
Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación
Laboratorio de Inteligencia Artificial
Opción profesional: Inteligencia Artificial

# UNA APLICACIÓN BASADA EN TÉCNICAS DE MINERÍA DE DATOS PARA LA EVALUACIÓN DE LA MORFOLOGÍA DE CABEZA DE ESPERMATOZOIDE HUMANO

Trabajo Especial de Grado presentado ante la Ilustre Universidad Central de Venezuela para optar al título de Licenciada en Computación, por la Bachiller:

Grindeliz A. Altuve F.

Tutora: Profa. Haydemar Núñez

Caracas, Abril de 2015.

## RESUMEN

El Análisis de Líquido Seminal (ALS), también llamado espermograma, es un examen diagnóstico que permite evaluar la fertilidad masculina. El ALS se basa en un conjunto de pruebas (macroscópicas y microscópicas) que proporcionan información sobre la calidad de una muestra de semen humana. Entre las evaluaciones microscópicas resalta la morfología espermática, la cual permite determinar el estado general de los espermatozoides.

La realización de esta prueba es compleja y vulnerable a la subjetividad del experto. Con el fin de apoyar este proceso, se desarrolló una aplicación que realiza la evaluación morfológica de cabeza de espermatozoide humano. Para esto se definió un modelo de clasificación basado en minería de datos, específicamente utilizando la técnica de Combinación de Clasificadores (*Ensemble Classifiers)*. Esta aplicación toma como entrada imágenes digitales de una muestra de semen, segmenta y extrae las características en cada uno de los espermatozoides presentes, para luego clasificar la muestra de acuerdo al protocolo que sigue el experto al realizar la evaluación.

Para la verificación y validación de la aplicación se clasificó un grupo de doce muestras de semen, para luego comparar los resultados obtenidos con la clasificación realizada por un experto al mismo grupo de muestras. Se pudo constatar que la cantidad de muestras clasificadas morfológicamente de forma correcta por la aplicación fue de un 75%.



**Palabras Claves:** Análisis del Líquido Seminal, Morfología espermática, Minería de datos, Combinación de Clasificadores.

# ÍNDICE

**Página**

# INTRODUCCIÓN

En los últimos años se ha producido un crecimiento exponencial de datos debido a la disponibilidad de almacenamiento y el gran nivel de procesamiento en las máquinas actuales. Datos como: textos, imágenes, videos, y cualquier otro tipo utilizado en diferentes aspectos de la vida diaria, pueden ser perfectamente digitalizados y almacenados en bases de datos.

Dichos datos poseen una gran cantidad de conocimiento útil, desconocido y no procesado. Es en este punto, donde la minería de datos tiene un papel protagónico al proporcionar distintas técnicas que pueden ser utilizadas para encontrar patrones, modelos y relaciones entre los datos, que ayuden en la toma de decisiones.

Existen múltiples áreas donde la minería de datos puede ser aplicada, pero es en el campo de la medicina donde este trabajo enfocará su atención. En muchas clínicas y hospitales, como resultado de atender a una gran cantidad de pacientes con distintas patologías, se genera una gran cantidad de información de gran valor para las investigaciones médicas y científicas. Entre estos estudios se encuentra el Análisis del Líquido Seminal (ALS) o también llamado espermograma, el cual se realiza en el área de la bioquímica. Este análisis provee información valiosa en procedimientos de reproducción asistida.

El espermograma se basa en un conjunto de pruebas macroscópicas y microscópicas que proporcionan información sobre la calidad de una muestra de semen humana, clasificándolo según criterios preestablecidos por la Organización Mundial de la Salud, OMS (2001). Las pruebas macroscópicas analizan características físico-químicas del semen, tales como aspecto, licuefacción, volumen, color, viscosidad y pH. Mientras que las pruebas microscópicas toman en cuenta aspectos como movilidad, vitalidad, morfología de los espermatozoides en la muestra, entre otros.

En el caso de la evaluación de la morfología, el experto debe contar manualmente 200 espermatozoides y clasificarlos de acuerdo a sus características morfológicas; de igual forma, debe poder diferenciarlos de otras estructuras presentes en la muestra, siendo vulnerable a la subjetividad del experto. Es importante destacar que los resultados obtenidos deben ser confiables y reproducibles para poder ser interpretados de la misma forma por distintos expertos en el área. No obstante, la realización de este estudio resulta compleja, debido al esfuerzo y tiempo que debe ser invertido.

En la Escuela de Computación de la Universidad Central de Venezuela se han realizado Trabajos de investigaciones cuyos objetivos se centran en la automatización de la evaluación morfológica de espermatozoides humanos. Uno de ellos es el presentado por Ferreira (2008), el cual consistió en la construcción de una herramienta de adquisición de conocimiento, la cual identificara y extrajera de forma automática características de células espermáticas humanas, y que funcionara a partir de la definición de un modelo basado en técnicas de procesamiento digital de imágenes. Sin embargo la clasificación

morfológica seguía dependiendo del conocimiento y subjetividad del experto. Es por esto que surge la investigación de Casañas, Ramos, y Núñez (2009), ésta se centró en la construcción de un modelo de clasificación morfológico de espermatozoides humanos, haciendo uso de distintos algoritmos de minería de datos y tomando como conjunto de datos los vectores de características extraídos de la investigación de Ferreira (2008).

Sin embargo, estas aproximaciones no se encuentran integradas en una única aplicación que sirva de apoyo al experto. Además, presentan ciertas características desfavorables que necesitan ser mejoradas, tal es el caso de los algoritmos de segmentación diseñados por Ferreira (2008), los cuales permiten identificar los espermatozoides presentes en una muestra, pero no se encuentran adaptados para tratar con imágenes de mayor resolución a la establecida en su aplicación (1600x1200 píxeles).

Con el fin de subsanar estos problemas, en este Trabajo Especial de Grado, se describe el desarrollo de una aplicación, basada en minería de datos, para apoyar al experto en la evaluación morfológica de cabeza de espermatozoides humanos, a partir de imágenes digitalizadas de las muestras de semen. La aplicación permitirá la carga de imágenes (de mayor resolución a 1600x1200 píxeles) correspondientes a láminas de muestras de semen, y estará en capacidad de realizar la evaluación de la morfología espermática de las muestras suministradas (de forma automática).

Este documento está estructurado de la siguiente manera. En el Capítulo 1, se presentan de manera general los conceptos básicos del Análisis de Líquido Seminal, así como los relacionados con la minería de datos, el proceso de descubrimiento de conocimiento a partir de datos y la combinación de modelos (*Ensemble Classifiers*). También se describen algunas aplicaciones de la minería de datos en el campo de la medicina.

En el Capítulo 2 se presenta la propuesta de Trabajo Especial de Grado, y se incluye un análisis detallado de los distintos módulos que conforman la aplicación desarrollada (Segmentación, Extracción de características y Clasificación). Para finalizar, se presenta el análisis de los resultados obtenidos al realizar la evaluación de la aplicación, así como las conclusiones y recomendaciones para trabajos futuros.

# CAPÍTULO 1. MARCO TEÓRICO

## 1.1. Análisis de líquido seminal

El Análisis del Líquido Seminal (ALS), también llamado espermograma, es un examen diagnóstico que permite evaluar la fertilidad masculina, así como también estimar el completo funcionamiento del aparato reproductor masculino (Ramos, Pereira, Núñez & others, 2007). También, el espermograma puede ser utilizado para uso posterior, luego de una intervención médico-quirúrgica como una vasectomía; de igual forma, es de gran utilidad en otros chequeos pre y post-operatorios.

El ALS está compuesto de un conjunto de pruebas macroscópicas y microscópicas (Mortimer, 1994). Las pruebas macroscópicas analizan características físico-químicas del semen, tales como aspecto, licuefacción, volumen, color, viscosidad y pH, mientras que las microscópicas evalúan concentración, aglutinación, motilidad, vialidad, morfología espermática, entre otras. (Padrón, Fernández & Gallardo, 1988; Cerezo, 2006). En algunos casos, es necesario aplicar un conjunto de evaluaciones adicionales llamadas complementarias con el fin de obtener más información acerca de la capacidad fertilizadora del semen (Ramos, 2009). A continuación, se resume el conjunto de pruebas que componen el ALS.

**- Evaluaciones Macroscópicas**

- **Aspecto:** Se analiza la homogeneidad o heterogeneidad de la muestra, color, presencia de cuerpos mucosos y fibrinas, opacidad o transparencia.

- **Licuefacción:** Se analiza la consistencia, puede ser líquida o con coágulos.

- **Volumen:** Debe ser medido utilizando una pipeta volumétrica, y el valor normal se encuentra entre los 2 y 6ml.

- **Viscosidad:** Es la propiedad de resistencia de un líquido al movimiento uniforme de su masa. Esta se reporta normal en caso de observar un filamento menor a 2cm y aumentado en caso que el filamento sea mayor a 2cm.

- **pH:** Este se mide utilizando un papel especial de rango de 2 a 9, al cabo de 30 segundos se debe observar el color de la zona impregnada y se compara con la cartilla de calibración. El valor de referencia considerado normal es mayor de 7.2.

**- Evaluaciones Microscópicas**

- **Aglutinación:** Es la unión entre espermatozoides móviles por cualquiera de sus partes.

- **Concentración:** Es la cantidad de espermatozoides presentes en la muestra por unidad de volumen.

- **Motilidad:** Permite contar los espermatozoides móviles e inmóviles en una muestra. Los espermatozoides móviles pueden clasificarse según su progresividad:

    - Móviles progresivos rápidos
    - Móviles progresivos lentos
    - Móviles no progresivos

- **Viabilidad o vitalidad espermática:** Es la proporción entre los espermatozoides inmóviles vivos y los inmóviles muertos.

- **Elementos celulares diferentes:** Se busca la presencia de células epiteliales e inmaduras, leucocitos, eritrocitos, bacterias y otros residuos.

- **Morfología Espermática:** Permite evaluar las partes o regiones que constituyen al espermatozoide humano. En la **Figura 1** se pueden apreciar las partes de un espermatozoide que son objeto de análisis: la cabeza, pieza intermedia y la cola. La cabeza se encuentra divida en dos regiones llamadas acrosoma y post-acrosoma.

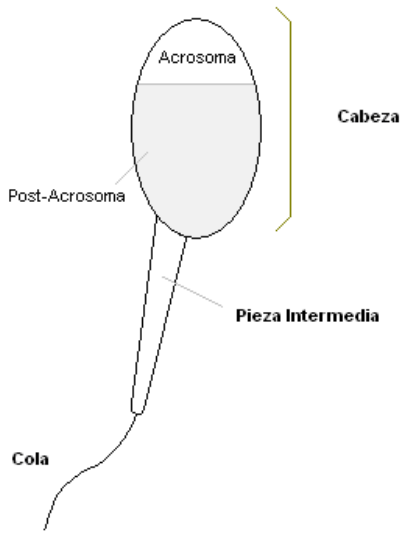



***Figura 1: Regiones del espermatozoide humano (Ferreira, 2008)***

### 1.1.1. Morfología Espermática

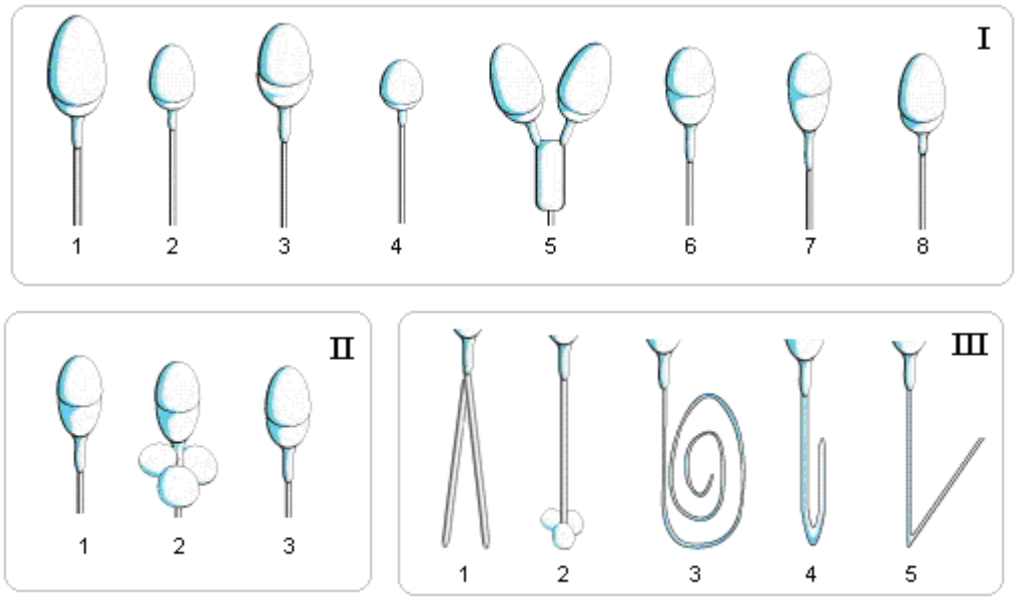
Al momento de realizar el análisis, el experto debe cumplir con las instrucciones que dicta la Organización Mundial de la Salud (OMS, 2001). Se debe contar de forma manual 200 espermatozoides y clasificarlos según sus características morfológicas, diferenciándolos de cualquier otra estructura presente en la muestra de semen (Katz, 1986). Dichos espermatozoides deben cumplir con las siguientes características morfológicas (de acuerdo a cada una de las regiones) para ser considerados normales:

- **Cabeza (Acrosoma y Post-Acrosoma):** Ovalada, única, lisa, acrosoma ocupa entre 40% y 70% del área. Longitud de 4-5 µm, ancho 2.5-3.5 µm. Si existen vacuolas, no deberán ocupar más de la mitad del área.

- **Pieza intermedia:** Estrecha, unida axialmente, alineada con la cabeza, sin restos citoplasmáticos, no debe exceder 1/3 del área de la cabeza. Longitud 6-10 µm.

- **Cola:** Uniforme, única, desenrollada, 45 µm de longitud, ligeramente más estrecha que la pieza intermedia, sin residuos citoplasmáticos.

Los espermatozoides que no cumplan con los parámetros anteriormente definidos son considerados como anormales, además que pueden presentar motilidad deficiente, al no contar con las condiciones aerodinámicas para desplazarse con la velocidad y orientación ideal (OMS, 2001). En la **Figura 2** se muestran las malformaciones de acuerdo a cada región del espermatozoide.



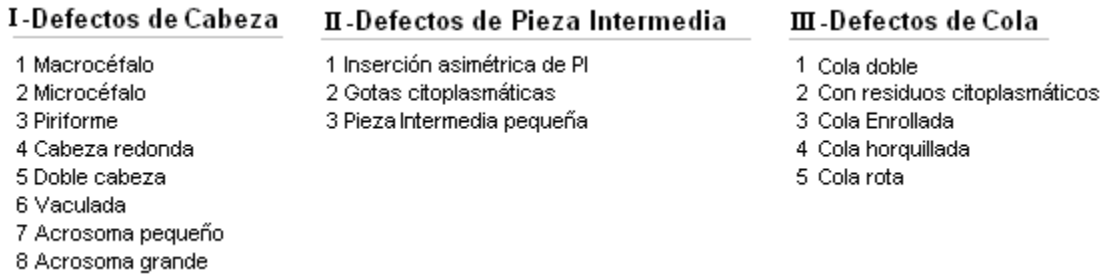



***Figura 2: Malformaciones morfológicas del espermatozoide humano (Ferreira, 2008)***

Para poder considerar una muestra morfológicamente normal, el porcentaje de espermatozoides normales debe ser mayor o igual a un 15%.

## 1.2. Minería de datos

El crecimiento desmedido del volumen de datos, generados por los distintos sistemas de gestión empresariales y la inadecuada exploración de los mismos, ha hecho necesario el uso de tecnologías que permitan su organización y procesamiento. Es por esto que surge la minería de datos como un enfoque distinto para el manejo de estos grandes repositorios de datos. (Obenshain, 2004).

De manera formal, la minería de datos se define como: "El proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos" (Witten, Frank & Hall, 2011). El objetivo de la minería de datos es transformar datos en conocimiento, haciendo uso de las técnicas más adecuadas.

Existen diversos campos, como se muestra en la **Tabla 1**, donde la minería de datos puede ser aplicada, para encontrar patrones ocultos que representen conocimiento. Este conocimiento obtenido puede generar grandes beneficios a las distintas organizaciones, al servir como ayuda en la futura toma de decisiones.

| Campo | Datos | Conocimientos |
|---|---|---|
| **Medicina** | Registros médicos asociados a pacientes | • Efectividad de los tratamientos<br>• Identificar riesgos de sufrir alguna patología<br>• Gestión hospitalaria |
| **Web** | Archivos de registros de servidores Web | • Análisis de comportamiento de usuarios<br>• Clasificación de sitios Web |
| **Mercado** | Variables financieras de compañías | • Análisis de riesgo de asignación de créditos<br>• Estimación de ventas<br>• Determinación de perfiles de compra |

| | | |
|---|---|---|
| **Telecomunicaciones** | Registros de llamadas | • Determinación de patrones de llamadas<br>• Detección de intrusos<br>• Uso de telefonía móvil |
| **Turismo** | Registro de reservaciones | • Identificación de patrones de reserva<br>• Segmentación de clientes |
| **Ingeniería de Software** | Valores de métrica de un proyecto | • Predicción y estimación de índices de calidad del software<br>• Duración y coste de proyectos |

***Tabla 1: Diferentes dominios donde la minería de datos puede ser aplicada***

### 1.2.1 Proceso de descubrimiento de conocimiento a partir de datos

El proceso de descubrimiento de conocimiento a partir de datos o KDD (*Knowledge Discovery In Databases)*, consta de diversas etapas, desde la obtención de los datos, hasta la aplicación del conocimiento adquirido.

Es común confundir los conceptos de minería de datos y KDD como equivalentes, pero existe una gran diferencia entre los dos: KDD es un proceso que consta de distintas etapas, y una de éstas es, justamente, la minería de datos (Hernández, Ramírez & Ferri, 2004).

Fayyad (1996) define KDD como "el proceso no trivial de identificar, a partir de datos, patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles". Según esta definición, Fayyad afirma que las propiedades que debería cumplir el conocimiento obtenido son:

- **Válido:** Los patrones encontrados deben permitir explicar o predecir datos nuevos.
- **Novedoso:** Deben aportar nuevos conocimientos.
- **Potencialmente útil:** Deben aportar algún beneficio, como la ayuda a la toma de decisiones.
- **Comprensible:** Los patrones deben facilitar su interpretación, revisión y validación en la toma de decisiones

Es importante analizar el dominio de la aplicación, lo que implica definir el problema a resolver y los objetivos a cumplir, para luego aplicar las cinco fases fundamentales del KDD, las cuales se explican a continuación **(ver Figura 3)**

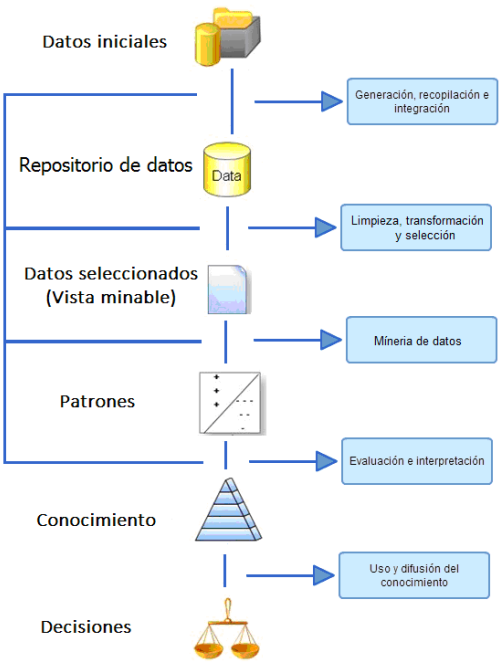



***Figura 3: Fases del proceso KDD (Hernández, Ramírez & Ferri, 2004)***

**- Generación, recopilación e integración de datos**

Los datos pueden ser obtenidos a través de múltiples experimentos o de aplicaciones para la adquisición de datos. Si éstos provienen de distintas fuentes de información (almacenes de datos, bases de datos relacionales, datos transaccionales, hojas de cálculo, entre otros), es necesario transformarlos a un formato común, para poder trabajarlos sin ninguna inconsistencia (Hernández, Ramírez & Ferri, 2004). Inclusive, algunos conjuntos de datos necesitarán ser preprocesados, tal es el caso de: conjuntos multimedia (imágenes, videos, audio), documentales (textos, páginas Web), entre otros.

Al finalizar esta fase se obtiene una estructura denominada “tabla atributo-valor”, la cual será utilizada en la próxima fase. En la tabla, las columnas son los atributos, variables o características, mientras que las filas son los distintos registros o instancias.

Los atributos pueden ser de distintos tipos:

- **Nominales o categóricos:** Variables cuyos valores representan una categoría, tales como, color, sexo, nacionalidad, entre otros.
- **Binarios:** Variables cuyos valores solo pueden representar dos categorías o estados, donde “0”significa ausencia y “1” significa presencia.
- **Ordinales:** Variables cuyos valores representan una jerarquía. Ejemplo grado académico: [Bachiller, Técnico, Licenciado, Postgrado].

- **Numéricos o cuantitavos:** Variables cuyos valores pueden ser expresados mediante números, éstos pueden ser discretos o continuos.

**- Preparación de los datos (Limpieza, Transformación y Selección)**

El objetivo de esta fase, es obtener la vista minable, por lo que los datos a utilizar deben cumplir con ciertas características que determinan la calidad de estos, tales como exactitud, completitud y consistencia (Ballou, & Tayi, 1999). Algunas de las tareas que se ejecutan en esta fase están dirigidas a:

- **Limpieza de los datos:** Se intenta eliminar o corregir datos que presentan características específicas, tales como, datos ausentes, valores fuera de rango, datos erróneos o con ruido, campos obsoletos o redundantes, entre otros.
- **Transformaciones:** En este caso, se aplican técnicas para modificar la forma de los datos y mejorar su representación, lo que permitirá que el proceso de minería sea más eficiente y los patrones más fáciles de entender. Por ejemplo, se puede cambiar el rango y tipo de las variables.
- **Selección de atributos:** Se escogen los atributos más importantes, con el fin de reducir la dimensionalidad de los datos, obtener modelos más comprensibles y reducir el tiempo de estimación de los modelos.

**- Minería de Datos**

Esta fase constituye el centro del KDD, y a veces se utiliza el término de minería de datos para referirse a todo el proceso (Witten, Frank & Hall, 2011). Su objetivo es producir conocimiento que pueda ser utilizado por el usuario. Para esto, es necesario la construcción de un modelo a partir de los datos, que pueda ser utilizado para realizar predicciones, o comprender mejor dichos datos (Hernández, Ramírez & Ferri, 2004). Antes de iniciar el proceso es necesario tomar ciertas decisiones como:

- **Determinar la tarea de minería de datos:** Éstas se pueden clasificar en dos tipos: tareas predictivas y tareas descriptivas.
- **Seleccionar el algoritmo o técnica:** Se busca un algoritmo que resuelva la tarea y que genere el tipo de patrón buscado.

En la Sección 1.2.2 se explicarán los tipos de tareas y modelos de la minería de datos. De igual forma, se explicarán las técnicas y algoritmos que son de importancia para este trabajo.

**- Evaluación e Interpretación**

En este paso, se interpretan y evalúan los patrones obtenidos en el proceso de minería de datos, lo que puede implicar repetir algunos de los pasos anteriores, realizar verificaciones con expertos, eliminar patrones redundantes, y utilizar medidas de evaluación de acuerdo a la tarea de minería de datos utilizada (Hernández, Ramírez & Ferri, 2004).

**- Uso y difusión del conocimiento**

En esta etapa, es común que se incorpore el conocimiento a un sistema que permita su uso. Generalmente, la difusión se logra a través de publicaciones, congresos, o estándares para el intercambio de conocimiento. El modelo debe ser revisado constantemente, para verificar su desempeño, ya que en algunos casos es necesario reevaluar, reentrenar y reconstruir de nuevo el modelo (Hernández, Ramírez & Ferri, 2004).

### 1.2.2. Tareas de la minería de datos

A la hora de aplicar la minería de datos, se debe tener en cuenta el tipo de problema que se intenta resolver. En general, existen dos enfoques:

**- Tareas predictivas:** En estas tareas los modelos predicen el valor de un atributo en particular (variable dependiente) basándose en los valores de otros atributos (variables independientes). Básicamente, existen dos tipos:

- **Clasificación:** En este caso, se tiene un atributo que representa una clase. El rango de valores para dicho atributo es un conjunto discreto de etiquetas simbólicas. El objetivo de esta tarea es generar un modelo que describa una relación de dependencia entre los demás atributos y la clase, que permita predecir la clase de futuras instancias.

- **Regresión:** En esta tarea el modelo resultante es una función que asigna a cada instancia un valor real. La diferencia entre la clasificación y la regresión es que en esta última, el valor a predecir para las nuevas instancias que se presenten es del tipo numérico.

**- Tareas Descriptivas:** En estos tipos de tareas no existe o no se toma en cuenta ningún atributo que represente la respuesta al problema. Los modelos identifican patrones que exploran las propiedades o las relaciones entre los datos. Entre las principales tareas descriptivas se encuentran:

- **Agrupación:** En este caso se buscan grupos en el conjunto de datos, tomando en cuenta la similitud de los valores de cada una de las instancias.

- **Asociación:** El objetivo de esta tarea es buscar relaciones frecuentes entre los distintos atributos que definen los datos. Dichas relaciones son representadas a través de reglas de asociación, las cuales son reglas de dependencias que predicen la ocurrencia de un atributo, basado en la ocurrencia de otros atributos.

- **Detección de anomalías:** Se buscan, entre las instancias, valores de atributos que sean muy distintos de los demás o que no entren entre los rangos esperados. Estos atributos son llamados “Outlier”. En general, a la hora de detectar estas anomalías, es necesario crear un perfil que refleje un comportamiento “Normal”, y a partir de éste, encontrar las características que sean distintas (Tan, Steinbach, & Kumar, 2005).

### 1.2.3. Técnicas de la minería de datos

En la minería de datos existen distintas técnicas que tienen como objetivo descubrir patrones a partir de datos. Las técnicas pueden ser implementadas por distintos algoritmos, por lo que es preciso comprender los parámetros y las características de estos, para determinar cuál es la técnica más apropiada a la hora de resolver un problema. A continuación, se explicarán las técnicas y algoritmos más relevantes para este trabajo, haciendo énfasis en la tarea a la cual están dirigidas, los tipos de atributos que soportan y las características más apreciables.

**A) Árboles de decisión:** Son utilizados en tareas de clasificación y regresión. Son estructuras formadas por un conjunto de nodos y de arcos dirigidos y organizados de forma jerárquica. Los nodos pueden ser clasificados en tres tipos, como se aprecia en la **Figura 4**. El nodo raíz, el nodo interno y el nodo hoja. Los dos primeros llevan a cabo pruebas o test sobre un atributo, donde los arcos de salida representan los distintos valores que el atributo puede tomar. Mientras que el nodo hoja representa el valor que retorna el árbol de decisión, es decir, la clase o valor numérico que se le asocia al registro (Hernández, Ramírez & Ferri, 2004).

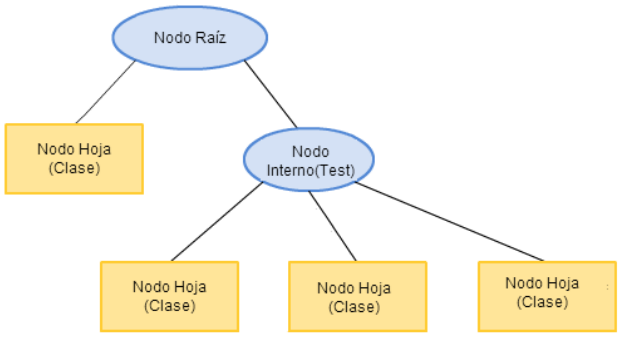



***Figura 4: Estructura de un árbol de decisión***

De forma general, en el proceso de construcción de un árbol de decisión es importante tener en cuenta ciertos aspectos:

**A.1)** En cada paso recursivo, el algoritmo debe seleccionar la mejor prueba de atributo para lograr dividir el conjunto de entrada en subconjuntos más pequeños, para esto existen distintos criterios de evaluación, tales como:

$$Entropía = \sum_{i=1}^{n} = P(i|t) * Log_2 P(i|t) \quad \textbf{(1)} \qquad Gini: 1 - \sum_{i=1}^{n} [P(i|t)]^2 \quad \textbf{(2)}$$

$$Error\ Esperado = 1 - Max_i[P(i|t)] \quad \textbf{(3)}$$

- En donde **n** es el número de clases

- **P (i|t)** = fracción de instancias pertenecientes a la clase i en el nodo t

Se debe seleccionar el nodo con el menor grado de impureza. El grado de impureza del padre, debe ser comparado contra el grado de impureza de los nodos hijos a través de la ganancia de información o promedio ponderado.

$$Ganancia = I(Nodo\ padre) - \sum_{j=1}^{k} \left(\frac{N_j}{N}\right) * I(NodoHijo_j) \quad \textbf{(4)}$$

**I** = Medida de impureza del nodo
**N** = Número de instancias en el nodo padre
**K** = Número de valores del atributo
**Nj** = Número de registros asignados al hijo

A la hora de seleccionar entre las posibles pruebas, se debe escoger la que tenga el índice más bajo.

**A.2)** A la hora de detener la expansión de los nodos, es necesario utilizar alguna condición de parada, tales como:

- Número máximo de iteraciones.
- Número definido de niveles o de hojas.

También se puede emplear una técnica de poda en el árbol:

- **Pre- poda:** Detener la generación de nodos cuando la medida de impureza es menor a la de un límite establecido.
- **Post- poda:** Se completa todo el árbol y luego se aplican técnicas para eliminar nodos que no sean necesarios.

Existen distintos algoritmos de árboles de decisión, pero en general se derivan del algoritmo básico de Hunt, el cual se muestra en la **Figura 5**:

```
Algoritmo_Hunt (D: Conjunto de datos, N: Nodo):

    Si los registros del conjunto D pertenecen a la misma clase C  Entonces
        Retornar N como un nodo hoja etiquetado con la clase C;

    Si _No el conjunto D es vacío Entonces
        Retornar N como un nodo hoja etiquetado con una clase por defecto;

    Si_ No los registros del conjunto D posee distintas clases Entonces
        Aplicar un Test de atributo, para dividir el conjunto D;
        Para cada resultado del test, crear un nodo hijo;
        Distribuir los registros del conjunto D, a los nodos hijos creados;
        Aplicar de forma recursiva Algoritmo_Hunt () a cada nodo hijo;

    Fin _Si

Fin _ Algoritmo_Hunt
```

***Figura 5: Algoritmo de Hunt (Hunt, 1966)***

A partir de este algoritmo que se ha presentado, se han propuesto otros que siguen la misma estructura base y que son muy utilizados, sin embargo, en este documento solo se explicarán los que siguen a continuación:

- **ID3 (Induction Decision Trees):** Al igual que el algoritmo de Hunt, cada vez que se crea un nuevo camino, se debe seleccionar un test que realice una mejor división de las instancias dependiendo de las clases. Para esto se utiliza como criterio de selección la Entropía (ver fórmula 3) mientras el valor sea menor, mejor será el test para realizar la clasificación. Por otro lado, el ID3 solo se puede aplicar a variables discretas, ya que no puede trabajar con atributos continuos debido a la Entropía (Quinlan, 1986). Otra desventaja de este algoritmo es la extensión de los árboles, lo que evita la fácil interpretación del mismo.

- **C4.5:** Debido a que está basado en ID3, su metodología es la misma, a excepción de que permite tratar con valores continuos, al realizar una discretización a los mismos. Por lo tanto se tienen dos tipos de prueba; si el atributo es discreto, se realiza el mismo tratamiento que en ID3, caso contrario si es continuo se debe comparar el valor del atributo contra un umbral X (Quinlan, 1993). Utiliza la razón de ganancia (Gain ratio) (ver fórmula 5) a la hora seleccionar un test, esta medida hace uso de la información de separación (Split Information) que es la entropía del conjunto de datos respecto al atributo a clasificar, así como también emplea la ganancia de información. Por lo tanto, la razón de ganancia es la división entre la ganancia de información y la información de separación.

$$Razón\ de\ ganancia = \frac{GananciadeInformación()}{SplitInfo()} \quad \textbf{(5)}$$

Donde $SplitInfo = \sum_{i=1}^{k} P(V_i) * Log_2(V_i)$ **(6)**

Emplea el método Post poda, con el algoritmo C4.5 se inicia en los nodos, hojas, hasta llegar al nodo raíz.

**B) Reglas de Clasificación:** Las reglas de clasificación son una alternativa a los árboles de decisión. Utilizan expresiones de la forma **SI** <*condición*> **Entonces** <*acción*>*,* donde la *Condición o Antecedente* está compuesto de una serie de test, tal como los test evaluados en los nodos de un árbol de decisión. Mientras que la *Acción o consecuente* determina la clase de una determinada instancia cubierta por dicha regla. Se dice que una regla cubre una instancia si la *Condición* corresponde con los valores de los atributos de la instancia (Witten, Frank & Hall, 2011).

Al momento de generar las reglas de clasificación es importante tener en cuenta ciertos aspectos:

**B.1)** Una regla R puede ser evaluada por su *Exactitud* y por su *Cobertura.* Dada una instancia X, de un conjunto de instancias D, $N_{cubiertos}$ es el número de instancias cubiertas por la regla R. $N_{correctos}$ es el número de instancias correctamente clasificadas por la regla R y |D| es el número de instancias en D. (fórmulas 7 y 8)

$$Cobertura(R) = \frac{Ncubiertos}{|D|} \quad \textbf{(7)} \qquad Exactitud(R) = \frac{Ncorrectos}{Ncubiertos} \quad \textbf{(8)}$$

La cobertura de una regla es el porcentaje de instancias que son cubiertas por dicha regla. Mientras que para la exactitud de una regla se toma en cuenta las instancias que son cubiertas y que porcentajes de estas son correctamente clasificadas. (Witten, Frank & Hall, 2011).

Los algoritmos de cobertura generan las reglas de forma secuencial (una a la vez), e idealmente, cada regla de la clase especificada deberá cubrir muchas de las instancias pertenecientes a esa clase y ninguna de las clases restantes (Han, Kamber & Pei, 2006). En la **Figura 6** se muestra un algoritmo básico de cobertura.

```
Algoritmo_Cobertura (D: Conjunto de instancias, Att_vals: Conjunto de atributos):

        conjunto_reglas= {} // El conjunto de reglas está vacío al inicio
        Para cada clase Ci:
                Repetir
                        regla= Crear_una_regla (D, Att_vals, Ci);
                        Eliminar (Instancias cubiertas por regla)
                Hasta Condición_de_parada
                conjunto_reglas= conjunto_reglas + regla;
        Fin_Para
        Retornar conjunto_reglas;

Fin_Algoritmo_ Cobertura
```

***Figura 6: Algoritmo básico de cobertura (Han, Kamber & Pei, 2006)***

La función **Crear_una_regla** encuentra la mejor regla para la clase actual (Ci), dado un conjunto de instancias. Las reglas pueden ser creadas de dos formas:

- De lo general a lo específico: Se inicia con una regla general que cubra todas las instancias y se van añadiendo test de forma iterativa para mejorar la calidad de la regla, excluyendo instancias que no pertenezcan a la clase (Ci).

- De lo específico a lo general: Se selecciona una instancia perteneciente a la clase actual (Ci) de forma aleatoria. Luego en cada iteración, se van eliminando test para generalizar la regla.

A partir de este algoritmo que se ha presentado, se han propuesto otros que siguen la misma estructura base y que son muy utilizados, sin embargo, en este documento solo se explicarán los que siguen a continuación:

- **CN2:** Este algoritmo crea una lista de reglas de clasificación ordenada, en donde la última regla es la "Regla por Defecto", la cual predice la clase más común en el conjunto de instancias. (Clark & Niblett, 1989). En este caso el criterio de selección utilizado para los test es la Entropía (Ver fórmula 3). Por otro lado CN2 utiliza una sustitución simple a la hora de tratar con valores desconocidos, simplemente reemplaza dicho valor con el más frecuente del atributo en cuestión. En el caso de los atributos numéricos utiliza el valor medio del subrango más frecuente (Clark & Niblett, 1989).

- **RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*):** Si el conjunto de datos solo presenta dos clases, el algoritmo selecciona una como la clase positiva y la otra como la clase negativa, creando así reglas con la clase positiva, y la clase negativa será la clase por defecto. En el caso de tener múltiples clases, estas son ordenadas de acuerdo al incremento de prevalencia de la clase

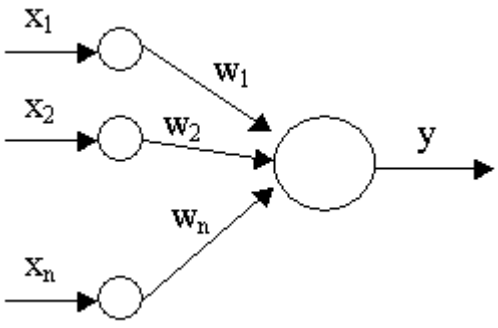
(fracción de instancias que pertenecen a una clase en particular). Utiliza como métrica de evaluación la ganancia FOIL, la cual se basa en el número de instancias positivas y negativas cubiertas, antes y después de agregar un test a la regla. Esta métrica selecciona los test que cubran el mayor número de instancias positivas y el menor número de instancias negativas (Han, Kamber & Pei, 2006) (Ver Fórmula 9)

$$FOIL\ GANANCIA(L,R) = T(\log_2(\frac{P1}{P1+N1}) - \log_2(\frac{P0}{P0+N0})) \quad \textbf{(9)}$$

**R:** Regla en definición
**L:** Nuevo test
**R':** La regla obtenida al agregarle el test a la regla original
**P0:** Número de instancias positivos cubiertos por R
**N0:** Número de instancias negativos cubiertos por R
**P1:** Número de instancias positivos cubiertos por R'
**N1:** Número de instancias negativos cubiertos por R'
**T:** Número de instancias positivos que siguen cubiertos por la regla R luego de añadir L.

**C) Redes Neuronales Artificiales**: Son modelos computacionales basados en las estructuras biológicas que componen el sistema nervioso central de los seres vivos. Su unidad básica de procesamiento es la neurona, la cual es capaz de recibir información proveniente del exterior, o de otras neuronas pertenecientes a la misma red. Esta información es procesada y transmitida a otras neuronas mediante las conexiones existentes entre ellas (Acosta, Salazar, & Zuluaga, 2000).

En otras palabras, una red neuronal es capaz de percibir estímulos (entrada), activar o inhibir neuronas (procesamiento), y generar resultados en base a las entradas de la red (salida), emulando así el comportamiento de una red neuronal biológica (Acosta, Salazar, & Zuluaga, 2000). En la **Figura 7** se aprecia la estructura básica de una red neuronal artificial




**Xi**: Canales de entrada donde se reciben los estímulos.

**Wi**: Pesos sinápticos asociados a cada conexión.

**Y:** Canal de salida, donde se transmite el valor generado por la función de salida.

***Figura 7: Estructura de una Red Neuronal Artificial***

Su utilidad radica en su capacidad para procesar información de forma paralela, logrando resolver problemas en los cuales los modelos de computación secuenciales son poco eficientes. Entre las áreas donde se aplican las redes neuronales artificiales se encuentran: reconocimiento de patrones, predicción de modelos, procesamiento de imágenes, filtrado de señales, entre otros (Acosta, Salazar, & Zuluaga, 2000).

A la hora de crear una red neuronal artificial es necesario tener en cuenta ciertos aspectos:

**C.1)** La información que ingresa a la red neuronal es procesada mediante tres funciones:

- Función de Entrada: Calcula los valores de entrada de la neurona.

- Función de activación: Calcula el valor que la neurona genera para ser enviado por el canal de salida, tomando en cuenta las entradas y los pesos sinápticos.

- Función de salida: La mayoría de los casos la función de salida es la función identidad F(X)=X, siendo la salida la que calcula la función de activación.

**C.2)** La estructura de una red neuronal artificial puede estar compuesta por una o muchas capas dependiendo del problema a resolver.

- La primera capa puede contener de una a N neuronas en paralelo por donde se recibe la información. (Garrido, Latorre, 2001).

- Luego se puede tener una o varias capas “ocultas” cuyo objetivo es procesar la información recibida a mediante el uso de las funciones de entrada, activación y de salida.
- Finalmente se tiene una capa de salida que proporciona los resultados de la red.

**C.3)** A la hora de resolver un problema de clasificación, en general, existen distintos tipos de redes neuronales, tales como:

- **Perceptrón Multicapa:** Se basa en la red neuronal “Perceptrón” (también denominada Perceptrón simple). Compuesto de una capa de entrada, una o más capas ocultas, y una capa de salida. El método de aprendizaje que suele usar este tipo de red es llamado Retropropagación del error (*Backpropagation*), el cual es un método de aprendizaje supervisado basado en la técnica del descenso por gradiente. Este funciona de la siguiente manera: Se inicializan los pesos y los umbrales, luego se presenta un patrón “N” de entrenamiento y se propaga hacia la salida, obteniendo la salida de la red Y(N). Se procede a evaluar el error cuadrático e(N), cometido para cada patrón. Luego, es necesario aplicar la regla “Delta generalizada”, la cual, modifica los pesos y los umbrales calculando el valor delta de atrás hacia adelante. Es decir, de la capa de salida, a la capa de entrada. Por último se calcula el error actual y de acuerdo al criterio de parada se detiene o se repite el procedimiento.

- **Redes neuronales de base radial o *Radial Basis Function Network* (RBFN):** Compuesta de tres capas, capa de entrada, una o más capas ocultas, y la capa de salida. Estas redes, a diferencia de otras, utilizan funciones de base radial como función de activación de la neurona. Estas funciones reciben dos parámetros, el centro que define un vector de la misma dimensión del vector de entrada. El otro parámetro es el ancho o Desviación estándar, el cual es el término empleado para identificar a la amplitud de la campana de Gauss originada por la función radial.

  La función de base radial comúnmente utilizada es la función Gaussiana (ver fórmula 10).

$$\varphi(X) = e^{-\frac{||x_i - c_i||^2}{2\sigma^2}} \qquad \textbf{(10)}$$

  Donde $X_i$ y $C_i$ Son los componentes de n dimensión del vector de entrada X y el vector centro C, respectivamente. Finalmente **σ** es la desviación estándar.

  Este tipo de redes neuronales requiere un proceso de entrenamiento de dos etapas, también llamado aprendizaje hibrido. En la primera etapa se aplica un aprendizaje no supervisado en la capa oculta (por ejemplo se aplica el algoritmo de agrupación o *Clustering*). En la segunda etapa se aplica un aprendizaje supervisado en la capa de salida (como el algoritmo Gradiente estocástico: LMS).

### 1.2.4. Evaluación de modelos de minería de datos

En la evaluación de un modelo de minería de datos, se pretende obtener una estimación de su rendimiento antes de ser aplicado a un nuevo conjunto de datos. A la hora de realizar la evaluación de los modelos obtenidos, es necesario utilizar una técnica y una medida de evaluación dependiendo de la tarea utilizada.

- **Técnicas de evaluación:** Son utilizadas para estimar el error de predicción del modelo.
  - **Selección aleatoria (*Holdout Method*):** En esta técnica se divide el conjunto de datos D en dos conjuntos de manera aleatoria: uno para el entrenamiento (70-80%), y otro para el test (20-30%). El modelo se construye utilizando el conjunto de entrenamiento y se determina el valor de la medida de evaluación sobre el conjunto test; este proceso se repite varias veces. Luego, se promedian todo los valores (Witten, Frank & Hall, 2011).
  - **Validación Cruzada (*Cross Validation*):** En esta técnica se divide el conjunto de datos (D) en (N) particiones disjuntas de forma aleatoria. Dichas particiones serán utilizadas de forma individual para realizar la

evaluación, y con la parte restante se realiza el entrenamiento. Luego de realizar el proceso de entrenamiento y la evaluación (N) veces, se promedia el valor de la medida de evaluación de cada iteración (Witten, Frank & Hall, 2011).

- **Medidas de evaluación:** Estas dependen de la tarea que se haya seleccionado. A continuación, se detallarán las medidas comúnmente utilizadas para evaluar la tarea de clasificación utilizada en la investigación.

  A la hora de evaluar un clasificador, se debe tomar en cuenta la cantidad de registros predichos de forma correcta o incorrecta. Para esto se puede hacer uso de una Matriz de confusión**,** tal como se muestra en la **Figura 8**, donde

  **VP:** Es el número de predicciones correctas cuando la instancia es positiva.
  **FP:** Es el número de predicciones incorrectas cuando la instancia es positiva.
  **FN:** Es el número de predicciones incorrectas cuando la instancia es negativa.
  **VN:** Es el número de predicciones correcta cuando la instancia es negativa.

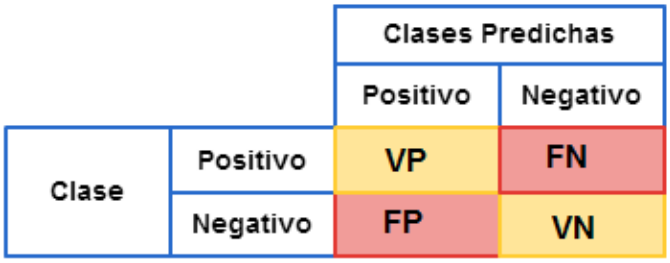



***Figura 8: Matriz de confusión para la evaluación de un clasificador***

La cantidad de predicciones correctas realizadas por el modelo, se determina sumando todos los valores que se encuentran en la diagonal de la matriz (VP+VN), mientras que la cantidad de predicciones incorrectas, se calcula sumando los valores que no formen parte de la diagonal (FP+FN).

Existen dos métricas de rendimiento que pueden ser calculadas, a partir de la matriz de confusión, como son **la exactitud y el error**. Lo ideal, es que los resultados arrojados por estas medidas sean altos en el caso de la exactitud, y mínimos en el caso del error.

$$Exactitud = \frac{Número\ de\ predicciones\ correctas\ (VP+VN)}{Número\ Total\ de\ predicciones\ (VP+FN+FP+VN)} * 100 \quad \textbf{(11)}$$

$$Error = \frac{Número\ de\ predicciones\ incorrectas\ (FP+FN)}{Número\ Total\ de\ predicciones\ (VP+FN+FP+VN)} * 100 \quad \textbf{(12)}$$

Existen otras medidas, que determinan el rendimiento de cada una de las clases:

- **Precisión**: Indica cuantas instancias clasificadas como positivas fueron clasificadas correctamente.

$$Presición = \frac{VP}{VP+FP} \quad \textbf{(13)}$$

- **Sensibilidad o Recall**: Indica cuantas instancias de la clase positiva fueron clasificadas correctamente

$$Sensibilidad\ o\ Recall = \frac{VP}{VP+FN} \quad \textbf{(14)}$$

- **Especificidad**: Indica cuantas instancias clasificadas como negativas fueron clasificadas correctamente.

$$Especificidad = \frac{VN}{VN+FP} \quad \textbf{(15)}$$

Existen medidas que son utilizadas en problemas con clases no balanceadas tal es el caso de la Media Geométrica (Guo et al., 2008; He et al., 2009).

$$MediaGeometrica = \sqrt{Sensibilidad * Especificidad} \quad \textbf{(16)}$$

### 1.2.5. Combinación de modelos de minería de datos

Cuando es necesario tomar una decisión importante, usualmente se toma en cuenta la opinión de varios expertos en el tema en lugar de confiar en un único juicio. En minería de datos un modelo puede ser considerado como un experto. Para esto se deben tomar en cuenta ciertos factores, tales como la cantidad y calidad de los datos de entrenamiento, o si la técnica implementada fue la apropiada para resolver el problema. Un enfoque más exacto para una toma de decisiones más fiable seria combinar la salida de diferentes modelos de datos (Witten, Frank & Hall, 2011). Dicho enfoque es conocido como combinación de clasificadores (*Ensemble Classifiers)* el cual consiste en la integración de múltiples clasificadores con el fin de obtener un modelo predictivo más preciso (Polikar, 2006).

Entre las ventajas de emplear este enfoque, se tiene que los resultados son menos dependientes en cuanto a peculiaridades del conjunto de entrenamiento (Rokach, 2010). Por otro lado, la desventaja radica en la posible complicación del modelo, perdiendo comprensibilidad. Es decir, no sería fácil determinar cuáles factores contribuyen a la mejora de las decisiones. (Witten, Frank & Hall, 2011).

A continuación se explicarán algunos de los métodos más conocidos para la combinación de modelos.

- **Bagging (*Bootstrap Aggregation*):** Este método intenta neutralizar la inestabilidad de los métodos de aprendizaje, alterando el conjunto de entrenamiento mediante la eliminación o duplicación de algunas instancias. En la **Figura 9** se aprecian las dos etapas del algoritmo, la generación del modelo y la clasificación. En la generación del modelo las instancias se muestrean al azar, remplazando el conjunto original para crear uno nuevo del mismo tamaño (esto es conocido como *Bootstrap Sample*). Dichos conjuntos de datos son diferentes entre sí, pero no son independientes ya que todos provienen del conjunto de datos original. En la etapa de la clasificación, por cada conjunto de datos generado es creado un clasificador, y luego se realiza una votación para seleccionar la clase más frecuente (Witten, Frank & Hall, 2011).

**Generación de Modelos** (**N**: número de instancias en el conjunto de entrenamiento):
**Para** cada iteración **T**:
Tomar N instancias del conjunto de entrenamiento.
Aplicar el algoritmo de aprendizaje a la nueva muestra.
Almacenar el modelo resultante.
**Clasificación**:
**Para** cada modelo **T**:
predecir clase de la instancia utilizando el modelo
**Retornar** clase que ha sido seleccionada como más frecuente.

***Figura 9: Algoritmo para Bagging (Witten, Frank & Hall, 2011)***

- **Boosting:** Con el método *Bagging* se obtienen resultados positivos cuando los modelos son significativamente diferentes entre sí, y cuando cada uno trata un porcentaje razonable del conjunto de datos correctamente. Lo ideal es que los modelos se complementen entre sí, y que cada uno sea un especialista en un área del dominio en donde los demás no tienen un buen desempeño. El método *Boosting* aprovecha este conocimiento al buscar explícitamente modelos que se complementen los unos con los otros (Witten, Frank & Hall, 2011).

  Mientras que con *Bagging* la creación de los modelos se realiza de forma individual, con *Boosting* la creación de un modelo es influenciado por el rendimiento del generado anteriormente. El algoritmo expuesto en la **Figura 10** asigna el mismo peso a cada una de las instancias del conjunto de datos, los cuales serán rebalanceados de acuerdo a la salida de los clasificadores. El peso de las instancias correctamente clasificadas disminuye, mientras que el de las clasificadas erróneamente aumenta, esto ocurre con el fin de que en la siguiente interacción el nuevo modelo se centre en las instancias con mayor peso (Witten, Frank & Hall, 2011).

**Generación de Modelos**

Asignar mismo peso a cada instancia del conjunto de entrenamiento
**Para** cada iteración **T**:
Aplicar el algoritmo de aprendizaje al conjunto ponderado.
Almacenar el modelo resultante.
Calcular el error **E** del modelo sobre el conjunto de datos y almacenar el error
Si **E** igual a cero, o mayor o igual a 0.5:
Terminar generación del modelo.
**Para** cada instancia en el conjunto de datos:
**Si** instancia clasificada correctamente por el modelo
Multiplicar peso de la instancia por (**E**/ (1-**E**))
Normalizar peso de todas las instancias.

**Clasificación**:

Asignar peso igual a cero a todas las clases:
**Para** cada T modelo:
Sumar –log (**E**/ (1-**E**)) a los pesos de las clases predichas por el modelo.
**Retornar** clase con el peso más alto.

***Figura 10: Algoritmo para Boosting (Witten, Frank & Hall, 2011)***

### 1.2.6. Aplicación de la minería de datos en medicina y áreas relacionadas

En la mayoría de los centros hospitalarios, es normal que se generen grandes cantidades de información de gran valor para las investigaciones médicas y científicas. Al mismo tiempo, la evolución de la tecnología ha permitido el estudio y la clasificación de dicha información, específicamente, haciendo uso de la minería de datos, la cual, como se explicó previamente, hace posible obtener patrones o modelos significativos que permiten generar nuevos conocimientos. Con este fin, se presentan a continuación algunas investigaciones en las cuales se aplicaron técnicas de minería de datos para resolver problemas en el área de la medicina que tienen que ver con el diagnóstico y la clasificación.

Una primera corriente de investigación ha consistido en aplicar técnicas de clasificación de datos para obtener un diagnóstico médico. Se toma como primer ejemplo la investigación realizada por Vega, Sánchez, y Cortiñas (2012), la cual se planteó como objetivo principal aplicar la técnica de árboles de decisión para hallar reglas que permitan clasificar un paciente con dengue en cualquiera de sus dos estados (FD) ó (FHD/SCD), a partir de las características clínicas y de laboratorio. La técnica empleada fue árboles de decisión y el algoritmo CART. El desempeño del algoritmo se evaluó sobre la capacidad del método de reducir la tasa de error global y su habilidad para clasificar correctamente a los pacientes con FHD/SCD. Las conclusiones de esta investigación apuntan a que la implementación de árboles de decisión puede constituir un método capaz de proporcionar

reglas de decisión útiles en la práctica clínica y en la elaboración de sistemas expertos para diagnosticar un paciente.

Siguiendo esta estrategia se ha implementado, por ejemplo, el uso de clasificadores para contribuir al diagnóstico de la Hipertensión arterial. Dávila y Sánchez (2012) seleccionaron para el desarrollo de la investigación el algoritmo J48 (Versión de WEKA) dentro de la técnica supervisada árboles de decisión. Sin embargo, se debe tener en cuenta que un problema de este tipo puede ser resuelto haciendo uso de distintas técnicas, por lo que Dávila y Sánchez (2012) decidieron utilizar una técnica adicional con el fin de proponer distintos modelos cuya combinación pueda mejorar el diagnóstico de la Hipertensión arterial. Para esto decidieron utilizar el algoritmo simple K- *means* para el desarrollo de la técnica no supervisada (agrupación).

En el mismo orden de ideas, la investigación realizada por Lorca, Arzola, y Pereira, (2010), se planteó como objetivo principal segmentar imágenes médicas para reconstruir modelos anatómicos 3D, haciendo uso de algoritmos de agrupación (*Clustering*). En esta investigación se implementaron dos algoritmos de agrupación, K- *means*, y *Fuzzy* K- *means*. Este último es una extensión del K- *means*, la diferencia radica en que el algoritmo K- *means* encuentra particiones para las que un punto pertenece a un solo cluster, mientras que *Fuzzy* K- *means* permite que un punto pueda pertenecer a más de un cluster, tomando en cuenta un grado o intervalo de pertenencia entre los cluster, el cual indica que un elemento puede pertenecer totalmente a una clase, a todas, o a ninguna. Los autores afirman que para el perfeccionamiento de las técnicas de segmentación de imágenes médicas mediante técnicas de agrupación es necesaria la búsqueda de procedimientos más avanzados de inicialización de los centroides, y la utilización de algoritmos que permitan alcanzar el óptimo global o aproximarse al mismo, tanto para el método K-*means* como para el *Fuzzy* K- *means*.

En ocasiones se necesitan técnicas para analizar problemas en los cuales se quiere predecir un valor real. En este caso es imprescindible utilizar técnicas de regresión, como por ejemplo la investigación realizada por Raghavendra y Jay (2011). Los algoritmos empleados fueron regresión logística, regresión logística con *Backward Elimination* (procedimiento de selección de variables en el que se incluyen todas las variables a la ecuación y se van eliminando aquellas que cumplan con el criterio de exclusión, es decir las que tengan la menor relación con la variable dependiente) y regresión logística con F*orward Selection* (procedimiento de selección de variables en el que se incluyen una por una las variables a la ecuación, siempre que cumplan con el criterio de entrada, es decir la que tenga mayor relación con la variable dependiente). La técnica de evaluación empleada fue la validación cruzada, mientras que las medidas de evaluación fueron la Exactitud (CA), la raíz cuadrada media estandarizada (RMSE) y el error absoluto medio (MAE). Los autores afirman que los resultados fueron concluyentes sobre la efectividad de la regresión logística con métodos de selección, ya que la exactitud (CA), el error absoluto medio (MAE), y la raíz cuadrada media estandarizada (RMSE) son más eficientes cuando los métodos de selección son empleados lo que valida la hipótesis de la investigación.

Por otra parte, en la investigación realizada por Delen, Walker, y Kadam (2004), se usó la minería de datos para desarrollar modelos de predicción del cáncer de mama. En esta oportunidad se utilizaron dos tipos de tareas, regresión y clasificación. Las técnicas empleadas fueron, árboles de decisión, redes neuronales, y regresión logística. En el caso de los árboles de decisión se implementó el algoritmo C5.0 y para las redes neuronales se empleó la arquitectura Perceptrón multicapas, (MLP por sus siglas en inglés). De los tres algoritmos empleados, el árbol de decisión tuvo mejores resultados con una exactitud de 93.6%, sensibilidad de 96%, y especificidad del 90.6%. En segundo lugar se encuentran las redes neuronales con las cuales se obtuvo una exactitud del 91.2%, sensibilidad del 94.3 %, y especificidad del 87.4%. En la última posición, la regresión logística con una exactitud del 89,2%, sensibilidad de 90.1%, y especificidad del 87.8%, en base a estos resultados los autores afirman que con los algoritmos de clasificación se obtuvieron mejores resultados en comparación con el algoritmo de regresión.

Algunas investigaciones emplean técnicas de asociación; tal es el caso de la presentada en Antonie, Zaïane y Coman (2001), la cual se centró en clasificar mamografías digitalizadas en dos grandes grupos, normal y anormal utilizando técnicas de asociación. La muestra consistió de 322 imágenes pertenecientes a tres grandes categorías: normal, benigno y maligno, mientras que la técnica empleada fue el algoritmo Apriori. Al analizar los resultados se observó que el conjunto de reglas era sensible al conjunto de datos desequilibrado que contiene alrededor de 70% de casos normales y solo el 30% de casos anómalos, por lo que se decidió repetir todo el proceso utilizando una distribución equilibrada de casos normales y anómalos. La tasa de éxito en este segundo intento fue mayor que el caso anterior, sin embargo, debido al desequilibro entre las imágenes malignas y las benignas, el número de reglas generadas para las imágenes malignas fueron extremadamente reducidas, por lo tanto, tres de cada cuatro imágenes malignas fueron clasificadas como anormales benignas. En resumen, los resultados obtenidos en el segundo intento presentaron una menor tasa de reconocimiento de grasa en mamografías anormales y una mayor tasa de reconocimiento para todos los casos normales.

Una investigación realizada por Pereira y Yépez (2012) destinada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino, empleó técnicas de asociación y clasificación para la búsqueda y descubrimientos de patrones insospechados y de interés en un conjunto de datos con características específicas de este grupo poblacional. El algoritmo utilizado para obtener las reglas de clasificación con árboles de decisión fue J48 (en WEKA), mientras que algoritmo utilizado para obtener las reglas de asociación fue el algoritmo Apriori. Los resultados obtenidos a través de las técnicas clasificación y asociación indican que estas son capaces de generar modelos consistentes, además aplicando la tarea de Asociación se conocieron los principales factores socioeconómicos y clínicos asociados a la supervivencia de este grupo poblacional.

El objetivo de la investigación de Wilford, Rosete y Rodríguez (2009), fue analizar la información que provee un conjunto de coronariografías realizadas a pacientes con cardiopatía isquémica. En este trabajo se utilizaron dos tipos de tareas: clasificación y

asociación. Originalmente se consideraron 12 variables y se utilizaron 3 conjuntos de entrenamiento D, D1, y D2, donde estos últimos dos son subconjuntos disjuntos del primero. A los tres conjuntos se les aplicaron las mismas técnicas, esto con el fin de realizar análisis comparativos entre los resultados de los datos centralizados (Conjunto D) y los datos separados en conjuntos disjuntos (conjuntos D1, D2). La técnica empleada, para la tarea de clasificación fue árboles de decisión, y el algoritmo utilizado fue C4.5. El algoritmo utilizado en la tarea de asociación fue el algoritmo Apriori. En los resultados obtenidos se analizó la semejanza entre los conocimientos obtenidos de los distintos modelos generados con el objetivo de integrar los modelos generados por los conjuntos D1 y D2, y que el resultado fuese similar al modelo del conjunto D. Los autores afirman que con el algoritmo Apriori, fue posible lograr la integración de distintos modelos de datos con distintas fuentes homogéneas.

Por su parte, la investigación realizada por Soni, Ansari, Sharma y Soni (2011) planteó como objetivo emplear distintos algoritmos de minería de datos para la predicción efectiva de distintas enfermedades del corazón. Entre las técnicas aplicadas se encuentran, árboles de decisión, clasificadores bayesianos, clasificación vía algoritmos de agrupación, así como también algoritmos genéticos utilizados para reducir el tamaño de los datos y obtener un conjunto óptimo de atributos. El resultado de aplicar distintas técnicas predictivas en el mismo conjunto de datos revela que el árbol de decisión obtuvo un mayor desempeño (89%), y la clasificación bayesiana obtuvo (86.53%), mientras que la clasificación vía algoritmos de agrupación obtuvo un menor rendimiento. Sin embargo, luego de aplicar los algoritmos genéticos, la exactitud de los algoritmos mejoró, el árbol de decisión obtuvo (99,2%), la clasificación bayesiana (96,5%) y la clasificación vía *clustering* obtuvo (88,3%).

Por último, se presenta la investigación realizada por Casañas, Ramos y Núñez (2011) cuyo objetivo se basa en la construcción de un modelo de clasificación morfológica de espermatozoides humanos a partir de imágenes. La evaluación morfológica es de gran utilidad ya que provee información útil en la toma de decisiones en procedimientos de reproducción. Pero la realización de dicha evaluación resulta una tarea compleja para el experto debido a todos los factores que se deben tener en cuenta, por lo que es deseable la automatización de dicho proceso haciendo uso de imágenes digitales correspondientes a láminas de muestras de semen. Para generar el conjunto de datos se utilizó una herramienta de adquisición de conocimiento creada por Ferreira (2008), esta aplicación permite cargar imágenes digitales para realizar la segmentación de los espermatozoides y extraer las características deseables de cada muestra. El uso de esta herramienta permitió generar un conjunto de datos a través del análisis de un grupo de muestras de semen. Entre las técnicas implementadas se encuentran árboles de decisión, reglas de clasificación y redes neuronales. En el caso de los árboles de decisión se aplicó el algoritmo C4.5, para las reglas de clasificación se utilizó el algoritmo RIPPER y por último se utilizaron redes neuronales de base radial (RBFNN). Los autores afirman que los mejores modelos son los estimados con las redes neuronales y con el algoritmo C4.5 en comparación con el algoritmo RIPPER.

# CAPÍTULO 2. MARCO APLICATIVO

## 2.1. Planteamiento del problema

En la Escuela de Computación de la Universidad Central de Venezuela se han realizado Trabajos de investigaciones cuyos objetivos se centran en la automatización de la evaluación morfológica de espermatozoides humanos. Uno de ellos es el presentado por Ferreira (2008), el cual consistió en la construcción de una herramienta de adquisición de conocimiento, la cual identificara y extrajera de forma automática características de células espermáticas humanas, y que funcionara a partir de la definición de un modelo basado en técnicas de procesamiento digital de imágenes. Esta aplicación requirió el diseño de un algoritmo de segmentación que permitiera identificar, con el menor margen de error posible, los espermatozoides presentes así como también el diseño de un algoritmo de extracción de características que midiera y codificara las distintas estructuras presentes en las regiones resultantes del procedimiento de segmentación. Esta herramienta, además de facilitar a los expertos el procesamiento de las imágenes de las muestras de espermatozoides, permitió introducir etiquetas de clasificación con el objetivo de construir vectores de características para estimar modelos de clasificación. Sin embargo, la clasificación morfológica seguía dependiendo del conocimiento y subjetividad del experto. Es por esto que la investigación de Casañas, Ramos, y Núñez (2011) se centró en la construcción de un modelo de clasificación morfológico de espermatozoides humanos, haciendo uso de distintos algoritmos de minería de datos y tomando como conjunto de datos los vectores de características extraídos de la investigación de Ferreira (2008).

Sin embargo, estas aproximaciones no se encuentran integradas en una única aplicación que sirva de apoyo al experto. Además, presentan ciertas características desfavorables que necesitan ser mejoradas, tal es el caso de los algoritmos de segmentación diseñados por Ferreira (2008), los cuales no se encuentran adaptados para tratar con imágenes de mayor resolución a la establecida en su aplicación (1600x1200 píxeles). En el caso de la investigación realizada por Casañas, Ramos, y Núñez (2011) se ve atada directamente a un único modelo de clasificación, sin ninguna posibilidad de modificarlo de acuerdo a las necesidades del experto.

Con el fin de subsanar estos problemas, se propone desarrollar una aplicación que de forma automatizada, permita la realización de evaluaciones morfológicas de cabezas de espermatozoides a partir de imágenes digitales, sirviendo de apoyo a los expertos en el área. La aplicación permitirá la carga de imágenes (de mayor resolución a 1600x1200px) correspondientes a láminas de muestras de semen, de igual forma permitirá establecer valores de referencia y estará en capacidad de realizar la evaluación de la morfología espermática de las muestras suministradas (de forma automática). Es importante mencionar que el experto contará con un módulo que generará un nuevo modelo de

clasificación en base a muestras clasificadas previamente basándose en su conocimiento. En la **Figura 11** se muestra un esquema general de la solución.

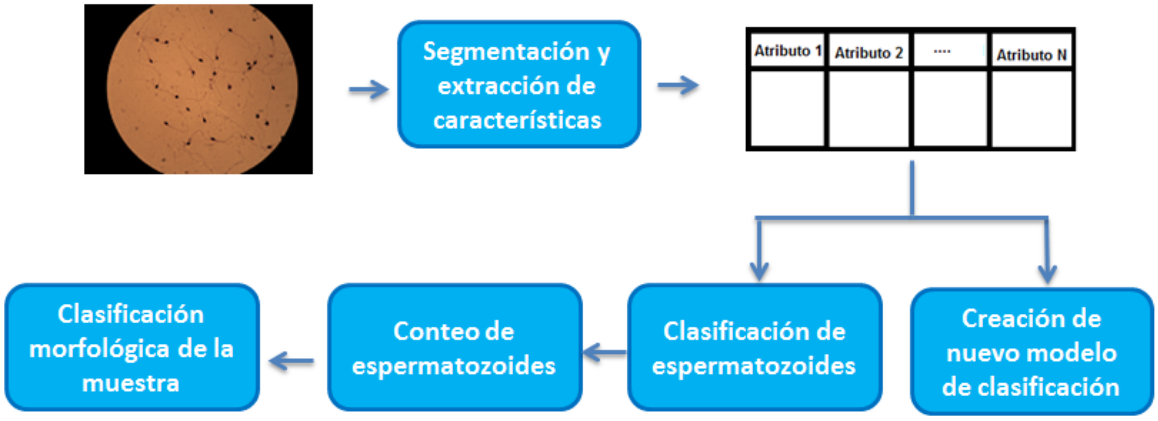



***Figura 11: Vista general de la aplicación***

- **Segmentación y extracción de características:** En este módulo, se procesarán las imágenes utilizando los algoritmos de segmentación y extracción de características diseñados por Ferreira (2008). Éstos permitirán identificar los espermatozoides que se encuentran en la muestra, así como también extraer las características presentes en cada uno de los espermatozoides. Además, se realizaran modificaciones para que se puedan procesar imágenes con resoluciones mayores a 1600x1200 píxeles.

- **Clasificación de espermatozoides:** En este módulo se clasificarán cada uno de los espermatozoides extraídos en la fase anterior, haciendo uso de un modelo construido aplicando el proceso de minería de datos. Como paradigma de clasificación se propone utilizar el esquema de combinación de clasificadores (*Ensemble Classifiers)*.

- **Creación de nuevo modelo de clasificación**: En este módulo el experto podrá clasificar muestras de forma manual basándose en su conocimiento, generando así vectores de características que serán almacenados y que podrán ser utilizados como conjunto de datos al momento de generar el nuevo modelo de clasificación. De igual forma, el experto podrá cargar un archivo con el conjunto de datos previamente clasificado.

- **Contabilización de espermatozoides:** Se llevará el conteo de espermatozoides por muestra, hasta alcanzar los 200 establecidos como límite, siguiendo el protocolo que utilizan los expertos.

- **Clasificación morfológica de la muestra:** Una vez clasificados 200 espermatozoides, se realizará la evaluación morfológica de la muestra, tomando como base las reglas que utilizan los expertos y los valores de referencia.

## 2.2. Objetivos

El objetivo de la presente investigación consiste en:

- **Objetivo general:** Desarrollar una aplicación, basada en minería de datos, que de forma automatizada realice la evaluación morfológica de cabeza de espermatozoides a partir de imágenes digitales.

- **Objetivos específicos**

  - Desarrollar el módulo de segmentación y extracción de características morfológicas a partir de imágenes digitales de muestras de células espermáticas humanas, tomando como base los algoritmos diseñados por Ferreira (2008).

  - Generar un modelo de clasificación de la morfología de cabeza de espermatozoide, aplicando el proceso de minería de datos, y utilizando el enfoque *Ensemble Classifiers*.

  - Desarrollar el módulo de conteo de espermatozoides y generación de la evaluación morfológica siguiendo las reglas de los expertos.

  - Desarrollar un módulo que permita al experto generar un nuevo modelo de datos a partir de la clasificación previa y bajo su criterio de las muestras de cabezas de espermatozoides.

  - Integrar los módulos en una aplicación que tome como entrada las imágenes correspondientes a una muestra de semen.

  - Evaluar el desempeño de la aplicación, a través de pruebas de rendimiento y de fidelidad.

## 2.3. Módulo de segmentación y extracción de características

En este módulo, se emplearon los algoritmos de segmentación y extracción de características diseñados por Ferreira (2008). Éstos permitieron identificar los espermatozoides presentes en una determinada muestra, así como también extraer las características de cada uno de estos espermatozoides.

- **Segmentación de imágenes:** Los algoritmos de segmentación hacen uso de imágenes de tamaño fijo (1600x1200 píxeles) las cuales fueron capturadas bajo las mismas condiciones que aplicaría un experto al momento de analizar una muestra. Es decir, la muestra debió ser posicionada en las regiones con menor densidad de espermatozoides con el fin de evitar agrupamientos. Luego, se

empleó la técnica de recorrido de escena en forma de “S” y por último se llevó un conteo de espermatozoides, hasta alcanzar los 200, para poder calcular la concentración espermática.

El próximo paso fue registrar las imágenes en el espacio de color RGB, por lo que se debió convertir cada imagen a escala de grises, evitando la perdida de información de las mismas. La estrategia utilizada para realizar la conversión, consistió en analizar el histograma de cada canal y determinar, cuál de estos aportaba mayor información de contraste a la escena. Los resultados indican que los canales rojo y verde aportan gran cantidad de información, a diferencia del canal azul, por lo que se decidió prescindir de este último. Luego se aplicó una binarización a la imagen, aplicando técnicas de umbralización de dos niveles sobre la imagen de canal rojo. La imagen resultante, permite construir dos matrices, la **MAP** (Máscara de área de procesamiento), y la **MCE** (Máscara de cabezas de espermatozoides) Ferreira (2008).

La MAP es utilizada para delimitar la región a segmentar indicando que píxel debe ser o no procesado. Por otro lado la MCE almacena las posibles áreas de la imagen que corresponden a las cabezas de los espermatozoides. Está ultima estructura se obtiene aplicando el algoritmo de búsqueda de profundidad o DFS sobre la imagen binarizada y descartando las regiones con dimensiones inferiores a 18x18 píxeles y superiores a 55x55 (Heurísticamente calculados), por lo que se debe cambiar el valor de estos a cero. Ferreira (2008).

Existen ocasiones, en donde las cabezas de espermatozoides cercanas entre sí, suelen fusionarse en una misma región, produciendo un inconveniente a la hora de realizar la segmentación de las imágenes. Por esta razón, se diseñó un mecanismo capaz de separar las regiones mediante la intensidad del píxel actual y sus vecinos llamada Núcleo de Círculo y Anillo (NCA).

La idea consiste en tener dos regiones. La región A está conformada por un círculo centrado en el píxel de interés y radio R. La región B consiste en un anillo con radio interno R y radio externo S, como se puede apreciar en la **figura 12.**

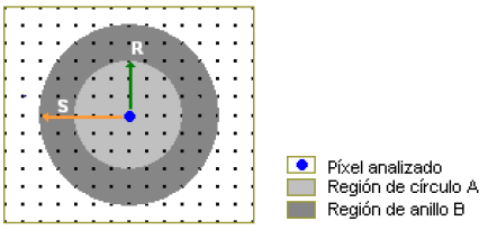



***Figura 12: Núcleo de Círculo y Anillo (NCA), Ferreira (2008)***

Se compara el promedio de las intensidades de gris de la región A, con el promedio de intensidades de gris de la región B, esta segmentación se logra teniendo en cuenta que los espermatozoides son más oscuros que el fondo, por lo que, si el promedio de A es menor al de B, al píxel central se le asigna el valor 1, caso contrario se le asigna valor 0. En la **figura 12** se observa la representación del NCA.

El siguiente paso consiste en construir la Máscara de Núcleo de Círculo y Anillo **(MNCA)** cuyo objetivo es almacenar la estructura de los espermatozoides como un conjunto, incluyendo cabeza, pieza intermedia y cola (en el caso de tenerla). Para poder construir esta matriz se deben realizar dos operaciones. Primero binarizar la imagen utilizando NCA con radio R=3 y S=5, y luego eliminar mediante DFS las regiones cuya dimensión sea inferior a 25x25 píxeles, Ferreira (2008).

Luego se debe construir la Máscara de Contornos de Espermatozoides **(MCOE)** cuya funcionalidad es corregir los errores de la MCE, haciendo uso de la MNCA. Al disponer de la matriz MCOE, es posible remover las regiones cuya área sea menor a 200 píxeles (haciendo uso de DFS), así como también las regiones que no sean intersectadas por la MNCA (correspondientes a manchas u otros elementos).

La efectividad de todas estas funciones (cuyo objetivo es crear estructuras matriciales para delimitar las regiones a procesar), puede verse afectada al utilizar una imagen de mayor o menor resolución, ya que las regiones a descartar podrían presentar una proporción distinta a la establecida por las heurísticas definidas previamente, lo que traería como consecuencia posibles errores de segmentación.

Con el fin de evitar discrepancias en el procesamiento de las imágenes debido a sus dimensiones, se modificaron todas estas funciones para que puedan realizar los mismos procesamientos haciendo uso de las proporciones de la imagen (porcentaje total del tamaño de la imagen). Para calcular las dimensiones de las regiones basadas en la proporción de la imagen, se determinó la relación entre el área total de la imagen utilizada por Ferreira (2008) y las dimensiones de cada región (calculadas heurísticamente) **(Ver fórmula 17).**

$$P_i = \frac{W_i}{A} \quad \textbf{(17)}$$

Sea $P_{i:}$ el factor de proporción de la región $R_i$

Sea $W_i$ el ancho de la región $R_i$

Sea A: el área de la imagen utilizada por Ferreira (2008).

Tomando como ejemplo el cálculo de la Máscara de Cabezas de Espermatozoides (MCE), la cual descarta las regiones con dimensiones inferiores a 18x18 píxeles o superiores a 55x55 píxeles, los valores resultantes serían los siguientes: $R_1$=18x18px y $R_2$= 55x55px. Así, el cálculo del factor de proporción produciría los siguientes resultados:

$$P_1 = \frac{18}{1600 * 1200} = 0.00000937$$

$$P_2 = \frac{55}{1600 * 1200} = 0.00002864583333$$

Una vez obtenido este factor, es posible su uso para el cálculo de las nuevas Regiones de delimitación de acuerdo con las dimensiones de la imagen a procesar **(Fórmula 18).**

$$R_i = A * P_i \quad \textbf{(18)}$$

Por ejemplo, si se procesa una imagen cuya área es el doble de las utilizadas por Ferreira (2008), las regiones obtenidas serían las siguientes:

$$R_1 = ((1600 * 1200)*2)*0.00000937 = 36$$

$$R_2 = ((1600 * 1200)*2)*0.00002864583333 = 110$$

Por lo tanto, al momento de calcular la MCE de la imagen cuya área es el doble de la original, se deberán descartar las regiones con dimensiones inferiores a 36x36 píxeles o superiores a 110x110 píxeles.

Como se mencionó anteriormente, este procedimiento se aplicó en todas las funciones cuyo objetivo era crear estructuras matriciales para realizar la segmentación de imágenes:

- Creación de la Máscara de Cabezas de Espermatozoide (MCE), donde se remueven las regiones con dimensiones inferiores a 18x18 píxeles o superiores a 55x55 píxeles.
- Creación de la Máscara de Núcleo de Círculo y Anillo (MNCA), donde se remueven mediante DFS, las regiones con dimensiones inferiores a 25x25 píxeles.
- Creación de la Máscara de Contornos de Espermatozoides (MCOE), donde se remueven mediante DFS, las regiones con dimensiones inferiores a 200 píxeles.

- **Extracción de características:** Una vez culminado el proceso de segmentación, e identificado cada una de las regiones en las imágenes, es necesario extraer las características de cada una de estas regiones, por lo que se deben seguir los criterios definidos por la OMS a la hora de clasificar morfológicamente cabezas de espermatozoides, Ferreira (2008).

  Se decidió utilizar como variables el área de la región acrosómica, y post-acrosómica, así como el área total de la cabeza. Estos valores fueron cálculados utilizando DFS y permitieron determinar la presencia de:

  - **Defectos de forma**: Estimando que tan ovalada resulta la cabeza del espermatozoide a través del cálculo del tamaño de sus ejes de orientación, o verificando la presencia de anormalidades calculando el perímetro de la cabeza, contando los píxeles del borde y comparándolos con el perímetro de la elipse definida a partir de los ejes.
  - **Defectos de inserción asimétrica de pieza intermedia:** Calculando el ángulo de inserción entre el eje horizontal de la cabeza y la pieza intermedia.
  - **Gota citoplasmática y cuello engrosado:** En este caso, es necesario calcular el ancho máximo de la pieza intermedia.

  Al finalizar esta fase, se obtuvo un conjunto de 13 variables que conformaron el vector de características; algunas de éstas surgieron de la combinación entre otras, en la **Tabla 2** se describe el conjunto completo y en la **Figura 13** se aprecian los aspectos morfológicos evaluados.

| **Parámetro** | **Descripción** |
|---|---|
| **Área cabeza** | Número de píxeles contenidos en la región de la cabeza. |
| **Área acrosoma** | Número de píxeles contenidos en la región acrosómica. |
| **Área post-acrosoma** | Número de píxeles contenidos en la región post-acrosómica. |
| **Eje horizontal cabeza** | Longitud en píxeles del eje horizontal de la cabeza. |
| **Eje vertical cabeza** | Longitud en píxeles del eje vertical de la cabeza. |
| **Diferencia Ei-Cr** | Diferencia en píxeles entre contorno real y contorno de elipse ideal calculada a partir de los ejes de la cabeza. |
| **Elipticidad** | División entre eje horizontal y eje vertical de la cabeza. |
| **Perímetro real** | Número de píxeles que conforman el borde de la cabeza. |
| **Perímetro ideal** | Perímetro elíptico definido por: $p = 2\pi\sqrt{((a^2+b^2)/2)}$. |
| **Diferencia Pi - Pr** | Diferencia entre perímetro elíptico ideal y perímetro real. |
| **Presencia PI** | Indica si la pieza intermedia está o no presente. |
| **Ángulo inserción PI** | Ángulo en grados entre eje Horizontal de la cabeza y eje PI–cabeza. |
| **Ancho máximo PI** | Ancho máximo en píxeles de PI. |

***Tabla 2: Conjunto de características extraídas de la imagen, Ferreira (2008)***

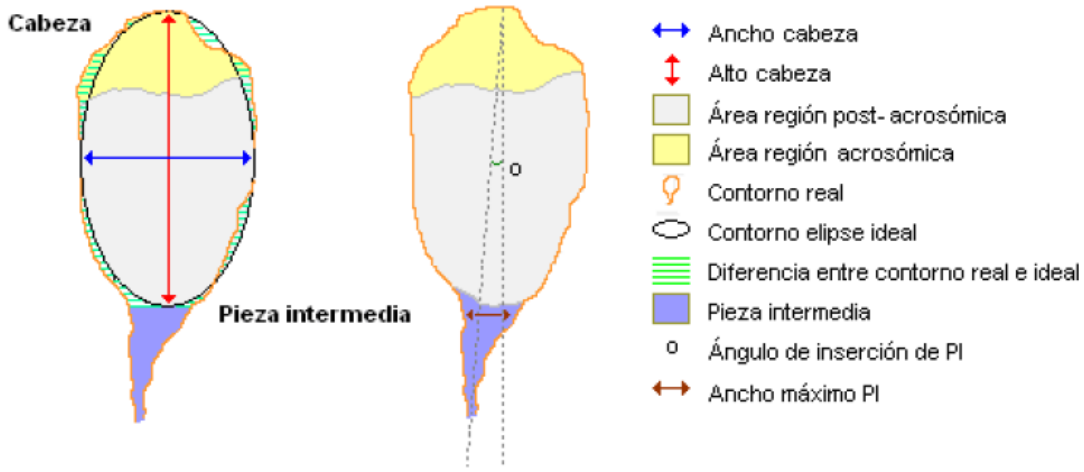



***Figura 13: Aspectos morfológicos evaluados, Ferreira (2008)***

### 2.4. Módulo de clasificación de espermatozoides

Como paradigma de clasificación se utilizó el esquema de combinación de clasificadores (*Ensemble Classifiers)*, aplicando el método *Bagging*. La finalidad de aplicar este paradigma, radica en mejorar los resultados de los clasificadores individuales.

Está técnica constó de dos etapas, la generación de los modelos, en donde se aplicó cada uno de los algoritmos seleccionados, al conjunto de entrenamiento:

- Árboles de decisión
- Reglas de Clasificación
- Redes Neuronales de base radial

La segunda etapa es la de clasificación, la cual emplea cada uno de los modelos generados para predecir la clase de cada una de las instancias del conjunto de datos.

Como se ve en la **Figura 14** una instancia puede pertenecer a distintas clases, pero aplicando un esquema de votación, dicha instancia es asignada a la clase más frecuente. La clasificación de la muestra se puede llevar a cabo mediante tres enfoques diferentes: unanimidad, mayoría simple y pluralidad. En el caso de la unanimidad, todos los clasificadores coinciden en su decisión; la mayoría simple se refiere a la coincidencia del 50% más uno; y finalmente la pluralidad la cual consiste en seleccionar la clase sobre la que coinciden como la correcta el mayor número de clasificadores (Campo, 2009). En este trabajo se implementó la pluralidad como esquema de votación.

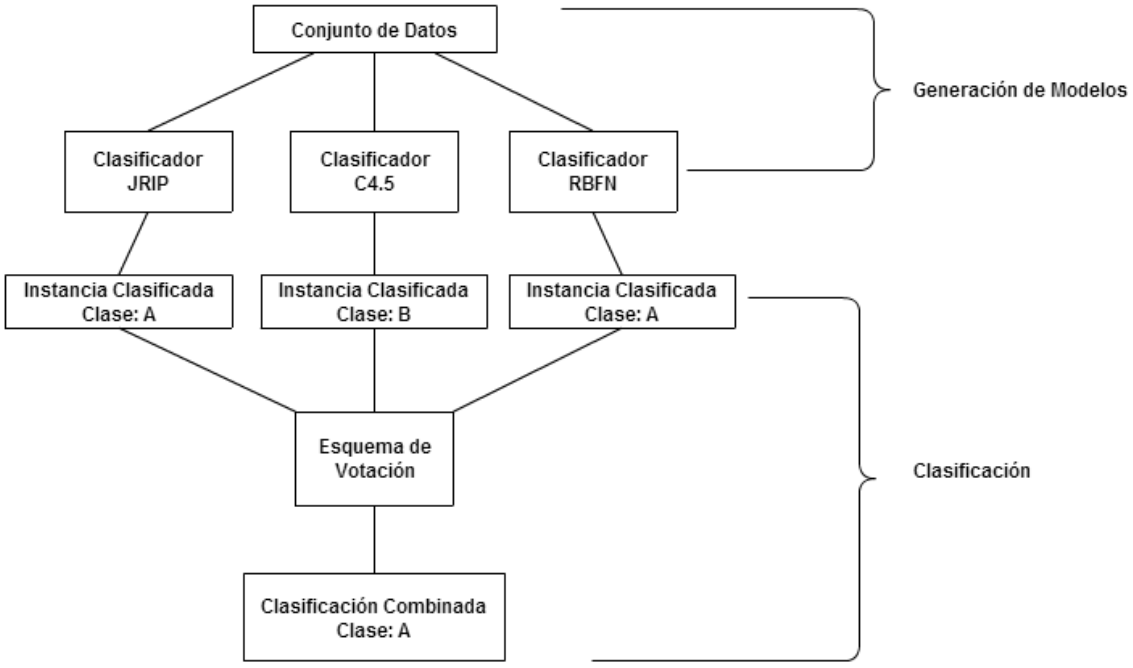



***Figura 14: Modelo de Combinación de Clasificadores***

Los **Anexos A y B** contienen la descripción de cada uno de los modelos obtenidos individualmente con los algoritmos J48 y JRIP respectivamente, recordando que los modelos generados por las redes neuronales no son humanamente entendibles.

### 2.5. Módulo de creación de un nuevo modelo de clasificación

Este módulo permite generar otro modelo de clasificación, utilizando un nuevo conjunto de datos que puede ser obtenido de dos formas:

- Utilizando el módulo de clasificación manual de las muestras, en donde el experto en base a su conocimiento determina la clase de cada uno de los espermatozoides presentes en la muestra, generando así vectores de clasificación (siguiendo el mismo lineamiento que Ferreira (2008) planteo en su aplicación). Estos vectores son almacenados en un archivo .arff (formato utilizado WEKA *Waikato Environment for Knowledge Analysis*, **ver anexo D**), listo para ser utilizado por los algoritmos de clasificación y generar un nuevo modelo de Clasificación, que puede ser utilizado en el módulo de clasificación automática.

- En este caso el experto ya tiene preparado los vectores de clasificación ordenados en un archivo .arff **(ver anexo D)**, los cuales pueden ser cargados directamente a la aplicación.

## 2.6. Estimación del modelo de clasificación

En esta sección se describe detalladamente, el proceso de descubrimiento de conocimiento a partir de datos (KDD) aplicado en esta investigación. Es importante mencionar que la calidad de las salidas obtenidas en cada fase influyó significativamente en los resultados de las fases posteriores.

- **Fase de generación, recopilación e integración de datos:** Para generar el conjunto de datos se procesaron 12 muestras de semen (ver figura 15). Por cada muestra se obtuvo una cantidad de imágenes distinta, ya que el número de imágenes depende directamente de la concentración espermática de la muestra. Es decir, si la concentración es baja la cantidad de imágenes es mayor, caso contrario si la concentración es alta la cantidad de imágenes disminuye. Esto es debido a los lineamientos definidos por la OMS (2001), los cuales especifican la cantidad total espermatozoides que debe ser analizada en una muestra, la cual no debe superar los 200.

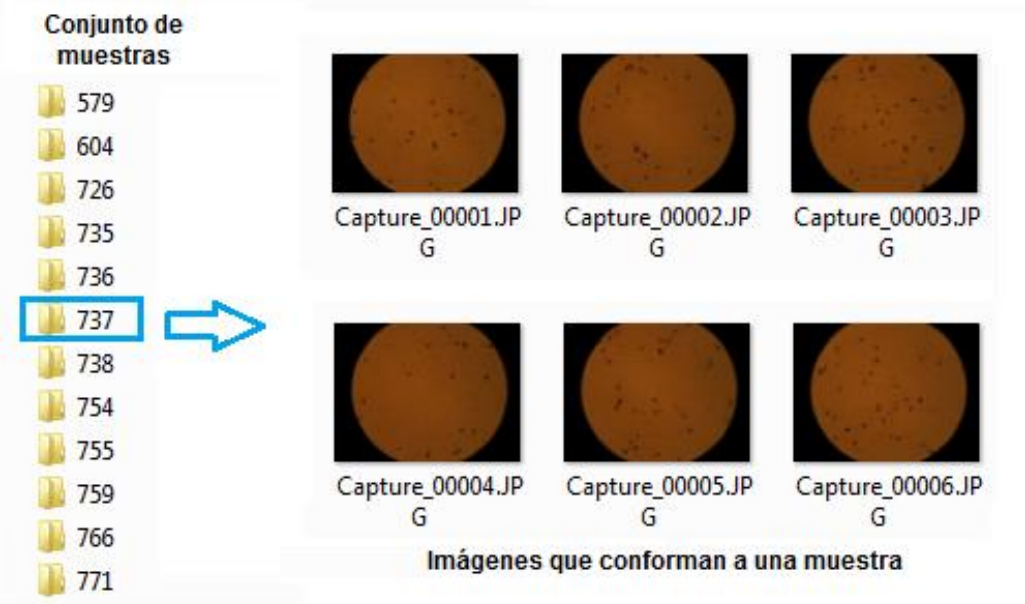



***Figura 15: Conjunto de imágenes que conforman a una muestra***

El siguiente paso fue utilizar la aplicación creada por Ferreira (2008), la cual permitió segmentar y extraer las características de las imágenes capturadas. Haciendo uso de esta herramienta y con la participación de expertos en la realización de este examen, se generó un conjunto de datos (También conocido como tabla-atributo-valor) compuesto de 868 vectores de características. El vector de característica está compuesto por 17 variables, descritas en la **Tabla 3.**

| Nombre | Tipo de dato |
|---|---|
| Área cabeza | Numérico |
| Área acrosoma | Numérico |
| Área post-acrosoma | Numérico |
| Eje horizontal cabeza | Numérico |
| Eje vertical cabeza | Numérico |
| Diferencia Ei-Cr | Numérico |
| Elipticidad | Numérico |
| Perímetro real | Numérico |
| Perímetro ideal | Numérico |
| Diferencia Pi - Pr | Numérico |
| Presencia PI | Numérico |
| Ángulo inserción PI | Numérico |
| Ancho máximo PI | Numérico |
| Tipo de cabeza | Nominal |
| Tipo de cuello | Nominal |
| Tipo de cola | Nominal |
| Clasificación | Nominal |

***Tabla 3: Variables que conforman el vector de características***

- **Fase de preparación de los datos (Limpieza, Transformación y Selección):** Una vez obtenidas y procesadas las imágenes en una tabla atributo valor, se procedió a realizar la fase de procesamiento de los datos para obtener la vista minable.
    - **Limpieza:** Mediante inspección visual, realizada por los expertos, se detectaron errores de segmentación en algunos registros, así como errores en el cálculo de algunas características, principalmente a la hora de calcular los ejes horizontales y verticales. Por otro lado, analizando los valores presentados en los histogramas de cada una de las variables, se localizaron valores extremos, los cuales fueron eliminados. Entonces el conjunto de datos fue reducido a 832 registros de los cuales 154 corresponden a cabezas normales y 678 a anormales.

    - **Transformaciones:** No fue necesario la modificación de la forma de los datos, es decir, cambiar el rango o tipo de dato de un atributo o un conjunto de atributos.

- **Creación de variables:** se generó una nueva variable llamada *Cociente_Acrosoma_Atotal,* el cual representa el cociente entre los *atributos Área Acrosoma* y *Área Cabeza*. Ya que la OMS (2001) especifica que el acrosoma de un espermatozoide normal debe ocupar entre el 40% y 70% del área de la cabeza.

- **Selección de variables:** Los atributos (Tipo de cuello, Tipo de cola, Clasificación) fueron eliminados previamente, ya que no son variables necesarias para la clasificación morfológica de cabezas. La variable Tipo de cabeza pasó a ser la variable Clasificación, dejando un total de 14 atributos presentes en la muestra. A este conjunto de datos se le aplicó diferentes criterios de selección de variables, tal como se muestra en la **tabla 4**

| Método de Búsqueda | Evaluador de Atributito | Atributos seleccionados |
|---|---|---|
| Ranking | Chi cuadrado | 4,2,3,7,5,6,1,8 |
| Ranking | GainRatio | 7,5,4,2,3,1,6,8 |
| Ranking | Ganancia de información | 4,2,7,3,5,1,6,8 |

***Tabla 4: Criterio de selección de variables***

El orden de selección de los atributos generados por los filtros, determina la importancia de los mismos, por lo que la variable número 8 (*Diferencia Ei – Cr*) puede ser descartada, al ser seleccionada como la última por todos los métodos de búsqueda.

Para validar que las variables descartadas no produjeron un cambio muy significativo se compara el porcentaje de instancias clasificadas correctamente tanto antes y después de la selección de variables (ver fase de Evaluación e Interpretación).

- **Generación de nuevos datos:** El conjunto de datos utilizado se encontraba desbalanceado ya que contenía más variables pertenecientes a la clase “Anormal”, en comparación con la clase “Normal”, lo que podía dificultar la generación de un buen modelo de clasificación. Es por esto que se utilizó el algoritmo SMOTE, aplicando sobre muestreo (*over-sampling*) de la clase minoritaria, con el fin de nivelar la proporción sobre los datos existentes. En la **Figura 16** se describe este algoritmo, el cual permitió generar cuatro vistas minables para el análisis de los resultados:

    - Una con los datos originales, 832 registros
    - Tres con la clase minoritaria aumentada en proporciones del (50%,100%, y 150%)

**Algoritmo_SMOTE (D: Conjunto de datos, P: proporción de datos a aumentar, K: número de vecinos):**

K_V_C []=**calcularVecinos(**Ejemplo_Clase_minoritaria**)** // se almacenan los K vecinos de la instancia perteneciente a la clase minoritaria
D_V_K []=**CaclularDiferencia(**D, K_V_C**)** // Se calcula diferencia entre el valor de muestra y el vecino
Elemento**=GenerarElemento(**D_V_K, Ejemplo_Clase_minoritaria**)** // Se genera un nuevo elemento, al sumar (la multiplicación de la diferencia con un aleatorio entre 0 y 1) con el valor original de la muestra

**Fin _ Algoritmo_SMOTE**

***Figura 16: Algoritmo SMOTE (Chawla et al., 2002)***

- **Fase de minería de datos:** En esta fase se seleccionaron las técnicas de minería de datos, de acuerdo a los objetivos específicos planteados para poder estimar el modelo de clasificación de la morfología de cabezas de espermatozoides humanos. Es importante mencionar que los resultados de la investigación realizada por Casañas, Ramos & Núñez (2011) fueron tomados en cuenta para la selección de las técnicas y algoritmos. En la **Tabla 5**, se especifican los algoritmos utilizados.

| Técnica de minería de datos | Algoritmo |
|---|---|
| Árboles de decisión | J48 (Implementación del algoritmo C4.5 en WEKA ) |
| Reglas de Clasificación | JRIP (Implementación del algoritmo RIPPER en WEKA) |
| Redes Neuronales | RBFNetwork (Implementación del algoritmo RBFN en WEKA) |

***Tabla 5: Técnicas y algoritmos empleados***

  - En esta fase se utilizaron los algoritmos que provee la plataforma para el aprendizaje automático y la minería de datos WEKA (Hall, Frank, Holmes & others, 2009). Esta plataforma es de software libre distribuido bajo la licencia GNU-GPL.

  - Como paradigma de clasificación se utilizó el esquema de combinación de clasificadores (*Ensemble Classifiers)*, explicado en la Sección 2.4.

- **Fase de Evaluación e Interpretación del modelo de clasificación:** En esta fase se utilizó como técnica de evaluación la validación cruzada de 10 particiones (*Cross Validation*). Mientras que las medidas de evaluación seleccionadas fueron la exactitud predictiva **(fórmula 11)** y la media geométrica **(fórmula 16).**

En la **tabla 6 y 7** se presentan los resultados obtenidos con cada una de las vistas minables utilizadas en las fases previas. La diferencia entre cada tabla radica en la variable eliminada en la fase de selección. La tabla 7 muestra los resultados antes de la eliminación de variables, mientras que la tabla 8 refleja los resultados después de eliminar la variable *Diferencia Ei – Cr*.

| Proporción de datos aumentada | RBFNetwork | | JRIP | | J48 | | Combinación de clasificadores | |
|---|---|---|---|---|---|---|---|---|
| | Exactitud | G-media | Exactitud | G-media | Exactitud | G-media | Exactitud | G-media |
| 0% | 82.932 | 68.365 | 81.850 | 64.404 | 81.009 | 75.644 | 82.692 | 71.111 |
| 50% | 81.848 | 81.258 | 80.638 | 75.453 | 80.638 | 78.649 | 82.178 | 79.626 |
| 100% | 82.657 | 83.625 | 80.020 | 78.058 | 80.020 | 81.468 | 81.440 | 82.140 |
| 150% | 82.878 | 83.944 | 80.244 | 79.283 | 82.690 | 83.104 | 83.72 | 84.495 |

***Tabla 6: Resultados obtenidos previo a la selección de variables***

| Proporción de datos aumentada | RBFNetwork | | JRIP | | J48 | | Combinación de clasificadores | |
|---|---|---|---|---|---|---|---|---|
| | Exactitud | G-media | Exactitud | G-media | Exactitud | G-media | Exactitud | G-media |
| 0% | 83.293 | 72.094 | 82.211 | 61.646 | 81.129 | 75.710 | 82.692 | 70.765 |
| 50% | 81.738 | 79.665 | 80.198 | 73.720 | 80.858 | 79.865 | 82.838 | 80.682 |
| 100% | 82.555 | 82.967 | 79.310 | 76.294 | 81.135 | 81.407 | 82.251 | 82.310 |
| 150% | 83.066 | 83.910 | 81.467 | 80.331 | 83.819 | 85.083 | 84.477 | 85.398 |

***Tabla 7: Resultados obtenidos después de la selección de variables***

- Los valores arrojados por la medida de Exactitud y la medida Geométrica reflejan que el modelo obtenido con la técnica de combinación de clasificadores es el mejor en comparación con los modelos obtenidos de forma individual, siempre que la proporción este aumentada al 150%. Sin embargo, luego de hacer la selección de variables, la tabla 8 refleja que en la mayoría de los casos, la combinación de clasificadores es mejor que los modelos individuales.

- El incremento de la media Geométrica está sujeto al aumento de la proporción de la clase minoritaria (“Normal”), debido a la reducción del desbalanceo de las clases, implicando una mejora en la clasificación de los datos.

- Los porcentajes de clasificación reflejados en las Tabla 6 y 7 tienen una variación, no muy significativa después de la selección de las variables. Sin embargo, descartar la variable *Diferencia Ei – Cr* tuvo repercusión en el proceso, al darle mayor ventaja a la combinación de clasificadores.

- Los resultados permiten concluir que el conjunto de datos cuya proporción fue aumentada al 150% y se le aplico selección de variables, es el más propicio para obtener mejores resultados al momento der clasificar una instancia.

### 2.7. Desarrollo de la aplicación

A continuación, se describen los casos de usos, la interfaz de usuario y posteriormente las tecnologías utilizadas.

**2.7.1 Casos de uso:** En esta sección se detallan los casos de usos básicos definidos para satisfacer los requerimientos de la aplicación.

**Nivel 0**

- **Actores:** (Usuarios) Bioanalistas encargados de realizar la clasificación morfológica de cabeza de espermatozoide humano.
- **Casos de uso:**

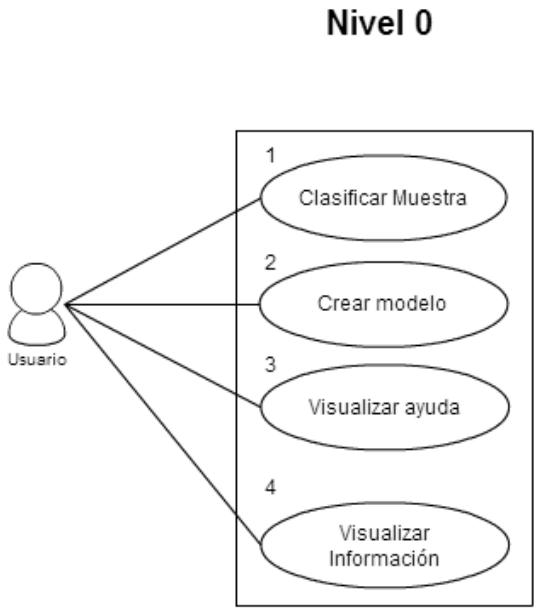



***Figura 17: Modelo de casos de uso Nivel 0 del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Clasificar Muestra |
| **ID** | 1 |
| **Pre-Condición** | Tener lista el grupo de imágenes que representa la muestra. |
| **Descripción:** | Esta actividad permite a los usuarios cargar un conjunto de imágenes para poder realizar la clasificación espermática de cabeza. |
| **Post-Condición** | Se despliega la pantalla de clasificación con distintas opciones para el usuario. |

| | |
|---|---|
| **Nombre** | Crear modelo |
| **ID** | 2 |
| **Pre-Condición** | Tener lista el grupo de imágenes o conjunto de datos (previamente clasificado) necesarios para la generación de un nuevo modelo de datos. |
| **Descripción:** | Esta actividad permite a los usuarios crear un nuevo modelo de datos a partir de dos opciones, cargar un conjunto de datos previamente clasificado, o cargar un conjunto de imágenes que serán clasificadas por el usuario. |
| **Post-Condición** | Se despliega una pantalla de acuerdo a la opción seleccionada por el usuario. |

| | |
|---|---|
| **Nombre** | Visualizar ayuda |
| **ID** | 3 |
| **Pre-Condición** | - |
| **Descripción:** | Esta actividad le muestra al usuario una guía básica, explicando las distintas funcionalidades de la aplicación |
| **Post-Condición** | Se despliega una pantalla de ayuda al usuario. |

| | |
|---|---|
| **Nombre** | Visualizar Información del desarrollador |
| **ID** | 4 |
| **Pre-Condición** | - |
| **Descripción:** | Esta actividad le muestra al usuario una pantalla de información acerca el desarrollador y el Tutor. |
| **Post-Condición** | Se despliega una pantalla de información al usuario. |

**Nivel 1**

- **Actores:** (Usuarios) Bioanalistas encargados de realizar la clasificación morfológica de cabeza de espermatozoide humano.
- **Casos de uso:**

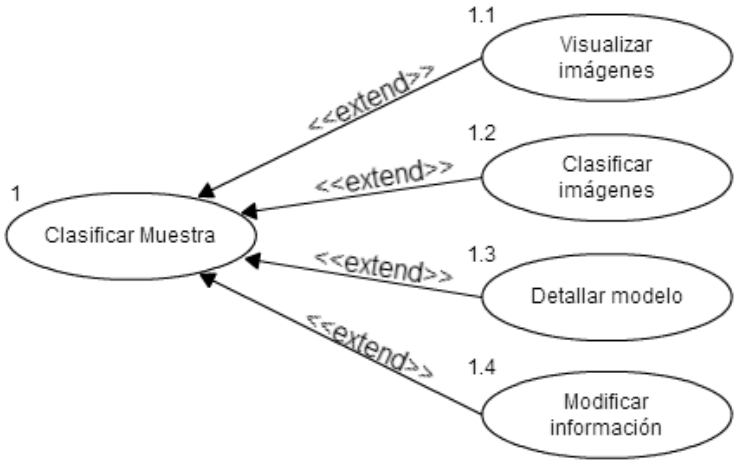



***Figura 18: Modelo de casos de uso Nivel 1 (ID: 1) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Visualizar imágenes |
| **ID** | 1.1 |
| **Pre-Condición** | Haber seleccionado un conjunto de imágenes a clasificar |
| **Descripción:** | Esta actividad permite a los usuarios visualizar todas las imágenes que conforman la muestra. |
| **Post-Condición** | Se identifican los espermatozoides reconocidos con un id numérico ascendente que inicia en 0. En el panel lateral se carga la imagen y características del primer espermatozoide reconocido. |

| | |
|---|---|
| **Nombre** | Clasificar imágenes |
| **ID** | 1.2 |
| **Pre-Condición** | Haber seleccionado un conjunto de imágenes a clasificar. |
| **Descripción:** | Esta actividad permite a los usuarios visualizar información detallada al momento de realizar la clasificación |
| **Post-Condición** | Se muestra en la interfaz una vista en miniatura del conjunto de imágenes |

| | |
|---|---|
| **Nombre** | Detallar modelo |
| **ID** | 1.3 |
| **Pre-Condición** | - |
| **Descripción:** | Esta actividad permite a los usuarios visualizar toda la información referente al modelo de clasificación utilizado para clasificar la muestra. |
| **Post-Condición** | Se muestra en la interfaz información referente a las variables utilizadas en el proceso de clasificación. |

| | |
|---|---|
| **Nombre** | Modificar información |
| **ID** | 1.4 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede modificar los datos que identifican la muestra (nombre, fecha, edad del paciente, nombre Bioanalista) |
| **Post-Condición** | Se muestra en la interfaz, los cambios realizados. |

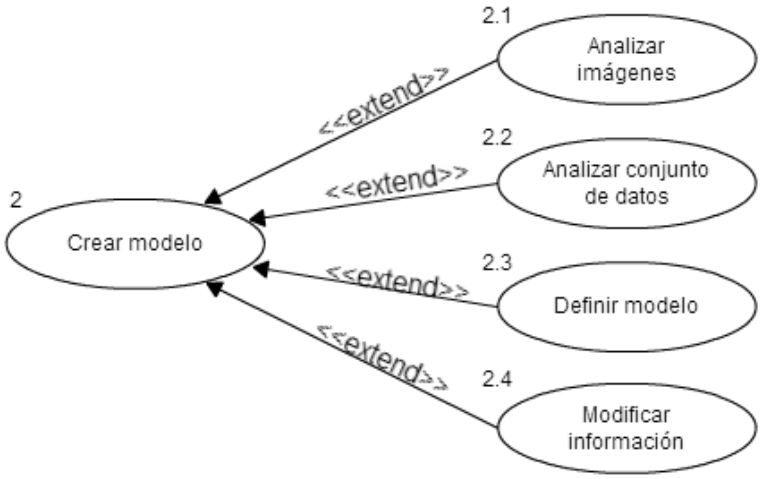



***Figura 19: Modelo de casos de uso Nivel 1 (ID: 2) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Analizar imágenes |
| **ID** | 2.1 |
| **Pre-Condición** | Haber seleccionado un conjunto de imágenes para generar un nuevo modelo de datos. |
| **Descripción:** | Esta actividad permite a los usuarios analizar todas las imágenes que conforman la muestra. |
| **Post-Condición** | Se identifican los espermatozoides reconocidos con un id numérico ascendente que inicia en 0. En el panel lateral se carga la imagen y características del primer espermatozoide reconocido. |

| | |
|---|---|
| **Nombre** | Analizar conjunto de datos |
| **ID** | 2.2 |
| **Pre-Condición** | Haber seleccionado un conjunto de datos (previamente clasificado) para generar un nuevo modelo de clasificación. |
| **Descripción:** | Esta actividad permite a los usuarios analizar el conjunto de datos |
| **Post-Condición** | Se muestra una tabla atributo valor, que el usuario puede editar de acuerdo a sus necesidades |

| | |
|---|---|
| **Nombre** | Definir modelo |
| **ID** | 2.3 |
| **Pre-Condición** | Haber generado el modelo de clasificación |
| **Descripción:** | Esta actividad permite a los usuarios visualizar toda la información referente al modelo de clasificación utilizado para clasificar la muestra. |
| **Post-Condición** | Se muestra en la interfaz información referente a las variables utilizadas en el proceso de clasificación. |

| | |
|---|---|
| **Nombre** | Modificar información |
| **ID** | 2.4 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede modificar los datos que identifican el conjunto de datos (nombre, fecha, edad del paciente, nombre Bioanalista) |
| **Post-Condición** | Se muestra en la interfaz, los cambios realizados. |

**Nivel 2**

- **Actores:** (Usuarios) Bioanalistas encargados de realizar la clasificación morfológica de cabeza de espermatozoide humano.
- **Casos de uso:**

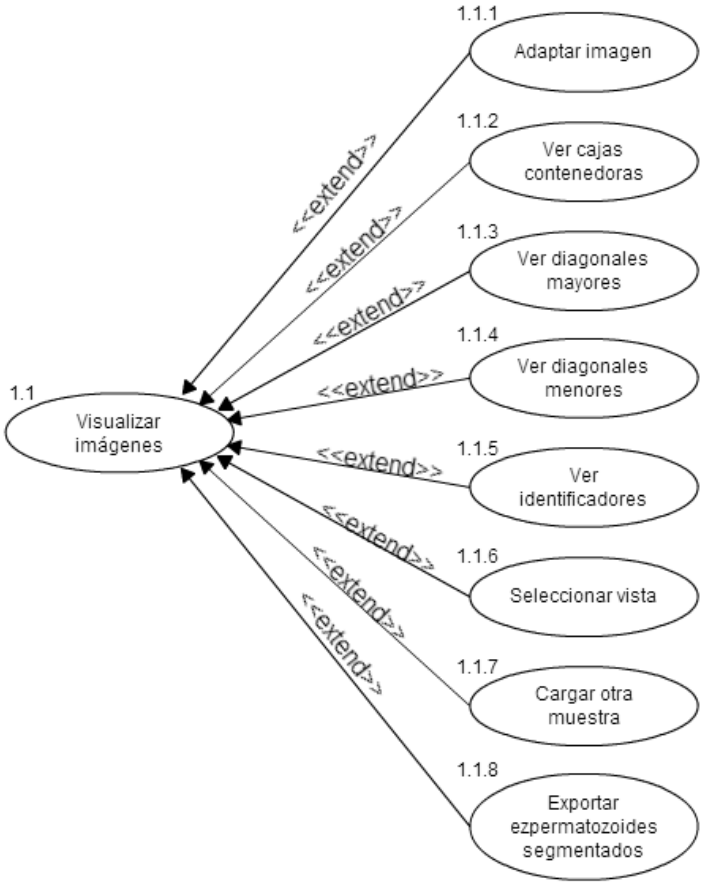



***Figura 20: Modelo de casos de uso Nivel 2 (ID: 1.1) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Adaptar imagen |
| **ID** | 1.1.1 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede amplificar o estabilizar el tamaño de la imagen de acuerdo a sus necesidades |
| **Post-Condición** | Se muestra en la interfaz, la imagen procesa amplificada |

| | |
|---|---|
| **Nombre** | Ver cajas contenedoras |
| **ID** | 1.1.2 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar las cajas contenedoras orientadas de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, las cajas contendoras. |

| | |
|---|---|
| **Nombre** | Ver diagonales mayores |
| **ID** | 1.1.3 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar las diagonales mayores de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, las diagonales mayores. |

| | |
|---|---|
| **Nombre** | Ver diagonales menores |
| **ID** | 1.1.4 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar las diagonales menores de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, las diagonales menores. |

| | |
|---|---|
| **Nombre** | Ver identificadores |
| **ID** | 1.1.5 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar los identificadores de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, los identificadores. |

| | |
|---|---|
| **Nombre** | Seleccionar vista |
| **ID** | 1.1.6 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario dispone de varias formas de visualizar los resultados (RGB, canal rojo, canal azul, canal verde, segmentada, mascara de contornos) |
| **Post-Condición** | Se muestra en la interfaz, la opción selecciona por el usuario. |

| | |
|---|---|
| **Nombre** | Cargar otra muestra |
| **ID** | 1.1.7 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede seleccionar otra muestra para clasificar si la que eligió previamente no es de su agrado. |
| **Post-Condición** | Se muestra en la interfaz, la opción selecciona por el usuario. |

| | |
|---|---|
| **Nombre** | Exportar espermatozoides segmentados |
| **ID** | 1.1.8 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede exportar como imágenes individuales los espermatozoides reconocidos por el módulo de segmentación. |
| **Post-Condición** | Se muestra en la interfaz, la opción selecciona por el usuario. |

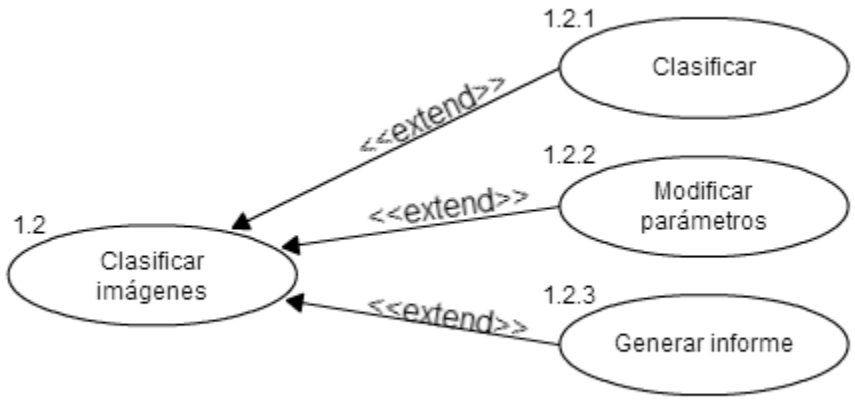



***Figura 21: Modelo de casos de uso Nivel 2 (ID: 1.2) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Clasificar |
| **ID** | 1.2.1 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede iniciar el proceso de clasificación de la muestra cuando desee. |
| **Post-Condición** | Se muestra en la interfaz, los resultados al nivel general obtenidos al clasificar la muestra. |

| | |
|---|---|
| **Nombre** | Modificar parámetros |
| **ID** | 1.2.2 |
| **Pre-Condición** | Haber clasificado la muestra |
| **Descripción:** | El usuario puede modificar el porcentaje normal morfológico utilizado como referencia. |
| **Post-Condición** | Se muestra en la interfaz, la nueva clasificación de la muestra en base al nuevo porcentaje |

| | |
|---|---|
| **Nombre** | Generar informe |
| **ID** | 1.2.3 |
| **Pre-Condición** | Haber clasificado la muestra |
| **Descripción:** | Permite descargar un informe con toda la información referente a la clasificación de la muestra. |
| **Post-Condición** | Se genera un informe en formato PDF |

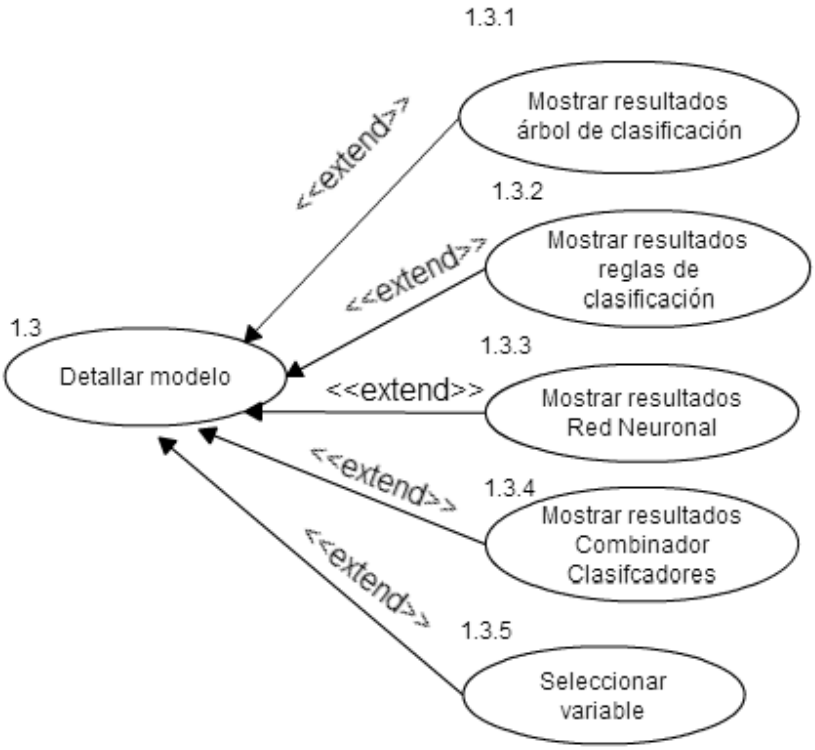



***Figura 22: Modelo de casos de uso Nivel 2 (ID: 1.3) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Mostrar resultados árbol de clasificación |
| **ID** | 1.3.1 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede visualizar el árbol de clasificación y los resultados de la evaluación realizada al modelo |
| **Post-Condición** | Se muestra en la interfaz, el árbol de clasificación y los resultados |

| | |
|---|---|
| **Nombre** | Mostrar resultados reglas de clasificación |
| **ID** | 1.3.2 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede visualizar el conjunto de reglas de clasificación y los resultados de la evaluación realizada al modelo |
| **Post-Condición** | Se muestra en la interfaz, el conjunto de reglas y los resultados |

| | |
|---|---|
| **Nombre** | Mostrar resultados Red Neuronal |
| **ID** | 1.3.3 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede visualizar los resultados de la evaluación realizada al modelo. |
| **Post-Condición** | Se muestra en la interfaz, los resultados de la evaluación realizada al modelo. |

| | |
|---|---|
| **Nombre** | Mostrar resultados combinación de clasificadores |
| **ID** | 1.3.4 |
| **Pre-Condición** | -- |
| **Descripción:** | El usuario puede visualizar los resultados de la evaluación realizada al modelo. |
| **Post-Condición** | Se muestra en la interfaz, los resultados de la evaluación realizada al modelo. |

| | |
|---|---|
| **Nombre** | Seleccionar variable |
| **ID** | 1.3.5 |
| **Pre-Condición** | -- |
| **Descripción:** | El usuario puede seleccionar cualquiera de las variables que conforma el conjunto de datos para obtener información más detallada de cada una. |
| **Post-Condición** | Se muestra en la interfaz, la información de la variable seleccionada |

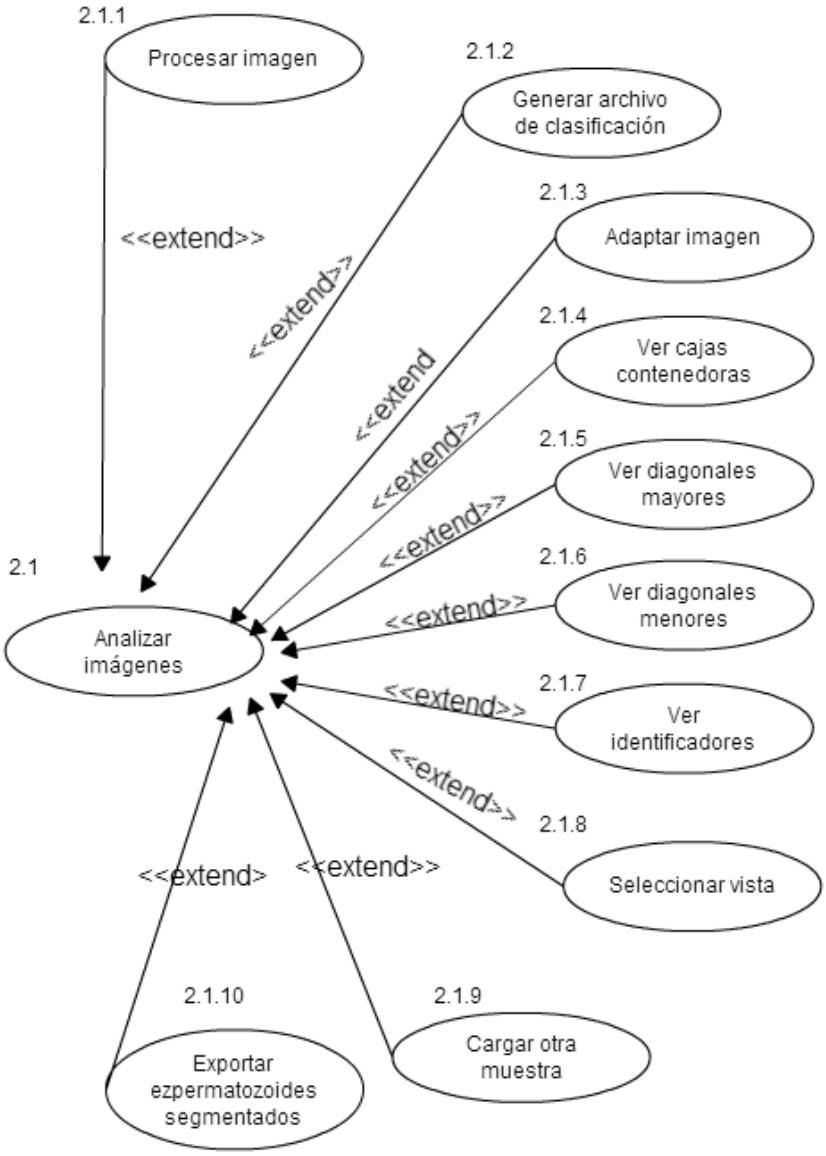



***Figura 23: Modelo de casos de uso Nivel 2 (ID: 2.1) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Procesar imagen |
| **ID** | 2.1.1 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede iniciar la secuencia de procesamiento que se ocupa de la segmentación y extracción de características de los espermatozoides |
| **Post-Condición** | Se identifican los espermatozoides reconocidos con un id numérico ascendente que inicia en 0. En el panel lateral se carga la clasificación del primer espermatozoide reconocido, así como las etiquetas que le hayan sido asignadas en procesamientos previos |

| | |
|---|---|
| **Nombre** | Generar archivo de clasificación |
| **ID** | 2.1.2 |
| **Pre-Condición** | Haber procesado todas las imágenes de la muestra |
| **Descripción:** | El usuario puede especificar al sistema el momento en el que desee generar el archivo de clasificación |
| **Post-Condición** | El sistema escribe un archivo de clasificación llamado 'espermatozoides.txt' |

| | |
|---|---|
| **Nombre** | Adaptar imagen |
| **ID** | 2.1.3 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede amplificar o estabilizar el tamaño de la imagen de acuerdo a sus necesidades |
| **Post-Condición** | Se muestra en la interfaz, la imagen procesa amplificada |

| | |
|---|---|
| **Nombre** | Ver cajas contenedoras |
| **ID** | 2.1.4 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar las cajas contenedoras orientadas de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, las cajas contendoras. |

| | |
|---|---|
| **Nombre** | Ver diagonales mayores |
| **ID** | 2.1.5 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar las diagonales mayores de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, las diagonales mayores. |

| | |
|---|---|
| **Nombre** | Ver diagonales menores |
| **ID** | 2.1.6 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar las diagonales menores de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, las diagonales menores. |

| | |
|---|---|
| **Nombre** | Ver identificadores |
| **ID** | 2.1.7 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede visualizar los identificadores de los espermatozoides reconocidos. |
| **Post-Condición** | Se muestra en la interfaz, los identificadores. |

| | |
|---|---|
| **Nombre** | Seleccionar vista |
| **ID** | 2.1.8 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario dispone de varias formas de visualizar los resultados (RGB, canal rojo, canal azul, canal verde, segmentada, mascara de contornos) |
| **Post-Condición** | Se muestra en la interfaz, la opción selecciona por el usuario. |

| | |
|---|---|
| **Nombre** | Cargar otra muestra |
| **ID** | 2.1.9 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede seleccionar otra muestra para clasificar si la que eligió previamente no es de su agrado. |
| **Post-Condición** | Se muestra en la interfaz, la opción selecciona por el usuario. |

| | |
|---|---|
| **Nombre** | Exportar espermatozoides segmentados |
| **ID** | 2.1.10 |
| **Pre-Condición** | Cargar un conjunto de imágenes |
| **Descripción:** | El usuario puede exportar como imágenes individuales los espermatozoides reconocidos por el módulo de segmentación. |
| **Post-Condición** | Se muestra en la interfaz, la opción selecciona por el usuario. |

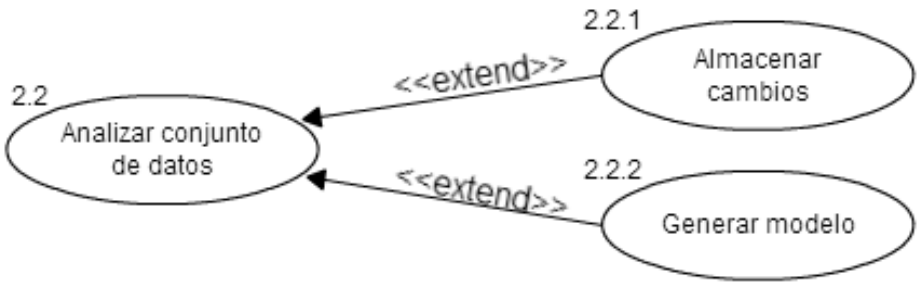



***Figura 24: Modelo de casos de uso Nivel 2 (ID: 2.2) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Almacenar cambios |
| **ID** | 2.2.1 |
| **Pre-Condición** | Cargar un conjunto de datos previamente clasificado |
| **Descripción:** | El usuario luego de modificar los valores del conjunto de datos cargado puede almacenar los cambios realizados. |
| **Post-Condición** | Se muestra en la interfaz los cambios realizados. |

| | |
|---|---|
| **Nombre** | Generar modelo |
| **ID** | 2.2.2 |
| **Pre-Condición** | Cargar un conjunto de datos previamente clasificado |
| **Descripción:** | El usuario puede generar cuando desee el nuevo modelo de clasificación |
| **Post-Condición** | Se genera el nuevo modelo de datos. |

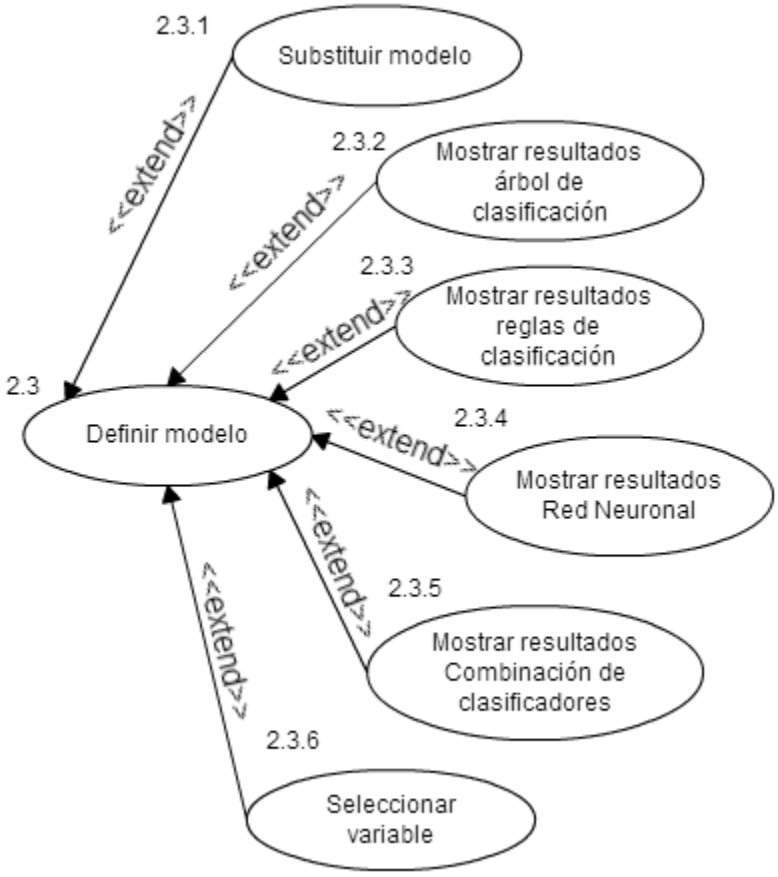



***Figura 25: Modelo de casos de uso Nivel 2 (ID: 2.3) del clasificador morfológico de cabezas***

| | |
|---|---|
| **Nombre** | Sustituir modelo |
| **ID** | 2.3.1 |
| **Pre-Condición** | Generar un modelo de clasificación |
| **Descripción:** | El usuario puede sustituir el nuevo modelo de clasificación en el sistema. |
| **Post-Condición** | Se sustituye el nuevo modelo de datos en el sistema. |

| | |
|---|---|
| **Nombre** | Mostrar resultados árbol de clasificación |
| **ID** | 2.3.2 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede visualizar el árbol de clasificación y los resultados de la evaluación realizada al modelo |
| **Post-Condición** | Se muestra en la interfaz, el árbol de clasificación y los resultados |

| | |
|---|---|
| **Nombre** | Mostrar resultados reglas de clasificación |
| **ID** | 2.3.3 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede visualizar el conjunto de reglas de clasificación y los resultados de la evaluación realizada al modelo |
| **Post-Condición** | Se muestra en la interfaz, el conjunto de reglas y los resultados |

| | |
|---|---|
| **Nombre** | Mostrar resultados Red Neuronal |
| **ID** | 2.3.4 |
| **Pre-Condición** | - |
| **Descripción:** | El usuario puede visualizar los resultados de la evaluación realizada al modelo. |
| **Post-Condición** | Se muestra en la interfaz, los resultados de la evaluación realizada al modelo. |

| | |
|---|---|
| **Nombre** | Mostrar resultados combinación de clasificadores |
| **ID** | 2.3.5 |
| **Pre-Condición** | -- |
| **Descripción:** | El usuario puede visualizar los resultados de la evaluación realizada al modelo. |
| **Post-Condición** | Se muestra en la interfaz, los resultados de la evaluación realizada al modelo. |

| | |
|---|---|
| **Nombre** | Seleccionar variable |
| **ID** | 2.36 |
| **Pre-Condición** | -- |
| **Descripción:** | El usuario puede seleccionar cualquiera de las variables que conforma el conjunto de datos para obtener información más detallada de cada una. |
| **Post-Condición** | Se muestra en la interfaz, la información de la variable seleccionada |

### 2.7.2 Tecnologías utilizadas

- Especificaciones de Hardware
  - Procesador: Intel® Pentium® CPU 2.60GHz
  - 2.00 GB de Memoria RAM
  - Disco Duro de 250 GB

- Sistema Operativo: la aplicación se desarrolló bajo ambiente Microsoft® Windows 7 Professional.

- Herramientas de Software

  - Kit de Desarrollo Java (JDK): Java Platform (JDK) 8u25
  - Entorno de desarrollo integrado: NetBeans IDE 8.0
  - Weka Jar: versión 3.6

**2.7.3. Diseño de la interfaz:** La aplicación resultante se denominó CMCE- Clasificador Morfológico de Cabezas de espermatozoides. La aplicación sigue los mismos lineamientos definidos por la aplicación creada por Ferreira (2008), una ventana donde se despliegue la muestra y todos los controles necesarios para llevar a cabo la clasificación. Sin embargo, esta nueva aplicación presenta varias funcionalidades extras, las cuales se explicarán a continuación:

Al iniciar la aplicación **(Ver Figura 26),** el usuario dispone de dos opciones principales, la clasificación automática de una muestra de espermatozoides, o la creación de un nuevo modelo de datos para futuras clasificaciones.


***Figura 26: Pantalla inicial de la aplicación***

Al seleccionar la opción "**Clasificar Muestra**" se le desplegara al usuario, una pequeña pantalla **(Ver figura 27)** que le indicara los pasos a seguir.

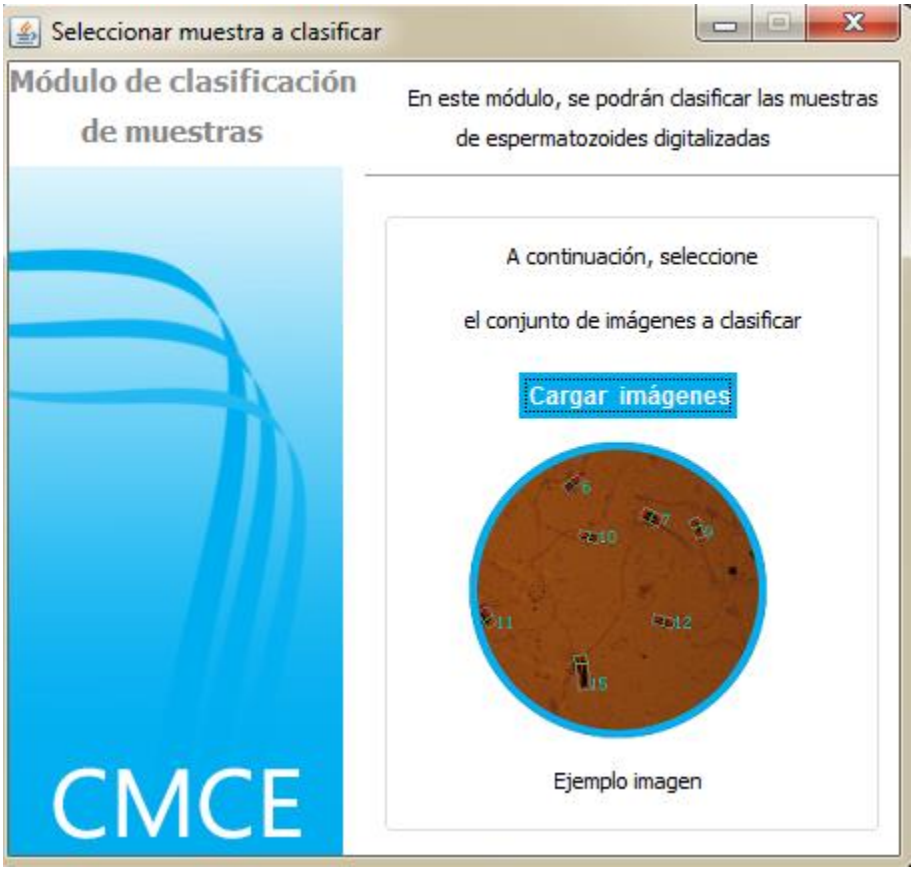



***Figura 27: Pantalla de indicaciones de la primera opción***

En este caso se le específica al usuario, que debe seleccionar la carpeta en donde se encuentra la muestra a clasificar. Luego de esto, se le despliega la pantalla principal, la cual tiene tres pestañas, cada una con un propósito distinto. En la **Figura 28** se muestra la primera pestaña, cuyo objetivo principal es permitir al usuario visualizar la muestra.

Entre las opciones que ofrece la pestaña de visualización se encuentran:

- Datos de la muestra: Permite identificar la muestra a través de un nombre o identificador para la misma, la fecha de clasificación, el nombre del paciente y el nombre del Bionalista que supervisa la clasificación.
- Opciones de visualización: Permite al usuario controlar distintas opciones que son dibujadas en las imágenes, tales como:
    1) Imágenes Ajustadas: Aumentar o disminuir el tamaño de la imagen actual
    2) Cajas Contenedoras: Dibuja las Cajas que delimitan a los espermatozoides
    3) Diagonales Mayores: Dibuja la diagonal mayor en cada uno de los espermatozoides
    4) Diagonales Menores: Dibuja la diagonal menor en cada uno de los espermatozoides
    5) ID de Elementos: Dibuja el identificador numérico de cada uno de los espermatozoides
- El cuadro de Características permite al usuario visualizar los valores extraídos de cada uno de los espermatozoides.
- Los cuadros “Espermatozoide” y “Segmentación” permiten al usuario visualizar el espermatozoide original, y la Segmentación del espermatozoide diferenciando entre Acrosoma, Post-Acrosoma y Pieza Intermedia.
- Vista: permite al usuario visualizar distintas formas en las que serán desplegadas las imágenes pertenecientes a la muestra:
    1) Imagen original RGB
    2) Cada uno de los canales RGB por separado (Rojo, Verde, Azul)
    3) Contornos dibujados sobre la imagen original RGB
    4) Máscara binaria con los contornos identificados
- Nueva muestra: permite al usuario seleccionar una nueva muestra para poder clasificarla.
- Descargar: permite al usuario exportar los espermatozoides segmentados de la imagen actual a ser procesada.

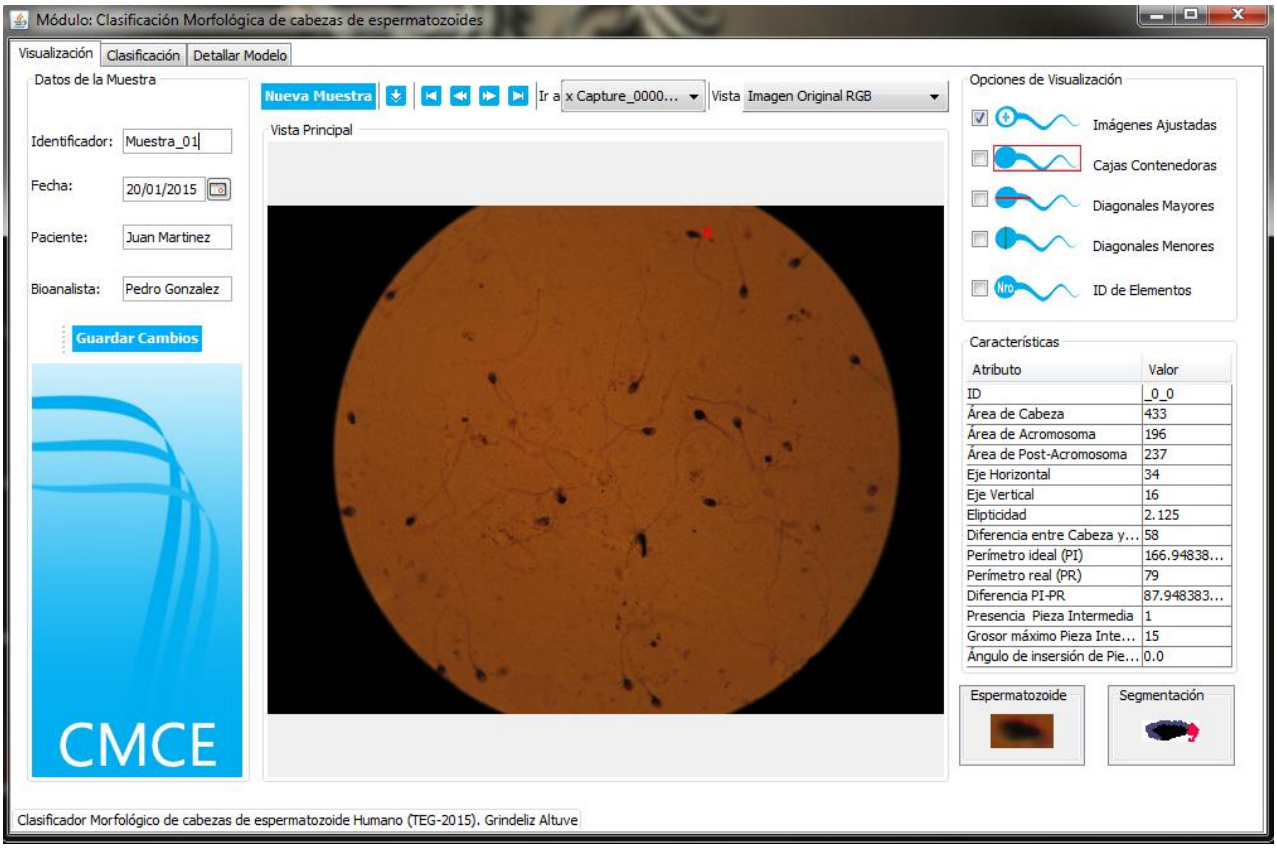



***Figura 28: Pantalla principal primera opción (Pestaña: Visualización)***

En la **figura 29**, se muestra la segunda pestaña “Clasificación”. Esta pestaña permite al usuario visualizar toda la información obtenido luego de realizar la clasificación a la muestra. Entre las opciones se encuentran:

- Datos de la muestra: al igual que la pestaña visualizar, permite desplegar información básica de la muestra
- Clasificar: opción que clasifica la muestra.
- Generar informe: esta opción permite al usuario descargar un informe con toda la información relacionada a la clasificación de la muestra en formato PDF. En el **Anexo C** se puede apreciar el modelo del informe.
- Resultados generales: Permite al usuario visualizar toda la información relacionada con la clasificación de la muestra
    1) Cantidad de espermatozoides extraídos de la muestra
    2) Cantidad de espermatozoides Normales
    3) Cantidad de espermatozoides Anormales
    4) Conteo de espermatozoides
    5) Porcentaje Normal Morfológico utilizado
    6) Porcentaje Normal Morfológico obtenido
    7) Clasificación general de la muestra

- Muestra: Permite al usuario seleccionar entre las distintas imagines pertenecientes a la muestra para ver sus características específicas, las cuales son desplegadas en el cuadro “Resultados por imagen”.
- Modificar Parámetros de clasificación: Permite al usuario modificar el porcentaje normal utilizado al momento de realizar la clasificación

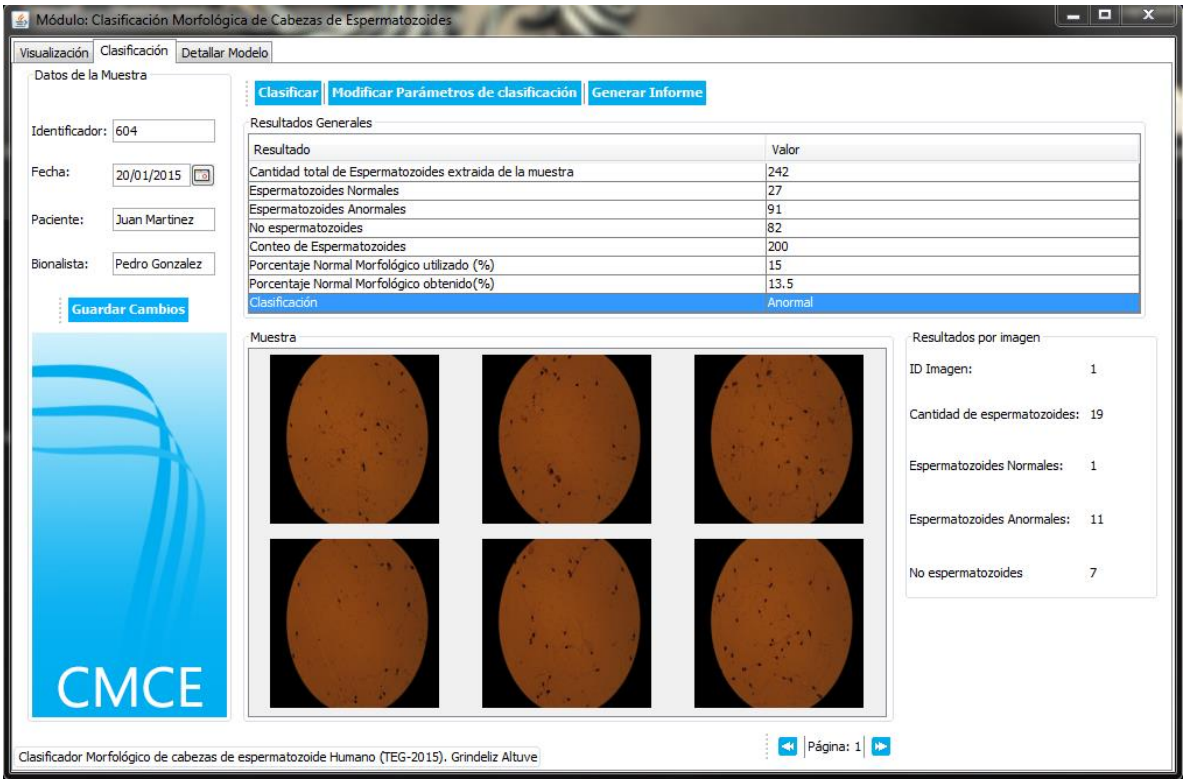



***Figura 29: Pantalla principal primera opción (Pestaña: Clasificación)***

En la **Figura 30**, se muestra la tercera pestaña “Detallar Modelo”. Esta pestaña permite al usuario visualizar toda la información referente al modelo de clasificación utilizado:

- Datos Modelo: al igual que las pestañas visualizar y clasificar, permite desplegar información básica pero esta vez del modelo de clasificación.
- Árbol de Clasificación: Permite al usuario graficar el árbol de clasificación y los valores de las medidas de evaluación generadas en la creación del modelo.
- Reglas de Clasificación: Permite al usuario graficar las reglas de clasificación y los valores de las medidas de evaluación generados en la creación del modelo.
- Red Neuronal: Permite al usuario generar los valores de las medidas de evaluación pertenecientes a la Red Neuronal.
- Combinación de clasificadores: Permite al usuario generar los valores de las medidas de evaluación pertenecientes a la Combinación de Clasificadores.
- Matriz de confusión: Permite al usuario visualizar la matriz de confusión generada en cada uno de los modelos (árbol de clasificación, reglas de clasificación, red neuronal, Combinación de clasificadores).

- Medidas de evaluación: Permite al usuario visualizar los valores de las medidas de evaluación pertenecientes a cada uno de los modelos.
- Conjunto de Variables: Permite al usuario visualizar las distintas variables pertenecientes al conjunto de datos utilizado.
- Variables Seleccionadas: Permite al usuario visualizar características específicas de la variable seleccionada.

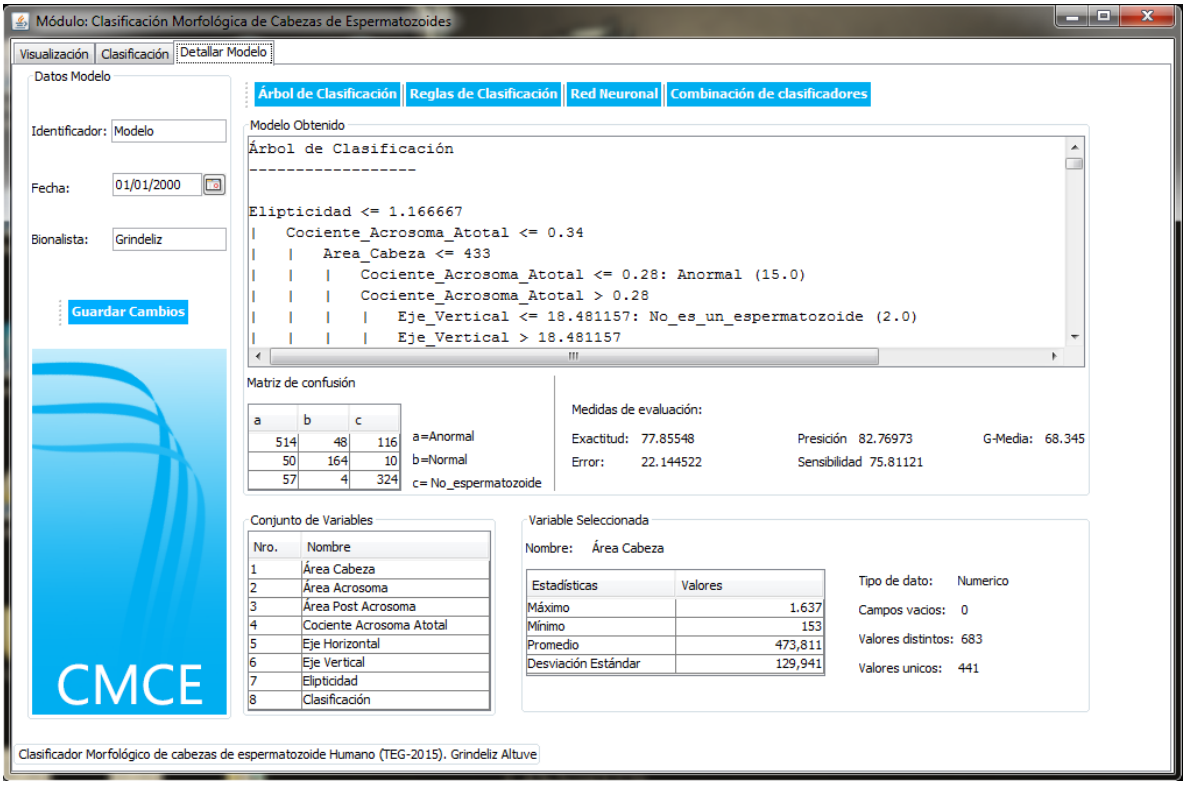



***Figura 30: Pantalla principal primera opción (Pestaña: Detallar Modelo)***

Volviendo a la pantalla inicial **(Ver figura 26)** al seleccionar la opción “Crear Modelo” se le desplegara al usuario, una pequeña pantalla **(Ver figura 31)** que le indicara los pasos a seguir.

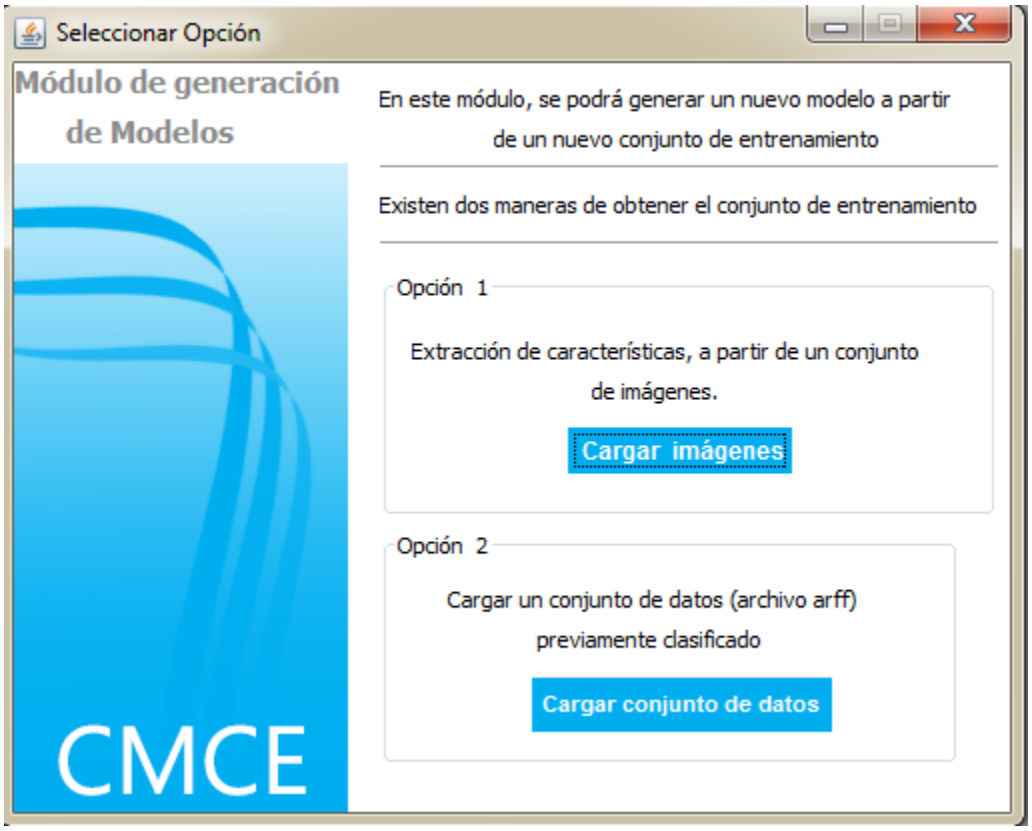



***Figura 31: Pantalla de indicaciones de la segunda opción***

En este caso se le específica al usuario que tiene dos opciones para poder generar el modelo. La primera le permite cargar una muestra y clasificarla manualmente en base a los conocimientos del experto (desplegándose la pantalla que se muestra en la **Figura 32**). La segunda opción le permite al usuario cargar una muestra previamente clasificada a través de un archivo arff (desplegándose la pantalla que se muestra en la **Figura 33**). Independientemente de la opción que elija, se desplegara una pestaña adicional **(Ver figura 34)** que permite visualizar el rendimiento del modelo de clasificación creado.

En la **Figura 32** se le presenta al usuario una pestaña (Clasificar Muestra) similar a la que se muestra en la **Figura 28**, la diferencia radica en las siguientes opciones:

- Clasificación: Permite al usuario realizar la clasificación manual de la muestra.
- Conteo: Permite al usuario controlar el conteo de espermatozoides, al momento de realizar la clasificación.

Una vez terminado la clasificación, se despliega la pantalla que se muestra en la **Figura 33.**

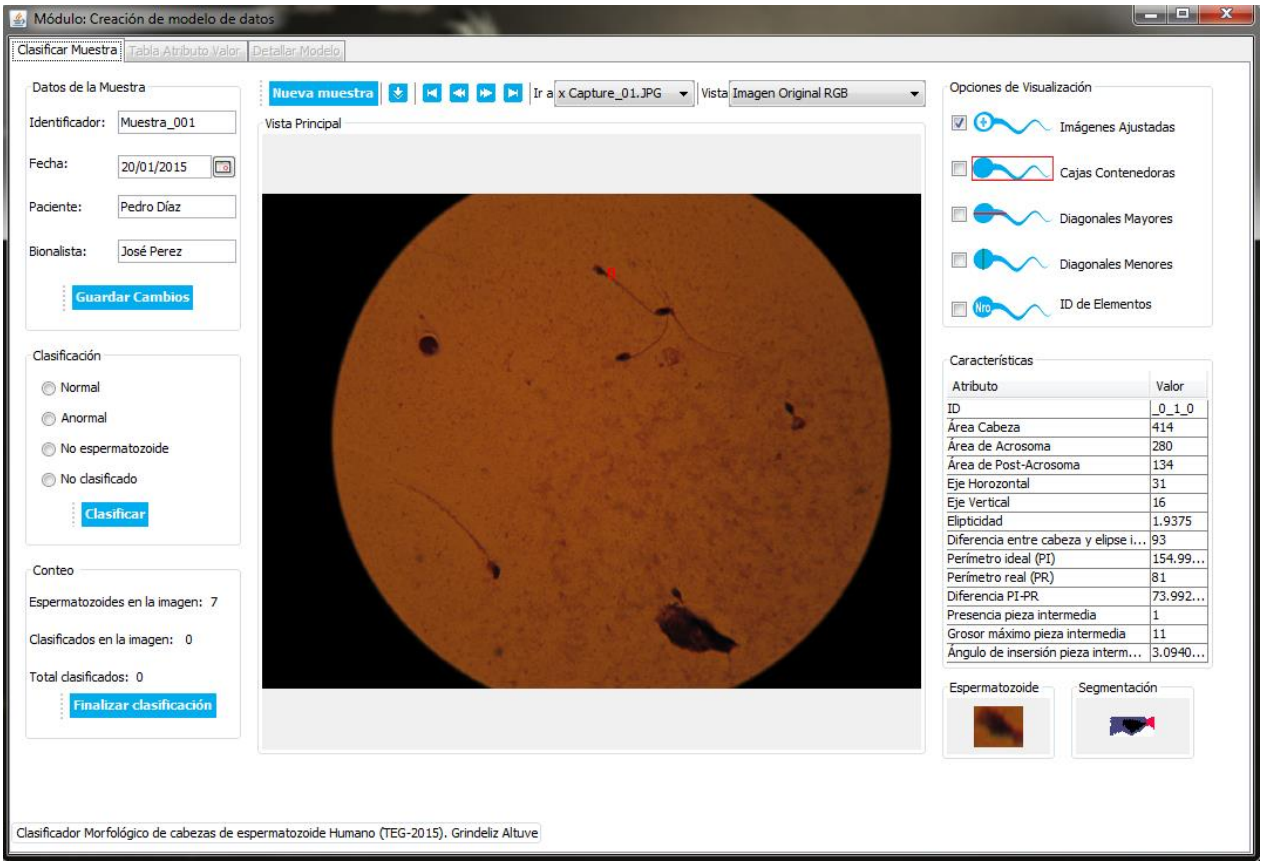



***Figura 32: Pantalla principal segunda opción (Pestaña: Clasificar Muestra)***

En la **Figura 33**, se muestra la segunda pestaña “Tabla Atributo Valor”. Esta pestaña permite al usuario visualizar todos los vectores de clasificación extraídos de la muestra (imágenes o archivo arff, **Anexo D**) desplegados en una tabla.

- Datos Modelo: al igual que la pestaña Clasificar Muestra, permite desplegar información básica del modelo de clasificación.
- Generar Modelo: Permite al usuario generar el nuevo modelo de clasificación.
- Conjunto Variables: Permite al usuario visualizar las distintas variables pertenecientes al conjunto de datos utilizado.
- Área cabeza: Permite al usuario visualizar características específicas de la variable seleccionada.

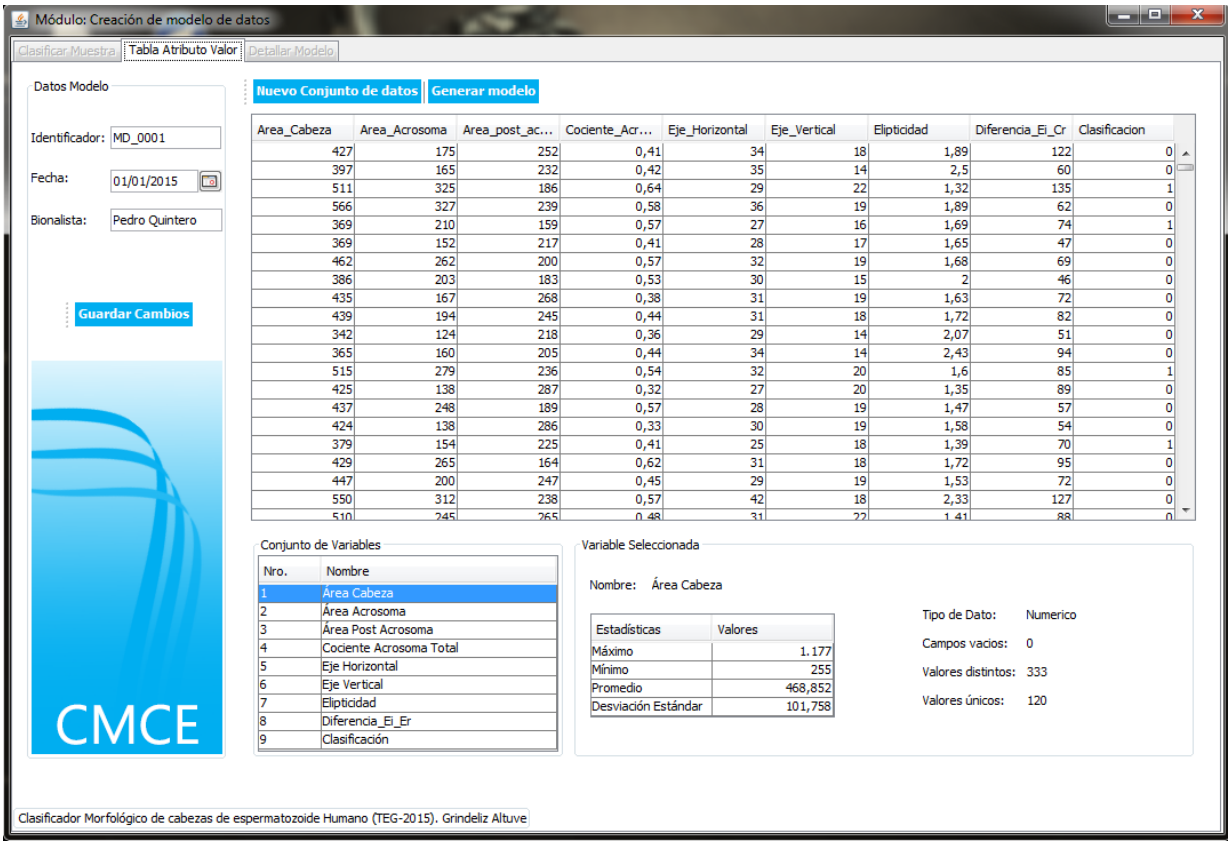



***Figura 33: Pantalla principal segunda opción (Pestaña: Tabla Atributo Valor)***

En la **Figura 34**, se muestra la tercera pestaña “Detallar Modelo”. Esta pestaña permite al usuario visualizar toda la información referente al modelo de clasificación creado:

- Datos Modelo: al igual que la pestaña Clasificar Muestra, permite desplegar información básica del modelo de clasificación.
- Árbol de Clasificación: Permite al usuario graficar el árbol de clasificación y los valores de las medidas de evaluación generadas en la creación del modelo.
- Reglas de Clasificación: Permite al usuario graficar las reglas de clasificación y los valores de las medidas de evaluación generados en la creación del modelo.
- Red Neuronal: Permite al usuario generar los valores de las medidas de evaluación pertenecientes a la Red Neuronal.
- Combinación de clasificadores: Permite al usuario generar los valores de las medidas de evaluación pertenecientes a la Combinación de Clasificadores.
- Matriz de confusión: Permite al usuario visualizar la matriz de confusión generada en cada uno de los modelos (árbol de clasificación, reglas de clasificación, red neuronal, Combinación de clasificadores).
- Medidas de evaluación: Permite al usuario visualizar los valores de las medidas de evaluación pertenecientes a cada uno de los modelos.
- Guardar Modelo: Permite al usuario almacenar en el sistema el nuevo modelo.

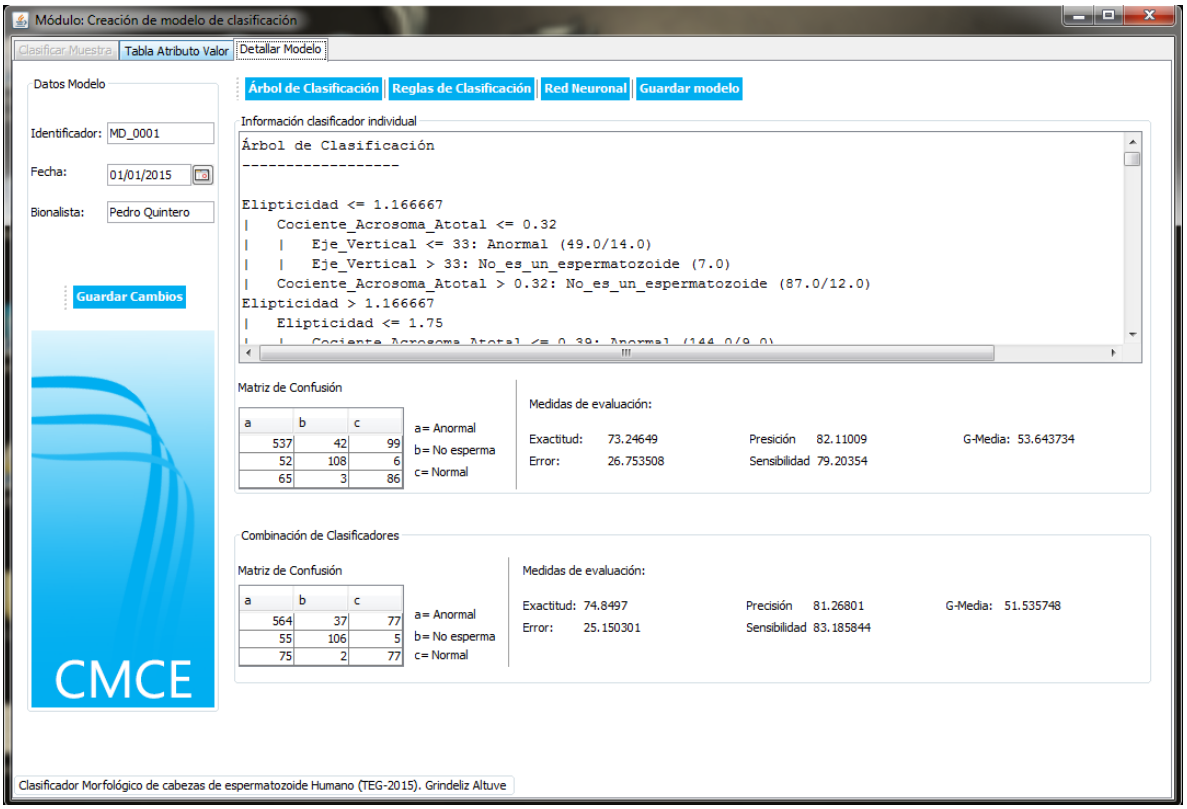



***Figura 34: Pantalla principal segunda opción (Pestaña: Detallar Modelo)***

Volviendo a la pantalla inicial **(Ver Figura 26)** se presentan dos opciones adicionales a las previamente explicadas. Ver Ayuda y Acerca de, la primera **(Ver Figura 35)** despliega una pantalla de ayuda para el usuario al momento de usar la aplicación, explicando cada una de las pestañas presentes en cada opción. La segunda opción **(Ver Figura 36)**, despliega al usuario una pantalla de información acerca del desarrollador y el tutor de la TEG.

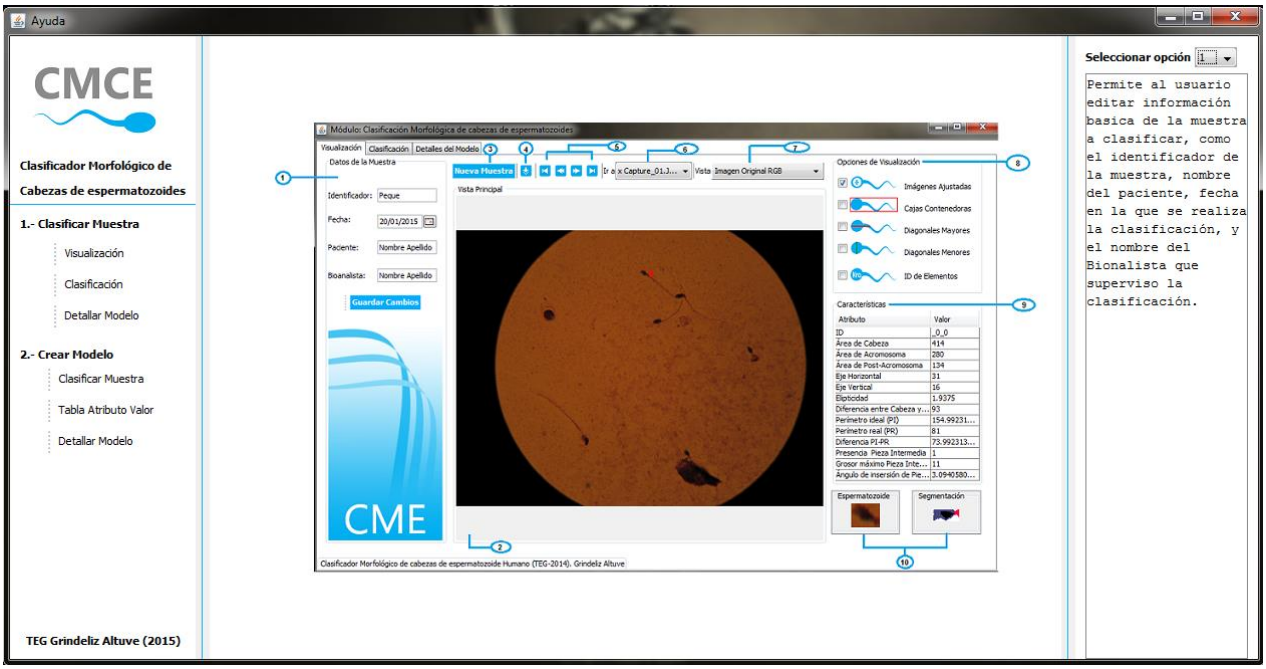



*Figura 35: Pantalla de ayuda para el usuario*

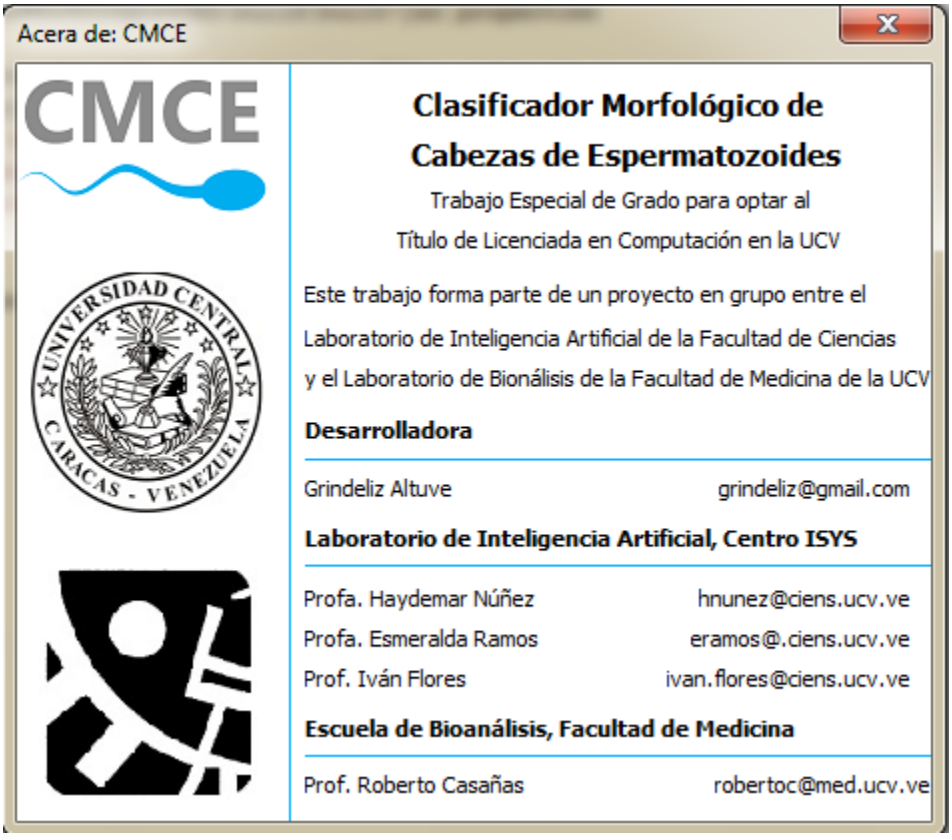



*Figura 36: Pantalla de información (Acerca de)*

## 2.8. Evaluación de la aplicación

Para la verificación y validación de la aplicación, se clasificó un grupo de doce muestras de semen, para luego comparar los resultados obtenidos con la clasificación realizada por un experto al mismo grupo de muestras. Los parámetros que se tomaron en cuenta fueron los siguientes:

- Cantidad total de estructuras presentes en la muestra (T_Estructuras)
- Cantidad total de espermatozoides clasificados, hasta llegar a 200, incluyendo Normales y Anormales (T_Espermatozoides)
- Cantidad de cabezas de espermatozoides normales (Normal)
- Cantidad de cabezas de espermatozoides anormales (Anormal)
- Porcentaje morfológico normal obtenido (% normal)
- Clasificación general de la muestra (Clase)

El porcentaje morfológico normal definido para la realización de la evaluación fue de 15%, es decir, si la muestra presenta un porcentaje de espermatozoides normales igual o mayor al 15% del total se consideró “Normal”, caso contrario la muestra se consideró “Anormal”. En la **Tabla 8**, se aprecian los resultados obtenidos.

| Parámetro | Valor |
|---|---|
| Estructuras identificadas | 2628 |
| Espermatozoides clasificados (Normal, Anormal) | 2628 |
| Muestras clasificadas que coinciden con el experto | 6 |
| Muestras clasificadas que no coinciden con el experto | 6 |

***Tabla 8: Resumen de valores obtenidos en la evaluación***

Los resultados arrojan una diferencia significativa en cuanto a la coincidencia en las clasificaciones realizadas por el experto y la aplicación. Se aprecia que mitad de las muestras coinciden en ambas clasificaciones (50%) y la otra mitad del conjunto de muestra no coincide (50%). La causa de estas diferencias puede deberse a los falsos reconocimientos obtenidos con los algoritmos de segmentación, es decir, elementos en la muestra que no son espermatozoides. El experto, al momento de clasificar la muestra tomó en cuenta estos casos, sin embargo, estos elementos son de igual forma procesados por la aplicación y clasificados como “Normales” o “Anormales” influyendo directamente en el aumento del porcentaje morfológico normal de la muestra, y por consiguiente en la clasificación de ésta.

Debido a los resultados obtenidos se decidió repetir todo el proceso de minería de datos, para así obtener un nuevo modelo de clasificación, pero esta vez utilizando un conjunto de datos que incluyera la clase “No_espermatozoides”, lo que permitió tomar en cuenta la presencia de estos elementos  al clasificar una instancia.

Para generar este nuevo modelo, se agregó al conjunto de datos original 166 instancias cuya clase era “No_espermatozoide”, formando un total de 998 instancias. Luego de esto se procedió a realizar de nuevo, la fase de preparación de los datos (limpieza, transformaciones, creación y selección de variables), obteniendo los mismo resultados que en el conjunto de datos previo.

Al igual que el conjunto de datos anterior, éste se encontraba desbalanceado ya que contenía más datos pertenecientes a la clase “Anormal”, por lo tanto se debió aumentar la proporción de las clases “No_espermatozoide” y “Normal”, utilizando el algoritmo SMOTE. Igual que en el caso anterior, se aumentaron las clases “Normal” y “No_espermatozoide” en proporciones del (50%,100% y 150%) cada una. Sin embargo, al aumentar la clase “No_espermatozoide” en estas proporciones, los resultados obtenidos no fueron los esperados, por lo que se decidió probar porcentajes menores a los anteriores para la clase “No_espermatozoide” (40%, 35% y 25%). De estos tres valores con los que se obtuvieron mejores resultados fue el caso del 35%, ya que al crear el modelo de clasificación con los otros valores (25%, 40%, 50%, 100%, 150%) la cantidad de muestras clasificadas por la aplicación que coincidían con el diagnóstico del experto disminuía en vez de aumentar. Por lo tanto, al momento de evaluar la data se tomaron en cuenta las siguientes vistas minables:

- Una con los datos originales, 998 registros.
- Tres con la clase “Normal” aumentada en proporciones del (50%,100% y 150%) y la clase “No espermatozoide” aumentada en un 35% en todos los casos.

Luego de esto, se aplicaron las técnicas de minería de datos, para obtener el nuevo modelo de clasificación.  En la **Tabla 9** se presentan los resultados obtenidos con cada una de las vistas minables.

| Proporción de datos aumentada | RBFNetwork | | JRIP | | J48 | | Combinación de clasificadores | |
|---|---|---|---|---|---|---|---|---|
| | Exactitud | G-media | Exactitud | G-media | Exactitud | G-media | Exactitud | G-media |
| 0% | 76.352 | 66.311 | 73.547 | 56.535 | 74.649 | 62.190 | 74.849 | 64.279 |
| 50% | 75.551 | 74.201 | 75.022 | 70.965 | 75.022 | 72.853 | 77.405 | 76.698 |
| 100% | 76.859 | 76.136 | 76.033 | 74.861 | 76.776 | 75.354 | 77.851 | 77.602 |
| 150% | 77.156 | 76.391 | 76.845 | 76.822 | 75.524 | 74.552 | 78.554 | 78.55 |

***Tabla 9: Resultados obtenidos con el nuevo modelo de clasificación***

Al comparar estos resultados con los reflejados en la **Tabla 7**, se puede constatar que la variación es muy sutil, es decir, a pesar de que el conjunto de datos posee una nueva clase, a la cual incluso se le aumento su proporción en un 35%, los resultados en ambos casos son muy parecidos, por esta razón, al igual que en el caso anterior los resultados permiten concluir que el conjunto de datos cuya proporción fue aumentada al 150% (Clase Normal) y 35% clase No_espermatozoide, es el más propicio para obtener una clasificación más exacta, teniendo en cuenta que la diferencia entre los valores obtenidos por los modelos de datos simples contra la combinación de clasificadores no es muy significativa.

A continuación, se presentan de forma detallada los resultados obtenidos al clasificar las doce muestras con el nuevo modelo de clasificación, además se toma en cuenta un nuevo parámetro (No_espermatozoides).

**Muestra #579**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 199 | 171 | 130 | 41 | 28 | 23.97% | Normal |
| Experto | 184 | 165 | 123 | 42 | 19 | 25.45% | Normal |

**Muestra #604**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 242 | 138 | 106 | 32 | 104 | 23.18% | Normal |
| Experto | 211 | 200 | 135 | 65 | 3 | 48.14% | Normal |

**Muestra #726**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 197 | 146 | 111 | 35 | 51 | 23.97% | Normal |
| Experto | 197 | 197 | 159 | 38 | 0 | 19.28% | Normal |

**Muestra #735**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 259 | 200 | 166 | 34 | 12 | 17% | Normal |
| Experto | 259 | 200 | 160 | 40 | 0 | 20% | Normal |

**Muestra #736**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 216 | 196 | 180 | 16 | 20 | 8% | Anormal |
| Experto | 216 | 200 | 180 | 20 | 0 | 10% | Anormal |

**Muestra #737**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 211 | 147 | 122 | 25 | 64 | 17.1% | Normal |
| Experto | 211 | 200 | 180 | 20 | 0 | 10% | Anormal |

**Muestra #738**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 237 | 200 | 160 | 40 | 9 | 20% | Normal |
| Experto | 237 | 200 | 166 | 34 | 0 | 17% | Normal |

**Muestra #754**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 238 | 200 | 118 | 82 | 24 | 41% | Normal |
| Experto | 168 | 168 | 135 | 33 | 0 | 19.64% | Normal |

**Muestra #755**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 231 | 168 | 152 | 16 | 63 | 9.52% | Anormal |
| Experto | 231 | 200 | 180 | 20 | 0 | 10% | Anormal |

**Muestra #759**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 216 | 153 | 138 | 15 | 63 | 9.80% | Anormal |
| Experto | 216 | 200 | 180 | 20 | 0 | 10% | Anormal |

**Muestra #766**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 189 | 118 | 92 | 26 | 71 | 22.03% | Normal |
| Experto | 189 | 189 | 171 | 18 | 0 | 9.52% | Anormal |

**Muestra #771**

| Clasificador | T_Estructuras | T_Espermatozoides | Anormal | Normal | No_espermatozoides | (%)normal | Clasificación |
|---|---|---|---|---|---|---|---|
| Aplicación | 193 | 173 | 115 | 58 | 20 | 33.52% | Normal |
| Experto | 193 | 193 | 169 | 24 | 0 | 12.43% | Anormal |

En la **Tabla 10** se muestran los resultados obtenidos de forma resumida

| Parámetro | Valor |
|---|---|
| Estructuras identificadas | 2628 |
| Espermatozoides clasificados (Normal, Anormal) | 2010 |
| No espermatozoides | 529 |
| Muestras clasificadas que coinciden con el experto | 9 |
| Muestras clasificadas que no coinciden con el experto | 3 |

***Tabla 10: Resumen de valores obtenidos, nueva muestra***

En este caso, los resultados arrojan una mejora significativa, en este segundo intento nueve de las doce muestras coincidieron con su correspondiente clasificación (75%), mientras que el resto de las muestras no coincidieron con la clasificación dada por el experto, lo que equivale a un 25% del total de las muestras. Dados estos resultados se puede concluir que, al agregar la clase “No espermatozoide” en el conjunto de datos, mejoraron los porcentajes de muestras correctamente clasificadas, disminuyendo así, los falsos reconocimientos por parte del modelo de clasificación al momento de procesar una estructura distinta a un espermatozoide.

Dado que, los valores obtenidos por los modelos de clasificación individuales contra el modelo de combinación de clasificadores no fueron muy significativos (Tabla 9), se realizó la clasificación de las doce muestras con cada uno de los modelos de forma individual (Tabla 12). De igual forma, para poder comparar los resultados, se procedió a calcular la precisión y el error de cada uno de los modelos (Tabla 11).

| Modelo | Precisión | Error |
|---|---|---|
| **RBFNetwork** | 82.5163% | 22.8438% |
| **JRIP** | 81.8471% | 23.1546% |
| **J48** | 81.7434% | 24.4755% |
| **Combinación de clasificadores** | 85.0921% | 21.4452% |

***Tabla 11: Precisión y Error de cada uno de los modelos obtenidos***

En la **Tabla 11**, se presentan los valores obtenidos al calcular la precisión y error de clasificación de cada uno de los modelos (valores que se calculan a partir de las matrices de confusión que se muestran en la **Figura 37**). Los resultados obtenidos para la combinación de clasificadores presentan un porcentaje de 85.0921% en cuanto a precisión, y menor a 22% en cuanto a error, los cuales son los valores promedio que generan los tres modelos restantes.

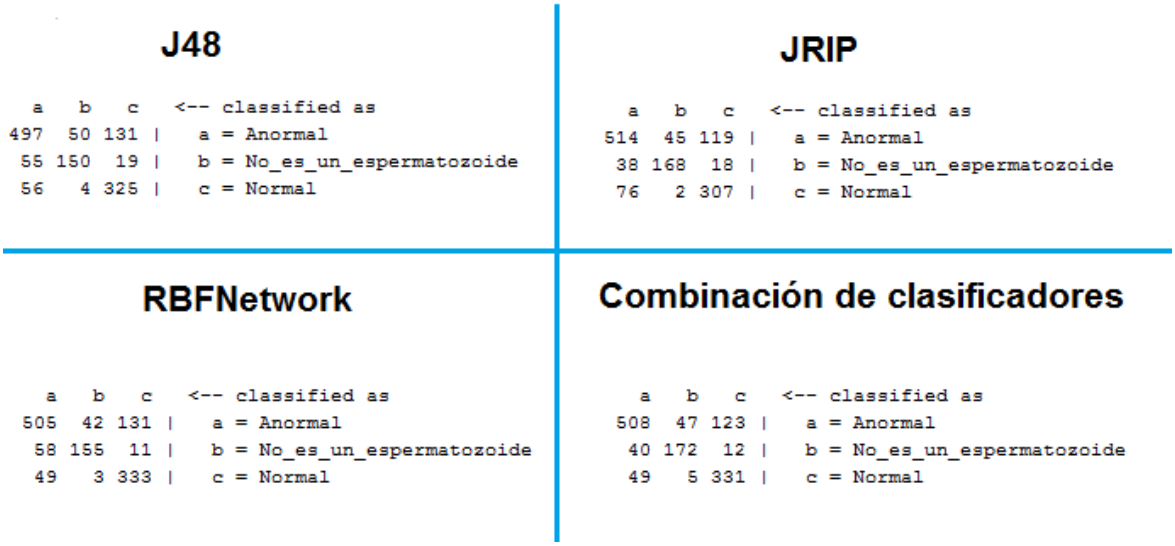



***Figura 37: Matrices de confusión de cada uno de los modelos obtenidos***

De las matrices de confusión que se presentan en la **Figura 37**, se puede observar que la cantidad de falsos reconocimientos es menor en el caso del modelo generado por la combinación de clasificadores (276). Mientras que los otros modelos presentan valores que se encuentran en un rango entre 315 y 298.

| Parámetro | RBFNetwork | JRIP | J48 |
|---|---|---|---|
| Estructuras identificadas | 2628 | 2628 | 2628 |
| Espermatozoides clasificados (Normal, Anormal) | 1971 | 2058 | 1982 |

| | | | |
|---|---|---|---|
| No espermatozoides | 570 | 451 | 558 |
| Muestras clasificadas que coinciden con el experto | 8 | 9 | 9 |
| Muestras clasificadas que no coinciden con el experto | 4 | 3 | 3 |

***Tabla 12: Valores obtenidos con los modelos individuales***

Los resultados reflejados en las tablas 10 y 12 indican que la cantidad de estructuras identificadas en los cuatro casos es la misma, debido a que el valor depende directamente de los algoritmos de segmentación. Por otro lado, los resultados arrojan una desventaja sobre el modelo de clasificación creado por la red neuronal RBFN, ya que éste aumenta la cantidad de muestras diagnosticadas de forma incorrecta. Sin embargo, la combinación de clasificadores, JRIP y J48 mantienen la misma cantidad de muestras diagnosticadas correctamente. Comparando los resultados en la **tabla 12** con los valores reflejados en la **tabla 11**, se muestra que si bien RBFNetwork tiene un porcentaje de error y de precisión similar al de JRIP y J48, existen algunas muestras que empeoran levemente sus resultados.

Los porcentajes de aciertos obtenidos con la combinación de clasificadores (Tabla 11) no constituyen el modelo óptimo. Sin embargo, calculando la media de todos los resultados obtenidos con los modelos individuales, se obtuvo una precisión de 85,092% y una tasa de error de 21.4452%, que en términos globales se considera aceptable.

Al momento de evaluar el rendimiento de la aplicación, se tomaron en cuenta otros aspectos como son el tiempo de ejecución de cada uno de los módulos de la aplicación (Segmentación, extracción de características, clasificación de espermatozoides, clasificación morfológica de la muestra). En la **Tabla 13** se muestran los tiempos promedios obtenidos en cada uno de los módulos de la aplicación, al procesar la muestra más pequeña (tres imágenes por muestra) y la más grande (40 imágenes por muestra).

| Módulo | Tiempo promedio |
|---|---|
| Segmentación de imágenes | 8,95 Segundos |
| Extracción de características | 0,0075 Segundos |
| Clasificación de espermatozoides | 6,335 Segundos |
| Clasificación morfológica de la muestra | 1,375 Segundos |
| Suma Total | 16,6675 Segundos |

***Tabla 13: Promedio de tiempos de ejecución de los módulos de la aplicación***

Los valores arrojados indican que los módulos con mayor tiempo de procesamiento son el de segmentación y clasificación de espermatozoides. Es necesario acotar que en el módulo de segmentación se realizan cálculos extras al momento de generar cualquiera de las estructuras matriciales necesarias para la detección de los espermatozoides presentes en la muestra. Estos cálculos aumentan el tiempo de ejecución al momento de segmentar una imagen. Por otro lado, se debe tomar en cuenta que el módulo de clasificación de datos, utiliza tres clasificadores para obtener una clasificación de una instancia. Cada una de las clasificaciones obtenidas por sus respectivos clasificadores, son procesadas por un esquema de votación. Todas estas consideraciones aumentan el tiempo de clasificación de una simple instancia.

Los tiempos de ejecución de los módulos están sujetos a la cantidad de imágenes y espermatozoides presentes en una muestra. Es por esto que mientras mayor sea la cantidad de imágenes y espermatozoides, mayor será el tiempo requerido por cada uno de los módulos de procesamiento.

## CONCLUSIONES Y RECOMENDACIONES

El objetivo principal de este trabajo consistió en la creación de una aplicación que de forma automatizada permitiera la realización de evaluaciones morfológicas de cabezas de espermatozoides a partir de imágenes digitales, evitando así que los resultados fueran expuestos a la subjetividad del experto.

La aplicación está compuesta de tres módulos principales. El primero se encarga de segmentar y extraer las características de cada uno de los espermatozoides. El segundo módulo tiene como objetivo clasificar cada uno de los espermatozoides extraídos, haciendo uso del modelo de clasificación creado. El último, permite generar un nuevo modelo de clasificación a partir de la creación de un conjunto de datos.

Los algoritmos de segmentación empleados en la aplicación fueron diseñados por Ferreira (2008), sin embargo estos fueron modificados para poder procesar imágenes de mayor resolución a la establecida en su trabajo de investigación, la cual era 1600x1200 píxeles. Las funciones modificadas fueron las que tenían como objetivo crear estructuras matriciales para delimitar las regiones a procesar (creación de la Máscara de Cabezas de Espermatozoide, creación de la Máscara de Núcleo de Círculo y Anillo, creación de la Máscara de Contornos de Espermatozoides). Para lograr esto se procedió a calcular la relación entre el área total de la imagen utilizada por Ferreira (2008) y las dimensiones de cada región. El cálculo del factor de proporción se obtuvo dividiendo el ancho de la región, entre el área de la imagen utilizada por Ferreira (2008). Una vez obtenido este factor, se multiplicó por el área de la nueva imagen a procesar, obteniéndose así la nueva región acorde a la resolución de la imagen. Los beneficios de introducir estas modificaciones, radica en que se podrá procesar imágenes de mayor resolución a la establecida por Ferreira (2008) y evaluar las heurísticas diseñadas para el reconocimiento de las estructuras en las muestras a evaluar.

Por otro lado, para la creación del módulo de clasificación de datos se utilizó el esquema de combinación de clasificadores *Ensemble Classifiers* aplicando el método *Baggins* o *Bootstrap Aggregation*. Este módulo consta de dos etapas, la generación de los modelos, en donde se aplicaron los algoritmos seleccionados (J48, JRIP, RBFN) al conjunto de entrenamiento (aumentando la clase minoritaria a un 150%, para disminuir el desbalance de la clase normal y 35% la clase No_Espermatozoide, así como también descartando la variable *Diferencia Ei – Cr*) y la etapa de clasificación, en donde se utilizan los modelos creados para clasificar la nueva instancia. Cada modelo determina la clase para la instancia, empleando un sistema de votación. Dicha instancia es asignada a la clase más frecuente.

En cuanto a los resultados obtenidos con el modelo, nueve de las doce muestras coincidieron con su correspondiente clasificación (75%), mientras que el resto de las

muestras no coincidieron con el diagnostico dado por el experto, lo que equivale a un 25% del total de las muestras. Si bien el porcentaje de aciertos obtenidos con la combinación de clasificadores no constituye el óptimo, al promediar todos los resultados generados con los modelos individuales, se obtienen resultados que en términos globales se consideran aceptables.

Mediante la combinación de clasificadores, se pretende obtener en todo momento el mejor resultado entre los distintos modelos utilizados, ayudando a que la clasificación realizada responda eficientemente a las expectativas deseadas. En trabajos futuros el objetivo es claro, se debe plantear el aumento de la cantidad de algoritmos de clasificación a emplear aprovechando la amplia variedad de algoritmos que ofrece la plataforma WEKA. Permitiendo así, obtener modelos cuyos resultados se complementen los unos con los otros y cuyos aciertos sean tan elevados como sea posible.

Otro aspecto que se debe tomar en cuenta es el rendimiento de los algoritmos de segmentación y extracción de datos. Es necesario optimizar estos algoritmos para disminuir el tiempo de ejecución de los mismos, ya que a diferencia del trabajo de Ferreira (2008), el cual segmentaba las imágenes de forma individual, la aplicación segmenta las imágenes de la muestra de forma continua, aumentado el tiempo de ejecución de los algoritmos. Por otro lado, no se debe olvidar que el módulo de clasificación de datos, utiliza tres modelos para obtener la clasificación de una instancia. Cada una de las clasificaciones obtenidas por sus respectivos modelos, son procesadas por un esquema de votación. Todas estas consideraciones aumentan el tiempo de clasificación de una simple instancia.

Para finalizar, se expone como recomendaciones e investigaciones futuras, hacer énfasis en la realización de pruebas utilizando imágenes de mayor resolución, para poder evaluar los cambios realizados a los algoritmos de segmentación y extracción de características de espermatozoides, esperando así, obtener mejores resultados al momento de extraer las características de cada una de las estructuras segmentadas.

## REFERENCIAS


Acosta, M., Salazar, H., & Zuluaga, C. (2000). Tutorial de Redes Neuronales. (Universidad Tecnológica de Pereira - Facultad de Ingeniería Eléctrica) Recuperado el 6 Enero de 2015, de http://goo.gl/hdtiyI

Antonie, M. L., Zaiane, O. R., & Coman, A. (2001). Application of Data Mining Techniques for Medical Image Classification. *MDM/KDD*, *2001*, 94-101.

Ballou, D.; & Tayi, G. (1999).*Enhancing Data Quality in Data Warehouse Environments. Communications of the ACM*, 42(1), 73-78. Bar-Ilan, J.

Broomhead, D. S., & Lowe, D. (1988). *Radial basis functions, multi-variable functional interpolation and adaptive networks* (No. RSRE-MEMO-4148). Royal signals and radar establishment Malvern (United Kingdom).

Campo, J; (2009). *Diseño de un nuevo clasificador supervisado para minería de datos.* Proyecto Fin de Máster en Sistemas Inteligentes para la Universidad Complutense. Madrid, España.

Casañas, R.; Ramos, E.; & Núñez, H. (2009). *Construcción de modelos de clasificación de la morfología de espermatozoide humano mediante técnicas de minería de datos.* Conferencia Latinoamericana de Informática, CLEI'2011. Quito, Ecuador

Casañas, R.; Castro, M.; Pereira, H.; Ramos, E.; & Núñez, H. (2007). *Aplicación de visualización de una ontología para el dominio del análisis del semen humano.* Ingeniería y Ciencia, ISSN 1794-9165, vol. 3, num. 5, pp. 43-66.

Cerezo, G. (2006). *El Análisis de Semen Humano: La Estandarización de su Evaluación.* Contacto Químico 1, vol. 3, oct.-dic. 6-8, 2006.

Clark, P., & Niblett, T. (1989). *The CN2 induction algorithm. Machine learning*, 3(4), 261-283.

Cohen, W., (1995). *Fact, Effective Rule Induction.* In ICML'95,12th International Conference on Machine .Learning, pp.115-123.

Colmenares G. (2004). *Funcion De Base Radial Radial Basis Function (RBF).* Recuperado el 6 de Enero de 2015, de http://goo.gl/NiRHPy

Dávila, F; Sánchez, Y. (2012) *Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas*. Universidad y Salud, 14(2) ,174-183. Recuperado el 15 de Febrero de 2015, de: http://goo.gl/s5wZRy

Delen, D.; Walker, G.; & Kadam, A. (2004) *Predicting breast cancer survivability a comparison of three data mining methods.* Artificial Intelligence in Medicine, 1-15.

Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases. AI Magazine*, 17(3), 37-54.

Ferreira, H.; (2008). *Herramienta para la extracción automática de características morfológicas de espermatozoides humanos a partir de imágenes*. Trabajo Especial de Grado para la Universidad Central de Venezuela. Caracas, Venezuela.

Garrido, L., & Latorre, J. I. (2001). *Aplicaciones empresariales de Data Mining*.

Guo, X.; Yin, X.; Dong, C.; Yang, G.; Zhou, G. (2008). On the Class Imbalance Problem. ICNC, Proceedings of the 2008 Fourth International Conference on Natural Computation. 4:192-201.

Hernández, J.; Ramírez, M. & Ferri, C. (2004). *Introducción a la minería de datos.* (3ra Edición) Madrid: Pearson - Prentice Hall.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.; (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.

Han, J., Kamber, M., & Pei, J. (2006). *Data Mining: Concepts and Techniques*, (The Morgan Kaufmann Series in Data Management Systems).

Hunt, E.; Marin, J. & Stone, P. (1966). *Experiments in induction.* New York: Academic Press.

Katz, D.; Overstreet, J.; Samuels, S.; Nis Wander, P.; Bloom, T.; Lewis, E. (1986) *Morphometric Analysis of Spermatozoa in the Assessment of Human Male Fertility*. Journal of Andrology. 7: 203-210.

Kruger, T.; Coetzee, K. (1999). *The role of sperm morphology in assisted reproduction. Human reproduction update.* 5(2):172-178.

Lorca, G.; Arzola, J. & Pereira, O. (2010). *Segmentación de Imágenes Médicas Digitales mediante Técnicas de Clustering.* Revista Aporte Santiaguino; 3(1): 1-9.

Mortimer, D. (1994). *Practical Laboratory Andrology*. Oxford University Press.

Obenshain, MK. (2004). *Application of data mining techniques to healthcare data*. Infect Control Hosp Epidemiol; 25(8):690-695.

OMS – Organización Mundial de la Salud (2001). *Manual de Laboratorio para el examen del semen humano y de la interacción entre el semen y el moco cervical.* Madrid, España. Editorial Médica Panamericana S.A. Cuarta Edición.

Padrón, R.; Fernández, G.; Gallardo, M. (1988) *Interpretación del Análisis Seminal.* Revista cubana Endocrinología, ISSN 0188-9796, Instituto Nacional de Endocrinología, Departamento de Reproducción Humana, 9(1), 81-90, referenciado en 47,50.

Pedraza, L.; Gutiérrez, L.; Duarte, S. & Niebles, F. (2012) *Aplicación de la Función de*

*Base Radial (RBF) en una red neuronal Artificial (RNA) para la predicción de series de tiempo en fenómenos físicos.*

Pereira, R. & Yépez, M. (2012). *La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino*. Universidad y Salud, 14(2), 117-129. Recuperado el 15 de Febrero de 2015, de: http://goo.gl/7KxFJe

Polikar, R. (2006) *Ensemble based systems in decision making*. IEEE Circuits Syst Mag 6(3):21–45.

Quinlan, J. (1986). *Induction of decision trees. Machine Learning*, 1:81–106.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann Publishers Inc.

Raghavendra, B. & Jay B. (2011). *Evaluation of logistic regression model with feature selection methods on medical dataset*. International Journal of Advanced Engineering Technology; 2(1): 228-233.

Ramos, E. (2009). *Sistema Basado en Agentes para Apoyar el Diagnostico de la Calidad del Semen Humano.* Tesis Doctoral para la Universidad Central de Venezuela. Caracas, Venezuela.

Rodríguez, J.; Rojas, E. & Franco, R. (2007). *Clasificación de datos usando el método K-nn*. Vínculos, volumen 4, primera edición.

Rokach, L. (2010). *Ensemble-based classifiers.* Journal Artificial Intelligence Review. Volume 33 Issue 1-2, Pages 1 – 39. February.

Soni, J.; Ansari, U.; Sharma, D. & Soni, S. (2011). *Predictive Data mining for medical diagnosis: An overview of Heart Disease prediction.* International Journal of computer Applications. Volumen 17, Octava edición.

Tan P.; Steinbach, M. & Kumar, V. (2005). *Introduction to Data Mining.* Boston, MA, USA: Addison-Wesley Longman Publishing Co.

Timarán, P; & Yépez, M. (2012). *La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino*. Universidad y Salud, 14(2), 117-129. Recuperado en 15 de febrero de 2015, de http://goo.gl/S1xZW1.

Vega, B.; Sánchez, L. & Cortiñas, J. (2012). *Clasificación de dengue hemorrágico utilizando árboles de decisión en la fase temprana de la enfermedad.* Rev Cubana Med Trop 2012.

Wilford, I.; Rosete, A. & Rodríguez, A. (2009). *Análisis de Información Clínica mediante técnicas de Minería de Datos.*

Witten, I; Frank, E. & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* (3ra Edición) Nueva Zelanda: Elsevier.

# ANEXOS

## Anexo A

A continuación se presenta el modelo de clasificación generado por el algoritmo J48 (Implementación del algoritmo C4.5 en WEKA)

```
Árbol de Clasificación
------------------
Elipticidad <= 1.166667
|  Cociente_Acrosoma_Atotal <= 0.34
|  |  Area_Cabeza <= 433
|  |  |  Cociente_Acrosoma_Atotal <= 0.28: Anormal (15.0)
|  |  |  Cociente_Acrosoma_Atotal > 0.28
|  |  |  |  Eje_Vertical <= 18.481157: No_es_un_espermatozoide (2.0)
|  |  |  |  Eje_Vertical > 18.481157
|  |  |  |  |  Eje_Horizontal <= 22.489767
|  |  |  |  |  |  Area_Cabeza <= 298: Anormal (3.0)
|  |  |  |  |  |  Area_Cabeza > 298: No_es_un_espermatozoide (2.0)
|  |  |  |  |  Eje_Horizontal > 22.489767: Anormal (5.0)
|  |  Area_Cabeza > 433
|  |  |  Area_post_acrosoma <= 504.307055
|  |  |  |  Elipticidad <= 0.96: Anormal (3.0)
|  |  |  |  Elipticidad > 0.96
|  |  |  |  |  Elipticidad <= 1.140515: No_es_un_espermatozoide (16.0)
|  |  |  |  |  Elipticidad > 1.140515: Anormal (2.0)
|  |  |  Area_post_acrosoma > 504.307055
|  |  |  |  Area_Cabeza <= 790: Anormal (8.0)
|  |  |  |  Area_Cabeza > 790: No_es_un_espermatozoide (14.0/1.0)
|  Cociente_Acrosoma_Atotal > 0.34: No_es_un_espermatozoide (104.0/10.0)
Elipticidad > 1.166667
|  Elipticidad <= 1.75
|  |  Cociente_Acrosoma_Atotal <= 0.4: Anormal (158.0/13.0)
|  |  Cociente_Acrosoma_Atotal > 0.4
|  |  |  Cociente_Acrosoma_Atotal <= 0.77
|  |  |  |  Area_Cabeza <= 585
|  |  |  |  |  Area_Cabeza <= 354
|  |  |  |  |  |  Cociente_Acrosoma_Atotal <= 0.5: Anormal (9.0)
|  |  |  |  |  |  Cociente_Acrosoma_Atotal > 0.5
|  |  |  |  |  |  |  Elipticidad <= 1.508861: No_es_un_espermatozoide (14.0)
|  |  |  |  |  |  |  Elipticidad > 1.508861: Anormal (2.0/1.0)
|  |  |  |  |  Area_Cabeza > 354
|  |  |  |  |  |  Eje_Horizontal <= 32.990144
|  |  |  |  |  |  |  Elipticidad <= 1.35
|  |  |  |  |  |  |  |  Eje_Vertical <= 19.94617: Normal (15.0/1.0)
|  |  |  |  |  |  |  |  Eje_Vertical > 19.94617
|  |  |  |  |  |  |  |  |  Area_Acrosoma <= 249
|  |  |  |  |  |  |  |  |  |  Area_post_acrosoma <= 190.16751: No_es_un_espermatozoide (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  Area_post_acrosoma > 190.16751
|  |  |  |  |  |  |  |  |  |  |  Cociente_Acrosoma_Atotal <= 0.415907: No_es_un_espermatozoide (2.0)
|  |  |  |  |  |  |  |  |  |  Cociente_Acrosoma_Atotal > 0.415907
|  |  |  |  |  |  |  |  |  |  |  Elipticidad <= 1.33: Anormal (19.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  Elipticidad > 1.33
|  |  |  |  |  |  |  |  |  |  |  |  Eje_Horizontal <= 27.177371: Anormal (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  Eje_Horizontal > 27.177371: Normal (3.0/1.0)
|  |  |  |  |  |  |  |  Area_Acrosoma > 249
```

```
| | | | | | | | | | Area_post_acrosoma <= 166: No_es_un_espermatozoide (4.0/1.0)
| | | | | | | | | | Area_post_acrosoma > 166
| | | | | | | | | | | Cociente_Acrosoma_Atotal <= 0.64
| | | | | | | | | | | | Elipticidad <= 1.283118
| | | | | | | | | | | | | Area_Cabeza <= 441.483746: Normal (3.0)
| | | | | | | | | | | | | Area_Cabeza > 441.483746
| | | | | | | | | | | | | | Area_Cabeza <= 545: Anormal (5.0)
| | | | | | | | | | | | | | Area_Cabeza > 545: Normal (7.0/2.0)
| | | | | | | | | | | | | Elipticidad > 1.283118: Normal (14.0/1.0)
| | | | | | | | | | | Cociente_Acrosoma_Atotal > 0.64: Normal (11.0)
| | | | | | | Elipticidad > 1.35
| | | | | | | | Area_Acrosoma <= 204
| | | | | | | | | Eje_Horizontal <= 27.864167
| | | | | | | | | Area_post_acrosoma <= 216.006683
| | | | | | | | | | Area_Cabeza <= 381
| | | | | | | | | | | Area_Acrosoma <= 161: Anormal (2.0)
| | | | | | | | | | | Area_Acrosoma > 161
| | | | | | | | | | | | Eje_Horizontal <= 26.976734: Normal (9.0)
| | | | | | | | | | | | Eje_Horizontal > 26.976734
| | | | | | | | | | | | | Area_Acrosoma <= 179.320388: Anormal (2.0)
| | | | | | | | | | | | | Area_Acrosoma > 179.320388: Normal (2.0)
| | | | | | | | | | Area_Cabeza > 381: Anormal (8.0/1.0)
| | | | | | | | | Area_post_acrosoma > 216.006683: Normal (9.0)
| | | | | | | | Eje_Horizontal > 27.864167: Anormal (24.0/3.0)
| | | | | | | Area_Acrosoma > 204
| | | | | | | | Cociente_Acrosoma_Atotal <= 0.687273: Normal (372.0/86.0)
| | | | | | | | Cociente_Acrosoma_Atotal > 0.687273
| | | | | | | | | Area_post_acrosoma <= 123: Anormal (7.0)
| | | | | | | | | Area_post_acrosoma > 123: Normal (16.0/6.0)
| | | | | | Eje_Horizontal > 32.990144
| | | | | | | Area_Acrosoma <= 216.266307: No_es_un_espermatozoide (4.0)
| | | | | | | Area_Acrosoma > 216.266307
| | | | | | | | Eje_Vertical <= 21.455163
| | | | | | | | | Cociente_Acrosoma_Atotal <= 0.58498: Anormal (24.0)
| | | | | | | | | Cociente_Acrosoma_Atotal > 0.58498
| | | | | | | | | | Eje_Horizontal <= 33.452236
| | | | | | | | | | | Area_post_acrosoma <= 202
| | | | | | | | | | | | Area_post_acrosoma <= 185: No_es_un_espermatozoide (2.0)
| | | | | | | | | | | | Area_post_acrosoma > 185: Anormal (2.0)
| | | | | | | | | | | Area_post_acrosoma > 202: Normal (4.0)
| | | | | | | | | | Eje_Horizontal > 33.452236: Anormal (6.0)
| | | | | | | | Eje_Vertical > 21.455163
| | | | | | | | | Eje_Vertical <= 22.338278
| | | | | | | | | | Area_Cabeza <= 565: Normal (7.0/1.0)
| | | | | | | | | | Area_Cabeza > 565: Anormal (3.0)
| | | | | | | | | Eje_Vertical > 22.338278: Anormal (2.0)
| | | | Area_Cabeza > 585
| | | | | Area_Cabeza <= 771
| | | | | | Eje_Horizontal <= 30.699341: No_es_un_espermatozoide (2.0)
| | | | | | Eje_Horizontal > 30.699341
| | | | | | | Elipticidad <= 1.32
| | | | | | | | Area_Cabeza <= 658: Anormal (5.0)
| | | | | | | | Area_Cabeza > 658: No_es_un_espermatozoide (3.0)
| | | | | | | Elipticidad > 1.32: Anormal (37.0)
| | | | | Area_Cabeza > 771: No_es_un_espermatozoide (4.0)
| | | Cociente_Acrosoma_Atotal > 0.77: No_es_un_espermatozoide (39.0/7.0)
| Elipticidad > 1.75
| | Cociente_Acrosoma_Atotal <= 0.71: Anormal (232.0/5.0)
| | Cociente_Acrosoma_Atotal > 0.71
| | | Cociente_Acrosoma_Atotal <= 0.82: Anormal (6.0/2.0)
| | | Cociente_Acrosoma_Atotal > 0.82: No_es_un_espermatozoide (4.0)

Número de hojas:  55
Tamaño del árbol:  109
```

## Anexo B

A continuación se presenta el modelo de clasificación generado por el algoritmo JRIP (Implementación del algoritmo RIPPER en WEKA)

Reglas de Clasificación:

==========

(Elipticidad <= 1.318182) and (Elipticidad <= 1.137214) and (Area_Acrosoma >= 132.490404) => Clasificacion=No_es_un_espermatozoide (105.0/9.0)

(Area_post_acrosoma <= 152) and (Elipticidad <= 1.345373) => Clasificacion=No_es_un_espermatozoide (38.0/6.0)

(Cociente_Acrosoma_Atotal >= 0.79) => Clasificacion=No_es_un_espermatozoide (29.0/5.0)

(Elipticidad <= 1.35) and (Eje_Horizontal <= 22.403622) and (Cociente_Acrosoma_Atotal >= 0.29) => Clasificacion=No_es_un_espermatozoide (19.0/6.0)

(Eje_Vertical >= 23.987788) and (Cociente_Acrosoma_Atotal <= 0.51) and (Area_Acrosoma >= 271) and (Elipticidad <= 1.484848) => Clasificacion=No_es_un_espermatozoide (11.0/0.0)

(Area_post_acrosoma >= 401) and (Eje_Horizontal <= 29) => Clasificacion=No_es_un_espermatozoide (6.0/1.0)

(Elipticidad <= 1.375) and (Area_Cabeza >= 675) and (Area_Acrosoma >= 183) => Clasificacion=No_es_un_espermatozoide (10.0/3.0)

(Eje_Vertical >= 20) and (Area_Acrosoma <= 245) and (Area_post_acrosoma <= 186.954377) => Clasificacion=No_es_un_espermatozoide (8.0/1.0)

(Area_post_acrosoma >= 433) and (Area_post_acrosoma <= 468) => Clasificacion=No_es_un_espermatozoide (5.0/1.0)

(Cociente_Acrosoma_Atotal >= 0.450479) and (Eje_Horizontal <= 32.990144) and (Eje_Vertical >= 19.001083) => Clasificacion=Normal (312.0/82.0)

(Area_post_acrosoma <= 217.382025) and (Elipticidad <= 1.681442) and (Area_Acrosoma >= 205) and (Area_Acrosoma <= 275.078228) => Clasificacion=Normal (97.0/17.0)

=> Clasificacion=Anormal (647.0/93.0)

Número de reglas: 12

**Anexo C**

A continuación se presenta el modelo del informe generado al momento de clasificar cualquier muestra de espermatozoides.


# Información de la muestra

Identificador de la muestra: Muestra_000

Fecha de clasificación: DD/MM/AAAA

Nombre del paciente: Nombre Apellido

Nombre del Bionalista: Nombre Apellido

# Resultados Generales

| **Resultado** | **Valor** |
|---|---|
| Cantidad Total de espermatozoides extraídos de la muestra | 00 |
| Espermatozoides Normales | 00 |
| Espermatozoides Anormales | 00 |
| No espermatozoides | 00 |
| Conteo de espermatozoides | 00 |
| Porcentaje Normal Morfológico utilizado (%) | 00 |
| Porcentaje Normal Morfológico Obtenido (%) | 00 |
| Clasificación | Clase |

# Resultados por imagen

| Imagen | Cantidad de espermatozoides | Espermatozoides Normales | Espermatozoides Anormales | No espermatozoides |
|---|---|---|---|---|
| 0 | 000 | 00 | 00 | 00 |
| 1 | 000 | 00 | 00 | 00 |
| 2 | 000 | 00 | 00 | 00 |

**Anexo D**

Los archivos con extensión arff (*Attribute Relation File Format*), son utilizados por WEKA para obtener los datos de entrada a procesar. Este formato está compuesto en tres partes:

- Cabecera: Todo archivo debe comenzar con esta declaración en su primera línea. **@relation <relation-name>**, donde <relation-name> no puede contener espacios en blancos.
- Declaraciones de atributos: En esta sección, se debe incluir una línea por cada atributo que compone al conjunto de datos. La sintaxis es la siguiente: **@attribute <attribute-name> <datatype>**, donde <attribute-name> indica el nombre del atributo (sin espacios), y <datatype> el tipo de dato del atributo (soportado por Weka: numérico, string, date, y nominal).
- Sección de datos: Inicia con la sintaxis: **@data**, se declaran los datos que componen la relación. Los atributos deben estar separados por comas y deberán tener el mismo número de atributos definidos en la sección anterior.

A continuación se presenta el formato utilizado en el archivo arff, al momento de crear un nuevo modelo de clasificación.

```
@relation Nombre_Relacion

@attribute Area_Cabeza numeric
@attribute Area_Acrosoma numeric
@attribute Area_post_acrosoma numeric
@attribute Cociente_Acrosoma_Atotal numeric
@attribute Eje_Horizontal numeric
@attribute Eje_Vertical numeric
@attribute Elipticidad numeric
@attribute Diferencia_Ei_Cr numeric
@attribute Clasificacion {Anormal,Normal,No_espermatozoide}

@data

414,280,134,0.676328,31,16,1.93,93,Normal
561,293,268,0.522281,39,17,2.29,134,Normal
511,205,306,0.401174,35,20,1.75,184,Normal
427,175,252,0.409836,34,18,1.88,122,Normal
487,217,270,0.445585,33,20,1.65,116, No_espermatozoide
```