Facultad de Ciencias Experimentales
Grado en Biología

**UNIVERSIDAD DE JAÉN**
*Facultad de Ciencias Experimentales*

Trabajo Fin de Grado

# Estudio y análisis del funcionamiento de técnicas de minería de datos en conjuntos de datos relacionados con la Biología

**Alumna: Fabiola Mesa Pérez**

**Julio, 2019**

**UNIVERSIDAD DE JAÉN**
*Facultad de Ciencias Experimentales*

Trabajo Fin de Grado

# Estudio y análisis del funcionamiento de técnicas de minería de datos en conjuntos de datos relacionados con la Biología

**Alumna: Fabiola Mesa Pérez**

**Julio, 2019**

**ÍNDICE**

**RESUMEN**

La minería de datos es un proceso computacional que permite la extracción de información a partir de un conjunto de datos (datasets) con el fin de transformarlo posteriormente para su uso. En el presente trabajo se analizan conjuntos de datos relacionados con la Biología y, para ello, se ha utilizado una serie de algoritmos que ayudan a determinar el funcionamiento de los datos.

De esta forma, se han analizado alrededor de 40 datasets y se ve cómo influyen el número de clases y el número de atributos sobre los algoritmos utilizados y cuáles de ellos han sido los más adecuados en el análisis de los datos. Se ha realizado un estudio experimental muy completo usando test estadísticos con los datasets. Se han aplicado estos algoritmos a un problema del campo de la Biología y de la Biomedicina, como por ejemplo el problema del cáncer, lo que llevará a determinar si el paciente necesita una exploración de forma invasiva o no, según las mediciones de los distintos parámetros influyentes.

**Palabras clave:** algoritmos de minería de datos, C4.5, problema del cáncer de mama, Chi-RW, FARCHD, FURIA, GAssist, métodos basados en reglas, PART, problemas de clasificación, RIPPER, test estadísticos no paramétricos, Wisconsin.

**ABSTRACT**

Data mining is a computational process that allows information to be extraction from a dataset in order to transform it later for use. In the present work, datasets related to Biology are analyzed and, for this purpose, a series of algorithms have been used to help determine the functioning of the data.

In this way, around 40 datasets have been analyzed and it is seen how the number of classes and the number of attributes influence the algorithms used and which of them have been the most appropriate in the analysis of the data. A very complete experimental study has been carried out using statistical tests with the datasets. These algorithms have been applied to a problem in the field of Biology and Biomedicine, such as cancer problem, which will lead to determining whether the patient needs an invasive exploration or not, depending on the measurements of the different influential parameters.

**Key words:** Breast Cancer Problem, C4.5, Chi-RW, classification problems, data mining algorithms, FARCHD, FURIA, GAssist, non-parametric statistical tests, PART, RIPPER, rule-based methods, Wisconsin.

## 1. INTRODUCCIÓN

En los últimos años, ha habido un aumento muy significativo de la cantidad de datos disponibles en el área de la Biología. Por ejemplo, existen alrededor de 130.000 estructuras en PDB, 58.000.000 de estructuras químicas en ChemSpider y 200.877.884 secuencias en GenBank, lo que ha supuesto un aumento por un factor de 3x10^6 desde sus inicios en 1982. El valor de esta gran cantidad de datos radica en la posibilidad de extraer información útil con la que poder modelar nuevas teorías, crear nuevas metodologías o realizar avances en diversas áreas para la Biología. El avance de la tecnología en los últimos años ha supuesto un gran avance en la recopilación de datos a procesar en aplicaciones de diversos campos.

Debido al gran volumen de datos, este análisis no puede ser manual. Por ello, surge la necesidad del proceso del descubrimiento de conocimiento en bases de datos (KDD, Knowledge Discovery in Databases), que incluye las etapas de extracción de datos, preprocesamiento, etc. La minería de datos (MD) (Han, 2011) es una de sus etapas y consiste en identificar patrones que sean válidos, novedosos, útiles y comprensibles en los datos. La minería de datos es un proceso esencial en el que se aplican métodos inteligentes para extraer patrones de datos. Las fuentes de datos pueden incluir bases de datos, almacenes de datos, la Web, otros depósitos de información o datos que se transmiten de forma dinámica al sistema (Han, 2011). Se trata de un área multidisciplinaria.

Relacionado con MD, el término aprendizaje automático, se refiere al diseño de nuevos algoritmos, lo que supone una parte de la Biología Computacional, además de su estudio y aplicación. Estos algoritmos se utilizan para abordar el problema de la clasificación (García, 2017), donde una de las opciones más prometedoras es el uso de los sistemas basados en reglas (Geva *et al.*, 1937), dado que representan el conocimiento en forma de regla, de tal manera que son comprensibles.

En este trabajo, primero se realiza una recopilación de conjuntos de datos de distintas bases de datos. Se editan hasta ponerlos en el formato específico para poder ser utilizados adecuadamente por la herramienta software Keel. Uno de los problemas es que cada conjunto está en un formato distinto y hay que conocerlos bien para

prepararlos, según las particularidades de cada uno, con el objetivo de que el proceso de minería de datos se pueda realizar de manera adecuada. El análisis se realiza con los algoritmos Chi-RW-C (Chi, 1996), C4.5 (Ross, 1993), PART (Holden & Freitas, 2008), RIPPER (Schultz, 2011), GAssist (Bacardit, 2004), FURIA (Cohen, 1995) y FARCHD (Alcalá-Fdez *et al.*, 2011). Además, se estudia la bondad de la aplicación de los diferentes algoritmos en cada caso, utilizado test estadísticos no paramétricos. Se valora, de esta manera, el comportamiento general de conjuntos de datos relacionados con Biología y se presenta un caso de estudio.

La memoria está organizada de la siguiente manera. En la primera sección, se presenta una breve revisión de las nociones generales para entender el uso de los algoritmos de clasificación. Se describen, posteriormente, los algoritmos objeto de estudio en este trabajo. En la segunda sección, se presentan los objetivos del trabajo. En la tercera sección, se incluyen una serie de experimentos y, en la cuarta sección, se realiza un análisis comparativo de los resultados obtenidos. Finalmente, se analizarán en profundidad un conjunto de datos y se presentará un caso de estudio. El trabajo terminará con unas conclusiones y perspectivas.

## 1.1. Aprendizaje automático y clasificación

En primer lugar, en esta sección se van a incluir definiciones de los conceptos que se van a manejar en este trabajo. Se comienza por la definición de aprendizaje automático, problemas de clasificación y, por último, el tipo de método que se utiliza, como los métodos basados en reglas, junto con los algoritmos usados.

### *1.1.1. Aprendizaje automático*

En el campo de la minería se pueden diferenciar dos tipos de aprendizajes automáticos: el aprendizaje supervisado y el aprendizaje no supervisado.

El método de aprendizaje supervisado es un método predictivo que se emplea para conseguir el descubrimiento de las relaciones entre las variables de entrada y la variable de salida (objetivo), cuyas relaciones buscadas son representadas en una estructura llamada modelo (García, 2017). Los principales problemas que pueden resolverse con el aprendizaje supervisado son la clasificación y la regresión. En la clasificación, hay un número finito de clases o categorías para predecir una muestra

y estos son conocidos por el algoritmo de aprendizaje (García, 2017). En cambio, la regresión es más compleja y requiere una necesidad computacional mayor. En ella, se ajusta un modelo para aprender el atributo objetivo de salida como función de los atributos de entrada. Así, el aprendizaje supervisado intenta elaborar la mejor metodología en la relación entre entradas y salidas de un objetivo.

El método de aprendizaje no supervisado presenta distintas categorías, siendo las más comunes el agrupamiento y la asociación. El agrupamiento o clustering es el proceso de particionar un conjunto de datos en un conjunto de subclases significativas denominadas grupos o clusters (Cáceres, 2007). En este método se desconocen las etiquetas de clases. Las reglas de asociación permiten descubrir hechos o relaciones que ocurren en común dentro de un determinado conjunto de datos.

### *1.1.2. Métodos de clasificación*

Este trabajo se centra en resolver problemas de clasificación, cuyo objetivo principal de la clasificación es obtener una predicción fiable, de forma que el modelo se ajuste a un conjunto de datos ya entrenados. Un ejemplo de problema de clasificación es el diagnóstico de la diabetes tipo 2. Este problema ha sido muy estudiado por investigadores, lo que ha motivado que exista una gran variedad de algoritmos para resolverlos (Almadni, 2011). En ese estudio, se evalúa el rendimiento del sistema difuso desarrollado para diagnosticar la diabetes tipo 2 comparando el sistema con dos modelos de clasificación, esto es, la regresión logística y técnicas de aprendizaje automático con el fin de predecir los resultados.

### *1.1.3. Métodos basados en reglas*

Los métodos basados en reglas son útiles y muy conocidos en el ámbito del aprendizaje automático debido a que son capaces de crear modelos interpretables (Geva *et al.*, 1997; Kotsiantis, 2007).

Hoy en día, con el auge de Internet en las cosas o IoT (Evans, 2011) y la explicabilidad de los resultados obtenidos (por ejemplo, coches automáticos), este tipo de técnicas son de gran utilidad, para, entre otras cosas, resolver el problema de la caja negra (Castelvecchi, 2016; Rudin, 2019), gracias a que obtienen modelos interpretables.

Estos métodos suelen emplearse para la filtración y reordenación de los resultados del modelo. Los métodos basados en reglas se incluyen en el aprendizaje automático de forma que se combina con un algoritmo de clasificación, siendo su unión en cascada. Tal y como se explica en el estudio de Villena *et al.* (2011), los métodos basados en reglas construyen modelos de precisión y cobertura sin mucho esfuerzo. La principal característica del método es utilizar reglas basadas en lenguaje natural, teniendo en cuenta el grado de complejidad. Las reglas pueden o no asociarse a cada categoría, de manera que validen, invaliden o incluyan la categoría en los resultados, siempre que cumplan con los requisitos de las reglas. También sirven para el reordenamiento de los resultados expresado en un lenguaje básico computacional. El reordenamiento viene dado por el algoritmo de aprendizaje. Además, en caso de no haber definido una regla para cada categoría, se utilizan dos reglas: accept (validar) y reject (invalidar). Se concluye, por tanto, que los métodos basados en reglas filtran los falsos positivos y resuelven los falsos negativos, tal y como se demuestra en el estudio mencionado con anterioridad.

### 1.2. Métodos de clasificación basados en reglas

En este trabajo, se van a utilizar algunos de los algoritmos más representativos disponibles en Keel (Knowledge Extraction based on Evolutionary Learning) software (Alcalá-Fdez *et al.*, 2009). Se usarán los métodos de clasificación basados en reglas siguientes: C4.5, PART, RIPPER, GAssist, Chi-RW, FURIA y FARCHD.

#### *1.2.1. C4.5*

C4.5 es un método de clasificación creado por Ross Quinlan en 1993 a modo de mejora de ID3 que desarrolló en 1986. El algoritmo C4.5 produce árboles de decisión pequeños y precisos, dando como resultado unos clasificadores rápidos y fiables. Estos clasificadores son los denominados árboles de decisión y son utilizados en numerosos dominios del mundo real, como, por ejemplo, para el diagnóstico de hipotiroidismo, diagnóstico de enfermedad de soja y aprobación de crédito (Ross, 1993).

C4.5 es, por tanto, un algoritmo muy popular y significativo en la minería de datos, ya que las reglas generadas son muy correctas, lo que a su vez da lugar a muchos descubrimientos en la investigación científica (Ngoc, 2017).

El algoritmo tiene en cuenta la división del conjunto de datos en una serie de pruebas, determinando la prueba más correcta para establecer una mayor información. Así, para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. En cambio, para cada atributo continuo, se realiza una prueba binaria sobre cada valor que toma el atributo en los datos. (López, 2005). De esta forma, el sistema decide la prueba más efectiva para dividir el conjunto de datos en cada nodo.

Las características que presenta C4.5 son las siguientes: recursividad, esto es, aplicándose de forma recursiva; utilización del criterio de proporción de ganancia (gain ratio), definido como I (Xi, C) /H (Xi), con esto se persigue evitar que las variables con mayor número de posibles valores salgan beneficiadas al ser seleccionadas; seleccionar un conjunto de datos de entrenamiento con el fin de generar el árbol de decisión inicial; menor frondosidad de los árboles de decisión y permite trabajar con valores continuos para los atributos, separando los resultados posibles en dos ramas (López, 2005).

### *1.2.2. PART*

PART es un algoritmo que extrae reglas de decisión combinando los descriptores PNF, son los denominados factor positivo - negativo (utilizado para la clasificación de fallas monofásicas y bifásicas que cumple con el punto de corte especificado de forma teórica) y CF, también llamado Coeficiente de Forma (posee un intervalo de corte capaz de identificar eventos por energización de transformadores), donde la calidad de la regla es alta y confiable (Velandia, 2010). Además, PART es el único algoritmo que presenta al descriptor *Rex*, una resistencia estimada a la parte imaginaria.

Este algoritmo, además, se considera un estándar en la industria como algoritmo de clasificación. Se trata de una implementación mejorada de reglas del algoritmo C4.5. Se considera un algoritmo muy mejorado en cuanto a la precisión en materia de predicción (Holden & Freitas, 2008).

### *1.2.3. RIPPER*

RIPPER es un algoritmo de aprendizaje que se basa en distintas reglas que utiliza para crear un conjunto de reglas que se encarga de identificar las clases posibles, mientras minimiza la cantidad de errores. El error se define por el número de ejemplos de formación mal clasificados por las reglas.

El algoritmo de aprendizaje inductivo (RIPPER) asume que los datos con los cuáles se ha entrenado previamente, son similares de alguna manera a los datos no vistos sobre los que realizara los cálculos para obtener las distintas reglas.

El interés en el uso del algoritmo RIPPER es que éste considera ambos ejemplos, tanto los positivos como los negativos, para poder generar un conjunto de teorías o hipótesis que se acercan más al concepto del objetivo que se plantea a este algoritmo previamente (Schultz, 2011). Se esperaría que RIPPER mostrara una mejora en la tasa de error en un conjunto de datos ruidoso (Cohen, 1996), lo que permite una mejor precisión a la hora de la realización del modelo construido por los datos usados, ya que, a mayor ruido, se obtendría una menor precisión. Sin embargo, la utilización de RIPPER nos otorga una mejora en la tasa de error, mejorando así la precisión.

### *1.2.4. GAssist*

GAssist es un sistema inspirado inicialmente en GABIL (De Jong *et al.*, 1993) y pertenece al modelo de Pittsburgh. Este algoritmo genético da lugar a un conjunto de reglas cuya ordenación se dispone en longitudes variables (Bacardit, 2004). Este sistema evoluciona a unos individuos representativos de una solución completa de problemas a los que se les aplica un algoritmo genético casi estandarizado, siendo un individuo los conjuntos de reglas ordenadas y de una longitud variada. El objetivo que se pretende es combinar la aptitud de una función basada en el principio de la longitud mínima de descripción (MDL) y en un operador que permita eliminar las reglas.

El conocimiento usado para los atributos reales se representa mediante una regla de intervalos de discretización adaptativa (ADI), la cual utiliza una semántica de reglas de GABIL aplicados a intervalos no estáticos, formados por el vínculo cercano de las discretizaciones. Éstos son los denominados intervalos de tiempo, los cuáles pueden

tomar dos caminos diferentes: 1) fusión entre ellos de forma potencial usando discretizadores en sincronía o 2) continuar su paso mediante la división del proceso de aprendizaje.

Asimismo, el sistema utiliza también un esquema de ventanas denominado aprendizaje incremental (ILAS) que se complementa con estratos alternados. El aprendizaje incremental estratifica la formación del conjunto en subconjuntos de tamaño uniforme y de distribución de clases similar. Cada repetición dada en algoritmo genético y usa cada estrato con el fin de calcular el estado físico. El método mencionado evidencia la introducción de una generalización implícita adicional a GAssist (Bacardit, 2004).

### *1.2.5. Chi-RW*

El algoritmo Chi-RW (Chi, 1996) y se caracteriza por la ponderación de las reglas. Esto es, determinar la relación existente entre las variables problema y establecer una asociación entre el espacio de rasgos y de espacios de las clases mediante una serie de pasos. Por ejemplo, se establece particiones lingüísticas y se procede al cálculo de partición difusa tras determinar el dominio de variación de una característica. A la partición difusa se le agrega cada ejemplo, el cual lleva asignada una regla difusa, con alto grado de relación y etiqueta de clase. Finalmente, se calcula el peso de la regla (Trawinski, 2010).

El algoritmo Chi-RW es un algoritmo de clasificación difusa que sirve para evaluar entornos difusos. Según Bardossy *et al.* (1995), este algoritmo se puede utilizar para el problema de modelación meteorológica.

### *1.2.6. FURIA*

El algoritmo FURIA (Cohen, 1995) es una amplitud de RIPPER (Schultz, 2011) y que se caracteriza por la utilización de reglas difusas. El algoritmo de aprendizaje de FURIA, mediante la aplicación de RIPPER, genera un conjunto de reglas crisp para después realizar el proceso de fuzzificación sin la utilización de t-normas (Elkano, M., 2015).

FURIA se caracteriza por ser un algoritmo de inducción de reglas desordenadas difusas, de ahí su abreviatura FURIA, siendo una ampliación y modificación de reglas de última generación RIPPER (Cohen, 1995). Particularmente, lo que hace es reconocer reglas difusas y desordenadas. En cambio, no permite aprender reglas convencionales ni listas de reglas. El tratamiento de ejemplos lo realiza por un método de estiramiento de reglas, el cual es lo suficiente eficiente.

La principal diferencia entre una regla difusa y una regla convencional es que la regla difusa suele abarcar más, por lo que juega con ventaja respecto a la regla convencional. Por ejemplo, en las reglas convencionales se producen una serie de modelos con agudos, lo que atrae a transiciones abruptas dadas entre distintas clases, siendo una peculiaridad no intuitiva y, además, puede cuestionarse. Se esperaría que una clase dada por una regla se disminuyera desde el núcleo a la frontera (de lleno a cero) de manera gradual.

### *1.2.7. FARCHD*

El algoritmo FARCHD es un método de clasificación basado en reglas de asociación difusa para los problemas de alta dimensión (Alcalá-Fdez *et al.,* 2011). Concretamente, el algoritmo obtiene distintos modelos que presentan una reducción en su número de reglas, siendo el promedio de reglas 39,2 y con pocos atributos en el antecedente.

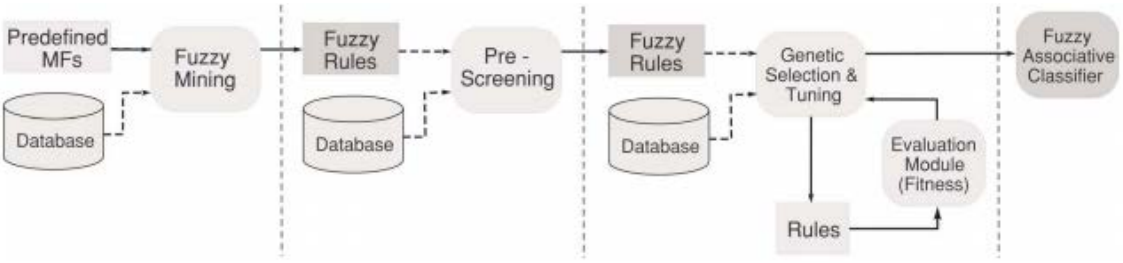


*Figura 1. Esquema del método FARCHD obtenido del estudio realizado por J. Alcalá-Fdez, R. Alcala y F. Herrera (2011).*

En la *Figura 1*., se muestra cómo funciona el algoritmo FARCHD sobre datasets de alta dimensión. En el estudio realizado por Alcalá-Fdez *et al.* (2011), el objetivo era crear un método de clasificación asociativos difusos que fuesen a la vez precisos y

compactos y que no supongan un alto coste computacional. Así, se intenta obtener reglas con un número menor de atributos en el antecedente.

## 2. OBJETIVOS

### 2.1. Objetivos generales

El principal objetivo es profundizar en el campo de la biología computacional, mediante el estudio de técnicas de clasificación basadas en reglas para clasificación de datos relacionados con la Biología con el fin de determinar cuál presenta un mejor funcionamiento. De esta forma, se pretende probar el funcionamiento de una serie de algoritmos, como C4.5 (Ross, 1993), PART (Holden & Freitas, 2008), RIPPER (Schultz, 2011), GAssist (Bacardit, 2004), Chi-RW (Chi, 1996), FURIA (Cohen, 1995) y FARCHD (Alcalá-Fdez *et al*., 2011), para el análisis de conjuntos de datos biológicos (datasets), entre los que destacan wdbc y wisconsin.

### 2.2. Objetivos específicos

En base al objetivo principal, se pretende alcanzar los siguientes subobjetivos:

- Adquirir conocimientos sobre distintas técnicas de clasificación basados en reglas, para su uso con datos relacionados con la Biología y Biomedicina.
- Aprender el manejo de herramientas software, en las que se incluyen el uso de test estadísticos, para aplicas dichas técnicas y contrastar su funcionamiento, comprendiendo sus ventajas y limitaciones.
- Asimilar los conceptos básicos para poder utilizar las técnicas aprendidas en futuras investigaciones.

## 3. MATERIAL Y MÉTODOS

### 3.1. Conjunto de datos

Los conjuntos de datos seleccionados provienen de “Irvine Machine Learning Repository” (UCI) (Dua & Graff, 2019) de “Knowledge Extraction based on Evolutionary Learning” (Keel) (Alcalá-Fdez *et al.*, 2011), Weka y Kaggle. Son repositorios de gran calidad, certificados y avalados por muchos otros trabajos de investigación. La variedad de los datasets con un número de variables que de 3 a 10000 y con un número de clases de 2 a 28, va a permitir realizar un estudio experimental muy completo.

En base a los objetivos descritos con anterioridad, se plantea el siguiente plan de trabajo. En primer lugar, se realiza una recopilación de conjuntos de datos en bases de datos, como Keel, Uci, etc. Estos datos necesitan una preparación, por lo que se editan hasta tener un formato específico para su utilización adecuada por la herramienta software Keel. Uno de los problemas que se presenta es que cada conjunto de datos se encuentra en un formato distinto a los demás y es necesario un buen conocimiento para su preparación, según las particularidades de cada uno con el objetivo de que el proceso de minería de datos se pueda llevar a cabo de la manera más adecuada. Así, se obtienen conjuntos de entrenamiento y test, para que los resultados no se vean influenciados por los valores y se pueda obtener conclusiones fiables.

Los datasets seleccionados están relacionados con la Biología, perteneciendo la mayoría de ellos al campo biomédico, pero también a otras disciplinas como la Zoología, la Botánica, la Evaluación de Ecosistemas y la Microbiología. Para llevar a cabo un buen proceso de experimentación es necesario utilizar varias pruebas, destacando test estadísticos no paramétricos. Los datasets que se incluyen en el presente trabajo deben incluir una serie de categorías que posibiliten su representación:

- Número de atributos
- Número de ejemplos o instancias
- Número de clases

Cada categoría se recoge en la *Tabla1*., en la que se incluye una breve descripción de cada dataset.

| **Datasets** | **Atributos** | **Ejemplos** | **Clases** | **Descripción** |
|---|---|---|---|---|
| Abalone | 8 | 4174 | 28 | Determinar la edad del molusco mediante mediciones físicas |
| Acute Inflammations | 6 | 120 | 2 | Diagnóstico de inflamación aguda urinaria |

| | | | | |
|---|---|---|---|---|
| Appendicitis | 7 | 106 | 2 | Inflamación del apéndice |
| Arcene | 10000 | 900 | | Distinción de patrones cancerígenos |
| Arrhythmia | 279 | 452 | 13 | Presencia y ausencia de arritmias cardíacas |
| Audiology | 69 | 226 | 24 | Audiología |
| Breast | 9 | 277 (286) | 2 | Cáncer de mama |
| Bupa | 6 | 345 | 2 | Trastornos del hígado |
| Cervical Cancer | 36 | 858 | 2 | Predicción de cáncer de cuello uterino |
| Cleveland | 13 | 297 (303) | 5 | Enfermedad del corazón |
| Contraceptive | 9 | 1473 | 3 | Método anticonceptivo |
| Demospongiae | 3 | 76 | 3 | Clasificación de esponjas marinas |
| Dermatology | 34 | 358 (3666) | 6 | Diagnóstico de dermatosis eritematoescamosas |
| Diabetic Retinopathy Debrecen | 20 | 1151 | 2 | Retinopatía diabética Debrecen |
| Echocardiogram | 12 | 132 | | Supervivencia de pacientes tras un ataque cardíaco |
| E. Coli | 7 | 336 | 8 | Localización de proteínas |
| Fertility Diagnosis | 10 | 100 | 2 | Diagnóstico de la fertilidad |
| Forest Types Mapping | 27 | 523 | 4 | Mapeo de tipos de bosques |
| Glass | 9 | | 7 | Cristal |
| Haberman | 3 | 306 | 2 | Supervivencia de pacientes con cáncer de mama mediante cirugía |
| HCC Survival | 49 | 165 | | Datos del carcinoma hepatocelular (HCC) |
| Heart | 13 | 270 | 2 | Enfermedad del corazón |

| | | | | |
|---|---|---|---|---|
| Hepatitis | 19 | 80 (155) | 2 | Diagnóstico de hepatitis |
| Horsecolic | 27 | 368 | 2 | Lesión con cirugía o sin cirugía |
| Ionosphere | 33 | 351 | 2 | Estado de propagación de las ondas |
| Iris | 4 | 150 | 3 | Tipo de planta de Iris |
| Lymphography | 18 | 148 | 4 | Radiografía de vasos y ganglios linfáticos |
| Mammographic | 5 | 830 (961) | 2 | Lesión masiva mamográfica |
| Micromass | 1300 | 931 | | Espectrometría de masas para bacterias gram+ y gram- |
| Monk-2 | 6 | 432 | 2 | Problemas de clasificación binaria artificial |
| Mushroom | 22 | 8124 | 2 | Identificación de especies de hongos según su comestibilidad |
| Newthyroid | 5 | 215 | 3 | Enfermedad de tiroides |
| Parkinsons | 23 | 195 | 2 | Detección de enfermedad de Parkinson |
| Pima | 8 | 768 | 2 | Relacionado con diabetes |
| Primary Tumor | 17 | 339 | 21 | Diagnóstico del tumor primario |
| Seeds | 7 | 210 | 3 | Propiedades geométricas de distintas variedades de trigo |
| Soy Bean Large | 35 | 683 | 19 | Enfermedad de la soja |
| Spectfheart | 44 | 267 | 2 | Diagnóstico de enfermedades cardíacas |
| Statlog (heart) | 13 | 270 | 2 | Registro estadístico del corazón |
| Wdbc | 30 | 569 | 2 | Diagnóstico de Wisconsin |
| Wine | 13 | 178 | 3 | Análisis químico del vino (acidez, etc) |
| Wisconsin | 9 | 699 | 2 | Cáncer de mama Wisconsin |

| | | | | |
|---|---|---|---|---|
| Yeast | 8 | 1484 | 10 | Predicción de localización celular de las proteínas |
| Zoo | 16 | 101 | 7 | Clasificación de animales |

*Tabla 1. Exposición detallada sobre los datasets seleccionados para la realización del análisis experimental. La tabla recoge los atributos, los ejemplos y las clases, así como una breve descripción de cada dataset.*

La investigación para la aplicación de problemas de clasificación a problemas del mundo real supone un gran avance, ya que se pueden desarrollar diferentes métodos de aplicación, así como un aumento en el conocimiento computacional.

### 3.2. Diseño de experimentos con Keel

En esta sección, se presenta el trabajo experimental realizado para analizar los algoritmos considerados. Este conjunto de experimentos se ha completado gracias a la herramienta de software KEEL (Knowledge Extraction based Evolutionary Learning) (Dua & Graff, 2019) que está programada en lenguaje Java en código abierto. KEEL puede ser utilizado para una gran cantidad de tareas del KDD, puesto que provee una interfaz fácil de utilizar basada en flujos de trabajo para diseñar experimentos con diferentes conjuntos de datos y algoritmos de inteligencia computacional. En la *Figura 2.,* aparece el flujo de trabajo utilizado para los experimentos en la interfaz de KEEL.

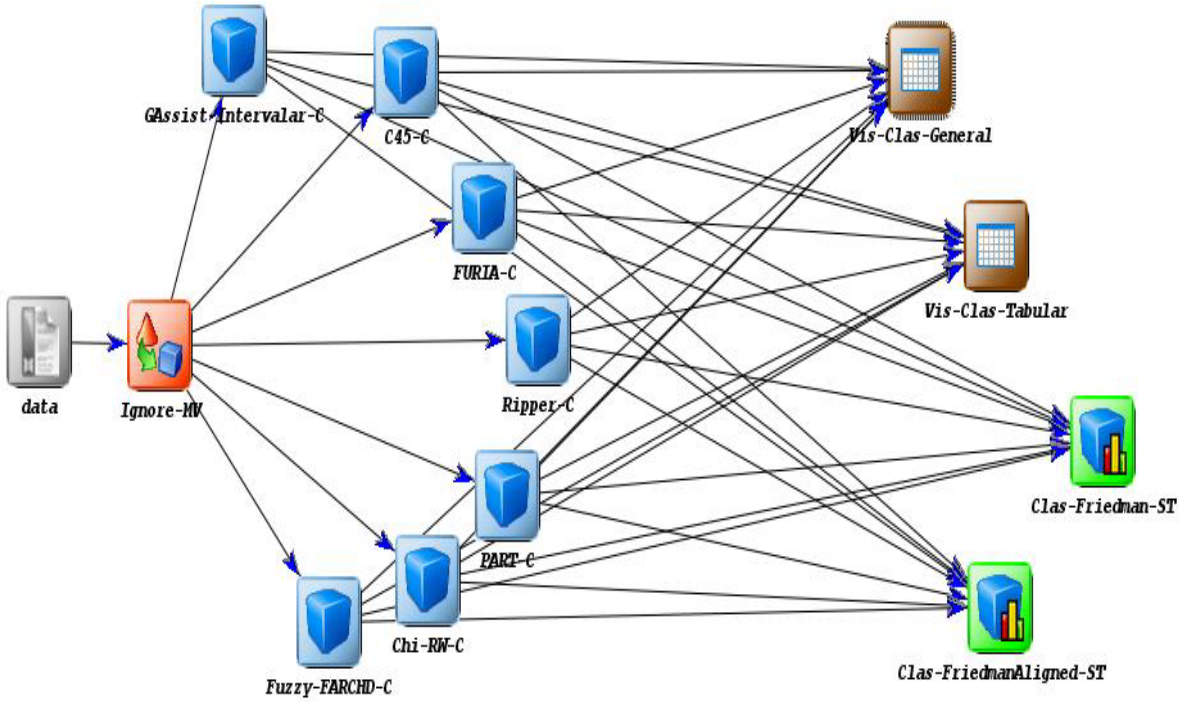


*Figura 2. Visualización en la interfaz de Keel del flujo de trabajo seguido para la realización de los experimentos.*

### 3.3. Parámetros de los algoritmos

Para todos los métodos que se usarán en la experimentación, se han considerado los parámetros por defecto que indicaban los autores de los métodos en sus artículos.

Además de lo explicado en el anterior apartado, se estudia bondad de la aplicación de los diferentes algoritmos en cada caso, utilizando test estadísticos no paramétricos. De esta forma, se valora el comportamiento general de conjuntos de datos relacionados con Biología y se presenta un caso de estudio.

Los algoritmos utilizados son FURIA, C4.5, Chi-RW, PART, RIPPER, GAssist y FARCHD, los cuales se explican en el apartado 1.2. Cada uno de estos algoritmos se analizará en conjunto con los datasets y test no paramétricos, de forma que se va a generar una serie de resultados de éstos a modo de porcentaje. Cuanto mayor sea el porcentaje generado, mayor fiabilidad de los resultados. Dependiendo de los resultados obtenidos, se podrá hacer un caso de estudio, el cual puede extrapolarse a la realidad.

### 3.4. Metodología de análisis

En la metodología de análisis se hace uso del método *k-fold*, siendo una manera de validar el modelo. El método *k-fold* presenta aumenta su rendimiento conforme el tamaño de conjuntos de datos disminuye (Yang & Huang, 2014). Los datos totales se dividen en *k* subconjuntos, se procede a utilizar diferentes subconjuntos para que el modelo entrenado se pueda validar con los otros subconjuntos *k-1* (Jung & Hu, 2015). La validez de este método se evalúa a través del error medio, obteniéndose así resultados que son representativos (Pérez-Planells *et al.*, 2015). Se evita perder muchos ejemplos para entrenar y poder así obtener un buen estimador del error. Lo habitual es usar *k=10* (usado en el estudio experimental) y repetir el proceso para obtener un estimador mejor. Por tanto, divide los *n* ejemplos en *k* conjuntos disjuntos. Para *i=1* hasta *k:* se entrena usando cada fold menos i y se emplea para estimar el error. Devuelve la media de los errores de los *k-fold.*

## 4. RESULTADOS

Aunque se han analizado un total de 40 datasets, finalmente después, junto al tutor, se han descartado algunos datasets. Esto se debe a que los algoritmos han sufrido algunos problemas, dando lugar a problemas de ejecución de ciertos datasets como mushroom, siendo el experimento final de 30 datasets. Al ejecutarse los algoritmos, éstos no son capaces de procesar. Motivos por los que han fallado: 1) problemas de demasiado número de atributos, como horsecolic; 2) problemas de demasiado número de ejemplos; 3) problemas de demasiadas clases; 4) falta de descripción de los datasets, etc.

| **Datasets admitidos** | **Datasets descartados** |
|---|---|
| Abalone | Acute Inflammations |
| Appendicitis | Arcene |
| Breast | Arrhythmia |
| Bupa | Audiology |
| Cleveland | Cervical Cancer |
| Contraceptive | Demospongiae |

| | |
|---|---|
| Dermatology | Echocardiogram |
| Diabetic Retinopathy Debrecen | Forest Type Mapping |
| E. Coli | Glass |
| Fertility Diagnosis | HCC Survival |
| Haberman | Horsecolic |
| Heart | Micromass |
| Hepatitis | Mushroom |
| Ionosphere | Newthyroid |
| Iris | |
| Lymphography | |
| Mammographic | |
| Monk-2 | |
| Parkinsons | |
| Pima | |
| Primary Tumor | |
| Seeds | |
| Soybean | |
| Spectfheart | |
| Statlog (heart) | |
| Wdbc | |
| Wine | |
| Wisconsin | |
| Yeast | |
| Zoo | |

*Tabla 2. Listado de datasets admitidos y descartados en el análisis experimental.*

Los resultados finales se recogen en la siguiente tabla:

*Tabla 3. Resultados obtenidos del análisis experimental realizado para determinar el funcionamiento de los distintos algoritmos en base a los datasets seleccionados.*

| Dataset | GAssist-Intervalar-C | | C45-C | | FURIA-C | | Ripper-C | | PART-C | | Chi-RW-C | | Fuzzy-FARCHD-C | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TRA | TST | TRA | TST | TRA | TST | TRA | TST | TRA | TST | TRA | TST | TRA | TST |
| abalone | 0,250 | **0,236** | 0,751 | 0,199 | 0,224 | 0,204 | 0,469 | 0,233 | 0,166 | 0,166 | 0,002 | 0,001 | 0,174 | 0,173 |
| appendicitis | 0,930 | 0,850 | 0,909 | 0,841 | 0,918 | 0,848 | 0,965 | 0,823 | 0,891 | 0,833 | 0,891 | **0,858** | 0,937 | 0,848 |
| breast | 0,864 | 0,718 | 0,771 | **0,765** | 0,775 | 0,732 | 0,882 | 0,675 | 0,711 | 0,711 | 0,857 | 0,678 | 0,917 | 0,718 |
| bupa | 0,799 | 0,637 | 0,859 | **0,670** | 0,806 | 0,692 | 0,873 | 0,615 | 0,616 | 0,590 | 0,599 | 0,579 | 0,785 | 0,669 |
| cleveland | 0,702 | 0,559 | 0,831 | 0,525 | 0,604 | 0,549 | 0,823 | 0,435 | 0,539 | 0,539 | 0,914 | 0,380 | 0,882 | 0,552 |
| contraceptive | 0,577 | 0,544 | 0,732 | 0,529 | 0,561 | 0,547 | 0,637 | 0,521 | 0,433 | 0,429 | 0,520 | 0,393 | 0,626 | 0,534 |
| dermatology | 0,983 | 0,955 | 0,982 | 0,955 | 0,988 | 0,958 | 0,994 | 0,922 | 0,718 | 0,704 | 1,000 | 0,215 | 0,999 | 0,924 |
| diabetesretinopat | 0,695 | 0,647 | 0,836 | 0,647 | 0,695 | 0,640 | 0,823 | 0,636 | 0,613 | 0,586 | 0,604 | 0,588 | 0,759 | 0,722 |
| ecoli | 0,826 | 0,748 | 0,917 | 0,795 | 0,904 | 0,822 | 0,925 | 0,730 | 0,443 | 0,447 | 0,758 | 0,720 | 0,925 | 0,798 |
| fertility-diagnosis | 0,944 | 0,780 | 0,893 | 0,870 | 0,899 | 0,820 | 0,952 | 0,720 | 0,884 | 0,890 | 0,963 | 0,350 | 0,969 | 0,800 |
| haberman | 0,815 | 0,722 | 0,759 | 0,732 | 0,771 | 0,738 | 0,586 | 0,510 | 0,735 | 0,735 | 0,743 | 0,732 | 0,809 | 0,722 |
| heart | 0,922 | 0,800 | 0,920 | 0,767 | 0,887 | 0,826 | 0,903 | 0,741 | 0,596 | 0,596 | 0,972 | 0,519 | 0,939 | 0,807 |
| heart-statlog | 0,928 | 0,796 | 0,930 | 0,770 | 0,882 | 0,781 | 0,919 | 0,748 | 0,596 | 0,596 | 0,969 | 0,522 | 0,944 | 0,826 |
| hepatitis | 0,997 | 0,800 | 0,947 | 0,825 | 0,971 | 0,850 | 0,965 | 0,863 | 0,900 | 0,825 | 0,989 | 0,238 | 1,000 | 0,875 |
| ionosphere | 0,984 | 0,900 | 0,986 | 0,889 | 0,979 | 0,900 | 0,977 | 0,877 | 0,783 | 0,752 | 0,976 | 0,653 | 0,984 | 0,923 |
| iris | 0,979 | 0,960 | 0,980 | 0,960 | 0,981 | 0,940 | 0,987 | 0,947 | 0,333 | 0,333 | 0,938 | 0,927 | 0,986 | 0,953 |
| lymphography | 0,936 | 0,745 | 0,919 | 0,763 | 0,952 | 0,866 | 0,955 | 0,770 | 0,655 | 0,655 | 1,000 | 0,127 | 1,000 | 0,817 |
| mammographic | 0,856 | 0,840 | 0,851 | 0,841 | 0,852 | 0,835 | 0,770 | 0,729 | 0,783 | 0,777 | 0,806 | 0,796 | 0,838 | 0,812 |
| monks | 0,990 | 0,982 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 0,528 | 0,527 | 0,972 | 0,429 | 0,999 | 0,995 |
| parkinsons | 0,933 | 0,835 | 0,974 | 0,846 | 0,992 | 0,896 | 0,957 | 0,846 | 0,831 | 0,779 | 0,884 | 0,846 | 0,972 | 0,897 |
| pima | 0,817 | 0,749 | 0,838 | 0,742 | 0,789 | 0,746 | 0,861 | 0,701 | 0,651 | 0,651 | 0,752 | 0,731 | 0,830 | 0,764 |
| primary-tumor | 0,492 | 0,377 | 0,648 | 0,427 | 0,526 | 0,394 | 0,757 | 0,287 | 0,351 | 0,331 | 0,891 | 0,108 | 0,839 | 0,379 |
| seeds | 0,968 | 0,905 | 0,977 | 0,924 | 0,986 | 0,933 | 0,978 | 0,914 | 0,832 | 0,810 | 0,925 | 0,919 | 0,980 | 0,924 |
| soybean | 0,733 | 0,694 | 0,965 | 0,913 | 0,953 | 0,897 | 0,981 | 0,888 | 0,253 | 0,251 | 0,992 | 0,262 | 0,978 | 0,906 |
| spectfheart | 0,824 | 0,802 | 0,981 | 0,768 | 0,945 | 0,780 | 0,918 | 0,700 | 0,794 | 0,794 | 0,984 | 0,663 | 0,921 | 0,787 |
| wine | 0,973 | 0,926 | 0,984 | 0,940 | 0,995 | 0,945 | 0,996 | 0,931 | 0,861 | 0,810 | 0,859 | 0,842 | 0,990 | 0,930 |
| wdbc | 0,977 | 0,937 | 0,991 | 0,928 | 0,992 | 0,953 | 0,986 | 0,930 | 0,897 | 0,882 | 0,961 | 0,930 | 0,988 | 0,963 |
| wisconsin | 0,989 | 0,955 | 0,981 | 0,956 | 0,989 | 0,965 | 0,983 | 0,961 | 0,703 | 0,698 | 0,980 | 0,913 | 0,987 | 0,969 |
| yeast | 0,535 | 0,525 | 0,811 | 0,581 | 0,639 | 0,593 | 0,722 | 0,497 | 0,346 | 0,334 | 0,296 | 0,291 | 0,641 | 0,591 |
| zoo | 0,972 | 0,931 | 0,986 | 0,940 | 0,970 | 0,940 | 0,999 | 0,910 | 0,844 | 0,820 | 1,000 | 0,653 | 1,000 | 0,930 |
| **PROMEDIO** | 0,840 | 0,762 | 0,897 | 0,777 | 0,848 | **0,786** | 0,885 | 0,735 | 0,643 | 0,628 | 0,833 | 0,562 | 0,887 | 0,784 |

Al observar la tabla, vemos los resultados mostrados como Porcentaje de Clasificación en Entrenamiento o Training (TRA) y Porcentaje de Clasificación en Prueba o Test (TST). El primero se indica para saber si con los datos reales son correctos o no, y, con el segundo, para testear, por ejemplo, para estimar si el paciente nuevo puede tener cáncer o no. Además, cuanto más se acerque a 1 el resultado, más fiable será. Por otro lado, los valores perdidos (missing values), en caso de que los datasets los tuvieran, no se han tenido en cuenta, por lo que no se han analizado en este experimento. Se han eliminado con el algoritmo Ignore-MV de eliminación en los datos de valores perdidos. La posibilidad de utilizar datos clasificados en problemas implica una simple correlación y así evitar suposiciones de normalidad (Friedman, 1937; Hotelling & Pabst, 1936).

Al realizar el Test de Friedman se obtiene un ranking de los algoritmos, mostrado en la tabla a continuación:

| **Método** | **Ranking Friedman** |
|---|---|
| FURIA-C | 2,2 |
| Fuzzy-FARCHD-C | 3 |
| C45-C | 3 |
| GAssist-Intervalar-C | 3,5 |
| Ripper-C | 5 |
| PART-C | 6 |
| Chi-RW-C | 6 |

*Tabla 4. Resultados del Test de Friedman. Se observan valores asociados a cada algoritmo relacionado con su funcionamiento.*

Los mejores algoritmos han sido FURIA, FUZZY y C4.5. Los diferentes métodos y el uso de la prueba de Friedman (Friedman, 1937) muestran un ranking, en el que las diferencias significativas existen entre los resultados observados en los conjuntos de datos. En la *Tabla 4.*, se observa que el mejor del ranking se obtiene por el método

FURIA. La posibilidad de utilizar datos de clasificación en problemas implica una correlación simple y evita suposiciones de normalidad (Friedman, 1937; Hotelling & Pabst, 1936). La prueba de Friedman consiste en sumar las veces que el mejor porcentaje se ha obtenido por dataset.

Analizando los resultados, vemos que el número de clases influye de forma negativa en el funcionamiento de los algoritmos, ya que todos ellos son genéricos y no están específicamente diseñados para resolver cada problema de clasificación. Por ejemplo, abalone y primary tumor, presentan un número de clases muy elevado, lo que ocasiona unos porcentajes de aciertos muy malos, por debajo del 50% en TST. En cambio, el número de atributos no influye tanto. El dataset soybean, en el que el número de clases también es elevado, no obtiene buenos resultados, pero solo en algunos de los algoritmos como PART-C y Chi-RW-C, y sí obtiene buenos resultados en RIPPER-C y FURIA-C.

Por lo general, cuando el número de clases es inferior a 20, se obtienen buenos resultados y cuando el número de clases es demasiado bajo, por ejemplo 2, los resultados de casi todos los algoritmos son bastante fiables por la no dificultad de clasificar las dos clases en comparación a cuando hay que clasificar un número superior. Por otra parte, se han obtenido una media de 15 reglas en los algoritmos de forma generalizada, debido a que la mayoría están pensados para obtener resultados interpretables.

### 4.1. Caso de estudio: cáncer

En cuanto a los datasets que han obtenido un alto porcentaje de clasificación de training y test son wisconsin y wdbc. Ambos son datasets destinados al diagnóstico de cáncer de mama.

Para wdbc (Wisconsin Diagnostic Breast Cancer), se va a estudiar los resultados obtenidos en una partición en la que el número de reglas es de 9: 4 malignos (M) y 5 benignos (B). Los parámetros de valor real para cada núcleo celular son: radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal.

| Atributo | Dominio |
|---|---|
| Área1 | [143.5, 2501.0] |
| Área2 | [6.802, 542.2] |
| Área3 | [185.2, 4254.0] |
| Compacidad1 | [0.019, 0.345] |
| Compacidad2 | [0.0020, 0.031] |
| Compacidad3 | [0.027, 1.058] |
| Concavidad1 | [0.0, 0.427] |
| Concavidad2 | [0.0, 0.396] |
| Concavidad3 | [0.0, 1.252] |
| Dimensión_fractal1 | [0.05, 0.097] |
| Dimensión_fractal2 | [0.0010, 0.03] |
| Dimensión_fractal3 | [0.055, 0.208] |
| Perímetro1 | [43.79, 188.5] |
| Perímetro2 | [0.757, 21.98] |
| Perímetro3 | [50.41, 251.2] |
| Puntos_cóncavos1 | [0.0, 0.201] |
| Puntos_cóncavos2 | [0.0, 0.053] |
| Puntos_cóncavos3 | [0.0, 0.291] |
| Radio1 | [6.981, 28.11] |
| Radio2 | [0.112, 2.873] |
| Radio3 | [7.93, 36.04] |
| Simetría1 | [0.106, 0.304] |
| Simetría2 | [0.0080, 0.079] |
| Simetría3 | [0.156, 0.664] |
| Suavidad1 | [0.053, 0.163] |
| Suavidad2 | [0.0020, 0.031] |

| | |
|---|---|
| Suavidad3 | [0.071, 0.223] |
| Textura1 | [9.71, 39.28] |
| Textura2 | [0.36, 4.885] |
| Textura3 | [12.02, 49.54] |
| **Clase** | {M, B} |

*Tabla 5. Datos numéricos de las mediciones de masa mamaria: dominio del atributo y tipos de clases.*

Estos parámetros fueron calculados a partir de una imagen procedente de un aspirado con aguja fina (Fine Needle Aspirate, FNA) de una mama mamaria por el Dr. William H. Wolberg, W. Nick Street y Olvi L. Mangasarian de la Universidad de Wisconsin en 1995. De esta forma, se puede dar un diagnóstico acerca del estado del tumor, esto es, saber si es benigno (B) o maligno (M).

Para ello, con el método FURIA, se han obtenido 9 reglas que indicarán si el tejido será Maligno (M) o Benigno (B), el cual presenta un nivel de confianza de la regla (CF) para saber su fiabilidad:

- R1: Si (Radio3 >= 16.82 (-> 16.35)) y (Perímetro3 >= 120.4 (-> 120.3)) entonces Clase = Maligno (CF = 0.99).
- R2: Si (Perímetro3 >= 102.5 (-> 100.9)) y (Textura3 >= 26.93 (-> 25.47)) y (Suavidad2 >= 0.006 (-> 0.005)) entonces Clase = Maligno (CF = 0.96).
- R3: Si (Perímetro3 >= 106.4 (-> 106)) y (Textura3 >= 20.24 (-> 19.58)) y (Puntos_cóncavos2 <= 0.009 (-> 0.01)) entonces Clase = Maligno (CF = 0.94).
- R4: Si (Puntos_cóncavos3 >= 0.146 (-> 0.143)) y (Suavidad3 >= 0.155 (->0.154)) entonces Clase = Maligno (CF = 0.98).
- R5: Si (Perímetro3 <= 106 (-> 127.3)) y (Puntos_cóncavos3 <= 0.11 (-> 0.111)) y Puntos_cóncavos1 <= 0.026 (-> 0.027)) entonces Clase = Benigno (CF = 1.0).
- R6: Si Perímetro3 <= 114.3 (-> 119.4)) y (Textura3 <= 25.94 (-> 26.38)) y (Simetría2 >= 0.017 (-> 0.016)) entonces Clase = Benigno (CF = 1.0).

- R7: Si (Puntos_cóncavos1 <= 0.051 (-> 0.054)) y (Perímetro3 <= 107.4 (-> 108.4)) y (Compacidad1 >= 0.065 (-> 0.06)) y (Suavidad3 <= 0.177 (-> 0.178)) entonces Clase = Benigno (CF = 1.0).
- R8: Si (Concavidad1 <= 0.086 (-> 0.099)) y (Textura1 <= 18.89 (-> 19.38)) y (Radio2 <= 0.257 (-> 0.271)) entonces Clase = Benigno (CF = 0.99).
- R9: Si (Puntos_cóncavos3 <= 0.146 (-> 0.156)) y (Área3 <= 867.1 (-> 876.5)) y (Perímetro1 >= 94.29 (-> 94.25)) entonces Clase = Benigno (CF = 0.97).

Observando las reglas, se ve que, si el radio de la masa del tumor aumenta y/o el perímetro de esta es grande, son algunos de los motivos por los cuáles se considere que el cáncer es de tipo Maligno (M). Por lo que, las reglas son bastante interpretables y muy útiles para realizar un diagnóstico.

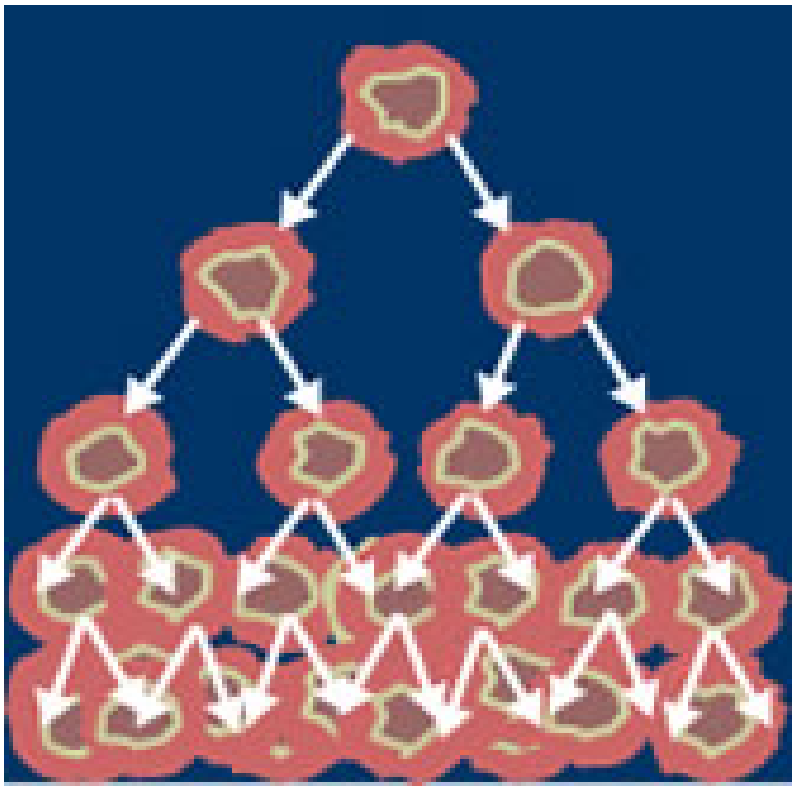

*Figura 3. Imagen digitalizada de un aspirado con aguja fina (FNA) de una mama mamaria. En la imagen se pueden ver las características que presentan los núcleos celulares (Universidad de Wisconsin, 1992).*

Con todo lo anterior se puede estimar las mediciones de los distintos parámetros registrados de forma científica y poderse evaluar, posteriormente, en pacientes. Por lo que, con datos reales de pacientes se puede realizar un diagnóstico aproximado, es por ello por lo que se necesitan diferentes mediciones, como las mostradas en la tabla 5. De esta forma, se hace una estimación de M o B sin tener que hacer una prueba de biopsia. Por tanto, la importancia del funcionamiento de los algoritmos radica en obtener unos resultados más fiables que permitan establecer una valoración real y verídica en el diagnóstico de la enfermedad, como puede ser, en este caso, el algoritmo FURIA.

## 5. DISCUSIÓN

Para determinar cuál o cuáles son los algoritmos que presentan un mejor funcionamiento de forma individualizada, se han utilizado un conjunto de datasets. Con el fin de validar si este conjunto de datos puede servir en el análisis (previo cambio de formato para su preparación), se ha probado, en primer lugar, si daba algún tipo de error, comprobando así que no todos han servido de utilidad para el experimento. Para solucionar este problema, se han descartado 14 datasets, quedando al final 30 datasets para el análisis. Posteriormente, los datasets se han testeado con los algoritmos disponibles en el programa Keel para evaluar el comportamiento para problemas de minería de datos, como la clasificación. Los conjuntos de datos se han analizado junto a una serie de algoritmos, como C4.5, Chi-RW, FARCHD, FURIA, GAssist y RIPPER y se han utilizado test estadísticos no paramétricos.

Teniendo en cuenta el principal objetivo, que es el análisis y estudio de técnicas de clasificación para un conjunto de datos relacionados con Biología, y, además de los resultados obtenidos tras analizar el funcionamiento de los distintos algoritmos, se puede concluir que el mejor algoritmo es FURIA, ya que, atendiendo a los altos porcentajes alcanzados para cada dataset, es el que ha mostrado unos mejores resultados para cada uno de ellos. En cuanto a los conjuntos de datos que mejor se han expresado, han sido wdbc y wisconsin, pudiendo dar lugar al estudio de un caso fiable por su viabilidad.

Por último, hay que destacar que ambos datasets (wdbc y wisconsin) están destinados al diagnóstico de cáncer de mama, de forma que, con el algoritmo FURIA y, en su defecto, C4.5 (segundo algoritmo que mejor funcionamiento ha mostrado en el análisis), tiene dos clases para el diagnóstico en pacientes: benigno y maligno. El uso de algoritmos fiables puede dar lugar a su implementación en el campo biomédico, implicando que el paciente con posibilidad patológica no tenga por qué someterse a diversas pruebas médicas que supongan algún tipo de invasión en su organismo. Para ello, ambos datasets presentan diferentes mediciones, como el área, la compacidad, la concavidad, la dimensión fractal, el perímetro y los puntos cóncavos, así como también el radio, la simetría, la suavidad y la textura, que llevarán a la determinación del estado de la masa mamaria, esto es, saber si la proliferación celular es de tipo benigno o maligno.

## 6. CONCLUSIÓN

En este trabajo se han estudiado diferentes sistemas basados en reglas, que son herramientas muy importantes en el campo de la minería de datos, ya que extraen conocimiento de conjuntos de datos. Estas herramientas están suponiendo una verdadera solución para el gran volumen de datos manejados en biología que parece que no va a hacer más que aumentar en los próximos años, debido al avance de nuevas estrategias y tecnologías para la obtención de nueva información. Según los resultados obtenidos, se ha podido observar que la información extraída puede ser válida, útil y comprensible. Los sistemas basados en reglas permiten extraer información de una manera comprensible.

Tras analizar los algoritmos de regla que hay y observar el que presenta un mejor comportamiento, se concluye que no todos los algoritmos sirven para todos los problemas, sino que, para algunos problemas, sería necesario diseñar un algoritmo específico para su resolución. Es el caso del dataset abalone, el cual era muy grande y ningún algoritmo funcionó.

En general, el algoritmo que mejor ha funcionado ha sido FURIA, ya que ha obtenido mejores resultados, a pesar de que, teóricamente, uno de los mejores algoritmos en minería de datos de clasificación es C4.5. En cuanto a las clases y atributos, decir que

las clases no se pueden eliminar porque son necesarias a la hora de cuantificar el problema, ya que un número elevado de clases dificulta la obtención de resultados, pero los atributos sí se pueden desechar aplicando preprocesamiento.

Con los problemas de clasificación se puede saber, en materia biomédica, si el paciente tiene una enfermedad o no sin necesidad de invadir, como sería mediante una biopsia en el caso del cáncer de mama. Esto es de vital importancia a la hora de diagnosticar una enfermedad, ya que hay un porcentaje de fallos que puede dar lugar a que el paciente asuma riesgos cuando no hay necesidad de invadir, pudiéndose realizar otro tipo de pruebas médicas.

La realización de este trabajo ha supuesto un primer paso para entender estas técnicas de minería de datos y así poder aplicar dichas técnicas en mi futuro profesional.

En resumen, viendo cómo aumenta la cantidad de datos en el área de la Biología y lo útiles que son estas técnicas, cabe esperar que, en un futuro próximo, la minería de datos y, concretamente, los sistemas basados en reglas sean herramientas suplementarias habituales de cualquier investigador en este campo.

## 7. BIBLIOGRAFÍA

Alcalá-Fdez, J., Alcalá, R. and Herrera, F. (2011). A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems with Genetic Rule Selection and Lateral Tuning. IEEE Transactions on Fuzzy Systems, 19: 857-872.

Alcalá-Fdez, J. *et al.* (2011). KEEL data-mining software tool: Dataset repository, integration of algorithms and experimental analysis framework. J. Mult. Log. Soft. Comput., 17: 255-287.

Almadni, D. (2011). Comparative analysis of classification models for diagnosis type 2 diabetes. King Abdulaziz University, Saudi Arabia.

Bacardit, J. (2004). Pittsburgh genetic-based machine learning in the data mining era: representations and generalization and run-time [Informe]. Department of Computer Science, University Ramon Llull, Barcelona.

Bardossy, A., Duckstein, L. and Bogardi, I. (1995). Fuzzy rule-based classification of atmospheric circulation patterns. International Journal of Climatology, 15: 1087-1097.

Cáceres, J. (2007). Reconocimiento de patrones y el aprendizaje no supervisado. Universidad de Alcalá, Madrid.

Castelvecchi, D. (2016). Can we open the black box of AI? Nature, 538: 20-23.

Cohen, W. W. (1996). Learning Trees and Rules with Set-valued Features. Proceedings of the Thirteenth National Conference on Artificial Intelligence, 709-716, Portland, OR.

De Jong, K. A., Spears, W. M. and Gordon, D. F. (1993). Using genetic algorithms for concept learning. Machine Learning, 13: 161-188.

Dua, D. y Graff, C. (2019). Repositorio de aprendizaje automático de la UCI. Irvine, CA: Universidad de California, Escuela de Información y Ciencias de la Computación.

Elkano, M., Galar, M., Sanz, J. and Bustince, H. (2015). Estudio de funciones de overlap n-dimensionales en Sistemas de Clasificación Basados en Reglas Difusas. Dpto. de Automática y Computación, Universidad Pública de Navarra, Pamplona.

Evans, E. (2011). Internet de las cosas – Cómo la próxima evolución de Internet lo cambia todo. Cisco Internet Business Solutions Group (IBSG), 2.

Fernández-Delgado, M., Cernadas, E. and Barro, S. (2014). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15: 3133-3181.

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: 675-701.

García, J. (2017). Modelos híbridos de aprendizaje basados en instancias y reglas para Clasificación Monotónica (Tesis Doctoral). Universidad de Jaén, Jaén.

Han, J. and Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kauffmann Publishers Inc.

Holden, N. and Freitas, A. A. (2008). A Hybrid PSO/ACO Algorithm for Discovering Classification Rules in Data Mining. Journal of Artificial Evolution and Applications, 8: 1-11.

Hotelling, H. and Pabst, M. R. (1936). Rank Correlation and Tests of Significance Involving No Assumption of Normality. The Annals of Mathematical Statistics, 7: 29-43.

Jung, Y. and Hu, J. (2015). A k-fold averaging cross validation procedure. Journal of Nonparametric Statistics, 27: 167-179.

López, B. (2005). Inteligencia artificial: algoritmo C4.5. Instituto Tecnológico de Nuevo Laredo, Tamaulipas.

Ngoc, P.V., Ngoc, C.V.T, Ngoc, T.V.T *et al.* (2017). A C4.5 algorithm for English emotional classification. Evolving Systems.

Pérez-Planells, Ll., Delegido, J., Rivera-Caicedo, J.P. and Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. Asociación Española de Teledetección.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 5: 206-215.

Salzberg, S.L. Mach. Learn. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kauffmann Publishers Inc.

Schultz, M. G., Eskin, E., Zadok, F. and Stolfo, S. J. (2001). Data mining methods for detection of new malicious executables. IEEE Symposium on Security and Privacy.

Trawinski, K. (2010). A fuzzy classification system for prediction of the results of the basketball games. International Conference on Fuzzy Systems.

Velandia, R. A. and Hernández, F. L. (2010). Evaluación de algoritmos de extracción de reglas de decisión para el diagnóstico de huecos de tensión (Trabajo de Grado). Universidad Industrial de Santander, Bucaramanga.

Villena-Román, J., Collada-Pérez, S., Lana-Serrano, S. and González-Cristóbal, J. C. (2011). Método híbrido para categorización de texto basado en aprendizaje y reglas. Procesamiento del Lenguaje Natural, 46: 35-42.

Yang, Y. and Huang, S. (2014). Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models- a case study. Forestry, 87: 654-662.