Análisis de bancos de datos médicos y financieros mediante algoritmos de cómputo inteligente

Analysis of medical and financial dataset through intelligent computing algorithms

FRANCISCO-DE LA CRUZ, María Mercedes†*, REYES-LEÓN, Patricio y SALGADO-RAMÍREZ, Julio César

ID 1er Autor: *María Mercedes, Francisco-de-la-Cruz* / **ORC ID:** 0000-0002-6000-0518, **CVU CONACYT ID:** 927575

ID 1er Coautor: *Patricio, Reyes-León* / **ORC ID:** 0000-0001-8521-9236, **CVU CONACYT ID** 928413

ID 2do Coautor: *Julio César, Salgado-Ramírez* / **ORC ID:** 0000-0003-1666-9924, **CVU CONACYT ID** 88983

DOI: 10.35429/P.2020.4.73.88

M. Francisco, P. Reyes y J. Salgado

Universidad Politécnica de Pachuca

*maria_francisco@micorreo.upp.edu.mx

F. Trejo (Coord.). Ciencias Multidisciplinarias. Proceedings-©ECORFAN-México, Pachuca, 2020.

Resumen

Los modelos de cómputo inteligente aplicados en áreas como finanzas y medicina, se han convertido en un área creciente para el interés científico. En el presente documento se muestra un estudio documental y experimental de la aplicación de los modelos de cómputo inteligente con mayor presencia en la literatura, realizando una investigación de las aplicaciones más importantes que se han hecho hasta el momento para la clasificación de patrones de bancos de datos financieros y médicos. Así mismo mediante el uso de la plataforma WEKA y de los repositorios KEEL y UCI, se estudia el desempeño que exhiben esos clasificadores de patrones al ser aplicados en la clasificación de patrones de bancos de datos financieros y médicos.

Algoritmos inteligentes, Enfermedades, Finanzas, Minería de datos

Abstract

Intelligent computing models applied in areas such as finance and medicine have become an area of growing scientific interest. This document shows a documentary and experimental study of the application of intelligent computing models with the greatest presence in the literature, carrying out an investigation of the most important applications that have been carried out so far for the classification of patterns of medical and finance datasets. Likewise, through the use of the WEKA platform and the KEEL and UCI repositories, the performance exhibited by these pattern classifiers is studied when applied in the classification of patterns from financial and medical datasets.

Intelligent algorithms, Diseases, Finance, Data mining

1. Introducción

Los algoritmos de cómputo inteligente son programas computacionales que son capaces de procesar grandes volúmenes de información y su razón de ser es el detectar patrones y sus comportamientos. El saber cómo se comportan los patrones permite a las computadoras sugerir decisiones y el usuario de estos algoritmos decidirá si la decisión sugerida es la esperada, que en la mayoría de los casos así sucede, lo que permite que se desarrollen herramientas computacionales basadas en algoritmos de cómputo inteligente en áreas sensibles del quehacer humano.

Toda actividad del ser humano se basa en la toma de decisiones lo que implica el hacer un análisis formal de la información del contexto en el que se encuentra éste para tomar una decisión responsable que lo beneficie. La información del contexto se basa en una serie de características agrupadas que definen un comportamiento, la agrupación de estas características es a lo que se le llama patrón y cada patrón tiene características bien definidas que le permiten diferenciarse de otros patrones, en esta parte, los algoritmos de cómputo inteligente pueden ser una herramienta útil para el ser humano.

El ser humano está muy interesado en la predicción de eventos que le inquietan y no se está hablando de cuestiones zodiacales o esotéricas sino del análisis de patrones de comportamientos. Por ejemplo, una persona programa su dispositivo para que suene una alarma y lo despierte a las 7:00 am, pero le indica al dispositivo que la alarma suene a las 7:15 de nuevo. Esto lo hace porque sabe que entra a trabajar a las 9:00 y que su trabajo queda a 15 minutos de su hogar. La pregunta que surge es ¿Por qué le indica al dispositivo que suene 15 minutos después de la primera hora que le indicó? La respuesta es simple, él conoce su rutina (patrón de comportamiento) y sabe que es muy probable que no se levantará a esa hora y que eso no le afectará en sus actividades. Esto es un claro ejemplo de conocer la información de su contexto y tomar una decisión que no le afecta. La realidad es que no toda decisión es tan fácil para el ser humano.

Hay dos campos muy sensibles para el ser humano que históricamente han sido fuentes de preocupación en las tomas de decisiones, uno es la salud y otro son los riesgos financieros. Amable lector piense en lo siguiente, si una persona pudiera saber si tiene un padecimiento crónico degenerativo antes siquiera comience a manifestarse ¿Qué decisión tomaría? Ahora bien ¿Qué pensaría amable lector, si supiera una persona cómo determinar si existe riesgo financiero con su dinero? ¿Cómo lo manejaría? Son interesantes estas preguntas y más interesante sería saber que existen herramientas útiles que pueden ayudar a responder las preguntas antes mencionadas.

Antes de tener respuestas a las preguntas mencionadas en el párrafo anterior es necesario dar un contexto de la información (patrones) que usan los algoritmos de cómputo inteligente. Los patrones debieron ser analizados previamente para determinar los rasgos o características que son suficientes y necesarios para conformarlos. Para probar la eficiencia de los algoritmos de cómputo inteligente, hay dos posibles caminos. En el primero los patrones son tomados de repositorios de datos de información real que la comunidad científica ha analizado y ha procesado, y en el segundo, es haciendo análisis estadístico, minería de datos, etc. ¿Cuál de estos dos métodos de uso de patrones se debe tomar en cuenta? La respuesta es simple. Si se requiere determinar que está sucediendo en un caso de uso real es necesario hacer minería de datos, análisis estadístico y otros procesos. Si se requiere analizar el rendimiento o performance de un algoritmo se hace uso de repositorios de datos reales que ya han sido procesado durante años por la comunidad científica.

La presente investigación tomará la información de repositorios de la comunidad científica para analizar el rendimiento en la clasificación de patrones con los algoritmos más usados en el estado del arte, bajo un ambiente de aprendizaje supervisado. ¿Qué se logra al saber el rendimiento o performance de los algoritmos de cómputo más usados en el estado del arte? Se logra proporcionar información categórica de qué algoritmo es más eficiente en algún área del quehacer humano, lo cual, permitirá desarrollar aplicaciones inteligentes donde computadoras u otros dispositivos al presentarles patrones de comportamiento predecirán situaciones que le permitirán a sus usuarios decidir en situaciones sensibles o permitir que los dispositivos o computadores tome decisiones.

Esta investigación se centra principalmente en dos rubros uno la salud y la otra en riegos financieros. En las siguientes secciones se hablará más detalladamente de los repositorios de datos, así como de qué situaciones como de enfermedad o finanzas son usados en la clasificación y de los algoritmos de cómputo inteligente más usados en el estado del arte.

En la sección 2 se muestran los trabajos relacionados a uso de modelos inteligentes en el área financiera y médica mencionando el trabajo realizado por investigadores para estas áreas. En la Sección 3 se muestran las herramientas utilizadas para el desarrollo de este trabajo, mostrando los bancos de datos y algoritmos seleccionados. En la sección 3 se muestran las conclusiones que se deducen de la fase experimental de este documento.

2. Trabajos relacionados

A continuación, se describen brevemente los trabajos relacionados analizados para el desarrollo de esta investigación.

Kim, M. propone un método de minería de datos basado en algoritmos genéticos para descubrir reglas de decisión de quiebra financiera a partir de decisiones cualitativas de expertos, mostrando que el algoritmo genético genera las reglas que tienen mayor precisión y mayor cobertura que los métodos de aprendizaje inductivo y las redes neuronales (Kim & Han, n.d.).

Peng, Y. propone desarrollar un enfoque de dos pasos para evaluar los algoritmos de clasificación para la predicción del riesgo financiero; construyendo una puntuación de rendimiento para medir los algoritmos de clasificación e introduciendo tres métodos de toma de decisiones de criterios múltiples (MCDM) éstos son los métodos TOPSIS, PROMETHEE y VIKOR para proporcionar una clasificación final. Los resultados mostraron que la regresión logística, la red bayesiana y los métodos de conjunto se colocaron como los tres clasificadores principales por TOPSIS, PROMETHEE y VIKOR (Peng, Y., *et al*., 2011).

Moloud Abdar en (Abdar *et al*., 2017), utilizó dos modelos de árbol de decisión llamados C5.0 mejorado y CHAID, para identificar con mayor precisión enfermedades del hígado. La precisión obtenida por C5.0 mejorado fue de 93.7 % siendo mejor comparado con CHAID el cual obtuvo 65.00 %, se observó que el género es un factor importante para el diagnóstico de la enfermedad, y se comparó el rendimiento obtenido por los algoritmos propuestos con el obtenido por los modelos de la literatura, concluyendo que C5.0 mejorado y CHAID tienen mejor desempeño en conjunto, en comparación con los modelos contra los que se comparó.

En un estudio de clasificación de regiones codificantes de proteínas en genomas procariotas realizado por Al Bataineh (Al Bataineh & Al-qudah, 2017), proponen un algoritmo para la clasificación de genes utilizando un clasificador bayesiano, obteniendo resultados competitivos, en comparación con los métodos de detección de genes conocidos en procariotas como GLIMMER y GeneMark dando oportunidad de utilizar su algoritmo para la detección de enfermedades.

Por su parte Chang, C. propone en su investigación (Chang, 2006), un modelo para el apoyo al personal médico en la toma de decisiones, para el diagnóstico de enfermedades crónicas, basado en un clasificador bayesiano. Dicho modelo resulta efectivo como apoyo en la toma de decisiones para el diagnóstico de enfermedades crónicas.

Vyas, R. menciona en su investigación (Vyas *et al*., 2016) que para entender una enfermedad es necesario comprender los mecanismos moleculares subyacentes, como el número de interacciones proteína-proteína el cual es muy limitado en comparación con las secuencias de proteínas disponibles, enfocando su investigación a la enfermedad de diabetes mellitus, propuso un modelo basado en Máquinas de Soporte Vectorial (MSV), para clasificar las huellas estructurales y genómicas correspondientes a la enfermedad y las que no lo son, obteniendo una exactitud de clasificación del 78,2 %.

Para la enfermedad de cáncer de mama, Mungle, T. proponen en su estudio (Mungle *et al*., 2017) un algoritmo de agrupamiento híbrido k-means para cuantificar el índice proliferativo de células de cáncer de mama basado en el conteo automático de núcleos Ki-67, por medio de imágenes RGB de cáncer de mama teñido con K-67.Obteniendo un buen desempeño en el modelo propuesto, donde destaca una taza de error de 6.84 %.

En un estudio realizado por Golub, T. (Golub *et al*., 1999), se propuso utilizar las expresiones genéticas de microarreglos de ADN, para la clasificación de cáncer en leucemias agudas, utilizando un clasificador automático, esta propuesta surge debido a que, en ese momento las técnicas para identificar nuevas clases de cáncer, a pesar de mejorar bastante, no contemplaban ampliamente el uso de clasificadores automáticos para esa tarea. Se lograron identificar nuevos tipos de cáncer de leucemia con el uso algoritmos de cómputo inteligente.

Para las enfermedades crónicas del riñón, Polat H. (Polat *et al*., 2017) menciona que la precisión de los algoritmos de clasificación, depende del uso correcto de técnicas de filtrado de características, para reducir la dimensión de los con- juntos de datos. Por lo que en su propuesta utilizó un algoritmo de MSV para el diagnóstico de enfermedades renales crónicas, empleando el método wrapper para evaluar subconjuntos filtrados de atributos, para reducir la dimensionalidad del dataset seleccionado. Se encontró que utilizando las técnicas de reducción de dimensionalidad y MSV, la exactitud para el diagnóstico fue de 98.5%.

En un estudio para realizar el diagnóstico de enfermedades crónicas Villuendas Rey (Villuendas-Rey *et al*., 2018), propuso un modelo de Clasificación Asistida para Datos Desequilibrados (ACID), en el cual se diseñó un nuevo algoritmo de clasificación, que trabaja con clases desbalanceadas, atributos mezclados y valores faltantes, reduce dimensionalidad, hace agrupamiento de subclases, y disminuye la influencia que tiene la superposición de clases, se comparó el desempeño del modelo con algoritmos como C4.5, KNN, MSV, entre otros para distintos conjuntos de datos de enfermedades crónicas. Con resultados TPR (taza de verdaderos positivos) promedio significativamente mejores que los clasificadores MLP, C4.5, 3-NN, SMO y logísticos.

Padmavathy, T. en su investigación para la detección de cáncer de mama por medio de imágenes (Padmavathy *et al*., 2019), propone un sistema de inferencia adaptativo neuro-difuso (ANFIS) para la clasificación de imágenes, se observó una exactitud del modelo de 98.02 % para la clasificación.

Guidi, G. en su investigación (Guidi *et al*., 2017), propone realizar la detección de cáncer de mama por medio de imágenes. Identificando automáticamente a los pacientes que pueden beneficiarse de un tratamiento adaptativo comparando la radioterapia administrada y la planificada. Utilizó máquinas de vectores de soporte y algoritmos de agrupamiento de K-means. La herramienta desarrollada pudo clasificar a los pacientes con diferentes niveles generales de variaciones morfológicas y predecir posibles problemas causados por diferencias relevantes entre la dosis planificada y la administrada.

González Patiño (González-Patiño *et al*., 2020), propone en su artículo probar metaheurísticas aplicadas a la segmentación de imágenes de mastografías, para la detección oportuna de cáncer de mama, la aplicación de estos algoritmos tiene una relación directa con problemas de optimización; sin embargo, en este estudio, su implementación está orientada a la segmentación de mastografías. Los resultados mostraron una menor tasa de error al utilizar estas metaheurísticas para la segmentación, en comparación con el método clásico conocido como Otsu.

Velázquez Rodríguez propone en su artículo (Velázquez-Rodríguez *et al*., 2020), una transformación matemática simple a el algoritmo asociativo Lernmatrix, dicha transformación elimina las alteraciones sustractivas entre patrones. Mejorando el rendimiento de la Lernmatrix significativamente para la clasificación de patrones, comparando el rendimiento con los algoritmos más significativos en la literatura, utilizando 20 conjuntos de datos con clases desbalanceadas, obteniendo buenos resultados, utilizando la métrica de rendimiento de exactitud balanceada.

Michael L. Raymer (Raymer *et al*., 2003), muestra en su investigación que utilizando un clasificador híbrido basado en la función discriminante de Bayes, que a su vez emplea la selección y extracción de características, para aislar características relevantes de grandes conjuntos de datos médicos. Se obtiene una buena discriminación de las características más relevantes de los conjuntos de datos médicos, para enfermedades, entre las que destacan hepatitis, diabetes y enfermedad de tiroides

Andrew I. Schein (Schein *et al*., 2004), propone un método novedoso para el aprendizaje activo de clasificadores de regresión logística basado en la función objetivo A-óptima, para la clasificación de los dataset: Wisconsin, thyroid, y splice junction gene sequence, del repositorio de la Universidad de Irvine California. Mostrando un desempeño para la métrica de exactitud de 57 %, para wisconsin, 91.8 % para splice junction gene sequence y 56 % para thyroid.

Lukasz A. Kurgan (Kurgan *et al*., 2001), describe un proceso computarizado de diagnóstico de perfusión miocárdica a partir de imágenes de tomografía computarizada por emisión de protón único (SPECT), para la detección de enfermedades cardíacas, utilizando técnicas de minería de datos, obteniendo como resultado, la creación de un dataset, que consta de 267 imágenes SPECT de pacientes previamente procesadas, y con información clínica e interpretación médica para enfermedades cardíacas.

H. Patel menciona en su investigación (Patel & Singh Thakur, 2017), que el uso de algoritmos para la clasificación de bancos de datos desbalanceados, muestran un sesgo que beneficia a la clase mayoritaria, lo que se convierte en un inconveniente al momento de reportar resultados. Por lo que propone un nuevo método de K vecino más cercano ponderado difuso, que maneja de manera óptima el problema del desequilibrio de los datos. Probando el modelo propuesto con los dataset: Spectfheart, ecoli1, vertebral, ionosphere y glass0. Obteniendo un buen rendimiento en comparación con los algoritmos KNN tradicionales.

Dheeba, J. propone en su investigación, realizar el diagnóstico asistido por computadora (CAD), para la detección oportuna de cáncer de mama mediante la detección de anomalías mamarias en mastografías digitales, utilizando un clasificador basado en redes neuronales llamado Red Neural Wavelet Optimizada de Enjambre de Partículas (PSOWNN), mostrando un desempeño de 93.67 % para la métrica de exactitud balanceada (Patel & Singh Thakur, 2017).

En el estudio realizado por Chen, H. L. y su grupo de trabajo, presentan un sistema experto de tres etapas basado en un enfoque de máquinas de soporte vectorial híbrido, para diagnosticar enfermedades de tiroides, centrándose en técnicas de selección de atributos, para mejorar el rendimiento del clasificador propuesto, realizando la clasificación del conjunto de datos, discriminando un atributo a la vez, obteniendo un desempeño máximo de exactitud de 98.59 % (Chen *et al*., 2012).

3 Métodos y materiales

En este capítulo se mencionan los algoritmos, conceptos y herramientas necesarias para la elaboración de esta investigación.

3.2 WEKA

Es importante mencionar que para la fase experimental de este proyecto, se hará uso de la plataforma de software Weka para el aprendizaje supervisado (Hall *et al*., 2014), la cual cuenta con un compendio de herramientas para la minería de datos, entre las que se encuentran diversos algoritmos de clasificación, los resultados obtenidos por esta herramienta de software, son perfectamente válidos ya que esta es ampliamente utilizada en investigaciones de este tipo (Soni et al., 2019).

3.3 Algoritmos de clasificación

A continuación, se describen brevemente los nueve algoritmos clasificadores de patrones que se han seleccionado en la plataforma WEKA.

NaïveBayes: este clasificador pertenece al enfoque probabilístico. Para la clasificación, utiliza el Teorema de Bayes de una forma ingenua (Naïve), ya que considera a todos los atributos probabilísticamente independientes (Bhargavi, *et al*., 2009) en lo sucesivo se utilizará la nomenclatura **NB** para referirse a este clasificador.

SVM: este algoritmo pertenece a las máquinas de soporte vectorial (SVM), las cuales son un conjunto de métodos para clasificación de patrones, para su funcionamiento utilizan la optimización de funciones y los llamados vectores de soporte. Este tipo de algoritmos tratan de encontrar un hiperplano que separe de la mejor manera a las clases; de no conseguirlo, se utiliza el kernel para transformar los patrones a un espacio de dimensión mayor a la original. Un kernel puede ser una función lineal, polinomial, de base radial o sigmoide, en lo sucesivo se utilizará la nomneclatura **SMO** para referirse a este clasificador ya que en la plataforma WEKA se le conoce con este nombre (Cortes, *et al*., 1995).

Logistic: este algoritmo es la regresión logística, la cual es una técnica estadística de aprendizaje automático. 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 sigmoidea, involucrando a la función exponencial para ello en lo sucesivo se utilizará la nomenclatura **RL** para referirse a este clasificador (Bootkrajang, *et al*., 2014).

Multilayer Perceptron: el MLP es una red neuronal artificial formada por múltiples capas (layers), con cuyo diseño se intenta resolver problemas de clasificación con clases que no son linealmente separables. El MLP es considerado por la comunidad científica como un excelente clasificador de patrones. El MLP está formado principalmente por tres capas: la capa de entrada, la capa oculta y la capa de salida: En esta última capa se encuentran las neuronas cuyos valores de salida corresponden a la etiqueta de clase, en lo sucesivo se utilizará la nomeclatura **MLP** para referirse a este clasificador (Rosenblatt, 1958).

Árboles de decisión: este tipo de algoritmos de clasificación son de los más usados en tareas de clasificación de patrones. Los árboles de decisión son apreciados porque son explicables, están basados en la teoría de grafos y permiten ver de forma estructurada cómo se clasifican las instancias de un conjunto de datos. La estructura contiene un nodo raíz en la parte superior del árbol, y los nodos intermedios llamados hojas que corresponden a los atributos. Visualizando en la parte inferior a las clases, **J48** es el nombre que se da en WEKA a un tipo de árbol de decisión y en lo sucesivo se utilizará esta nomenclatura para referirse a él (Quinlan, 1990).

RandomTree: es un algoritmo de clasificación de patrones que consiste en la construcción de un árbol de decisión de manera aleatoria se utilizará la nomenclatura **RT** para referiste a este algoritmo (Kalmegh, 2015).

RandomForest: este clasificador consiste en un bosque aleatorio; es una combinación de árboles de decisión. Es decir que se generan múltiples árboles de manera aleatoria, y cada árbol, emite un voto unitario para la clase más popular, y de esa manera le es posible clasificar un patrón de entrada dado. Se usará la nomenclatura **RF** para referirse a él (Zhang, *et al*., 2014).

IB1: este algoritmo es la versión que ofrece WEKA del clasificador 1-NN, el cual asigna a un patrón de prueba la clase a la que pertenece su vecino más cercano (nearest neighbor). El acrónimo IB proviene del inglés “Instance Based” (Haro-García, *et al*., 2019).

IB3: este algoritmo es la versión que ofrece WEKA del clasificador 3-NN, el cual asigna a un patrón de prueba la clase que resulta por votación en los tres vecinos más cercanos.

LM-τ[9]: La *Lernmatrix* propuesta por Karl Steinbuch (Steinbuch, 1961) desarrollada en 1961 es el primer modelo de memorias asociativas conocido para la clasificación de patrones, sin embargo, el desempeño mostrado por este clasificador resulta poco competitivo si se compara con los algoritmos antes mencionados, sin embargo en un estudio realizado en 2020 (Velázquez-Rodríguez et al., 2020), se propone realizar una modificación a las generalidades con las que trabaja el algoritmo original, respetando las base en las que se fundamenta. Mediante dos transformaciones hechas a los datos de entrada, que son Johnson-Möbius y Tau 9, se logra convertir cualquier número flotante a binario y hace que los patrones de entrada tengan características distintivas para el proceso de la clasificación. Este clasificador no es parte de WEKA por lo que fue implementado en el lenguaje de programación Python para evaluar su desempeño.

Consta de dos fases, la primera fase de entrenamiento o aprendizaje consta de entrenar una memoria M conocida como matriz bidimensional de entrenada, y se hace con patrones binarios de entrada o vectores unitarios $X^{\mu} \in A^{n}$, A= {0,1}, generando como salida $y^{\mu} \in A^{m}$ el cual corresponde a un número de clases distintas m, además se cumple que cada valor correspondiente de m_{ij} de M de la Lernmatix tiene un valor inicial de 0, y de acuerdo a (1) se actualiza siguiendo la regla $m_{ij} + \Delta m_{ij}$ (Uriarte-Arcia *et al*. 2014) donde:

$$\Delta m_{ij} = \begin{cases} +\varepsilon \text{ si } y_i^{\mu} = 1 = x_j^{\mu} \\ -\varepsilon \text{ si } y_i^{\mu} = 0 \text{ y } x_j^{\mu} = 1 \\ 0 \text{ en otro caso} \end{cases} \quad (1)$$

Se considera el valor de ε = 1.

La segunda fase, la de recuperación o recuerdo consiste en identificar las coordenadas del patrón de salida al que pertenece un patrón de entrada dado, es decir a qué clase pertenece $x^{\omega} \in A^{n}$, indicando las coordenadas de y^{ω} mediante expresión (2).

$$y_i^{\omega} = \begin{cases} 1 \text{ si } \sum_{i=1}^{n} m_{ij} x_i^{\omega} = \text{MAX}\left[\sum_{j=1}^{n} m_{hj} x_j^{\omega}\right] \\ 0 \text{ en otro caso} \end{cases} \quad (2)$$

Para realizar las transformaciones de patrones requeridas por el algoritmo LM-τ[9], primero se realiza la transformación de Johnson-Möbius, colocando ceros y unos siguiendo las siguientes reglas:

Encontrar el valor mínimo del dataset de patrones y sumarlo a cada valor del dataset.

1. Truncar los decimales multiplicando por una potencia de diez adecuada.

2. Identificar el valor del dataset mayor y concatenar con unos y ceros siguiendo la ecuación: $e_m - e_j$ ceros con e_j unos, donde e_m es el elemento mayor del dataset.

Posteriormente se aplica la transformada 9 la cual consiste en concatenar con 0 y 1 en casi de encontrar un 0 en cada elemento del dataset y con 1 y 0 en caso de encontrar un 1, con la finalidad de hacer cada patrón más distinto uno de otro.

3.4 Reconocimiento de patrones

El reconocimiento de patrones es un área de las ciencias computacionales que tiene como objetivo, generar algoritmos que reconozcan situaciones cotidianas tal como lo hace el ser humano de manera automática (Lindberg, 2020), la información que describe a esas situaciones o eventos, está representada por patrones, tupas, vectores o matrices de datos, para que de esta manera los modelos generados sean capaces de reconocer un determinado evento, dentro del área de reconocimiento de patrones se llevan a cabo las tareas de clasificación de patrones, recuperación, regresión y agrupamiento. Para ello existen paradigmas de aprendizaje, el supervisado, el no supervisado, el semi supervisado y el aprendizaje por refuerzo.

En este trabajo se utilizará el paradigma de aprendizaje supervisado, en el cual se realizan únicamente dos tareas: clasificación o entrenamiento y recuperación o aprendizaje. En este tipo de aprendizaje un experto proporciona una etiqueta a un patrón, en un conjunto de datos de entrenamiento, y el objetivo de los algoritmos es clasificar patrón de entrada válido dado, después de haber sido entrenado.

3.5 Conjunto de datos

Los patrones que se utilizan para el aprendizaje y la clasificación suelen ser representados en matrices de datos. Esta representación de los datos es conocida como dataset, o conjunto de datos. Dentro del dataset representado como una tabla, (ver tabla 1), se encuentran filas y columnas, las filas representan las instancias o patrones, para este ejemplo se muestran registros médicos, y cada renglón o patrón es el registro completo de un paciente. Las columnas del dataset representan a los atributos, que son las características que describen el banco de datos, lo cual representa a las características que describen a los patrones, en este caso son características de paciente como: Sexo, edad, dolor de pecho, colesterol, entre otras. en la última columna de la derecha, también llamada columna objetivo, se muestran las clases, las cuales describen lo que se está buscando clasificar, para este caso se consideran dos clases: enfermo o sano.

Tabla 1 Banco de datos

Atributos					
Edad	Sexo	Dolor_pecho	Colesterol	…	Clases
63	1	3	233	…	1
37	1	2	255	…	1
.	.	.	.	…	.
.	.	.	.		.
.	.	.	.		.
60			282	…	0

Fuente: Elaboración propia agosto 2020

En esta investigación, se analizarán conjuntos de datos de dos clases, en los que se observa que el número de patrones pertenecientes a una clase superan a los de la otra clase, a este fenómeno se le conoce como desbalanceo de clases, para conocer el índice de desbalance (en inglés Imbalance Ratio o IR) de un dataset se utiliza (3).

$$\boldsymbol{IR} = \frac{\boldsymbol{Clase\ mayoritaria}}{\boldsymbol{Clase\ minoritaria}} \tag{3}$$

Si IR >1.5, se considera que el dataset está desbalanceado.

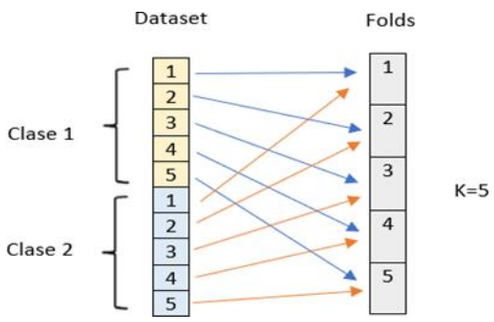
Es importante mencionar que existen repositorios de conjuntos de datos, que sirven como fuente de consulta, para la realización de investigaciones relacionadas al reconocimiento de patrones, los repositorios más sobresalientes son: el repositorio de aprendizaje automático de Irvine, http: //archive.ics.uci.edu/ml/datasets.php el cual, fue puesto a disposición de los grupos de investigación por la Universidad de California en Irvine de Estados Unidos. Además, se encuentra el repositorio de KEEL https://sci2s.ugr.es/keel/datasets.php, puesto a disposición para los grupos de investigación por la Universidad de Granada, en España.

3.6 Método de validación

Como se mencionó anteriormente un algoritmo de clasificación realiza dos tareas primordiales, el aprendizaje o recuperación y la clasificación. Para realizar estos procesos se utiliza el dataset, el cual se divide en dos subconjuntos de datos, un subconjunto para pruebas y otro para aprendizaje.

Esta división en subconjuntos se obtiene a partir de un método de validación, para esta investigación se utiliza el método de validación llamado k-fold coross-validation estratificado con k=5, el cual es el método de validación recomendado para datasets con un desbalanceo de clases (Lindberg, 2020). Este método consiste en dividir el dataset en 5 subconjuntos o folds, asegurándose de que todas las clases estén representadas en cada fold como muestra la figura 1 en la que se representa la división de un dataset con 2 clases.

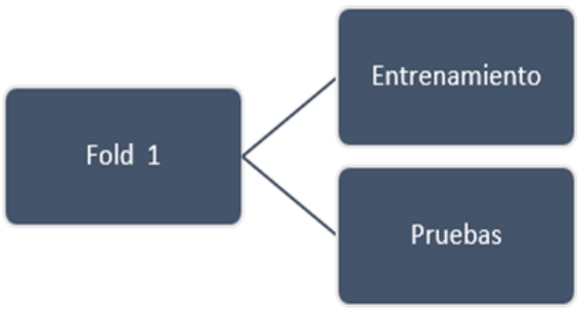
Figura 1 5-fold cross-validation



Fuente: Elaboración propia agosto 2020

Para los conjuntos de datos desbalanceados se recomienda utilizar una variante del método de validación 5-fold cross-validation, el cual consiste en dividir cada fold en dos conjuntos de datos, uno para entrenamiento y otro para pruebas, como muestra la figura 2.

Figura 2 División de fold.



Fuente: Elaboración propia agosto 2020

Realizando un total de cinco iteraciones, una para cada fold, este método es llamado 5 × 2 cross-validation.

3.7 Métricas de rendimiento

Ahora que se describió cómo se hace el entrenamiento y la clasificación para un algoritmo, se debe de explicar cómo se evalúa el rendimiento del algoritmo, es decir que tan bueno es clasificando los patrones de prueba, de cada iteración que se hace, para ello se utiliza la matriz de confusión que se muestra en la figura 4, la cual muestra cómo ha realizado la clasificación de patrones.

Figura 4 Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente: Elaboración propia agosto 2020

En donde:

VP = Verdaderos positivos: Muestra la cantidad de patrones positivos clasificados correctamente como positivos.

VN = Verdaderos negativos: Muestra la cantidad de patrones negativos clasificados correctamente como negativos.

FN = Falsos negativos: Muestra la cantidad de patrones positivos clasificados como negativos.

FP = Falsos positivos: Muestra la cantidad de patrones negativos clasificados como positivos.

De la matriz de confusión se obtienen las métricas de rendimiento, como la exactitud (en inglés accuracy), que se calcula como sigue:

$$\boldsymbol{Exactitud} = \frac{\boldsymbol{VP+VN}}{\boldsymbol{VP+FP+VN+FN}} \quad (4)$$

La cual muestra el porcentaje de patrones clasificados correctamente por el clasificador, respecto al total de patrones clasificados.

La sensibilidad (en inglés sensibility) la cual se calcula como sigue:

$$\boldsymbol{Sensibilidad} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FN}} \quad (5)$$

La cual muestra el porcentaje de patrones positivos clasificados correcta- mente como positivos. La especificidad (en inglés especificity), la cual se calcula con la siguiente ecuación:

$$\boldsymbol{Especificidad} = \frac{\boldsymbol{VN}}{\boldsymbol{VN+FP}} \quad (6)$$

Esta métrica muestra el porcentaje de patrones clasificados correctamente como negativos por el clasificador. Para un dataset con IR>1.5, la métrica de exactitud, no se toma como válida, debido a que el algoritmo tiende a favorecer a la clase mayoritaria en la clasificación, por lo que se sugiere utilizar la métrica de exactitud balanceada (en inglés Balanced Accuracy o BA) y se calcula como sigue:

$$\boldsymbol{BA} = \frac{\boldsymbol{Sensibilidad+Especificidad}}{\boldsymbol{2}} \quad (7)$$

Ahora que se conocen las métricas de rendimiento para evaluar a un algoritmo de clasificación, es importante mencionar que ningún clasificador obtendrá en todos los datasets posibles un 100 % de patrones clasificados correctamente, debido a la existencia del teorema de No-Free-Lunch, el cual fue propuesto en un principio para problemas de optimización (Wolpert, 1997), y posteriormente adaptado para el área de reconocimiento de patrones, específicamente para la clasificación, mencionando que el mejor clasificador no existe, ya que no es posible encontrar un clasificador que no cometa errores, y este teorema es válido para todos los algoritmos de clasificación (Adam *et al*., 2019).

3.8 Bancos de datos

En esta sección se muestra una breve descripción de los bancos de datos utilizados para la clasificación de patrones, obtenidos de los repositorios UCI (Frank & Asuncion, 2010) y KEEL (Alcalá-Fdez *et al*., 2011) con algoritmos inteligentes mencionados anteriormente, para bancos de datos médicos que toman como base a (León *et al*., 2020b) y la publicación de aplicación del modelo realzada en (León et al., 2020a) incluyendo un análisis para bancos de datos financieros.

El conjunto de datos **qualitative bankruptcy** perteneciente al repositorio de UCI, muestra 250 patrones descritos por 6 atributos correspondientes a parámetros cualitativos en quiebra distribuidos en dos clases quiebra denotado por B y no-quiebra denotado por NB, obtenidas del análisis de bases de datos financieras utilizadas en enfoques cuantitativos referentes a la estabilidad financiera y la tendencia de las mismas (Kim & Han, n.d.) se usará la nomenclatura **qbank,** para referirse a este banco de datos.

Wisconsin: este banco de datos fue donado al repositorio UCI por el Dr. WIlliam H. Wolberg del Hospital de la Universidad de Wisconsin. El datset contiene información de casos clínicos de pacientes con cáncer de mama que se sometieron a cirugía por dicho padecimiento; consta de 9 atributos, y un total de 683 patrones distribuidos en dos clases: tumores benignos y tumores malignos, con la finalidad de identificar patrones a pacientes sanos y enfermos, se usará la nomenclatura **wsc,** para referirse a este banco de datos.

El banco de datos New Thyroid, donado al repositorio UCI por Stefan Aberhard de la Universidad James Cook, en Australia. Contiene información de pacientes que padecen de problemas relacionados a la glándula tiroides; consta de 5 atributos numéricos y un total de 215 patrones. Las tres clases de la función tiroidea son: normal, hipertiroidismo e hipotiroidismo. Tomando como base el banco de datos New Thyroid, KEEL generó 2 bancos de datos de dos clases:

new-thyroid1: las dos clases son hipertiroidismo y el resto de los 215 patrones, se usará la nomenclatura **nwt1,** para referirse a este banco de datos.

new-thyroid2: las dos clases son hipotiroidismo y el resto de los 215 patrones, se usará la nomenclatura **nwt1,** para referirse a este banco de datos.

Haberman: este banco de datos contiene casos de un estudio realizado 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; consta de 3 atributos numéricos, y un total de 306 patrones distribuidos en dos clases: sobrevivió 5 años o más después de la cirugía (clase 1), o el paciente murió dentro de 5 años posterior a la cirugía (clase 2) se usará la nomenclatura **hbrm,** para referirse a este banco de datos.

Spectfheart: este banco de datos describe el diagnóstico de imágenes de tomografía computarizada SPECT, con el propósito de detectar anomalías cardiacas; consta de 44 atributos numéricos, y un total de 267 patrones distribuidos en dos clases: normal o anormal, el cual tiene como objetivo diagnosticar problemas en superficies de arterias, se usará la nomenclatura **spec,** para referirse a este banco de datos.

South German Credit: es un conjunto de datos del repositorio de UCI que cuenta con un total de 1000 patrones de datos, que describen a patrones crediticios alemanes de Statlog, para identificar el riesgo financiero, este conjunto de datos contiene un total de 20 atributos numéricos distribuidos en 2 clases, cero para malo (riesgo) y 1 para bueno (sin riesgo), se usará la nomenclatura **sgc,** para referirse a este banco de datos.

Se puede observar en la Tabla 2 un resumen de las características de los bancos de datos descritos previamente. En cada caso, se especifica el nombre del banco de datos, el número de atributos de que está formado cada patrón, el número de patrones que contiene el banco de datos y, finalmente, el número de clases en que se distribuyen todos los patrones de ese banco de datos.

Tabla 1 Bancos de datos

Banco	Atributos	Patrones	Clases
Qualitative bankruptcy	6	250	2
South German Credit	20	1000	2
Wisconsin	9	683	2
newt-thyroid1	5	215	2
newt-thyroid2	5	215	2
Haberman	3	306	2
Spectfheart	44	267	2

Fuente: Elaboración propia agosto 2020

3.9 Especificaciones del equipo

Para la implementación del algoritmo LM-τ[9] el cual fue programado en Python, se utilizó una computadora personal marca Dell, modelo E6430, con un procesador Intel Core I5-3210U a 2.5GHz, una memoria RAM de 8 GB, y un sistema operativo Windows 10.

4. Resultados

En esta sección se presenta a manera de ejemplo el cómo se forma la matriz de confusión de la LM-τ[9] dado que este algoritmo no se encuentra en Weka. Se usó el banco de datos de Wisconsin con un conjunto de 137 patrones a ser clasificados, de los cuales 90 son negativos y 47 positivos. Además, se mostrarán los resultados de los experimentos, usando el método de validación 5-fold-cross-validation y **BA** como métrica de desempeño de los 10 clasificadores. Para determinar el desempeño de los 10 clasificadores se tomó como métrica la expresión (7), que es la **Exactitud Balanceada**.

La Tabla 2 muestra la clasificación de los 137 patrones tomados del banco de datos de Wisconsin. Se observa que la LM-τ[9] clasificó 43 patrones de 47 como verdaderos positivos y clasificó 89 verdaderos negativos de 90. Generando una Exactitud del 96% en la clasificación. Además, la Tabla 2 muestra la Sensibilidad, la Especificidad y La Exactitud Balanceada que se calcularon con las expresiones (5), (6) y (7). Pero sólo se toma la **BA** como discriminante de comparación con los otros 9 algoritmos. En Weka también se genera la matriz de confusión del algoritmo que está clasificando y se toma la **BA** obtenida para compararse con los demás algoritmos.

Tabla 2 Matriz de confusión de la LM-τ[9]

	Base de datos Wisconsin	
	Verdaderos Positivos	Verdaderos Negativos
positivos	43	4
negativos	1	89
Exactitud	0,96350365	
Sensibilidad	0,977272727	
Especificidad	0,988888889	
Sensibilidad Balanceada	0,983080808	

Fuente: Elaboración propia agosto 2020

En la Tabla 3 se presentan los resultados obtenidos al aplicar los 10 algoritmos de clasificación descritos en la sección 5, a los 4 bancos de datos médicos descritos. Todos los resultados representan el valor de la **BA** con dos decimales. Se puede observar el desempeño obtenido por los clasificadores, mostrando en negritas a aquellos que obtuvieron el puntaje más alto para **BA**, además, se observa que el clasificador **LM-τ[9]** obtuvo buenos resultados con un 0.99 para **nwt1 y ntw2**. **SMO** obtuvo 0.96 para **wsc, IB1** obtuvo 0.64 para **hbrm** y **NB** obtuvo 0.77 para **spec**. Los algoritmos mencionados anteriormente son los que tuvieron los mejores porcentajes de clasificación en alguna de los 5 bancos de datos médicos. Se observa que a excepción del **SMO** en los bancos de datos **nwt1** y **nwt2** su desempeño difiere significativamente del resto de los demás clasificadores. Los resultados de clasificación hacen ver que, aunque algunos algoritmos obtienen porcentajes más altos que otros los resultados son muy competidos que no hay una ventaja clara de uno sobre otro.

Con respecto a los bancos de datos **hbrm** y **espec** aunque los clasificadores tienen porcentajes de recuperación muy aproximados los resultados son bastante malos y esto queda evidente al comparar los porcentajes de estos dos bancos con los bancos de datos **wsc**, **nwt1** y **nwt2**. Se concluye que los 10 algoritmos son eficientes en tan solo 3 de los 5 bancos de datos.

Tabla 3 Resultados de bancos de datos médicos

Bancos de datos médicos					
Algoritmos	**wsc**	**nwt1**	**nwt2**	**hbrm**	**spec**
NB	0.96	0.98	0.98	0.57	**0.77**
SMO	**0.97**	0.77	0.75	0.50	0.50
LR	0.96	0.96	0.96	0.54	0.62
MLP	0.95	0.95	0.95	0.58	0.66
J48	0.94	0.94	0.94	0.57	0.60
RT	0.92	0.98	0.91	0.58	0.65
RF	0.96	0.95	0.92	0.55	0.60
IB1	0.94	0.97	0.98	**0.64**	0.59
IB3	0.96	0.96	0.92	0.55	0.58
LM-t[9]	0.96	**0.99**	**0.99**	0.52	0.62

Fuente: Elaboración propia agosto 2020

La Tabla 4 muestra los resultados obtenidos al aplicar los 10 algoritmos en los 2 bancos de datos financieros. **IB1** obtiene 1, es decir el 100% de clasificación correcta de patrones, pero este resultado no es real, esto es considerado un error según lo afirma el teorema de No-Free Lunch (Wolpert, 1997) que menciona que no existe un clasificador que no cometa errores, el resultado observado se debe al desbalanceo de clase, lo que implica que el clasificador **IB1** solo es capaz de clasificar a la case mayoritaria del banco de datos. Es interesante ver que no existe un clasificador ideal dado que el balanceo de las clases favorece a un resultado perfecto. Por otro lado, los restantes algoritmos, a excepción del **J48** y **RT**, obtuvieron 0.99. Se puede deducir que, para este banco de datos, prácticamente 9 algoritmos son confiables en la clasificación de patrones financieros porque tiene al menos 0.98 en la clasificación correcta de patrones. Por el lado del banco de datos **sgc** la mejor clasificación la tienen los algoritmos **LM-t[9]** y **RF** con 0.68 de clasificación correcta. Los algoritmos **NB**, **SMO**, **LR**, **J48** obtuvieron en la clasificación correcta de patrones el 0.67. El **IB3** obtuvo 0.64 y el **RT** e **IB1** obtuvieron 0.62. El que tuvo peor desempeño fue el **MPL** con 0. Para este banco de datos se observan diferencias significativas entre algoritmos. Aunque 6 algoritmos obtienen al menos el 0.68 en la clasificación correcta de patrones para el **sgc** los resultados son malos. Al comparar el rendimiento de los 10 algoritmos obtenidos en el **sgc** con **qbank**, se evidencia este hecho.

Tabla 4 Resultados de bancos de datos financieros

Banco de datos financieros		
Algoritmos	**qbank**	**sgc**
NB	0.99	0.67
SMO	0.99	0.67
LR	0.99	0.67
MLP	0.99	0.00
J48	0.98	0.67
RT	0.98	0.62
RF	0.99	**0.68**
IB1	**1**	0.62
IB3	0.99	0.64
LM-t[9]	0.99	**0.68**

Fuente: Elaboración propia agosto 2020

Los 10 algoritmos analizados muestran ser competitivos entre ellos, pero lo que es una realidad es que, en los bancos de datos, **hbrm**, **spec** y **sgc**, los porcentajes de clasificación son muy malos. En diagnósticos médicos y financieros se debe garantizar que éstos se aproximen al 100% por razones obvias. Esta aseveración hace ver que los 10 algoritmos en los 3 bancos de datos mencionados anteriormente no sirven. Para aplicar estos algoritmos en estos bancos de datos es necesario analizar si los bancos de datos están correctamente conformados. Sin embargo, en los 4 bancos de datos restantes los 10 algoritmos resultaron ser clasificadores confiables.

Como dato adicional a los resultados mostrados anteriormente, es importante destacar que en (Velázquez-Rodríguez *et al*., 2020) se muestra una tabla comparativa de rendimiento del reciente algoritmo LM-τ[9] con bancos de datos diversos mostrando resultados donde supera a algunos algoritmos de clasificación que se incluyeron en este trabajo y donde no quedo en el primer lugar en la recuperación queda muy cerca de los algoritmos que lo superaron. Este dato es importante debido a que el presente trabajo confirma que la LM-τ[9] puede ser una opción confiable en la clasificación de patrones médicos y financieros.

5. Agradecimientos

Los autores agradecen a la Universidad Politécnica de Pachuca y al Consejo Nacional de Ciencia y Tecnología (CONACYT) bajo el número de becario 927575 y 928413, al SNI - Sistema Nacional de Investigadores por su apoyo en la realización del presente trabajo de investigación.

6. Conclusiones

El hecho que el presente trabajo se haya centrado en patrones médicos y financieros aporta al trabajo original de (Velázquez-Rodríguez et al., 2020). Si bien es cierto los algoritmos aplicados en los bancos de datos usados generaron resultados muy parecidos a la LM-τ[9] que tienen sus ventajas y desventajas, así como los algoritmos usados aquí las tienen. Los resultados obtenidos hacen que surja la pregunta ¿Qué clasificador usar en patrones médicos o financieros? El presente trabajo le proporciona al lector los elementos necesarios para que él decida qué clasificador es mejor.

En la introducción se plantearon algunas preguntas y estas fueron: si una persona pudiera saber si tiene un padecimiento crónico degenerativo antes siquiera comience a manifestarse ¿Qué decisión tomaría? y si supiera una persona cómo determinar si existe riesgo financiero con su dinero ¿Cómo lo manejaría? La respuesta es que tomaría la mejor decisión para su bienestar basado en un análisis serio y confiable de la información. En este tenor, los datos usados, son datos reales que han sido tomados de los repositorios de KEEL y UCI, por lo que se asegura que los resultados obtenidos en esta investigación son confiables. Se analizaron dos ámbitos, enfermedades y quiebras financieras. Fueron 7 bancos de datos usados: wsc, nwt1, nwt2, hbrm, spec, qbank y sgc que fueron clasificados usando el método del 5-fold-cross-validation con los siguientes algoritmos: NB, SMO, LR, MLP, J48, RT, RF, IB1, IB3 y LM-τ[9]. Los resultados muestran que el lector tendrá elementos necesarios para decidir qué clasificadores son útiles para el desarrollo de herramientas en el área médica y de finanzas para la predicción de eventos sea en salud o en el área de finanzas. Es natural pensar que toda vía hay un tramo que recorrer en análisis de comportamientos en estas áreas, pero estos resultados dan una perspectiva interesante para la elección de un clasificador que coadyuve al desarrollo de herramientas inteligentes en el rubro que se analizaron.

7. Referencias

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.

Adam, S. P., Alexandropoulos, S. A. N., Pardalos, P. M., & Vrahatis, M. N. (2019). No free lunch theorem: a review. In Approximation and Optimization (pp. 57-82). Springer, Cham.

Al Bataineh, M., & Al-qudah, Z. (2017). A novel gene identification algorithm with Bayesian classification. Biomedical Signal Processing and Control, 31, 6-15.

Berral-García, J. L. (2016, July). A quick view on current techniques and machine learning algorithms for big data analytics. In 2016 18th international conference on transparent optical networks (ICTON) (pp. 1-4). IEEE.

Bhargavi, P., & Jyothi, S. (2009). Applying naive bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security, 9(8), 117-122.

Bootkrajang, J., & Kabán, A. (2014). Learning kernel logistic regression in the presence of class label noise. Pattern Recognition, 47(11), 3641-3655.

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.

Chen, H. L., Yang, B., Wang, G., Liu, J., Chen, Y. D., & Liu, D. Y. (2012). A three-stage expert system based on support vector machines for thyroid disease diagnosis. Journal of medical systems, 36(3), 1953-1963.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Dheeba, J., Singh, N. A., & Selvi, S. T. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. Journal of biomedical informatics, 49, 45-52.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by geneexpression monitoring. science, 286(5439), 531-537.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*(5439), 531-537.

González-Patiño, D., Villuendas-Rey, Y., Argüelles-Cruz, A. J., & Karray, F. (2019). A novel bio-inspired method for early diagnosis of breast cancer through mammographic image analysis. Applied Sciences, 9(21), 4492.

Guidi, G., Maffei, N., Vecchi, C., Gottardi, G., Ciarmatori, A., Mistretta, G. M., ... & Costi, T. (2017). Expert system classifier for adaptive radiation therapy in prostate cancer. Australasian physical & engineering sciences in medicine, 40(2), 337-348.

Haro-García, A., Cerruela-García, G., & García-Pedrajas, N. (2019). Instance selection based on boosting for instance-based learners. Pattern Recognition, 96, 106959.

Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. International Journal of Innovative Science, Engineering & Technology, 2(2), 438-446.

Kim, M. J., & Han, I. (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. Expert Systems with Applications, 25(4), 637-646.

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial intelligence in medicine, 23(2), 149-169.

León, P. R., Salgado Ramírez, J. C., & Luis Velázquez Rodríguez, J. (2020). Application of the Lernmatrix tau[9] to the classification of patterns in medical datasets. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*, 8488–8497.

Lindberg, A. (2020). Developing theory through integrating human and machine pattern recognition. Journal of the Association for Information Systems, 21(1), 7.

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.

Mungle, T., Tewary, S., Arun, I., Basak, B., Agarwal, S., Ahmed, R., ... & Chakraborty, C. (2017). Automated characterization and counting of Ki-67 protein for breast cancer prognosis: A quantitative immunohistochemistry approach. Computer Methods and Programs in Biomedicine, 139, 149-161.

Padmavathy, T. V., Vimalkumar, M. N., & Bhargava, D. S. (2019). Adaptive clustering based breast cancer detection with ANFIS classifier using mammographic images. Cluster Computing, 22(6), 13975-13984.

Patel, H., & Thakur, G. S. (2017). Improved fuzzy-optimally weighted nearest neighbor strategy to classify imbalanced data. Int J Intell Eng Syst, 10, 156-162.

Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing, 11(2), 2906-2915.

Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of medical systems, 41(4), 55.

Quinlan, J. R. (1990). Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20(2), 339-346.

Raymer, M. L., Doom, T. E., Kuhn, L. A., & Punch, W. F. (2003). Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 33(5), 802-813.

Reyes León, P., Salgado Ramírez, J. C., & Velázquez Rodríguez, J. L. (2020). Pre-Diagnosis of Chronic Diseases by the Application of Computational Intelligence Models. Computación y Sistemas, 24(3).

Schein, A. I., & Ungar, L. (2004). A-optimality for active learning of logistic regression classifiers.

Soni, R., Kumar, B., & Chand, S. (2019). Optimal feature and classifier selection for text region classification in natural scene images using Weka tool. Multimedia Tools and Applications, 78(22), 31757-31791.

Steinbuch, K. (1961). Die lernmatrix. Kybernetik, vol. 1, no 1, p. 36-45. Springer.

Uriarte-Arcia, A. V., López-Yáñez, I., & Yáñez-Márquez, C. (2014). One-hot vector hybrid associative classifier for medical data classification. PloS one, 9(4), e95715.

Velázquez-Rodríguez, J. L., Villuendas-Rey, Y., Camacho-Nieto, O., & Yáñez-Márquez, C. (2020). A Novel and Simple Mathematical Transform Improves the Perfomance of Lernmatrix in Pattern Classification. Mathematics, 8(5), 732.

Villuendas-Rey, Y., Alanis-Tamez, M. D., Rey-Benguría, C., Yáñez-Márquez, C., & Nieto, O. C. (2018). Medical Diagnosis of Chronic Diseases Based on a Novel Computational Intelligence Algorithm. *J. UCS*, *24*(6), 775-796.

Vyas, R., Bapat, S., Goel, P., Karthikeyan, M., Tambe, S. S., & Kulkarni, B. D. (2016). Application of Genetic Programming (GP) formalism for building disease predictive models from protein-protein interactions (PPI) data. IEEE/ACM transactions on computational biology and bioinformatics, 15(1), 27-37.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization IEEE Transactions on Evolutionary Computation. E997.

Zhang, L., & Suganthan, P. N. (2014). Random forests with ensemble of feature spaces. Pattern Recognition, 47(10), 3429-3437.