INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

TESIS

Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente

PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

ING. MARIANA DAYANARA ALANIS TAMEZ

DIRECTORES DE TESIS:

DR. CORNELIO YÁÑEZ MÁRQUEZ

DRA. YENNY VILLUENDAS REY

Ciudad de México

Junio 2018

SIP-14 bis

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 14:00 horas del día 28 del mes de febrero de 2018 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente"

Presentada por el alumno:

ALANIS	**TAMEZ**	**MARIANA DAYANARA**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

B	1	6	0	6	2	6

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA
Directores de Tesis

Dr. Cornelio Yáñez Márquez

Dra. Yenny Villuendas Rey

Dr. Oleksiy Pogrebnyak

Dr. Marco Antonio Moreno Ibarra

Dra. Guohua Sun

Dr. Amadeo José Argüelles Cruz

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Marco Antonio Ramírez Salinas

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México el día 7 del mes Junio del año 2018, el (la) que suscribe Alanis Tamez Mariana Dayanara alumno (a) del Programa de Maestría en Ciencias de la Computación con número de registro B160626, adscrito a Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Dr. Cornelio Yáñez Márquez y Dra. Yenny Villuendas Rey y cede los derechos del trabajo intitulado Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección malanis93@hotmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Mariana Dayanara Alanis Tamez

Nombre y firma

Resumen

Las técnicas de computación inteligente aplicadas a la medicina se han convertido en un área creciente de investigación en todo el mundo. Entre ellos, la aplicación y desarrollo de nuevos modelos y algoritmos para el diagnóstico y predicción de enfermedades han sido un tema de investigación activo. En el presente trabajo de tesis se realiza un estudio del desempeño de los algoritmos de clasificación de patrones más utilizados de acuerdo a la literatura y su aplicación en el prediagnóstico de enfermedades crónicas.

La contribución a la investigación del presente trabajo es la propuesta de un nuevo modelo de clasificación y su aplicación al prediagnóstico de enfermedades crónicas. El nuevo modelo disminuye los efectos indeseados de la maldición de la dimensionalidad; trata la presencia de pequeños disjuntos detectando subclases en las clases; trata con datos desbalanceados y maneja datos mezclados y perdidos; además, de que trata con el solapamiento de clases. El modelo propuesto es interpretable y transparente, debido a que se sabe exactamente por qué una instancia pertenece a una determinada clase. Los resultados experimentales obtenidos fueron validados para identificar diferencias significativas en el rendimiento por medio de pruebas estadísticas y *post-hoc*.

Abstract

Classification models applied to medicine have become an increasing area of research worldwide. Such as, the application and development of known models and algorithms for disease diagnosis and prediction have been an active research topic.

The present work is a performance study of pattern classification algorithms most used in the literature, and its application to the pre-diagnosis of chronic diseases.

The contribution to the research of the present work is the proposal of a new classification model and its application to the diagnosis of chronic diseases. The new model dismisses the undesired effects of the curse of dimensionality; it deals with the presence of small disjoints detecting subclasses in classes; it deals with imbalanced data, and handles mixed as well as incomplete data; besides, it deals with class overlapping. In addition, the proposed model is interpretable and transparent it is so because we know exactly why an instance belong to a certain class. The experimental results obtained were validated to identify significant differences in performance through statistical and post-hoc tests.

Agradecimientos

Agradezco al Centro de Investigación en Computación, al Instituto Politécnico Nacional, CONACyT y a mi familia por el apoyo brindado en este proyecto.

A mis padres, porque sin su invaluable apoyo no estaría donde me encuentro ahora, muchas gracias.

A mis hermanas por su incondicional apoyo y amistad.

A mi asesor en la Universidad de Bristol, en Inglaterra, Reino Unido, el profesor Peter Flach, por su apoyo y enseñanzas.

Índice General

Índice de Tablas

Índice de Figuras

Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente

Capítulo 1. Introducción

En el presente capítulo da una introducción al presente trabajo en donde se sintetiza el contenido del documento, las razones que dieron origen al trabajo, el problema a resolver y su importancia; en él se incluyen los antecedentes, la justificación, el objetivo tanto general como los específicos, la contribución científica y la estructura general del documento.

Para el diagnóstico médico, las aplicaciones computacionales se han desarrollado principalmente en dos áreas: desarrollo de algoritmos y desarrollo de sistemas para dar soporte al diagnóstico médico. El impacto social de las enfermedades crónicas en la población es uno de los temas actuales de la investigación científica en todo el mundo [1].

Los esfuerzos realizados por los grupos de investigación son notables, con gran interés por minimizar los efectos negativos de este tipo de enfermedades. Para los profesionales implicados en cuestiones de salud pública, es evidente que el diagnóstico temprano de una enfermedad crónica en un paciente aumenta sus posibilidades de supervivencia o en su caso aumenta su calidad de vida [2].

En este contexto, las técnicas de Cómputo Inteligente aplicadas a la medicina se han convertido en un área de investigación cada vez mayor en todo el mundo y, la aplicación y el desarrollo de nuevos modelos y algoritmos para el diagnóstico o la predicción de enfermedades es un tema de investigación activo. Un análisis cuidadoso de estas técnicas, modelos y algoritmos permite darse cuenta de que tienen puntos débiles y fallas que vale la pena enfrentar. Por ejemplo, algunos modelos dependen de la disponibilidad de un cierto tipo de datos [3]; otros son modulares y tienen la desventaja de que si alguno de los módulos falla, el diagnóstico no se llevará a cabo [4]; algunos más se comportan como “cajas negras”, lo que hace imposible determinar qué instancias se clasificaron incorrectamente y por qué [5]; algunos, aunque útiles para ayudar al diagnóstico tempranos correctos, requieren datos de procesamiento previo, modificando así el banco de datos original [2].

En un esfuerzo por intentar minimizar las desventajas descritas anteriormente, en el presente trabajo se propone un nuevo modelo de clasificación, diseñado para el prediagnóstico de enfermedades

crónicas. El nuevo modelo disminuye los efectos indeseados de la maldición de la dimensionalidad; trata la presencia de pequeños disjuntos detectando subclases en las clases; trata con datos no balanceados y maneja datos mezclados y perdidos; además, trata con el solapamiento de clases.

El modelo propuesto es interpretable y transparente, debido a que se sabe exactamente por qué una instancia pertenece a una determinada clase. Esta es una clara ventaja sobre otros clasificadores del estado del arte los cuales serán mencionados en el capítulo 2.

1.1. Antecedentes

De acuerdo a la Organización Mundial de la Salud [6], las enfermedades crónicas son consideradas como enfermedades de larga duración y generalmente de progresión lenta, es decir, que se va desarrollando y avanzando la enfermedad lentamente a través del tiempo. Este tipo de enfermedades no se dan porque si, sino que se ven frecuentemente heredadas o dependiendo de la comunidad en la que se habite, esto como consecuencia de diversos factores ambientales o alimenticios que interactúan con un perfil genético vulnerable. Algunos ejemplos de estas enfermedades son: enfermedades cardíacas o respiratorias, cáncer, diabetes, entre otras; dichos ejemplos son las principales causas de mortalidad en el mundo, siendo responsables del 63% de las muertes. En 2008, 36 millones de personas en todo el mundo aproximadamente murieron a causa de una enfermedad crónica, de las cuales el 50% era de sexo femenino y el 29% era de menos de 60 años.

En perspectiva de las ciencias de la computación, diversos trabajos han abordado el tema en general de las enfermedades crónicas, entre los trabajos más relevantes se encuentran los siguientes.

Chang et al. [3] desarrollaron un sistema web de apoyo de decisiones que considera principalmente el análisis de sensibilidad, así como las decisiones previas y posteriores óptimas y necesarias para el diagnóstico de algunas enfermedades crónicas, en este trabajo se enfocan en la enfermedad granulomatosa crónica. Este sistema toma como base varios factores, entre ellos el Teorema de Bayes para integrar las opiniones de los expertos y la información obtenida por el sistema, esto proporcionará al personal experto una base para tomar de decisiones de calidad en cuanto al diagnóstico médico. Como ventaja se tienen los costos computacionales que rodean la toma de decisiones, a pesar de ello, una clara desventaja es el uso de un solo criterio, como el Teorema de Bayes, para el apoyo en la toma de decisiones, es necesario tener más opciones para finalmente determinar cuál es el más adecuado para este tipo de problemas.

Havlik et al. [4] proponen una solución para el desarrollo rápido de dispositivos para aplicaciones tele-médicas, monitoreo remoto y tecnologías de asistencia. El enfoque utilizado fue diseñar y realizar un sistema modular compuesto por módulos de entrada para adquisición de señales, configuración del medio de comunicación de datos y de control del sistema, unidad de control para preprocesamiento de señales y una interfaz de usuario; con lo anterior se puede tener una manera clara de cómo está desarrollado el sistema lo que evitaría ambigüedades y generaría más confianza tanto en los pacientes como en los expertos en el tema; sin embargo, si alguno de los módulos falla, el diagnóstico no podría llevarse a cabo con la ayuda de este sistema.

Rijo et al. [5] utilizaron un método de minería de texto para respaldar las decisiones médicas relacionadas con el diagnóstico de epilepsia y la clasificación basada en el código ICD-9 en niños, este código se usa para describir el diagnóstico de un paciente, incluidos síntomas, enfermedades o trastornos y contribuye para una comprensión común e interpretación de un diagnóstico. La clasificación se plantea a partir de un método de minería de texto utilizando registros médicos electrónicos y aplicando el modelo del “k” vecino más cercano, a pesar de la efectividad mostrada en el trabajo de Rijo no es posible determinar exactamente qué instancias se clasificaron incorrectamente y por qué; además de eso, el autor propone como único trabajo futuro ampliar la cantidad de datos analizar lo cual no garantiza mayor rendimiento en los resultados.

María Sanchez-Santana et al. [7] introdujeron otro enfoque computacional para el diagnóstico médico. Proponen un nuevo entorno de tele-diagnóstico para la detección de problemas cardiovasculares y permiten al personal médico identificar y cuantificar semiautomáticamente las posibles complicaciones cardiovasculares de un paciente. Esta herramienta utiliza principalmente análisis de imágenes para detectar posibles anomalías cardiovasculares, sin embargo, aunque esta herramienta proporciona valiosa información para los médicos especialistas, esta no realiza un diagnóstico como tal de una enfermedad en concreto y a su vez no menciona la efectividad que pudiera tener contra otros sistemas o herramientas similares.

Aiping Lu et al. [8] utilizaron como base la Medicina Tradicional China (*TCM*) como una guía en la clasificación y diagnóstico biomédico. En pocas palabras, la *TCM* significa que "la energía vital del cuerpo (*chi* o *qi*) circula a través de canales, llamados meridianos, que tienen ramas conectadas a órganos y funciones corporales" [9]. En el trabajo de Aiping [8], evalúa principalmente la eficacia de la práctica de *TCM* utilizando clasificación de patrones para el diagnóstico de enfermedades biomédicas, y la base

biológica del patrón *TCM*, aunque este trabajo asegura que puede conducir a nuevos hallazgos en ciencias biológicas, no se menciona que tan certero es el diagnóstico de enfermedades ni contra que otros sistemas fue comparado.

Cabe señalar que cada una de las soluciones descritas anteriormente tiene desventajas y limitaciones con las consecuencias negativas en los diagnósticos que producen. Por lo tanto, es necesario llevar a cabo investigaciones, como el presente trabajo de tesis, donde se proponen soluciones que superan las desventajas y limitaciones de los modelos actuales.

1.2. Justificación

El ritmo de vida de los seres humanos ha cambiado drásticamente en los últimos años; los niveles de contaminación ambiental y la alimentación han influido drásticamente en la salud de la población mundial. A pesar de que la esperanza de vida ha aumentado considerablemente en los últimos 50 años debido a los grandes avances de la medicina [10], un estudio denominado “Carga Global de Enfermedad de estudios” realizado en 2016 [11], indica que las enfermedades infecciosas, materno-infantiles y la desnutrición causan menos muertes y menos enfermedades; como resultado, menos niños mueren cada año, pero los adultos más jóvenes y de mediana edad están muriendo y sufriendo de enfermedades denominadas “crónicas”; estas se convierten en las causas predominantes de muerte en el mundo. Este estudio indica que, a partir de 1970, los hombres y las mujeres en todo el mundo han ganado un poco más de diez años de esperanza de vida en general, pero pasan más años viviendo enfermos.

El diagnóstico oportuno de este tipo de enfermedades, hasta ahora incurables, ayudará a mejorar la calidad de vida de los pacientes y aportará información útil para el desarrollo de una cura.

1.3. Objetivos

Los objetivos planteados para el presente trabajo son los siguientes:

Objetivo General

Proponer un modelo de clasificación inteligente de patrones capaz de lidiar con datos mezclados, valores perdidos, no balanceados, con clases solapadas y con presencia de pequeños disjuntos, para el prediagnóstico de enfermedades crónicas.

Objetivos Específicos

Los objetivos específicos son los siguientes:

- Recolectar bancos de datos que abarquen las enfermedades crónicas más comunes.
- Estudiar de manera experimental los algoritmos clásicos de clasificación inteligente de patrones.
- Evaluar si el nuevo modelo propuesto es adecuado para el prediagnóstico de enfermedades crónicas.
- Realizar pruebas de significancia estadística que permitan valorar el desempeño del modelo propuesto con respecto a los del estado del arte.

1.4. Contribuciones Científicas

Las contribuciones del presente trabajo son las siguientes:

- La evaluación del comportamiento diversos algoritmos de clasificación sobre datos médicos que abarquen las enfermedades crónicas más comunes.
- Un nuevo modelo de clasificación inteligente de patrones que compita fuertemente con los algoritmos clásicos de clasificación inteligente de patrones.
- Algoritmos evolutivos, como la Evolución Diferencial (*Differential Evolution*) aplicados al nuevo modelo para la selección de rasgos.

1.5. Estructura del documento

El presente trabajo está constituido por seis capítulos en los que se realiza un estudio del desempeño de los algoritmos de cómputo inteligente más utilizados de acuerdo a la literatura; y a partir de eso se propone un nuevo modelo de clasificación de patrones.

En el presente capítulo se presentó una introducción del presente trabajo en él se incluyen los antecedentes, justificación, objetivos y contribuciones de este trabajo de tesis, mientras que el resto del documento está organizado de la siguiente manera:

Estos algoritmos se aplicaron al prediagnóstico de enfermedades crónicas; más específicamente, se hizo una comparación del desempeño de algoritmos clásicos y del nuevo modelo propuesto sobre varios bancos de datos que abarcan las enfermedades crónicas más comunes. Los diferentes bancos de datos tienen principalmente: datos no balanceados, mezclados, categóricos, numéricos y valores faltantes (o perdidos).

El capítulo 2 contiene el estado del arte el cual incluye los más recientes trabajos e investigaciones relacionadas con el presente trabajo. El capítulo 3 contiene los materiales y métodos en dónde se destacan los algoritmos clásicos de cómputo inteligente, así como la construcción correcta de los conceptos que darán soporte al trabajo. En el capítulo 4 se da una propuesta de investigación en la cual se da una solución a la problemática principal. En el capítulo 5 se incluyen los resultados del objeto de estudio, procesos experimentales y productos finales y se comentan e interpretan los resultados con relación al objetivo propuesto. El capítulo 6 se da una conclusión experimental relacionada con los objetivos originales además de los trabajos a futuro que pudieran surgir a partir de este trabajo.

Finalmente se incluye el listado de las referencias bibliográficas correspondiente en su totalidad a fuentes de información utilizadas en el presente trabajo.

Capítulo 2. Estado del Arte

En este capítulo se incluyen los más recientes trabajos e investigaciones relacionadas con técnicas de computación inteligente aplicadas al diagnóstico médico.

Las enfermedades crónicas se han vuelto un problema creciente en los últimos años; un estudio reciente en China [12] indica que este tipo de enfermedades ahora representan aproximadamente el 80% de las muertes y el 70% de los años de vida perdidos por discapacidad en China. Las enfermedades cardiovasculares y el cáncer son las principales causas de muerte esto debido a la exposición a diferentes factores de riesgo como: la contaminación, el calentamiento global, los malos hábitos alimenticios, entre otros.

Actualmente, un diagnóstico temprano es necesario para atender la creciente carga de las enfermedades crónicas, además la mayoría de estas se desarrollan rápidamente causando muertes y mermando la calidad de vida de las personas que las padecen.

Para el diagnóstico médico, se han propuesto varias soluciones algorítmicas en los últimos años, entre las más destacadas se encuentran:

Fazekas [1] abordó la periodicidad de la leucemia infantil en Hungría utilizando series de tiempo estacionales. El análisis de la estacionalidad de la leucemia linfoide infantil en Hungría se realizó tanto en el número total de pacientes como en las series de datos divididas. Los autores encontraron una cierta periodicidad en las fechas del diagnóstico en pacientes con leucemia. Aunque hubo alguna diferencia en los patrones de los picos de componentes estacionales de las tres series de tiempo, la mayoría de los picos cayeron dentro de los meses de invierno en las tres series de tiempo evaluadas. Esto fue más significativo en el grupo de todos los pacientes y en el grupo de edad más joven. La ventaja de lo anterior es que los resultados de los análisis demostraron la ocurrencia estacional de la leucemia linfoide infantil en Hungría. Estos resultados se refieren a que en los meses de invierno ciertos efectos ambientales o ciertas infecciones virales, pueden provocar la manifestación de la enfermedad. El conjunto de datos fue del Grupo de Trabajo de Oncología Pediátrica Húngara y contenía los datos de todos los pacientes con leucemia linfoide diagnosticados entre 1988 y 2000. En este intervalo de tiempo se registraron un total de 814 niños húngaros (de los cuales 467 eran varones). Los pacientes tenían entre 0 y 18 años, con una edad media de 6 a 4 años y una mediana de 4 a 5 años.

Los componentes de las series de tiempo se identificaron y aislaron utilizando paquetes de programas estadísticos con sus respectivos valores preestablecidos.

Por otro lado, en los resultados no hubo evidencia de estacionalidad en el diagnóstico de la leucemia linfoblástica aguda. Como conclusión, los datos evaluados destacan el papel de los efectos ambientales, como infecciones virales, epidemias, entre otros; solamente en el inicio de la enfermedad y en ciertos periodos de tiempo lo cual la hace ineficiente si se planea evaluar en otras circunstancias diferentes a las ya mencionadas.

Por lo anterior, Fazekas [1] sugiere evaluar otro tipo de datos mediante otro tipo de estudios como trabajo a futuro con la finalidad de mejorar los resultados obtenidos.

Dada la importancia de las enfermedades crónicas y el número de personas que las padecen, Abdar et al. [2] estudió en la enfermedad hepática mediante el uso de dos métodos en el área de minería de datos. El hígado humano es uno de los principales órganos del cuerpo humano y la enfermedad hepática puede causar muchos problemas en la vida del ser humano. El hígado es uno de los órganos más grandes del cuerpo humano (LDCR, 2015). El hígado está en la parte superior derecha del abdomen y debajo del diafragma [13]. El hígado realiza la purificación de la sangre e identifica sustancias tóxicas como el alcohol y las excreta. Hay muchos tipos de problemas hepáticos que causan más de 100 enfermedades diversas, como: hepatitis neonatal, cirrosis biliar primaria, hígado graso, cirrosis, cáncer de hígado, colangitis esclerosante primaria, hepatitis, porfiria, Síndrome de Reye, enfermedad hepática en el embarazo, Sarcoidosis, Hepatitis Tóxica, Enfermedad de Almacenamiento de Glucógeno Tipo 1, Tirosinemia y Enfermedad de Wilson, entre otras [13]. La Fundación Canadiense del Hígado (*CLF*) publicó un informe en 2013 que muestra que, durante los últimos 8 años, aproximadamente el 30% de las muertes en ese país se deben a una enfermedad hepática. De acuerdo a las estimaciones de esta fundación, una persona de cada 10 sufre uno de los diversos tipos de enfermedad hepática, lo que significa que alrededor de 3 millones de personas en Canadá tienen algún tipo de enfermedad hepática [14].

Una predicción rápida y precisa de la enfermedad hepática permite tratamientos tempranos y efectivos. En este sentido, varias técnicas de minería de datos han ayudado a una mejor predicción de esta enfermedad. Debido a la importancia de la detección de la enfermedad hepática y al aumento del número de personas que la padecen, los autores estudian la enfermedad hepática mediante el uso de dos métodos bien conocidos en el área de minería de datos. En este estudio, se consideran 583 instancias de un banco de datos de enfermedades hepáticas del Repositorio de Aprendizaje Automático de la Universidad de California en Irvine (*UCI-MLR)* [15]. En este trabajo dos algoritmos llamados

"*Boosted* C5.0" [16] y "*CHAID*" [17] analizan el banco de datos, los resultados y sus respectivas comparaciones con otros métodos. Los resultados de esta evaluación muestran que los algoritmos "*Boosted* C5.0" y "*CHAID*", ambos basados en árboles de decisión, son capaces de producir reglas para la enfermedad hepática. Los resultados también muestran que "*C5.0*" considera en su análisis el género en la enfermedad hepática, un factor que falta en muchos otros estudios. Mientras tanto, utilizando las reglas generadas por el algoritmo "*C5.0*", obtuvieron un resultado importante sobre la baja susceptibilidad de la enfermedad hepática en mujeres.

Timothy J. W. Dawes et al. [18] evalúan y determinan si la supervivencia del paciente evaluado y los mecanismos de insuficiencia ventricular derecha en la hipertensión pulmonar podrían predecirse mediante el uso de aprendizaje automático supervisado de patrones tridimensionales del movimiento cardíaco sistólico. La ventaja de este trabajo es que utiliza como base otro tipo de factores, como el movimiento cardíaco sistólico, para identificar síntomas que pudiesen dar como resultado el diagnóstico de la hipertensión pulmonar; algo que amplía el panorama de síntomas y factores que puedan llevar a una persona a ser diagnosticada con esta enfermedad; sin embargo, utiliza únicamente datos de imágenes basadas en el funcionamiento cardiovascular y ninguna basada en los pulmones, que es el propósito original de estudio. Se considera que debería tomar ambas fuentes de información, tanto cardiovascular como pulmonar para tener un diagnóstico más certero y resultados más confiables.

Yu et al. [19] introdujeron un sistema de diagnóstico de osteoporosis con la ayuda de las redes neuronales artificiales (*RNA*). Las características se extrajeron mediante la observación de imágenes de rayos "X" y los principales síntomas clínicos de pacientes con osteoporosis por tres especialistas y tres radiólogos experimentados. Un porcentaje de los pacientes evaluados se seleccionó al azar como el conjunto de entrenamiento y el resto se considera como el conjunto de predicción. Los resultados están relacionados directamente con las características de la osteoporosis. Los resultados de diagnóstico del modelo de la red neuronal se comparan con los resultados de la regresión logística. A pesar de que en este estudio se observan resultados satisfactorios, no se especifica la medida de comparación, es decir, no queda claro que tomaron como base para afirmar porque un algoritmo sobresale con respecto al otro. En este estudio, todavía hay algunas limitaciones. Por un lado, el número de casos seleccionados es un pequeño y consecuente, además, con base a lo que mencionan los autores, el modelo establecido no es estable ya que la red neuronal fue entrenada pocas veces.

Imamura et al. [20] utilizaron un árbol de decisión para determinar el límite seguro de realizar una cirugía de cáncer de hígado basado en tres variables: si la ascitis (acumulación de líquido seroso) está presente, el nivel de bilirrubina total y la tasa de retención de indocianina a los 15 minutos. Este modelo se aplicó en pacientes que muestran un signo de cirrosis descompensada debido a un valor elevado de bilirrubina o ascitis incontrolable, en el banco de datos evaluado la hepatectomía no está indicada lo cual lo convierte en primera instancia en una desventaja ya que el modelo podría clasificar de manera errónea con más facilidad; a pesar de esto tuvo resultados satisfactorios en la detección de los valores elevados de bilirrubina y la detección de ascitis.

Kanas et al. [21] utilizaron modelos de predicción basados en métodos de aprendizaje automático y se probaron con datos de la base de datos del Atlas del Genoma del Cáncer (*TCGA*). Los datos se obtuvieron teniendo en cuenta la metilación del promotor "O6-metilguanina-ADN-metiltransferasa" (*MGMT*) y se ha demostrado que se asocia con mejores resultados en pacientes con glioblastoma (GBM) y estos pueden ser un marcador predictivo de sensibilidad de las personas a la quimioterapia. La determinación del estado de metilación del promotor de MGMT requiere tejido obtenido mediante resección quirúrgica o biopsia. El análisis realizado demostró que la proporción de volumen de edema-necrosis, relación de volumen tumor-necrosis, volumen de edema y localización del tumor y características de mejora se asociaron con el estado de la metilación del promotor de MGMT en el glioblastoma, lo cual indican los autores, proporciona evidencia adicional de una asociación entre las características estándar de *MRI* preoperatoria y el estado de metilación de MGMT en glioblastoma. La clasificación fue realizada mediante un ensamble de clasificadores con los siguientes algoritmos: *random forest, "k"* vecinos más cercanos (utilizando la distancia Euclidea), Gaussian Naive Bayes y el árbol de decisión J48. Los resultados, de acuerdo los autores, son mejores comparados con otros sistemas similares, pero la cantidad de datos analizados fue muy grande lo cual significó un costo computacional muy alto.

Yu Sun et al. [22] utilizaron máquinas de soporte de vectores y evaluaron su rendimiento para predecir la localización de tumores en la próstata utilizando imágenes de resonancia magnética multiparamétricas (*MRI*). Se recolectaron datos de dieciséis pacientes antes de la prostatectomía radical, utilizaron un *kernel gaussiano* que fue entrenado y probado en diferentes bancos de datos de pacientes que sufren de este tipo de tumores. Los parámetros se optimizaron utilizando la validación cruzada *leave one out*. Aunque los autores indican que los resultados fueron relevantes, los datos con

los que se evaluó el desempeño del clasificador fueron muy pocos (en total veintiuno) por lo que los mismos autores proponen como trabajo a futuro agregar datos adicionales del paciente para aumentar la sensibilidad y la especificidad del modelo y, agregando otras características biológicas del tumor que se utilizarán en la optimización de la radioterapia siendo esta una ventaja para un diagnóstico certero, sin embargo no se garantiza el éxito de su propuesta hasta ser evaluada. Como clara desventaja solo se evaluó un solo clasificador, además, la cantidad de pacientes enfermos (dieciséis) es muy grande comparada con la de pacientes sanos (cinco) una clara muestra de desbalance de clases (con una razón de desbalance de 3.2), lo que puede afectar al rendimiento del clasificador y la fiabilidad de los resultados obtenidos.

Para finalizar este capítulo, Zheng et al. [23] propusieron un sistema para identificar pacientes con y sin diabetes *mellitus* tipo 2, estos datos provinieron de *Electronic Health Records.* El objetivo de este trabajo es desarrollar un sistema semi-automático basado en el aprendizaje automático para evitar los falsos positivos. Se evalúa el desempeño de diferentes clasificadores tales como "k" vecino más cercano, *Naïve Bayes*, Árboles de decisión, *Random Forest*, Máquinas de Soporte de Vectores y Regresión logística. Los datos recabados están compuestos de 300 muestras de pacientes (161 enfermos, 60 sanos y 79 sujetos no confirmados), estos datos fueron recuperados de un repositorio del Registro electrónico de Salud (EHR), y fueron obtenidos desde el año 2012 hasta 2014. Como ventaja a este trabajo es la cantidad de clasificadores evaluados, ya que a diferencia de los trabajos de otros autores mencionados en este capítulo, al ser más de uno ayuda a determinar con mayor facilidad cual fue mejor entre ellos, a pesar de esto, el banco de datos evaluado contiene más pacientes no identificados que sanos lo cual el clasificador podría confundir uno con otro a la hora de asignarles una etiqueta de clase, lo anterior dificultaría el diagnóstico certero de esta enfermedad y la confianza y efectividad de este estudio.

En este capítulo, se presentó una compilación de resultados de otras investigaciones que abarcan el diagnóstico o prediagnóstico de enfermedades principalmente crónicas utilizando técnicas de computación inteligente o clasificación inteligente de patrones. También se presentan ventajas y desventajas además lo que puede aportar a la investigación actual.

Capítulo 3. Materiales y métodos

En este capítulo, se destacan los algoritmos de cómputo inteligente seleccionados, así como la construcción correcta de los conceptos que darán soporte al trabajo.

Los temas abordados en este capítulo son: Conceptos fundamentales del reconocimiento de patrones; Algoritmos de clasificación inteligente de patrones el cual incluye los árboles de decisión, "k" vecino más cercano, regresión logística, perceptrón multicapa y las máquinas de soporte vectorial; y finalmente el algoritmo metaheurístico de evolución diferencial.

El reconocimiento de patrones [24] es la disciplina científica cuyo objetivo es la clasificación de los objetos en una serie de categorías o clases. Dependiendo de la aplicación, estos objetos pueden ser imágenes, formas de onda de señal o cualquier tipo de medidas que necesiten clasificarse.

Existen diferentes tipos de aprendizaje automático en el reconocimiento de patrones, estos se describen brevemente en la figura 1.

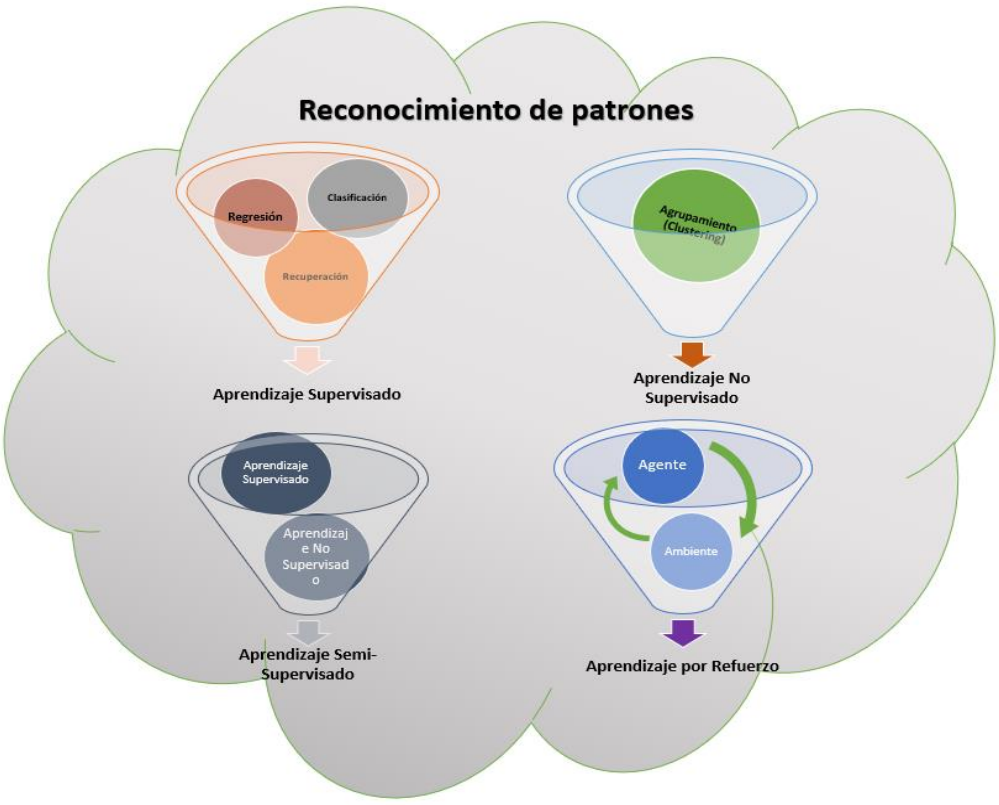

Figura 1. Tipos de aprendizaje automático

Aprendizaje supervisado [24]**:** En el aprendizaje supervisado, un experto proporciona una etiqueta para cada patrón en un conjunto de entrenamiento. El objetivo del aprendizaje supervisado es obtener el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos. De este tipo de aprendizaje es parte tres diferentes tareas.

Clasificación: Asigna a los patrones de entrada a una etiqueta de clase, basándose en las características extraídas.

Regresión: Se refiere al ajuste de curvas; típicamente se tiene una entrada y una salida. La diferencia de que la variable de salida es continua y no categórica como en el caso de la clasificación.

Recuperación: Es poco común, ya que sólo la realizan algunas redes neuronales y las memorias asociativas.

Aprendizaje no supervisado [24]**:** En este tipo de aprendizaje un modelo es ajustado a las observaciones; se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento previo de aprendizaje. En este tipo de aprendizaje el sistema forma “agrupaciones” dado un conjunto particular de patrones.

Agrupamiento (*Clustering*): Este procedimiento consiste en la agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud (generalmente la distancia euclídea).

Otros tipos de aprendizaje automático conocidos son los siguientes:

Aprendizaje semi-supervisado [24]**:** Toma las fortalezas del aprendizaje supervisado y no supervisado, además de que puede estar relacionado con cualquiera de las cuatro tareas del reconocimiento de patrones.

Aprendizaje por reforzamiento [24]**:** En el aprendizaje por reforzamiento, no da ninguna señal de categoría deseada; en cambio, indica si una categoría tentativa es correcta o incorrecta. Esto es análogo a que simplemente dice que algo está bien o mal, pero no dice específicamente porque está mal.

3.1. Algoritmos de clasificación inteligente de patrones

A continuación, se describen los algoritmos de clasificación inteligente de patrones propuestos para la evaluación de los bancos de datos del presente trabajo el cuál se enfoca en el aprendizaje supervisado y la tarea de clasificación. Estos algoritmos fueron seleccionados sobre la base de su capacidad de manejar datos mezclados o perdidos, esto con la finalidad de no alterar los datos proporcionados por los profesionales de la salud y no provocar confusión o desconfianza en cuanto a los resultados obtenidos.

Árboles de decisión

Los árboles de decisión [24] son capaces de clasificar un patrón a través de una secuencia de preguntas. Este enfoque de "preguntas" es particularmente útil para datos no métricos, ya que todas las preguntas pueden ser resueltas con un "sí/no", "verdadero/falso" o "valor (propiedad)", es decir, un conjunto de valores que no requieren ninguna noción de métrica.

Dicha secuencia de preguntas se muestra en el árbol de decisión direccionado o simplemente en árbol, donde por convención el nodo raíz se muestra en la parte superior, conectado por sucesivas (direcciones) enlaces o ramas a otros nodos. Estos están conectados de manera similar hasta enlazar a los nodos terminales o de hoja, que no tienen enlaces adicionales.

Para tareas de clasificación, se toma un patrón en particular y este comienza en el nodo raíz que, a su vez, solicita el valor de una propiedad particular del patrón.

Los diferentes enlaces desde el nodo raíz corresponden a los diferentes valores posibles. En los árboles que se encuentran en la literatura actual, todos los enlaces deben ser mutuamente distintos y exhaustivos, es decir, se seguirá un único enlace. El siguiente paso es tomar la decisión en el nodo subsecuente apropiado del subárbol, que puede considerarse la raíz de un subárbol. Cada hoja nodo lleva una etiqueta de categoría y al patrón de prueba se le asigna la categoría de la hoja nodo alcanzado.

Uno de los árboles de decisión y más utilizados es el *ID3*, este algoritmo recibió su nombre debido a que era el tercero en una serie de identificación o "*ID*" procedimientos. Está diseñado para ser utilizado solo con entradas nominales (no ordenadas). Si el problema involucra variables de valor real, primero se agrupan en intervalos, siendo cada intervalo tratado como un atributo nominal desordenado.

El árbol *C4.5*, el sucesor de ID3, es el más popular entre una serie de métodos de árbol de "clasificación". En él, se pueden tratar valores reales (a diferencia del ID3), este algoritmo usa heurística para "podar" basado en técnicas de significancia estadística.

"k" Vecino más cercano ("k" *Nearest Neighbour*)

El algoritmo del vecino más cercano (K *nearest neighbours*) clasifica nuevas instancias como la clase mayoritaria de entre los "k" vecinos más cercanos de entre el conjunto de entrenamiento.

En la fase de entrenamiento de este algoritmo este sólo guarda las instancias, no construye ningún modelo (a diferencia de los árboles de decisión), la clasificación se realiza en la fase de prueba.

En el presente trabajo y para este algoritmo, se utilizó la disimilitud de "HEOM" [25] cómo medida de distancia, esto debido a que maneja descripciones de datos mezclados y perdidos. La disimilitud de HEOM se calcula como:

$$HEOM(x,y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)}$$

$$d_a = \begin{cases} 1 \quad \text{if } x_a \text{ or } y_a \text{ si el dato es desconocido} \\ overlap(x_a, y_a) \text{ si el dato es categórico} \\ diff(x_a, y_a) \text{ si el dato es numérico} \end{cases}$$

$$overlap(x_a, y_a) = \begin{cases} 0 & \text{si } x_a = y_a \\ 1 & \text{en otro caso} \end{cases}$$

$$diff(x_a, y_a) = |x_a - y_a| / (max_a - min_a)$$

Nota: *Se considera que x_a, y_a son los valores de la característica a, para las instancias x y y, considera que max_a y min_a a son los valores máximo y mínimo de la característica a.*

Regresión Logística

La regresión logística [26] es una técnica de aprendizaje automático del campo de la estadística.

El modelo de regresión logística toma como entradas valores reales y hace una predicción sobre la probabilidad de que la entrada pertenezca a una clase determinada. Esta probabilidad es calculada con base en una función denominada "función logística".

La función logística, también llamada función sigmoidea, fue desarrollada por estadísticos para describir las propiedades del crecimiento de la población en la ecología, aumentando rápidamente y maximizando la capacidad de carga del medio ambiente; esta función está basada en la ecuación 1.

$$\frac{1}{1 + e^{-valor\ de\ entrada}} \quad \text{.....(1)}$$

Esta función resulta en una curva en forma de *S* que puede tomar cualquier número de valor real y asignarlo a un valor entre 0 y 1, tal como se muestra en la figura 2.

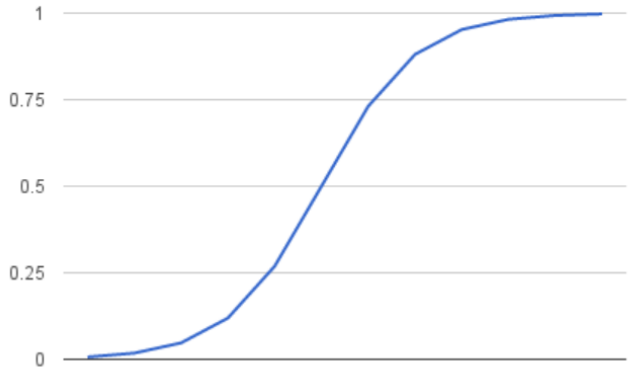

Figura 2. Función Logística

Este modelo tiene similitudes con otros modelos, ya que el resultado es real (como la regresión) pero limitado (como la clasificación).

La regresión logística usa una ecuación como representación. Los valores de entrada (x) se combinan linealmente utilizando pesos o valores de coeficientes (a los que se hace referencia como la letra mayúscula griega Beta) para predecir un valor de salida (y).

En la ecuación 2, se muestra un ejemplo de regresión logística:

$$y = e^{\frac{(b_0+b_1 \cdot x)}{(1+e^{(b_0+b_1 \cdot x)})}} \quad \text{.....(2)}$$

Donde y es el resultado predicho, b_0 es el sesgo y b_1 es el coeficiente para el valor de entrada único (x). Cada columna en sus datos de entrada tiene un coeficiente b asociado (un valor real constante) que debe aprenderse en la fase de entrenamiento.

La representación real del modelo que se almacenaría en la memoria o en un archivo son los coeficientes en la ecuación (el valor beta o b).

Perceptrón Multicapa (*Multilayer Perceptron*)

En 1969, Minsky y Papert [27], demostraron que tanto el perceptrón y el modelo “Adaline” no pueden resolver problemas no lineales (como la XOR). En 1986, Rumelhart [27] presenta la "Regla Delta Generalizada" para adaptar los pesos propagando los errores hacia atrás (*backpropagation)*, es decir, propagar los errores hacia las capas ocultas inferiores de la red. De esta forma se consigue trabajar con múltiples capas y con funciones de activación no lineales; en ese momento nace “el perceptrón multicapa” o *MLP* que es una red neuronal artificial (*RNA*) formada por múltiples capas, el *MLP* está

diseñado principalmente para la resolución de problemas que no son linealmente separables. Las capas pueden clasificarse en tres tipos:

- Capa de entrada: Son aquellas neuronas que contienen únicamente los patrones de entrada en la red.
- Capas ocultas: Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de siguientes capas.
- Capa de salida: Neuronas cuyos valores de salida corresponden a la etiqueta de clase.

Máquinas de soporte vectorial

Las máquinas de soporte vectorial (*SVM*) [24] son un conjunto de métodos de aprendizaje supervisados utilizados para la detección de clasificación, regresión y valores atípicos (o comúnmente llamados *outliers*).

Estos algoritmos son eficaces en espacios de alta dimensión incluso en casos donde el número de dimensiones es mayor que el número de instancias.

Las *SVM's* utilizan un subconjunto de patrones de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que también son eficientes desde el punto de vista de la memoria; se pueden especificar diferentes funciones del *kernel* para la función de decisión. Se proporcionan núcleos comunes, pero también es posible especificar *kernels* personalizados.

La función "*kernel*" puede ser cualquiera de las siguientes:

- **Lineal:** De la forma (x, x')
- **Polinomial:** $(\gamma(x, x') + r)^d \cdot d$ Siendo d el grado de la ecuación y r el coeficiente.
- **Base radial:** $(-\gamma||x, x'||^2) \cdot \gamma$ Gamma se especifica mediante γ, debe ser mayor que 0.
- **Sigmoide:** $(tanh(\gamma(x, x') + r))$, donde r es el coeficiente.

Esta sección, incorporó los algoritmos de clasificación inteligente de patrones seleccionados para la realización de este trabajo de investigación.

Se seleccionaron estos cinco algoritmos de clasificación de última generación. Todos ellos pueden tratar datos mezclados y perdidos.

3.2. Algoritmos Metaheurísticos

La computación inteligente "depende de datos numéricos suplidos por los fabricantes y no depende del conocimiento". Un sistema puede utilizar el término "computación inteligente" si trata datos de bajo nivel, como datos numéricos, si tiene un componente de reconocimiento de patrones y si no utiliza el conocimiento tan exacto y completo como la inteligencia artificial [28].

Existen cinco principios fundamentales [28] los cuáles son:

- **Lógica difusa:** Procesamiento del lenguaje.
- **Redes Neuronales:** Análisis de datos, clasificación, memorias asociativas, agrupamiento de datos y control. Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso humano. Se basa en el sistema de interconexión de "neuronas" que colaboran entre sí para producir un estímulo de salida.
- **Computación evolutiva:** Se basa principalmente en la optimización de una función dada. Se basa fundamentalmente en el comportamiento de sistemas biológicos o físicos
- **Teoría del aprendizaje:** Aprendizaje computacional. Se utilizan para lograr que los sistemas mejoren su desempeño al "aprender" de su entorno.
- **Métodos probabilísticos:** Evaluación de posibles soluciones, se utilizan para evaluar posibles soluciones y en muchos casos guían la toma de decisiones.

Para el presente trabajo, se utilizan algoritmos de computación evolutiva con la finalidad de optimizar los resultados obtenidos en la clasificación.

Evolución diferencial (*Differential Evolution - DE)*

La evolución diferencial (*Differential Evolution - DE*) [29] es un algoritmo perteneciente a la computación evolutiva, es decir, se utiliza principalmente para optimización de funciones. Este algoritmo, está diseñado de tal manera que pueda cubrir los siguientes requerimientos:

1. Capacidad para manejar funciones de costos no diferenciables, no lineales y multimodales.
2. Paralelizable, es decir, poder hacer frente al cálculo funciones de costo intensivo.
3. Facilidad de uso, es decir, pocas variables de control para dirigir la fase de optimización. Estas variables también deben ser robustas y fáciles de elegir.

4. Buenas propiedades de convergencia, es decir, convergencia consistente con el mínimo global en pruebas independientes consecutivas.

Los nuevos individuos de las poblaciones son generados a través de operadores de mutación, cruzamiento y selección siguiendo ese orden.

La Evolución diferencial es un algoritmo de búsqueda directa en el cual se utilizan n vectores de d dimensiones, donde n y d son números naturales fijos para algún problema en específico, por lo que dichos vectores para cada una de las diferentes generaciones G, se pueden denotar por la siguiente expresión:

$$Q_{i,j}^{G},$$

Donde:

$$i = 1,2,\cdots,n,$$
$$j = 1,2,\cdots,d$$

En la figura 3 se puede observar un ejemplo de $n = 5$ vectores pertenecientes a la generación $G = 1$, cada uno de los cuales tiene una dimensión $d = 6$.

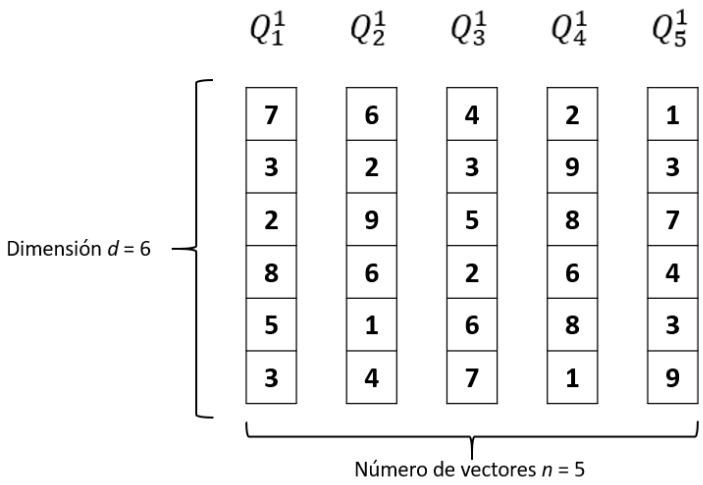

Figura 3. Ejemplo de Evolución Diferencial con la Generación 1

En el ejemplo anterior, cada uno de los vectores mostrados forma parte de una de las generaciones G, en las que, tanto los valores de n como los de d permanecen constantes durante todo el proceso, es decir, el número de atributos de cada uno de los vectores y los mismos vectores en cada una de las generaciones no se modifica en el transcurso del proceso.

La fase de inicialización con la que se llevará a cabo del proceso de optimización debe ser de manera aleatoria, posteriormente cada uno de los nuevos individuos de generaciones posteriores, se producirán por medio de la operación de mutación.

La mutación dentro de la Evolución Diferencial es un proceso en el que, a partir de una población existente de vectores, se producen nuevos individuos para generaciones posteriores. Es así que para cada uno de los vectores de la generación *G* denotados por Q_i^G, un vector mutante es generado de acuerdo con la siguiente ecuación:

$$v_i^{G+1} = Q_{r1}^G + F \cdot (Q_{r2}^G - Q_{r3}^G)$$

Donde $r1$, $r2$ y $r3$ son índices enteros seleccionados de forma aleatoria $r1, r2, r3 \in \{1,2,3,\dots,n\}$ los cuales deben de ser diferentes tanto diferentes entre sí como diferentes entre sus índices i; para poder cumplir con dicha condición, el valor de n debe de ser mayor o igual al número de vectores (n).

En la figura 4, se muestra un ejemplo de lo anterior; en la primera iteración se toma el primer vector (Q_1^1), para la asignación de r de este vector se descarta el primero (Q_1^1) se descarta, y se selecciona alguno de los restantes (Q_2^1 , Q_3^1 , Q_4^1 , Q_5^1)

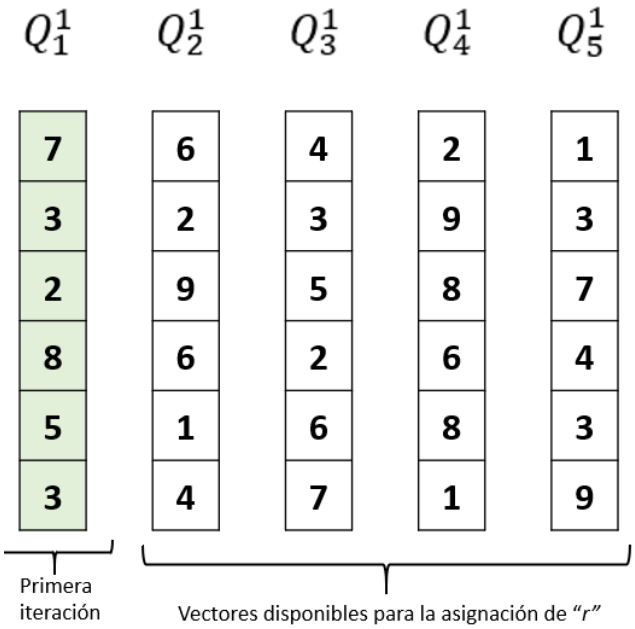

Figura 4. Ejemplo de asignación de r en DE

Como se puede observar en la figura anterior, para la primera iteración el vector objetivo será Q_1^1, por lo cual los índices para $r_1, r_2\ y\ r_3$, podrán tomar los valores de los índices restantes de forma aleatoria.

Por último, con el fin de evitar caer en mínimos o máximos locales (según sea el caso de optimización), el valor de F es una constante real ∈ $(0, 2]$, la cual controla la amplificación de la variación diferencial. Una vez realizado el proceso de mutación, se lleva a cabo la operación de cruzamiento, la cual tiene como objetivo incrementar la diversidad de los individuos de la siguiente generación y así poder elegir al mejor de ellos. En esta operación se lleva a cabo la recombinación de las d componentes de los n vectores objetivo Q_i^G, para así generar un individuo intermedio u_i^{G+1}, denotado por la siguiente expresión:

$$u_i^{G+1} = \left[(u_{i,1}^{G+1}, u_{i,2}^{G+1}, \dots, u_{i,d}^{G+1})\right]^T$$

Esta operación recombina los elementos tanto del vector $Q_{i,j}^G$, como del vector $v_{i,j}^{G+1}$, siguiendo una constante de cruzamiento (*CR* ∈ $[0,1]$), la cual determinará si la componente que se preservará vendrá del vector mutado o del vector objetivo y será determinada por el usuario. Esta operación está descrita de la siguiente forma:

$$u_{i,j}^{G+1} = \begin{cases} v_{i,j}^{G+1}, \ si\ (rand(0,1) \leq CR)\ \text{ó}\ j = rand(0,d) \\ z_{i,j}^{G}, \ si\ (rand(0,1) > CR)\ y\ j \neq rand(0,d) \end{cases}'$$

Donde:

$$i = 1,2,\cdots,n,$$
$$j = 1,2,\cdots,d$$

- El valor de $rand(0,1)$ es un número aleatorio en el intervalo $[0, 1]$
- CR es la constante de cruzamiento
- El valor de $rand(0, d) \in \{1,2, \dots, d\}$ es un índice elegido aleatoriamente, este garantiza que al menos uno de los elementos del vector mutado esté presente dentro del nuevo vector (vector intermedio).

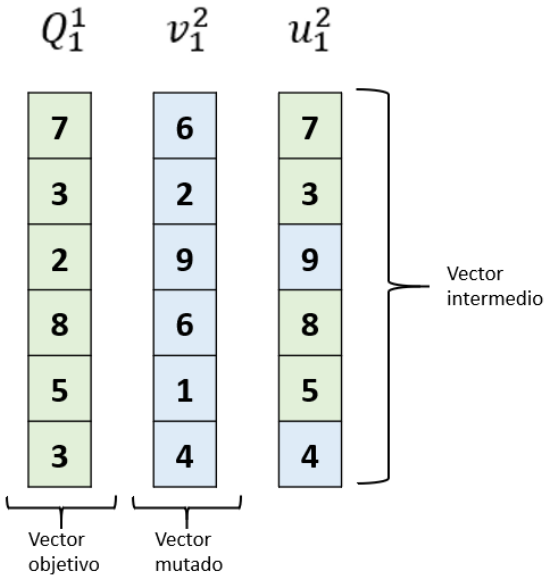

Figura 5. Ejemplo de cruzamiento DE

Como se puede observar en la figura 5, cada una de las componentes del nuevo vector intermedio proviene de los vectores objetivo y mutado. La constante de cruzamiento es lo que determina de cuál de los dos vectores se obtendrá el valor para el nuevo vector.

Finalmente, la operación de selección determina los nuevos individuos que formarán parte de la siguiente generación dentro del proceso de evolución. En esta operación se decide si un vector intermedio formará parte de la siguiente generación o no. Para determinar esto, el vector intermedio es comparado con el vector objetivo. Si el vector intermedio generado durante este proceso tiene un mejor valor (maximización o minimización según sea el caso) que el vector objetivo de la generación actual, entonces el nuevo vector de la siguiente generación será el vector intermedio, de lo contrario el vector objetivo se queda para la siguiente generación.

Este algoritmo, ya fue usado con éxito aplicado al aprendizaje supervisado, tal es el caso del trabajo de Ramírez et.al. [30] en el cual presenta un método efectivo para mejorar algunos de los parámetros en un clasificador asociativo (el clasificador *Gamma*), aumentando así su rendimiento.

3.3. Algoritmo de agrupamiento propuesto

Los conjuntos compactos [31] son los componentes conexos de un grafo de similitud máxima (*Maximum Similarity Graph - MSG*).

Un *MSG* es un grafo dirigido que conecta cada instancia con sus instancias más similares. Formalmente, sea $G = (X, \theta)$ un *MSG* para un conjunto de instancias X, donde θ es el conjunto de arcos. En este grafo, dos instancias $x, y \in X$ forman un arco $(x, y) \in \theta$ si $\max_{z \in X}\{sim(x, z)\} = sim(x, y)$, donde $sim(x, y)$ es una función de similitud. Por lo general, $sim(x, y) = 1 - d(x, y)$, donde $d(x, y)$ es una función de disimilitud. En caso de conexiones, el *MSG* establece una conexión entre la instancia y cada una de sus instancias más cercanas. Los conjuntos compactos son las componentes conexas de dicho grafo. Formalmente, un subconjunto $N \neq \emptyset$ de X es un conjunto compacto si y solo si:

a) $$\forall y \in X \left[x \in N \wedge \left(\begin{array}{l} \max\limits_{\substack{x \in X \\ x_i \neq y}} \{sim(x, y)\} = sim(x, y) \\ \vee \max\limits_{\substack{x \in X \\ x \neq y}} \{sim(y, x)\} = sim(y, x) \end{array} \right) \right] \Rightarrow y \in N$$

b) b) $$\forall x, y \in N, \exists x_1, \cdots, x_q \in N \left[\left[\begin{array}{l} x = x_1 \wedge y = x_q \wedge \forall p \{1, \cdots, q-1\} \\ \left[\begin{array}{l} \max\limits_{\substack{z \in X \\ z \neq x_p}} \{sim(x_p, z)\} = sim(x_p, x_{p+1}) \\ \vee \max\limits_{\substack{z \in X \\ z \neq x_p}} \{sim(x_{p+1}, z)\} = sim(x_{p+1}, x_p) \end{array} \right] \end{array} \right] \right]$$

c) Cada objeto aislado es un conjunto compacto, que está degenerado.

Los conjuntos compactos tienen varias ventajas para el análisis de datos, tales como: No asumen ninguna propiedad de datos, no necesitan ningún parámetro para su construcción (excepto la función de similitud para comparar dos instancias y manejar directamente datos mezclados y perdidos). Además, las instancias se conectan solo a sus instancias más similares en el conjunto de entrenamiento, que es información valiosa particularmente en las zonas de alto riesgo de Bayes.

3.4. Bancos de datos seleccionados

En esta sección, se muestra un resumen de los bancos de datos seleccionados, estos incluyen información sobre las diferentes y más comunes enfermedades crónicas [6], como cáncer de mama, enfermedades de la tiroides, enfermedades cardíacas, trastornos hepáticos, diabetes, entre otras.

Los bancos de datos utilizados en este documento se tomaron del repositorio de datos de Extracción de Conocimiento Aprendizaje Evolutivo (*KEEL)* [32] *y* del Repositorio de Aprendizaje Automático de la Universidad de California en Irvine (*UCI-MLR)* [15] . Se seleccionaron bancos de datos de clasificación relacionados con la medicina.

A continuación, se describe cada uno de los bancos de datos seleccionados.

Breast Cancer dataset (breast) [33] ***:*** es uno de los tres dominios proporcionados por el Instituto de Oncología, que ha aparecido repetidamente en la literatura de aprendizaje automático. Este banco de datos tiene dos clases, 201 instancias de una clase y 85 instancias de la otra. Las instancias se describen mediante nueve atributos, algunos de los cuales son numéricos y otros son nominales.

Liver Disorders dataset (bupa) [34]***:*** analiza algunos trastornos hepáticos que podrían ser causados por el consumo excesivo de alcohol.

Heart Disease dataset (cleveland) [35]***:*** es un subconjunto de 14 atributos del banco de datos de enfermedades del corazón (la parte obtenida del *V.A. Medical Center*, *Long Beach* y *Cleveland Clinic Foundation*). El objetivo es detectar la presencia de una enfermedad cardíaca en el paciente.

Haberman's Survival dataset (haberman) [36]***:*** contiene casos de un estudio realizados entre 1958 y 1970 en el Hospital *Billings* de la Universidad de Chicago sobre la supervivencia de pacientes que se habían sometido a cirugía por cáncer de mama. El objetivo es determinar si el paciente sobrevivió 5 años o más (positivo) o si el paciente murió dentro de 5 años (negativo).

Statlog (Heart) dataset (heart) [37]***:*** El objetivo es detectar la ausencia o presencia de una enfermedad cardíaca.

Hepatitis dataset (hepatitis) [38]***:*** contiene una mezcla de atributos numéricos y valores reales, con información sobre pacientes afectados por hepatitis. El objetivo es predecir si estos pacientes morirán o sobrevivirán.

Mammographic Mass dataset (mammographic) [39]***:*** se puede utilizar para predecir la gravedad (benigna o maligna) de una lesión mamográfica a partir de atributos llamados "BI-RADS" y la edad del paciente. Los datos fueron recolectados en el Instituto de Radiología de la Universidad *Erlangen-Nuremberg* en Alemania entre 2003 y 2006.

Thyroid Disease dataset (newthyroid) [40]**:** es uno de los varios bancos de datos sobre tiroides disponibles en el repositorio de *KEEL* y *UCI.* El objetivo es detectar si un paciente determinado es normal o si padece hipertiroidismo o hipotiroidismo.

Pima Indians Diabetes dataset (pima) [41]***:*** el objetivo de este banco de datos es determinar si el paciente muestra signos de diabetes de acuerdo con los criterios de la Organización Mundial de la Salud.

Post-Operative dataset (post-operative) [42]***:*** el objetivo de la clasificación de este banco de datos es determinar a dónde deben enviarse los pacientes en un área de recuperación postoperatoria para evitar se desarrollen enfermedades. Debido a que la hipotermia es una preocupación importante después de la cirugía, los atributos corresponden aproximadamente a las mediciones de temperatura corporal.

South African Hearth dataset (saheart) [43]***:*** consiste en diversos datos de hombres en una región de alto riesgo (*Western Cape, Sudáfrica*) donde abundan enfermedades del corazón. La clase indica si la persona tiene una enfermedad coronaria: negativa o positiva.

SPECTF Heart dataset (spectfheart) [44]***:*** describe el diagnóstico de imágenes de tomografía computarizada de emisión monofotónica (por sus siglas en inglés *Single Photon Emission Computed Tomography SPECT*), esto con el fin de detectar anomalías cardiacas. Cada uno de los pacientes se clasifica en dos categorías: normal o anormal.

Thyroid Disease dataset (thyroid) [45]***:*** El objetivo es detectar si un paciente determinado es normal o sufre de hipertiroidismo o hipotiroidismo. Este banco de datos es uno de los varios bancos de datos sobre tiroides disponibles en el repositorio de *KEEL* y *UCI*.

Breast Cancer Wisconsin dataset (wdbc) [46]: contiene 30 atributos obtenidos a partir de una imagen digitalizada de una punción aspiración con aguja fina (*PAAF*) de una masa mamaria. El objetivo es determinar si un tumor encontrado es benigno o maligno.

Breast Cancer Wisconsin dataset (wisconsin) [47]***:*** contiene casos de un estudio realizados en la Universidad de *Wisconsin, Madison* sobre pacientes que se sometieron a cirugía para el cáncer de mama. El objetivo es determinar si el tumor detectado es benigno o maligno.

En la tabla 1, se indican a detalle las características de cada uno de los bancos de datos mencionados.

Tipo de enfermedad	No	Bancos de datos	Atributos		Análisis de desbalance		Valores Perdidos	Clases
			Numéricos	Categóricos	Instancias	IR		
Cáncer	1	breast	0	9	277	2.420	Sí	2
	2	mammographic	6	0	830	1.060	Sí	2
	3	haberman	3	0	306	2.778	No	2
	4	wisconsin	9	0	683	1.858	Sí	2
	5	wdbc	30	0	569	1.684	No	2
Enfermedad respiratoria	6	post-operative	0	8	87	62.000	Sí	3
Enfermedad cardiovascular	7	heart	13	0	270	1.250	No	2
	8	saheart	8	1	462	1.888	No	2
	9	spectfheart	44	0	267	3.855	No	2
	10	cleveland	13	0	297	12.308	Sí	5
Diabetes	11	pima	8	0	768	1.866	No	2
Otras enfermedades crónicas	12	thyroid	21	0	7200	40.157	No	3
	13	newthyroid	5	0	215	5.000	No	3
	14	hepatitis	19	0	80	5.154	Sí	2
	15	bupa	7	0	345	1.379	No	2

Tabla 1. Descripción de los bancos de datos utilizados

El resumen de la tabla 1 incluye la cantidad de atributos numéricos y categóricos, el número de instancias, la relación de desbalance (en inglés *Imbalance* Ratio o *IR*) entre la clase mayoritaria y minoritaria, la presencia o no de valores perdidos y el número de clases.

Capítulo 4. Propuesta de investigación

En este capítulo se describen los métodos y técnicas utilizadas para la validación de los resultados, la evaluación del rendimiento de los algoritmos seleccionados y las pruebas estadísticas. Además, se propone un nuevo algoritmo de clasificación inteligente de patrones.

4.1. Algoritmo de clasificación propuesto (*Assisted Classification for Imbalance Data-ACID*)

Esta sección está dedicada a la explicación del modelo propuesto para el prediagnóstico médico: el modelo de Clasificación Asistida para Datos No Balanceados por sus siglas en inglés (*Assisted Classification for Imbalance Data-ACID).* Se explican las ideas principales del modelo, así como su funcionamiento, también se detalla la fase de entrenamiento de *ACID* y, finalmente, se aborda la fase de clasificación.

El modelo *ACID* está diseñado principalmente para tratar datos no balanceados, muy comunes en los dominios médicos (por lo general, la cantidad de personas enfermas es mucho menor que la cantidad de personas sanas). Además, está destinado a la manipulación de datos mezclados y perdidos, que también es una situación común en escenarios médicos. Como *ACID* es un modelo supervisado, requiere como entrada un conjunto T de instancias clasificadas, descritas por un conjunto de características $A = \{A_1, \dots, A_n\}$. El valor de la característica i-ésima de una instancia $x \in T$ se denota por x_i. Si este valor es desconocido, es decir, faltante $x_i = '?'$. Cada instancia debe pertenecer solo a una clase de un conjunto de clases de clasificación $K = \{K_1, \dots, K_p\}$.

Funcionamiento

Wolpert demuestra, mediante el teorema de *no-free-lunch* que ningún algoritmo de clasificación puede superar a todos los demás de acuerdo con todas las medidas de rendimiento, en todos los dominios [48]. Sin embargo, algunas características deseadas incluidas en el modelo propuesto hacen de *ACID* un buen modelo de clasificación para el prediagnóstico médico.

Uno de los aspectos clave de los conjuntos de datos de enfermedades médicas es que a menudo están no balanceados. Es decir, el número de instancias en la clase mayoritaria es mayor que el número de instancias en la clase minoritaria, que generalmente es la clase de interés. La clasificación de los datos no balanceados es un desafío y varios factores influyen en ella. A continuación, se discuten algunos de ellos [49]:

1. **Sesgo hacia la clase mayoritaria.** Algunos clasificadores se inclinan por la clase mayoritaria, debido a su funcionamiento. Por ejemplo, las redes neuronales hacen un ajuste de pesos considerando el error general, no el error de cada una de las clases. De forma similar, los árboles de decisión generalmente tienen criterios de poda o de abandono basados en la homogeneidad general de las hojas, mientras que las máquinas de vectores de soporte ajustan sus parámetros teniendo en cuenta el rendimiento general.
2. **Presencia de pequeños disjuntos.** Se encuentran pequeños disjuntos cuando una clase ocupa una región diferente en el espacio de características. Por lo tanto, emergen las llamadas subclases o subtipos. Las subclases contienen instancias de la misma clase, pero muy diferentes entre sí. La presencia de pequeños disjuntos afecta mucho a los clasificadores.
3. **Solapamiento de clases.** En tales escenarios, el riesgo de Bayes aumenta y el rendimiento de los clasificadores disminuye.
4. **La maldición de la dimensionalidad.** Afecta a los clasificadores basados en similitud, debido a que a medida que aumenta el número de dimensiones, las instancias tienden a parecer "más similares".

En la figura 6, se muestra un ejemplo del conjunto de datos de trébol (*clover dataset)* [50], que ejemplifica algunos de los factores antes mencionados que influyen en la clasificación de datos no balanceada. Los conjuntos de datos médicos generalmente tienen esos factores. El autor muestra en [50] una representación de dos conjuntos de datos médicos, en los que está clara la presencia de pequeños disjuntos y solapamiento de clases. Además, en dicho estudio, los autores mencionan la maldición de la dimensionalidad.

Las descripciones de los pacientes pueden estar dispersas de alguna manera, es decir, pacientes muy diferentes tienen las mismas enfermedades. Por ejemplo, si el valor de un determinado examen es externo a algunos límites normales, los resultados indican una enfermedad; un paciente adulto que tiene un nivel de glucosa en ayuno inferior a 4,0 mmol / L (72 mg / dL) o superior a 6,0 mmol / L (108 mg / dL), se considera enfermo, mientras que un paciente con un nivel de glucosa entre 4.0 a 6.0 mmol / L (72 a 108 mg / dL) se considera saludable.

En el ejemplo de la siguiente figura, se utilizó el clasificador de los “k” vecinos más cercanos (kNN con k=5) y se observa lo siguiente:

a) Presencia de pequeños disjuntos (subclases) dentro de una clase.

b) Solapamiento de clases.

c) Regiones originales de decisión.

d) Sesgo (*bias*) hacia la clase mayoritaria (con k=5).

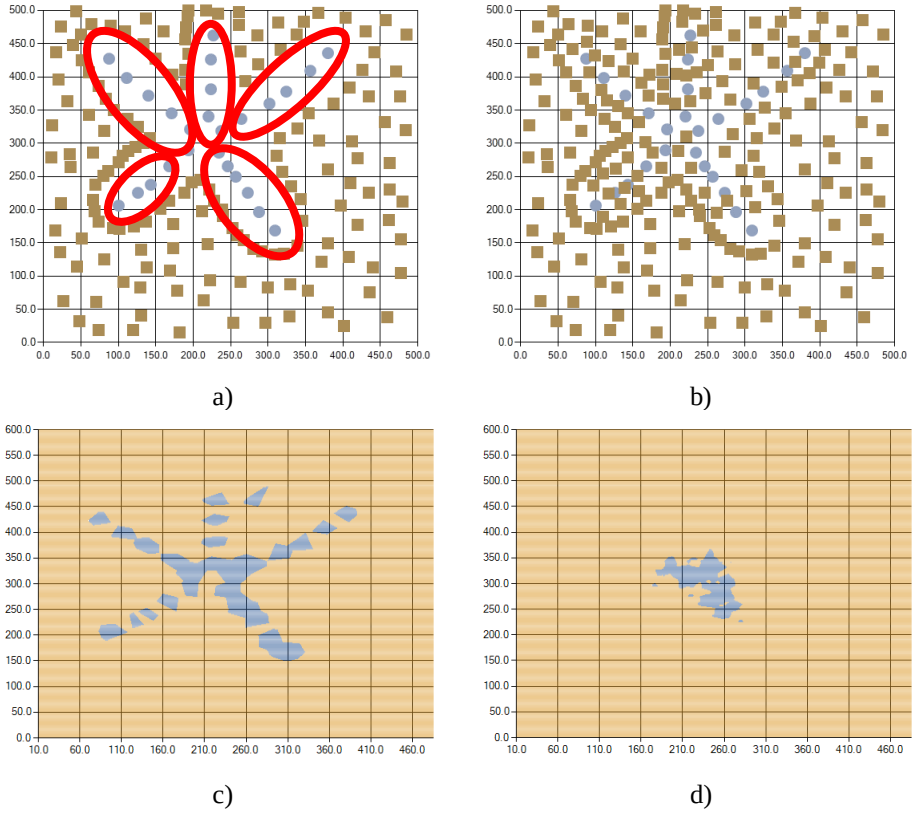

Figura 6. Ejemplo de factores que influyen en la clasificación desequilibrada en el "clover dataset".

En la figura 7, se observa otro ejemplo de factores que influyen en la clasificación no balanceada.

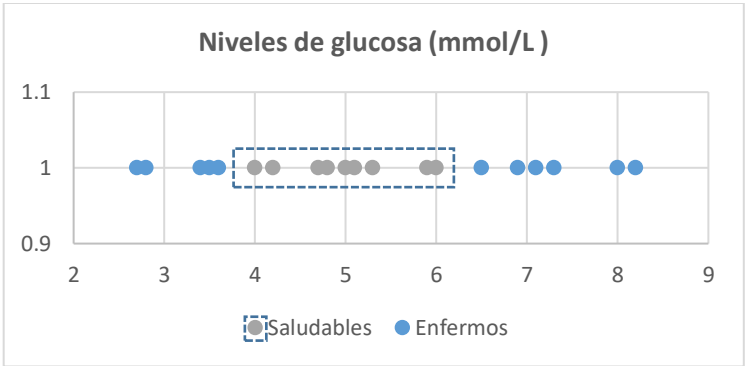

Figura 7. Ejemplo de una clase (enferma) que ocupa dos regiones en el espacio de características.

Para tratar con datos no balanceados afectados por los factores previamente explicados, se diseñó un nuevo algoritmo de clasificación. Las ideas que dieron origen al clasificador propuesto son las siguientes:

1. **Disminuir los efectos de la maldición de la dimensionalidad.** Para hacerlo, se propone utilizar un algoritmo metaheurístico para poder asignar ponderaciones a los atributos. Además, se consideraron como irrelevantes los atributos que tienen menos peso. Al hacerlo, se asegura que el clasificador solo tenga en cuenta las características relevantes.
2. **Tratar los disjuntos**. Se busca la presencia de pequeños disjuntos en las clases. Para hacerlo, se estructura cada clase por separado, para encontrar las diferentes subclases (disjuntos) de la clase. Se considera que una subclase tiene instancias muy similares, siendo diferente de las instancias de otras subclases. Para encontrar estas subclases se utilizan técnicas de agrupamiento.
3. **Sesgo (*bias*) a la clase mayoritaria**. Para garantizar que el clasificador propuesto sea imparcial hacia la clase mayoritaria, se utiliza una estrategia basada en la similitud, que ofrece las mismas posibilidades de semejanza para cada clase. Esta estrategia consiste en obtener la similitud con cada subclase y luego devolver la subclase más similar de cada clase. Es decir, cada clase tiene solo una subclase más similar. Luego, se le asigna una etiqueta a la instancia de acuerdo con la etiqueta de la subclase más similar, entre todas las clases. Garantiza que cada clase tenga un único valor de similitud para fines de comparación, siendo imparcial al clasificar con cualquiera de las clases.
4. **Tratar con el solapamiento de clases.** Para disminuir la influencia del solapamiento de clases, se agrega la similitud de la instancia para clasificar con respecto a una subclase. Es decir, en lugar de considerar la instancia más similar, se utilizan operadores de agregación para comparar una instancia con un conjunto de instancias en una subclase. Lo anterior garantiza que el solapamiento de clase y las instancias ruidosas o mal etiquetadas influyan menos en el proceso de clasificación.

Figura 8. Fases de entrenamiento y clasificación de ACID

A continuación, se detallan las fases de entrenamiento y clasificación de *ACID*. La fase de entrenamiento aborda los puntos anteriores uno y dos, mientras que la fase de clasificación aborda los puntos tres y cuatro; tal como se muestra en la figura 8.

Fase de Entrenamiento

Con base en los cuatro puntos mencionados en la sección anterior se propone lo siguiente:

Disminuir los efectos de la maldición de la dimensionalidad

La fase de entrenamiento de *ACID* comienza tratando con la "maldición de la dimensionalidad". Para hacerlo, se les asignan ponderaciones a los atributos usando un algoritmo evolutivo llamado Evolución Diferencial (*Differential Evolution - DE*). Se utiliza la Evolución Diferencial debido a que este algoritmo demostró haber tenido buenos desempeños en experimentos con otros clasificadores supervisados [30]. Primero, se determina la estrategia de codificación adecuada para el cálculo del peso. El cálculo de características y pesos es un tipo de problema de optimización continua, debido a que las características son valores reales.

Para aplicar *DE*, se codifican los pesos de entidad como un vector de valores reales en el intervalo [0,1]. Lo anterior permite analizar la importancia relativa de las características en el proceso de clasificación. Teniendo en cuenta esta codificación, cero significa que la característica es completamente irrelevante; mientras que uno significa que esta característica es la más relevante para la tarea de clasificación. Además, la codificación en el intervalo [0,1] facilitaría el desarrollo posterior de una versión difusa del modelo *ACID.*

El vector de valores tendrá como longitud la cantidad de características que describen el problema. La figura 9 muestra un ejemplo de un individuo de "*DE*", que codifica los pesos de las características asociadas a un conjunto de datos con cuatro atributos.

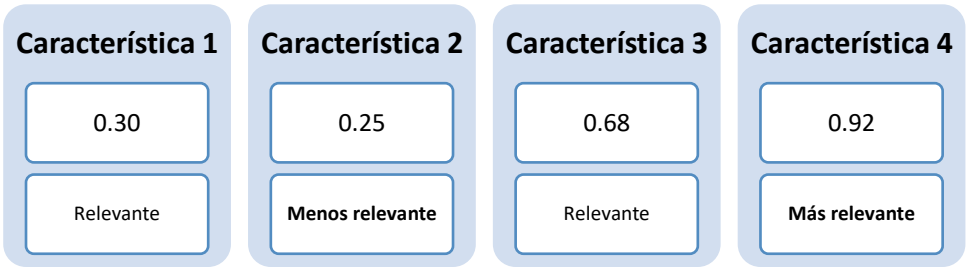

Figura 9. Codificación de un individuo de Evolución Diferencial en el modelo ACID

La definición de la función de aptitud constituye un aspecto clave en los algoritmos de optimización. En esta investigación, se utiliza la Tasa de Verdaderos Positivos promedio (*True Positive Rate* - *TPR*) [51] como función de aptitud y, para obtenerla, es necesario tener un conjunto de validación. Se dividió el conjunto de entrenamiento en dos subconjuntos: entrenamiento y validación, mediante el procedimiento de *Hold-Out*, con un 70% de instancias para entrenamiento y un 30% de instancias para el cálculo de aptitud. La figura 10 ilustra el proceso de cálculo de la función de aptitud en el modelo *ACID.*

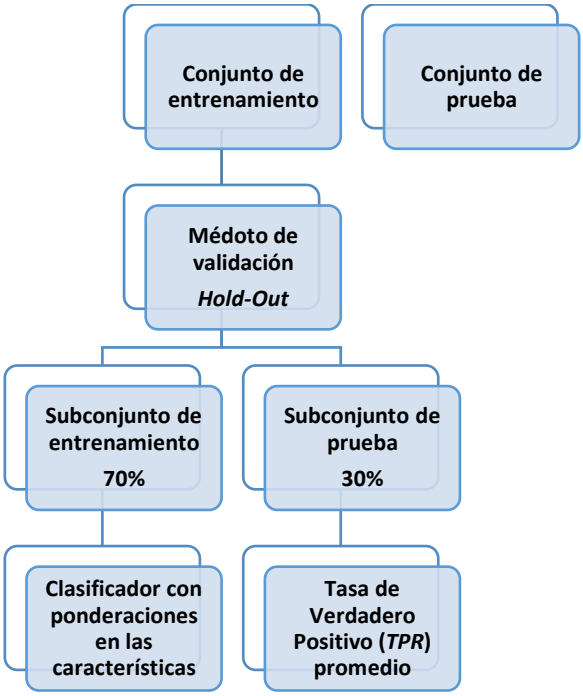

Figura 10. Cálculo de la aptitud en el modelo ACID

Tratar los disjuntos

Para hacer frente a los pequeños disjuntos, se divide el conjunto de entrenamiento en clases, y luego, se estructura cada clase. Los algoritmos de agrupación apuntan a obtener "la estructura natural de la relación de los datos" [52]; por lo tanto, son una opción útil para detectar la presencia de pequeños disjuntos en una clase. Sin embargo, los datos médicos a menudo contienen datos mezclados (es decir, contienen atributos numéricos y categóricos) y con valores perdidos (algunas instancias tienen valores de atributo faltantes). Además, no se sabe con exactitud cuántos disjuntos puede tener una clase en particular.

Para hacer frente a este escenario, se requiere de un algoritmo de agrupamiento capaz de manejar datos mezclados y perdidos, y sin un número predefinido de grupos. En el Enfoque Lógico Combinatorio para el Reconocimiento de Patrones, hay varios procedimientos de agrupamiento [53]. Uno de ellos es la estructuración en conjuntos compactos (*Compact Sets Structuralization*).

Se propone uno de los conjuntos compactos [53] para encontrar pequeños disjuntos en las clases debido a que maneja directamente datos mezclados y perdidos y no necesitan ningún parámetro para su obtención, es decir, no requieren la cantidad de grupos para obtener. Después de obtener los grupos de cada clase, se almacenan para un uso posterior.

En resumen, la fase de entrenamiento de *ACID* consta de dos fases: disminuir la maldición de la dimensionalidad y detectar subclases en las clases (tratar los disjuntos).

El proceso devuelve dos tipos de datos: los pesos que fueron asignados a cada una de las características y las subclases de cada clase. La figura 11 ilustra la fase de entrenamiento y sus datos de retorno en *ACID*.

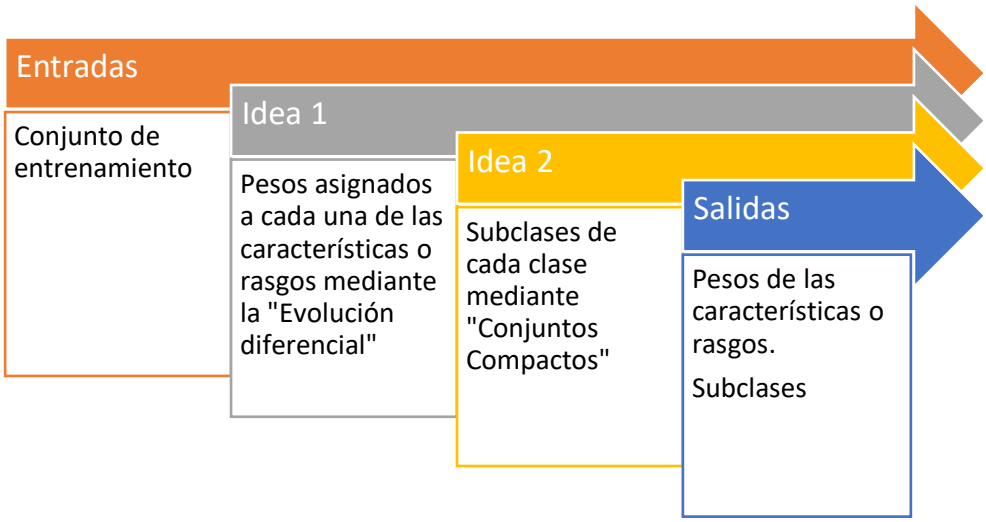

Figura 11. Fase de entrenamiento del modelo ACID

Como se muestra en la figura anterior, *ACID* calcula un conjunto de pesos $W = \{w_1, \dots, w_n\}$ asociados al conjunto de características. Para esta tarea, *ACID* usa la Evolución Diferencial (*Differential Evolution*) [22]. Además, en la fase de entrenamiento, *ACID* obtiene un conjunto de pequeños disjuntos existentes de cada clase, que se usarán más adelante en la fase de clasificación. Las instancias ruidosas y atípicas (*outliers*) están en las subclases; sin embargo, contribuirán menos a la clasificación de instancias.

Fase de Clasificación

El principal objetivo del modelo *ACID* es tratar con datos no balanceados y con el solapamiento de clases en la fase de clasificación. Para hacerlo, usa una estrategia de cálculo de la similitud para garantizar que no haya sesgo en ninguna de las clases. Además, usa la agregación para disminuir la influencia del solapamiento de clases, así como la de las instancias ruidosas o mal etiquetadas.

Las siguientes subsecciones explican estas ideas a detalle.

Sesgo (*bias*) a la clase mayoritaria.

Para garantizar un rendimiento imparcial, el modelo *ACID* utiliza el cálculo de la similitud. A diferencia de otros clasificadores como el vecino más cercano (*Nearest Neighbor*) [54], *ACID* calcula la similitud de la instancia para clasificar con respecto a cada subclase de cada clase. Luego, almacena la subclase más similar de cada una de las clases. Después de eso, *ACID* compara los valores de similitud y etiqueta la instancia con la etiqueta de la clase que tiene el mayor valor de similitud.

Al hacerlo, se garantiza que la probabilidad de selección sea la misma para todas las clases. Esto hace que *ACID* sea imparcial en cualquiera de las clases en el conjunto de entrenamiento. En la figura 12 se muestra un ejemplo de cómo *ACID* clasifica una instancia.

En dicho ejemplo, se representan los datos de entrenamiento en 2D y los pequeños disjuntos encontrados en la fase de entrenamiento. También se muestra la instancia para clasificar y las disjuntos más similares para la instancia.

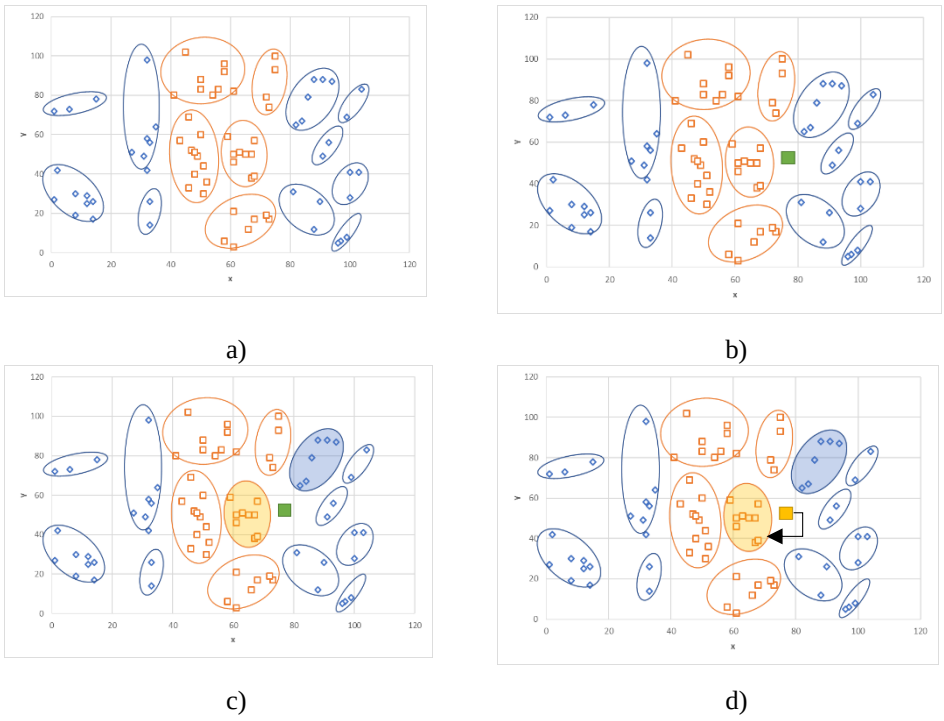

Figura 12. Ejemplo de fase de clasificación de ACID

En la figura 12 se ejemplifica lo siguiente:

a) Subclases (pequeños disjuntos) obtenidas en la fase de entrenamiento.
b) Instancia desconocida para clasificar.
c) Subclases más similares de cada clase.
d) Etiqueta asignada de acuerdo a la subclase más similar.

La figura 13, indica un resumen del proceso de la fase de clasificación.

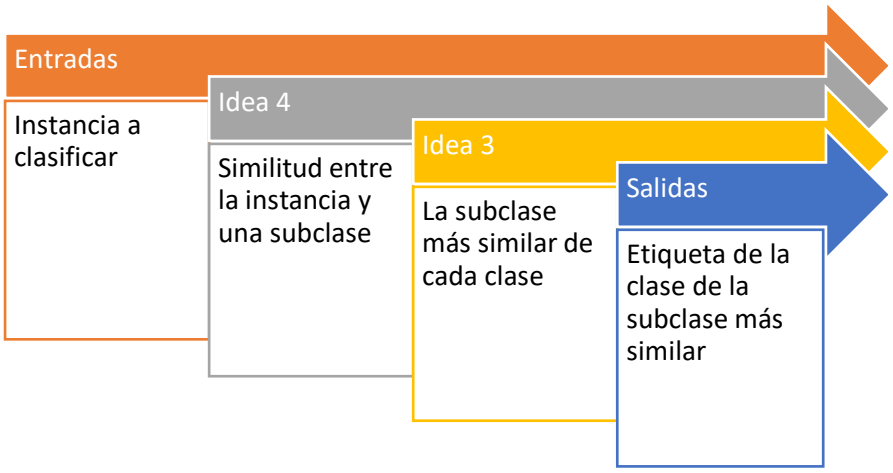

Figura 13. Fase de clasificación del modelo ACID

El modelo *ACID* permite usar cualquier función de similitud o cualquier función de disimilitud (el modelo en lugar de considerar la subclase más similar considera que la subclase menos disímil es la que contiene la etiqueta correcta).

Tratar con el solapamiento de clases

Para disminuir la influencia del solapamiento de clases, el modelo *ACID* considera la disimilitud general del patrón para clasificar o, para cada subclase obtenida en la fase de entrenamiento. Es decir, en lugar de considerar la instancia más similar, utiliza operadores de agregación para comparar una instancia con un conjunto de instancias en una subclase. Existen varias medidas para calcular la disimilitud entre una instancia y un conjunto de instancias; entre las más comunes están: enlace único (*Single-Linkage*), enlace completo (Complete-Linkage), enlace promedio (Average-Linkage) y enlace centroide (Centroid-Linkage) [52]. En la tabla 2 se resumen estas medidas, siendo $\bar{x}$ el centroide del conjunto $C_i y C_i$ y $d(x, y)$ es la función de disimilitud entre instancias.

Tipo de enlace	Función de disimilitud
Enlace único	$D(C_i, o) = \min_{x \in C_i}\{d(x, o)\}$
Enlace completo	$D(C_i, o) = \max_{x \in C_i}\{d(x, o)\}$
Enlace promedio	$D(C_i, o) = \sum_{x \in C_i} d(x, o)/\lvert C_i \rvert$
Enlace centroide	$D(C_i, o) = d(\bar{x}, \bar{o})$

Tabla 2. Medidas de disimilitud entre una instancia y un conjunto de instancias

Para representar un grupo, el uso del centroide es el esquema más popular [52]. Funciona bien cuando los grupos son compactos. Sin embargo, cuando los grupos son muy grandes, este esquema no los representa adecuadamente. En tal caso, el uso de una colección de puntos límite o de holotipos como

representantes de un grupo es una buena práctica. El número de puntos utilizados para representar un grupo debe aumentar a medida que aumenta la complejidad de su forma.

En la fase experimental del presente trabajo, se utilizó un enlace promedio para determinar la disimilitud general de una instancia a una subclase.

El enlace promedio permite disminuir la influencia de datos ruidosos o atípicos en el proceso de clasificación. Además, disminuye la influencia del solapamiento de clases, en la figura 14 y la tabla 3 se ejemplifica lo anterior.

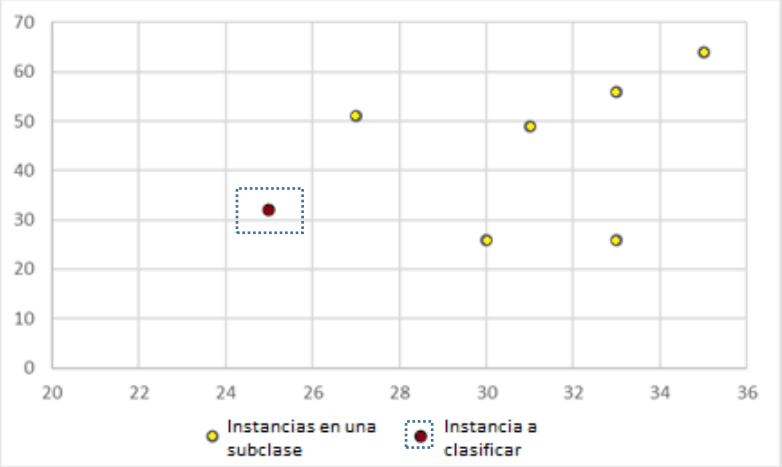

Figura 14. Ejemplo del enlace promedio en el modelo ACID

Instancia por clasificar: $\sigma = (25{,}32)$

Instancias en una subclase:

X	**27**	**35**	**33**	**31**	**33**	**30**
Y	51	64	56	49	26	26
$\boldsymbol{d(x,o)}$	19.10	33.53	25.30	18.03	10.00	7.81

Enlace promedio:

$$\boldsymbol{D(C_i,o) = \sum_{x \in C_i} d(x,o)/|C_i| = 113.77/6 = 18.96}$$

Tabla 3. Ejemplo de enlace promedio

Para concluir este capítulo, *ACID* es un clasificador novedoso diseñado para tratar con datos no balanceados. Maneja con éxito conjuntos de datos mezclados y perdidos, y aborda algunos de los factores más relevantes que influyen en la clasificación de datos no balanceada. Trata la maldición de la dimensionalidad y la presencia de pequeños disjuntos; es imparcial para cualquiera de las clases y maneja datos solapados, ruidosos, atípicos o mal etiquetados.

Además, *ACID* es un modelo interpretable; es decir, *ACID* es transparente, debido a que se sabe exactamente porque una instancia pertenece a una determinada clase. Esta es una clara ventaja sobre otros clasificadores del estado del arte.

Capítulo 5. Resultados y su discusión

En este capítulo, se presentan los resultados experimentales obtenidos para el diagnóstico médico, utilizando el nuevo modelo de clasificación propuesto, así como otros modelos de estado del arte.

En este capítulo se abordan los métodos de validación, medidas de rendimiento y las pruebas estadísticas empleadas, además de los resultados finales obtenidos.

La figura 15 ilustra los esquemas del diseño del experimento llevado a cabo.

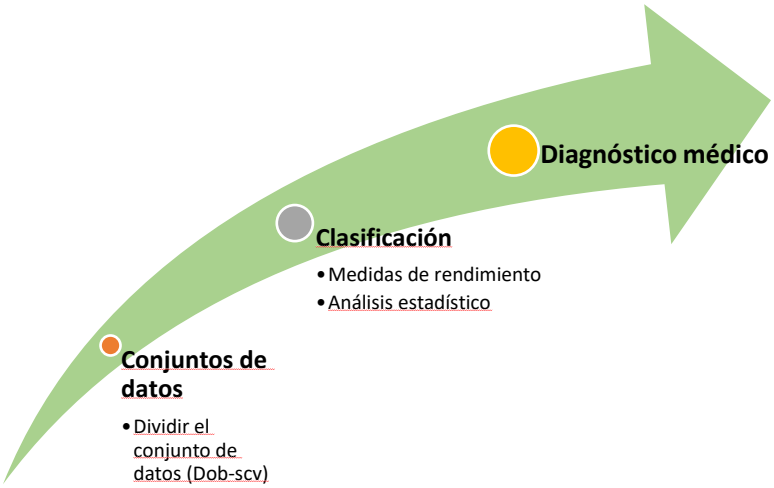

Figura 15. Esquema del diseño experimental

Se seleccionaron cinco algoritmos de clasificación de última generación. Todos ellos pueden tratar datos mezclados y perdidos. Esta selección incluye "k" vecino más cercano (kNN) (con k=3) [44], Perceptrón multicapa (*MLP*) [45], Árboles de decisión (*C4.5*) [46], Máquinas de soporte vectorial (*SMO*) [47] y Regresión logística (*Logistic)* [48]. Para *MLP*, *C4.5*, *Logistic* y *SMO*, se utilizaron los valores de parámetros predeterminados ofrecidos en el paquete de software KEEL [42] [43].

Para el vecino más cercano y el clasificador *ACID*, se utilizó la disimilitud de *HEOM* [16], que maneja conjuntos de datos mezclados y perdidos. *HEOM* utiliza dos enfoques diferentes para calcular la disimilitud sobre los atributos numéricos y categóricos. Se considera que x_a, y_a son los valores del atributo a, para las instancias x y y, se considera que max_a y min_a son los valores máximo y mínimo del atributo a.

5.1. Métodos de validación

En reconocimiento de patrones, los bancos de datos que sufren de desbalance de clases son un problema importante para la clasificación de datos. La principal propiedad de este tipo de problema de clasificación es que los ejemplos de una clase superan significativamente en número a los ejemplos de la otra. En medicina, la clase minoritaria generalmente representa el concepto más importante que debe aprenderse, y es difícil de identificar, ya que puede estar asociado con casos excepcionales y significativos, o porque la adquisición de datos de estos ejemplos es costosa. En la mayoría de los casos, el problema de clase no balanceado está asociado a la clasificación binaria, pero a menudo ocurre el problema de clase múltiple y, dado que puede haber varias clases minoritarias, este problema es más difícil de resolver.

Para fines de validación, se utilizó el procedimiento de *Distributed optimally balanced stratified cross validation (Dob-scv)* con cinco *folds* (u hojas) ya que este es recomendado para escenarios no balanceados [49]. En este método se divide el banco de datos en cinco *folds*, cada uno contiene el 20% de los patrones del banco de datos.

Para cada *fold*, el algoritmo fue entrenado con los ejemplos contenidos en los *folds* restantes y luego probado con el *fold* actual. Este valor tiene el objetivo de tener suficientes instancias de clases positivas en los diferentes *folds*, evitando así problemas adicionales en la distribución de datos, especialmente para conjuntos de datos altamente no balanceados [55].

5.2. Medidas de rendimiento

Tradicionalmente, la tasa de precisión o *accuracy rate* (ver ecuación 1) ha sido la medida empírica más comúnmente utilizada. Sin embargo, en el marco de bancos de datos no balanceados, la precisión ya no es una medida adecuada, ya que no distingue entre el número de ejemplos correctamente clasificados de diferentes clases. Por lo tanto, este método puede llevar a conclusiones erróneas, por ejemplo, que un clasificador indique una precisión del 90% en un banco de datos con un valor *IR* de 9 (es decir, el 90% de los patrones pertenece a una clase y el 10% pertenece a la otra); este porcentaje de precisión no es confiable tomando en cuenta que el clasificador clasifica todos los ejemplos como negativos.

$$Tasa\ de\ precisión\ = Accuracy\ rate = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

Para obtener la tasa de precisión, se toman en consideración los verdaderos positivos (*True Positives* o *TP*), verdaderos negativos (*True Negatives* o *TN*), falsos positivos (*False Positives* o *FP*) y los falsos negativos (*False Negatives* o *FN*). También es tomada en cuenta la matriz de confusión, ver figura 16:

	Positivo	**Negativo**
Positivo	TP	FP
Negativo	FN	TN

Figura 16. Ejemplo de una Matriz de confusión

En escenarios donde hay bancos de datos no balanceados, la evaluación del rendimiento de los clasificadores se debe llevar a cabo utilizando medidas específicas tomando en cuenta la distribución de clase; las medidas de rendimiento habituales se vuelven inadecuadas [51] esto se debe al sesgo que tales medidas tienen hacia la clase mayoritaria, que a su vez puede dar lugar a conclusiones engañosas. Para evaluar el rendimiento en bancos de datos no balanceados con múltiples clases, se propone el uso del promedio de la tasa verdadera positiva (*True Positive Rate* o *TPR*) [51].

En un problema de dos clases, el *TPR* (también conocido como *recall* o *sensibilidad*) considera el total de instancias positivas correctamente clasificadas, en relación con el total de instancias de la clase positiva, considerando los verdaderos positivos (*True Positives* o *TP*), verdaderos negativos (*True Negatives* o *TN*), falsos positivos (*False Positives* o *FP*) y los falsos negativos (*False Negatives* o *FN*). Ver ecuación (2).

$$Sensibilidad = TPR = Recall = \frac{TP}{TP + FN} \tag{2}$$

En un problema con k clases, la sensibilidad de clasificación toma en consideración el total de instancias correctamente clasificadas de la clase j, relativa al total de instancias de la clase j-ésima. Por lo tanto, la sensibilidad de clasificación para la clase j calcula la probabilidad de clasificar correctamente una instancia de la clase j. Para el cálculo de dicha sensibilidad de clasificación, sea n_j el número de instancias correctamente clasificadas (en una matriz de confusión de k clases), y sea t_j el total de instancias pertenecientes a la clase j. Por lo tanto, por esta razón, la sensibilidad de clasificación (también factor de recuerdo o promedio de la tasa verdadera positiva) de la clase j, indicada por S_j, se calcula de la siguiente manera:

$$S_j = \ Recall_j = TPR_j \ = \frac{n_j}{t_j} \tag{3}$$

Aunque la sensibilidad mínima de clasificación permite manejar múltiples clases (ver ecuación 4), solo considera la menor de las tasas clasificadas correctamente entre las clases.

$$\text{Sensibilidad mínima} = Mínimo = \min_{j=1..k}\{S_j\} \tag{4}$$

En relación con este tema, nuestro objetivo es hacer uso de una métrica de rendimiento que le dé el mismo peso a cada una de las clases del problema, independientemente de la cantidad de ejemplos que tenga, en este caso se propone la sensibilidad promedio. Ver ecuación 5.

$$Sensibilidad\ promedio = Promedio\ de\ TPR\ = \frac{1}{k}\sum_{i=1}^{k} S_j \tag{5}$$

En la ecuación anterior, k es el número de clases y S_j es el *TPR* para la clase j-ésima. Esta medida de rendimiento permite evaluar el rendimiento global de los algoritmos de clasificación sobre todas las clases del problema, no solo sobre la clase minoritaria. El uso del *TPR* promedio por clase permite tener en cuenta todas las clases, sin sesgo hacia ninguna clase en particular. En la figura 17 se presenta un ejemplo del cálculo de medidas de rendimiento dada una matriz de confusión teniendo en cuenta el Promedio de *TPR* y el Mínimo, con k = 3 clases.

		Clase Predicha		
		A	**B**	**C**
Clase Real	**A**	7	2	1
	B	5	3	2
	C	3	1	6

$$S_A = \frac{7}{10} = 0.7, S_B = \frac{3}{10} = 0.3, S_C = \frac{6}{10} = 0.6$$

a) $Promedio\ de\ TPR\ = \frac{0.7+0.3+0.6}{3} = \frac{1.6}{3}$ b) $Mínimo = min\{0.7, 0.3, 0.6\} = 0.3$

Figura 17. Ejemplo del cálculo de medidas de rendimiento.

Es por eso que, en este trabajo, se eligió una medida del rendimiento que otorga el mismo peso a cada una de las clases, independientemente del número de muestras que tenga cada una.

5.3. Pruebas estadísticas

Para determinar qué algoritmos de clasificación obtuvieron los mejores resultados experimentales se utilizaron pruebas de hipótesis. Las pruebas de hipótesis estadísticas evalúan si existe una diferencia significativa en el rendimiento de los diferentes algoritmos de clasificación. Se seleccionaron pruebas no paramétricas para la investigación actual [56]. Particularmente, se seleccionó la prueba de Friedman ya que es ampliamente recomendada para comparar múltiples clasificadores sobre múltiples bancos de datos.

La prueba de Friedman [57] [58] consiste en ordenar las muestras y reemplazarlas por sus rangos respectivos de la siguiente manera: el mejor resultado corresponde al rango 1, el segundo mejor al rango 2, el tercero al rango 3 y así sucesivamente. Después de eso, se toma en consideración la existencia de muestras idénticas; en ese caso, se les asigna un rango promediado.

Si la prueba de Friedman rechaza la hipótesis nula de igualdad de rendimiento, se debe aplicar una prueba *post-hoc* para determinar entre qué algoritmos hay diferencias significativas [56]. Entre las diferentes pruebas *post-hoc* recomendadas para el análisis de rendimiento en múltiples conjuntos de datos [56] [59] [60] se utiliza la prueba de Holm [61]. Esta prueba usa un procedimiento descendente para ajustar el valor de significancia α. Para esto, los valores de p se ordenan de forma ascendente (es decir, del más significativo al menos significativo). Si $p_1 < \frac{\alpha}{l-1}$, la prueba rechaza la hipótesis nula y la prueba continúa la comparación con el siguiente valor p, considerando si $p_2 < \frac{\alpha}{l-2}$. Esta prueba continúa este proceso hasta que no puede rechazar una de las hipótesis, dado que $p_i \geq \frac{\alpha}{l-i}$. En este punto, la prueba tampoco rechazó las hipótesis restantes.

Existen muchas herramientas automatizadas especializadas para el cálculo de la prueba de Friedman y las pruebas *post-hoc*. En esta investigación, se utiliza el software KEEL [62] [63].

En la tabla 4 se muestran los resultados obtenidos por los algoritmos de clasificación analizados, para los problemas de diagnóstico médico considerados. Los mejores resultados se destacan en letras negritas. Se utilizaron los valores de parámetros predeterminados ofrecidos en el paquete de software KEEL [62] [63] .

Para el algoritmo del “k” más cercano y el clasificador *ACID,* se utilizó la disimilitud de HEOM [25], que maneja datos mezclados y perdidos.

Bancos de datos	C4.5	kNN	Logistic	MLP	SMO	ACID
breast	0.591	0.605	0.595	0.659	0.632	**0.710**
bupa	0.614	0.652	0.659	0.535	0.500	**0.654**
cleveland	0.292	0.297	0.319	0.298	0.310	**0.405**
haberman	0.578	0.583	0.564	**0.649**	0.500	0.591
heart	0.775	0.803	0.835	0.833	0.833	**0.847**
hepatitis	0.679	0.732	0.641	0.820	0.693	**0.841**
mammographic	**0.838**	0.818	0.828	0.459	0.824	0.753
newthyroid	0.894	0.914	0.956	0.695	0.767	**0.986**
pima	0.687	0.690	**0.730**	0.708	0.714	0.700
post-operative	0.328	0.343	0.326	**0.641**	0.336	0.539
saheart	0.618	0.607	0.669	0.643	**0.688**	0.646
spectfheart	0.565	0.701	0.606	0.579	0.509	**0.762**
thyroid	**0.976**	0.593	0.724	0.447	0.518	0.756
wdbc	0.479	0.475	0.487	0.500	0.477	**0.970**
wisconsin	0.502	0.511	0.512	0.510	0.503	**0.979**
Total de bancos de datos en los que fue mejor	2	0	1	2	1	**9**

Tabla 4. Tasa promedio verdadera positiva obtenida por los algoritmos de clasificación

Los peores resultados se obtuvieron en el conjunto de datos de *Cleveland*, donde el mejor resultado de diagnóstico fue una *TPR* promedio de 0.45. Otros conjuntos de datos difíciles de diagnosticar son *bupa*, *haberman* y *saheart*, con resultados de *TPR* promedio inferiores a 0.70 (se recomienda un valor superior a este).

El algoritmo propuesto obtuvo muy buenos resultados, superando a otros clasificadores en nueve de los 15 conjuntos de datos.

El algoritmo *ACID* obtuvo buenos resultados para los conjuntos de datos *new-thyroid, wdbc* y *wisconsin*, con valores de *TPR* promedio por encima de 0.97. Además, obtuvo un aumento en el rendimiento de hasta el 47%, considerado como el segundo mejor algoritmo de clasificación.

En la figura 18 se muestra la diferencia en el rendimiento con respecto al algoritmo propuesto y el segundo mejor algoritmo (rendimiento positivo); y con respecto a los mejores algoritmos, en 6 de 15 de los conjuntos de datos evaluados el algoritmo propuesto no obtuvo los mejores resultados (rendimiento negativo).

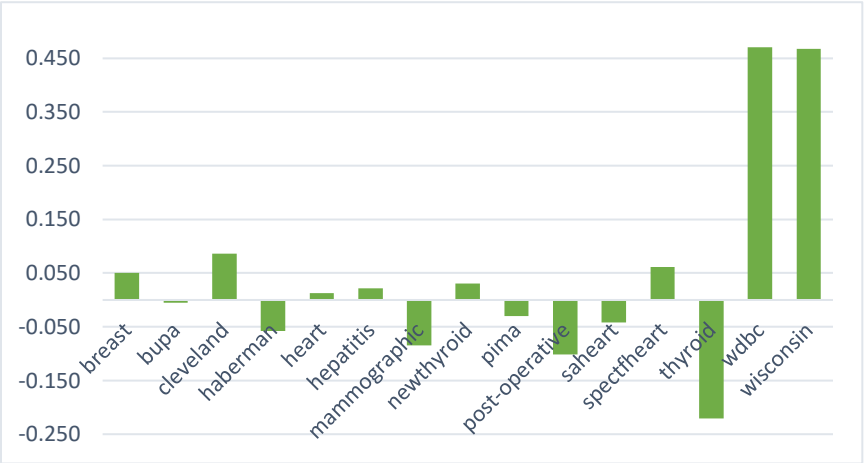

Figura 18. Diferencia en el rendimiento de ACID frente a otros algoritmos

Para saber si el algoritmo propuesto es el más apropiado para el prediagnóstico médico de enfermedades crónicas, se aplicó la prueba de Friedman [57] [58], dando un valor de probabilidad de $p = 0.001586$, que está en gran medida por debajo del nivel de significancia establecido de $\alpha = 0.05$ para una confianza del 95%.

Los posicionamientos (*rankings)* de algoritmos de acuerdo con la prueba de Friedman se muestran en la tabla 5, donde el mejor clasificador para esta tarea es claramente *ACID*. Teniendo en cuenta los resultados de la prueba de Friedman, se aplicó una prueba *post-hoc*, la prueba de Holm [61].

No.	Algoritmo	Posicionamiento *(Ranking)*
1	ACID	1.867
2	Logistic	3.000
3	MLP	3.567
4	kNN	3.933
5	SMO	4.100
6	C4.5	4.533

Tabla 5. Mejores algoritmos de acuerdo a Friedman: el mejor intérprete es ACID

La prueba rechaza la hipótesis cuyo valor ajustado es inferior o igual a *0.05*. Es decir, que existen diferencias significativas en el *TPR* promedio obtenido por la propuesta y por cualquier otro clasificador, esto se puede observar en la tabla 6.

i.	Algoritmo	z	p	Prueba de Holm
5	C4.5	3.9036	0.000095	0.01
4	SMO	3.269265	0.001078	0.0125
3	kNN	3.02529	0.002484	0.016667
2	MLP	2.488545	0.012827	0.025
1	Logistic	1.65903	0.09711	0.05

Tabla 6. Comparación post-hoc obtenida por la prueba Holm

Estos resultados confirman que el clasificador *ACID* es adecuado para el diagnóstico médico, con resultados de *TPR* promedio significativamente mejores que *MLP, C4.5, 3-NN, SMO* y clasificadores logísticos.

En este capítulo se concluye que en el análisis del prediagnóstico de varias enfermedades, el modelo *ACID* obtuvo muy buenos resultados; debido a que supera significativamente a otros clasificadores en escenarios médicos. Dichos resultados respaldan la afirmación de que la propuesta es muy útil para el prediagnóstico de enfermedades crónicas.

Capítulo 6. Conclusiones y Trabajo a Futuro

En este capítulo se presenta una conclusión general del trabajo de investigación y los posibles trabajos a futuro que podrían surgir a partir de este.

En el presente trabajo de tesis, se propuso y presentó un nuevo modelo de clasificación, diseñado para el diagnóstico médico, llamado *ACID* (clasificación asistida para el modelo de datos de desequilibrio o por sus siglas en inglés *Assisted Classification for Imbalance Data model*), que es capaz de manejar datos no balanceados, con atributos mezclados categóricos y numéricos y valores perdidos, además de que este modelo se ocupa de la presencia de pequeños disjuntos en datos no balanceados y el solapamiento de clases.

La idea principal de *ACID* es estructurar datos y encontrar la estructura más similar a la instancia para clasificar. Este peculiar funcionamiento permite que los problemas de manejo tengan la misma clase en diferentes regiones de decisión. Además, al considerar la estructura más cercana a cada clase, *ACID* maneja con éxito datos no balanceados, debido a que cada clase tiene la misma representación para propósitos de clasificación por lo que lo convierte en un algoritmo imparcial.

Por otro lado, por datos de estructura, *ACID* reduce la influencia de datos ruidosos y atípicos, lo que facilita la clasificación correcta de instancias.

Un aspecto clave de *ACID* es que es un modelo interpretable; es decir, *ACID* es transparente, debido a que sabemos exactamente porqué una instancia pertenece a una determinada clase.

Los resultados experimentales ilustran el buen rendimiento de *ACID*, debido a que supera a varios clasificadores de estado del arte, en nueve de 15 conjuntos de datos de las enfermedades crónicas más comunes.

Además, de acuerdo con la prueba de Friedman, el mejor clasificador en los experimentos llevados a cabo es *ACID*; por otra parte, la prueba Holm *post-hoc* concluye que existen diferencias significativas en la Tasa de Verdadero Promedio (*TPR*) obtenida por el algoritmo propuesto y por cualquier otro clasificador.

Lo anterior confirma que el clasificador *ACID* es adecuado para el prediagnóstico de enfermedades crónicas, con mejores resultados que los clasificadores *MLP, C4.5, 3-NN, SMO* y *Logistic*; lo anterior

cumple con el objetivo general de proponer un modelo de clasificación inteligente de patrones capaz de lidiar con datos mezclados, perdidos, no balanceados, con clases solapadas, con presencia de pequeños disjuntos, para el prediagnóstico de enfermedades crónicas.

En el proceso, se cumplieron los objetivos específicos que fueron:

- Recolectar bancos de datos que abarquen las enfermedades crónicas más comunes.
- Estudiar de manera experimental los algoritmos clásicos de clasificación inteligente de patrones.
- Evaluar si el nuevo modelo propuesto es el adecuado para el prediagnóstico de enfermedades crónicas.
- Realizar pruebas de significancia estadística que permitan identificar que algoritmo es el que sobresale de entre los otros.

Como trabajo futuro, se propone aplicar el modelo *ACID* sobre otros conjuntos de datos y evaluar si este algoritmo tiene buenos resultados con otro tipo de aplicaciones que no sean médicas, además se propone realizar más comparaciones con respecto a otros clasificadores supervisado.

Se propone el desarrollo posterior de una versión difusa del modelo *ACID,* esto se facilitaría en un futuro debido a que se codificaron los pesos en el intervalo [0,1] en el algoritmo de Evolución Diferencial (*Differential Evolution - DE*).

También se propone considerar los problemas multiclase y las versiones difusas de *ACID* aplicadas a otras líneas de investigación.

Referencias

[1] Fazekas, M. (2006). Analysing Data of Childhood Acute Lymphoid Leukaemia by Seasonal Time Series Methods. J. UCS, 12(9), 1190-1195.

[2] Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251.

[3] Chang, C. C., Cheng, C. S., & Huang, Y. S. (2006). A Web-Based Decision Support System for Chronic Diseases. J. UCS, 12(1), 115-125.

[4] Havlik, J., Lhotska, L., Parak, J., Dvorak, J., Horcik, Z., & Pokorny, M. (2013). A Modular System for Rapid Development of Telemedical Devices. J. UCS, 19(9), 1242-1256.

[5] Rijo, R., Silva, C., Pereira, L., Gonçalves, D., & Agostinho, M. (2014). Decision Support System to Diagnosis and Classification of Epilepsy in Children. J. UCS, 20(6), 907-923.

[6] World Health Organization. (1973). Chronic diseases. Chronic diseases.

[7] Sanchez-Santana, M. A., Aupet, J. B., Betbeder, M. L., Lapayre, J. C., & Camarena-Ibarrola, A. (2013). A tool for telediagnosis of cardiovascular diseases in a collaborative and adaptive approach. Journal of Universal Computer Science, 19(9), 1275-1294.

[8] Lu, A., Jiang, M., Zhang, C., & Chan, K. (2012). An integrative approach of linking traditional Chinese medicine pattern classification and biomedicine diagnosis. Journal of Ethnopharmacology, 141(2), 549-556.

[9] Barrett, S. (2011). Be Wary of Acupuncture, Qigong, and" Chinese Medicine.

[10] Who.int. (2018). OMS Capítulo 1: Salud mundial: retos actuales. [En línea] Disponible en: http://www.who.int/whr/2003/chapter1/es/index1.html [Fecha de acceso: Diciembre 2017].

[11] Schutte, A. E. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016:, a systematic analysis for the Global Burden of Disease Study 2016.

[12] Wang, L., Kong, L., Wu, F., Bai, Y., & Burton, R. (2005). Preventing chronic diseases in China. The lancet, 366(9499), 1821-1824.

[13] Abdel-Misih, S. R., & Bloomston, M. (2010). Liver anatomy. Surgical Clinics, 90(4), 643-653.

[14] Canadian Liver Foundation, 2013. [En línea]. Disponible en: http://ps70.sb.marqui.com/support-liver-foundation/advocate/Liver_Disease_in_Canada_Report.aspx. [Fecha de acceso: Febrero 2018].

[15] UC Irvine Machine Learning Repository. [En línea] Disponible en: http://archive.ics.uci.edu/ml/ [Fecha de acceso: Octubre 2017].

[16] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[17] Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. SpringerPlus, 5(1), 89.

[18] Dawes, T. J., de Marvao, A., Shi, W., Fletcher, T., Watson, G. M., Wharton, J., ... & Cook, S. A. (2017). Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study, Radiology, 283(2), 381-390.

[19] Yu, X., Ye, C., & Xiang, L. (2016). Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing, 214, 376-381.

[20] Imamura, H., Sano, K., Sugawara, Y., Kokudo, N., & Makuuchi, M. (2005). Assessment of hepatic reserve for indication of hepatic resection: decision tree incorporating indocyanine green test. Journal of hepato-biliary-pancreatic surgery, 12(1), 16-22.

[21] Kanas, V. G., Zacharaki, E. I., Thomas, G. A., Zinn, P. O., Megalooikonomou, V., & Colen, R. R. (2017). Learning MRI-based classification models for MGMT methylation status prediction in glioblastoma. Computer Methods and Programs in Biomedicine, 140, 249.

[22] Sun, Y., Reynolds, H., Wraith, D., Williams, S., Finnegan, M. E., Mitchell, C., ... & Haworth, A. (2017). Predicting prostate tumour location from multiparametric MRI using Gaussian kernel support vector machines: a preliminary study, Australasian physical & engineering sciences in medicine, 40(1), 39-49.

[23] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.

[24] Duda, R. O., Hart, P. E., & Stork, D. G. (1973). Pattern classification (pp. 526-528). Wiley, New York.

[25] Wilson, D. R., & Martinez, T. R.: "Improved heterogeneous distance functions"; Journal of Artifficial Intelligent Research, (1997), 1-34.

[26] Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). Learning from data (Vol. 4). New York, NY, USA:AMLBook.

[27] Minsky, M. (5). Paper, S.(1969). Perceptrons.

[28] Bezdek, J. C. (1992). On the relationship between neural networks, pattern recognition and intelligence. International journal of approximate reasoning, 6(2), 85-107.

[29] Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4), 341-359.

[30] Ramirez, A., Lopez, I., Villuendas, Y., & Yanez, C. (2015). Evolutive improvement of parameters in an associative classifier. IEEE Latin America Transactions, 13(5), 1550-1555.

[31] Ruiz-Shulcloper, J., & Abidi, M. A. (2002). Logical combinatorial pattern recognition: A Review.

[32] KEEL dataset repository. [En línea] Disponible en: https://www.keel.es. [Fecha de acceso: Octubre 2017].

[33] KEEL dataset repository. Breast Cancer data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=97. [Fecha de acceso: Septiembre 2017].

[34] KEEL dataset repository. Liver Disorders (BUPA) data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=55. [Fecha de acceso: Septiembre 2017].

[35] KEEL dataset repository. Heart Disease (Cleveland) data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=57. [Fecha de acceso: Septiembre 2017].

[36] UCI Machine Learning repository. Haberman's Survival Data Set. [En línea] Disponible en: https://archive.ics.uci.edu/ml/machine-learning-databases/haberman/. [Fecha de acceso: Septiembre 2017].

[37] KEEL dataset repository. Statlog (Heart) data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=99. [Fecha de acceso: Septiembre 2017].

[38] UCI Machine Learning repository. Hepatitis Data Set. [En línea] Disponible en: https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/. [Fecha de acceso: Septiembre 2017].

[39] UCI Machine Learning repository. Mammographic Mass Data Set. [En línea] Disponible en: http://archive.ics.uci.edu/ml/machine-learning-databases/mammographic-masses/. [Fecha de acceso: Septiembre 2017].

[40] KEEL dataset repository. Thyroid Disease (New Thyroid) Multi-class Imbalanced data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=1072. [Fecha de acceso: Septiembre 2017].

[41] UCI Machine Learning repository. Pima Indians Diabetes Data Set. [En línea] Disponible en: https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/. [Fecha de acceso: Septiembre 2017].

[42] UCI Machine Learning repository. Post-Operative Patient Data Set. [En línea] Disponible en: https://archive.ics.uci.edu/ml/machine-learning-databases/postoperative-patient-data/. [Fecha de acceso: Septiembre 2017].

[43] KEEL dataset repository. South African Hearth data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=184. [Fecha de acceso: Septiembre 2017].

[44] KEEL dataset repository. SPECTF Heart data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=185. [Fecha de acceso: Septiembre 2017].

[45] UCI Machine Learning repository. Thyroid Disease Data Set. [En línea] Disponible en: http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/. [Fecha de acceso: Septiembre 2017].

[46] UCI Machine Learning repository. Breast Cancer Wisconsin (Diagnostic) Data Set. [En línea] Disponible en: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/. [Fecha de acceso: Septiembre 2017].

[47] KEEL dataset repository. Breast Cancer Wisconsin (Original) data set. [En línea] Disponible en: http://sci2s.ugr.es/keel/dataset.php?cod=73. [Fecha de acceso: Septiembre 2017].

[48] Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In Soft computing and industry (pp. 25-42). Springer, London.

[49] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.

[50] Napierała, K., Stefanowski, J., & Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In Rough sets and current trends in computing (pp. 158-167). Springer Berlin/Heidelberg.

[51] Fernández, A., López, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-based systems, 42, 97-110.

[52] Dubes, R. C., & Jain, A. K. (1988). Algorithms for clustering data.

[53] Trinidad, J. F. M., Shulcloper, J. R., & Cortés, M. S. L. (2000). Structuralization of universes. Fuzzy sets and systems, 112(3), 485-500.

[54] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[55] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.

[56] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), 1-30.

[57] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the american statistical association, 32(200), 675-701.

[58] Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 11(1), 86-92.

[59] Garcia, S., & Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research, 9(Dec), 2677-2694.

[60] García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences, 180(10), 2044-2064.

[61] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65-70.

[62] Alcalá-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., ... & Fernández, J. C. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing, 13(3), 307-318.

[63] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic, & Soft Computing, 17.

[64] Yu, X., & Gen, M. (2010). Introduction to evolutionary algorithms. Springer Science & Business Media.