RF	80.15 %	88.39 %	44	6
Movement_libras	RF	83.89 %	86.11 %	90	24
Soybean	J48	97.87 %	100.00 %	35	2
Promedio		88.08 %	93.39 %	97.08	21.39

7. Conclusiones y trabajo a futuro

En este trabajo de tesis se presenta en forma detallada la información necesaria para el correcto entendimiento y aplicación de la metodología DAGA.

Se demostró la utilidad de la metodología DAGA a través de una serie de experimentos .

Como conclusiones particulares se obtuvo lo siguiente:

- Además de la metodología DAGA, se propuso un nuevo algoritmo genético el AGD, este algoritmo demostró ser un algoritmo competitivo comparado con algoritmos genéticos simples parecidos.
- La metodología DAGA probó de forma experimental ser una opción viable para proporcionar un subconjunto de características que ofrezcan una elevada eficiencia de clasificación.

En lo relacionado al AGD se obtuvieron las siguientes conclusiones:

- Al realizar la exploración del espacio de búsqueda en conjuntos de datos medianos, se pudo constatar que las soluciones entregadas por el AGD estuvieron muy cercas o fueron incluso las optimas en los diferentes conjuntos de datos.
- Al realizar la exploración del espacio de búsqueda se mostró que las soluciones entregadas por el AGD mejoran al explorar mayor espacio de búsqueda. Sin embargo esto no se cumple con todos los algoritmos, como muestra con el algoritmo genético AGS2 al 26 % no genera mejores soluciones a medida que el espacio de exploración se aumenta.

- En conjuntos de datos grandes el AGD al 24 % demostró ser superior encontrando en general soluciones con mayor eficiencia de clasificación e incluso soluciones con una mayor reducción de características.

Al finalizar el trabajo de tesis, diversas preguntas quedan abiertas para un posterior análisis. Algunos de los puntos aconsejados para continuar el presente trabajo de tesis son los siguientes:

- Se propone el utilizar la metodología DAGA con diferentes algoritmos de clasificación.
- Se propone el uso de la metodología DAGA con algún algoritmo diferente pasa seleccionar el clasificador a usar de una forma mas eficiente.
- Se propone el uso de diferentes métodos para probar la eficiencia de clasificación e integrarlos en la metodología DAGA.

8. Apéndice A: Uso de la herramienta

En la Figura 51 se muestra la interfaz principal de la aplicación. Esta interfaz se penso para ser intuitiva y fácil de manipular por cualquier persona.

Los patrones deben tener un ejemplo de clasificación por cada línea, una característica por columna y la última columna debe ser la clase a la que pertenece, todos los datos deben estar separados por comas, el programa solo acepta valores reales. Esto es de acuerdo con las restricciones del clasificador usado, el CHAT.

En el campo número de corridas, se debe ingresar la cantidad de veces que se quiere ejecutar el algoritmo.

Se pueden utilizar 5 tipos de algoritmos para realizar la selección de características:

- BTSP: Algoritmo genético simple con método de selección Binary Tournamet y método de reproducción Single point crossover
- BTDP: Algoritmo genético simple con método de selección Binary Tournamet y método de reproducción Double point crossover
- SRSSP: Algoritmo genético simple con método de selección Stochastic Reminder Sampling y método de reproducción Single point crossover
- SRSDP: Algoritmo genético simple con método de selección Stochastic Reminder Sampling y método de reproducción Double point crossover
- AGD: Algoritmo genético dominante de acuerdo a lo explicado en la sección 4.2.1 del presente trabajo

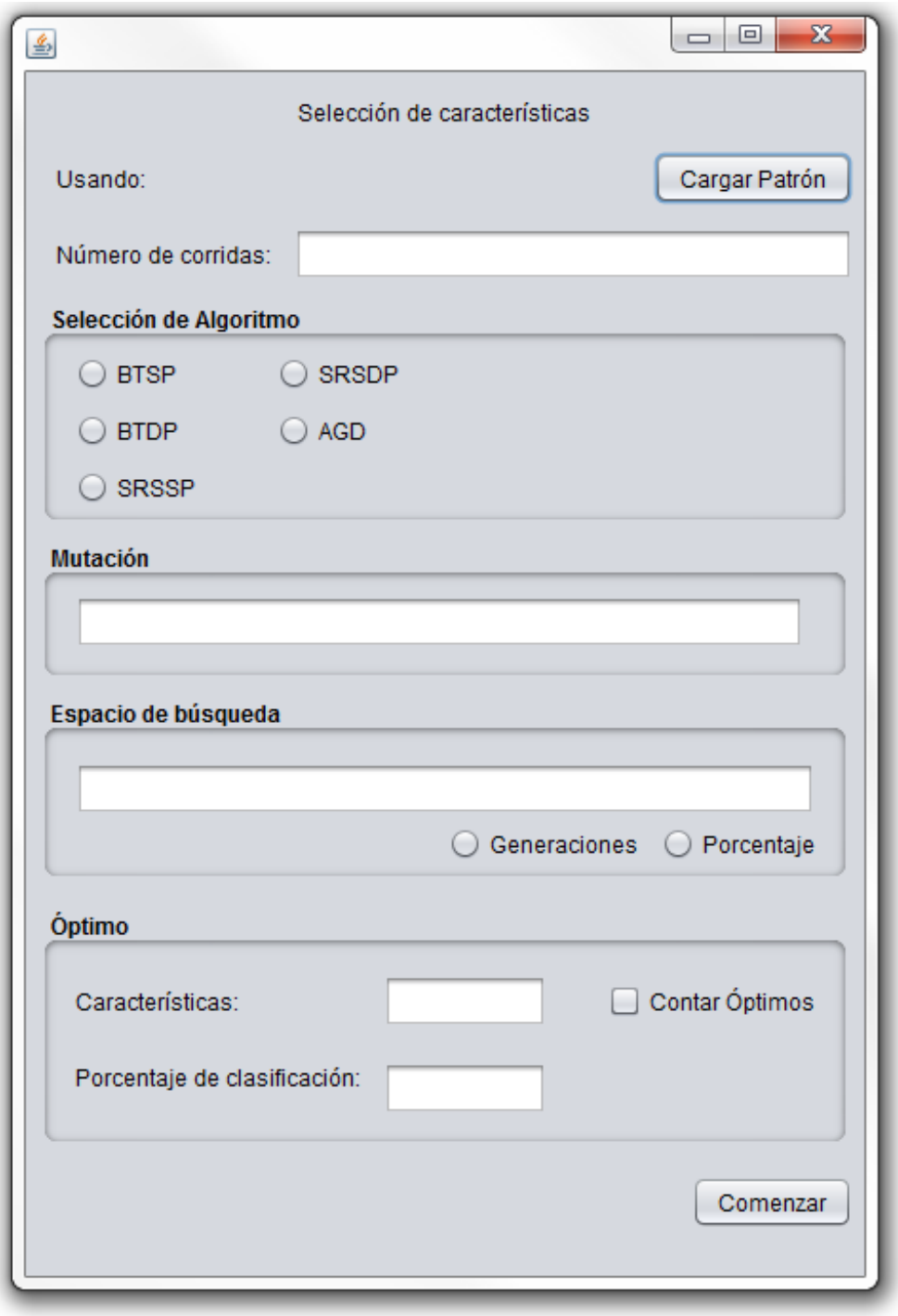

Figura 51: Interfaz principal

Se pueden seleccionar 2 tipos de limitación para el espacio de búsqueda, ya sea por número de generaciones o por el porcentaje de espacio de búsqueda requerido.

En el campo de mutación, se puede ingresar un único valor de mutación con el que se desea que trabaje el algoritmo, o se pueden ingresar diversas mutaciones, separadas por comas, el programa iterará sobre todas las mutaciones, presentando resultados para cada una.

En el campo de Espacio de búsqueda, se puede limitar la cantidad de ejemplos que el algoritmo probará antes de llegar a una solución, esta limitación se puede realizar ya sea a través de generaciones o de porcentajes. En caso de seleccionar generaciones el dato de entrada debe ser entero, en caso de porcentaje el dato de entrada puede ser flotante. En caso de analizar más de un valor ya sea en generaciones o porcentaje, se realizarán las iteraciones correspondientes, estos valores en caso de ser más de uno deben ser insertados separados por comas.

En caso de conocer los valores óptimos de la selección de características y querer saber cuántas veces se encuentran estos valores en las pruebas realizadas, se usa la sección de óptimos, aquí se ingresan los valores óptimos del porcentaje de clasificación y número de características.

Referencias

[1] I. Guyon, *Feature Extraction, Foundations and Applications*, I. Guyon and S. G. Masoud, Eds. Springer, 2006.

[2] M. Aldape-Pérez, C. Yáñez-Márquez, O. Camacho-Nieto, and Á. Ferreira-Santiago, "Feature selection using associative memory paradigm and parallel computing," *Computación y Sistemas*, vol. 17, no. 1, pp. 41–52, 2013.

[3] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Computers in Biology and Medicine*, vol. 39, no. 1, pp. 69 – 78, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010482508001674

[4] R. Jensen and Q. Shen, *Computational intelligence and feature selection: rough and fuzzy approaches*. John Wiley & Sons, 2008, vol. 8.

[5] H. Liu and H. Motoda, *Computational methods of feature selection*, H. Liu and H. Motoda, Eds. Taylor & Francis Group, 2008.

[6] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 1992, vol. 1.

[7] C. Darwin and W. F. Bynum, *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt, 2009.

[8] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[9] M. Mitchell, *An introduction to genetic algorithms.* MIT press, 1998.

[10] J. Lu, T. Zhao, and Y. Zhang, "Feature selection based-on genetic algorithm for image annotation," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 887–891, 2008.

[11] K. A. De Jong, *Evolutionary computation: a unified approach.* MIT press, 2006.

[12] R. Leardi and A. L. González, "Genetic algorithms applied to feature selection in {PLS} regression: how and when to use them," *Chemometrics and Intelligent Laboratory Systems*, vol. 41, no. 2, pp. 195 – 207, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169743998000513

[13] H. Handels, T. RoSS, J. Kreusch, H. Wolff, and S. Pöppl, "Feature selection for optimized skin tumor recognition using genetic algorithms," *Artificial Intelligence in Medicine*, vol. 16, no. 3, pp. 283 – 297, 1999, 1999.pdf. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365799000056

[14] J.-H. Hong and S.-B. Cho, "Efficient huge-scale feature selection with speciated genetic algorithm," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 143 – 150, 2006, 2006.pdf. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865505002035

[15] A. Özçift and A. Gülten, "Genetic algorithm wrapped bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases," *Digital Signal Processing*, vol. 23, no. 1, pp. 230 – 237, 2013, 2013 Erithematho.pdf. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200412001625

[16] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 2052 – 2064, 2014, 2014 Credit risk.pdf. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413007239

[17] C. D. Stefano, F. Fontanella, C. Marrocco, and A. S. di Freca, "A ga-based feature selection approach with an application to handwritten character recognition," *Pattern Recognition Letters*, vol. 35, no. 0, pp. 130 – 141, 2014, frontiers in Handwriting Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513000391

[18] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," *Urbana*, vol. 51, pp. 61 801–2996, 1991.

[19] R. Sivaraj and T. Ravichandran, "A review of selection methods in genetic algorithm," *International journal of engineering science and technology*, vol. 3, no. 5, pp. 3792–3797, 2011.

[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees (wadsworth, belmont, ca, 1984)," in *Proceedings of the Thirteenth International Conference, Bari, Italy*, 1996, p. 148.

[21] K. Bache and M. Lichman, "Uci machine learning repository," *URL http://archive. ics. uci. edu/ml*, vol. 19, 2013.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.