

Universidad Central "Marta Abreu" de Las Villas  
Facultad de Matemática, Física y Computación



# Clasificación de células cervicales en la prueba de Papanicolaou

Tesis para optar por el título de Máster en Ciencias de la Computación

**Autor**

Ing. Solangel Rodríguez Vázquez

**Tutor**

DrC. Juan V. Lorenzo Ginori

Santa Clara, 2015

*“Mark Twain dijo:*

*“Los dos días más importantes de tu vida son el día en que naces y el día en que descubres por qué.”*

# Dedicatoria

*A toda mi familia por su apoyo incondicional.*

*A los amigos que siempre estuvieron y a los no amigos que me dijeron que no podía.*

*A todas las personas que luchan por sus sueños sin importar los obstáculos.*

# Agradecimientos

Hoy es un día de nervios y felicidad, primero por lograr hacer lo que siempre anhelé graduarme en esta universidad y segundo porque no solo para mí es importante, sino para todos aquellos que de una forma u otra pusieron un poquito de sí para obtener este resultado. Muchas son las personas que colaboraron y que hoy no están conmigo, unos se fueron de mi lado, otros se fueron del país, en fin cada quién tomó su camino. A pesar de ello hoy les agradezco porque en su momento pusieron su corazón y apoyo en lo que hoy es un día feliz.

Mi mayor agradecimiento y aunque no empezara por ahí va para mi familia y mis padres, que siempre estuvieron y estarán conmigo apoyándome y aconsejándome así que aunque no es necesario, GRACIAS.

A aquellos amigos, que son parte de mi familia y que me dieron su mano cuando realmente me hacía falta un amigo y que siempre han estado conmigo.

A alguien especial para mí, que llegó inesperadamente a mi vida y que ha aguantado lo que no creo que haría yo. Que me ha demostrado mil veces su amor y quién si no fuera por él realmente no estuviera aquí, MI NOVIO. A él debo tanta felicidad el día de hoy.

Y bueno, quién sí no puede quedarse aunque sea el último de la lista, lo que no significa menos importante, para mí, mi TUTOR el Dr.C. Juan V. Lorenzo Ginori que me adentro y condujo con sabiduría por la bioinformática y que puso todo su tiempo y apoyo para que esta investigación saliera con el rigor científico y la calidad necesaria, además de un poquito de paciencia. Por su guía, enseñanza, motivación, apoyo constante y visión científica, GRACIAS.

Un agradecimiento especial para los profesores de la Maestría por el conocimiento aportado durante los cursos recibidos. A mis compañeros y amigos de la maestría que fuimos un grupo inseparable Genaro, Janet, Yoannys, Carlos Andrés, Carlos Herrera, Cocote, Maykel y César.

A todos aquellos que de una forma u otra pusieron un granito de arena para ayudarme, tanto en este trabajo como en mi formación personal y profesional. A todos ustedes les estaré eternamente agradecida...

GRACIAS.

# Resumen

Hace 40 años, el cáncer de cérvix uterino representaba una de las mayores amenazas de muerte por cáncer entre las mujeres. Con el avance continuo en la medicina y la tecnología, las muertes por esta enfermedad han disminuido significativamente. Las investigaciones referentes a este tema han podido determinar síntomas claves que permiten detectar a tiempo esta enfermedad para darle un tratamiento oportuno.

La citología convencional es una de las técnicas más utilizadas, siendo ampliamente aceptada, de bajo costo, y con mecanismos de control. Con el objetivo de aliviar la carga de trabajo a los especialistas, algunos investigadores han propuesto el desarrollo de herramientas de visión computacional para detectar y clasificar las transformaciones en las células de la región del cuello uterino.

El presente trabajo tiene como objetivo proveer a los investigadores de una herramienta de clasificación automática, aplicable a las condiciones existentes en los centros médicos y de investigación del país. Esta herramienta debe ser capaz de clasificar las células del cuello del útero, basándose solamente en las características extraídas de la región del núcleo y sin utilizar las características del citoplasma, de manera que se reduzca la tasa de falsos negativos en la prueba de Papanicolaou. A partir del estudio realizado, se obtuvo una herramienta haciendo uso de la técnica *k*-vecinos más cercanos con la distancia manhattan, el cual mostró un alto desempeño manteniendo valores de AUC superiores al 91% y llegando hasta un 97.1% con respecto a los clasificadores *SVM* y *RBF Network*, los que también fueron analizados.

**Palabras clave:** cáncer de cérvix uterino, células del cuello uterino, clasificación de células, *kNN*, núcleos celulares, *SVM*, distancias.

# Índice General

Introducción .....	1
Fundamento Teórico .....	5
1.1. Detección de células anómalas en el cérvix .....	5
1.1.1. Causa/Efecto de la presencia de células anómalas en el cérvix .....	5
1.1.2. Métodos de diagnóstico .....	9
1.1.3. Características de las imágenes celulares.....	12
1.1.4. Rasgos de las imágenes celulares.....	13
1.2. Procesamiento Digital de Imágenes: Clasificación de objetos en imágenes.....	15
1.2.1. Componentes de un sistema de clasificación .....	17
1.2.2. Clasificadores de datos, Técnicas de Clasificación y Reglas de clasificación .....	18
1.2.3. Medidas de evaluación.....	21
1.3. Herramientas y librerías de procesamiento de imágenes .....	23
1.3.1. ImageJ .....	23
1.3.2. Fiji .....	24
1.3.3. Matlab .....	24
1.3.4. WEKA.....	24
1.4. Conclusiones Parciales.....	25
Metodología de análisis, clasificación y diagnóstico .....	26
2.1. Descripción de la solución propuesta.....	26
2.1.1. Caracterización de las fases del proceso .....	27
2.2. Descripción y metodología de uso de los algoritmos propuestos .....	28
2.2.1. SVM (Descripción del algoritmo, Funciones utilizadas).....	30
2.2.2. <i>kNN</i> (Descripción del algoritmo, Funciones utilizadas) .....	33
2.2.3. Redes Neuronales Artificiales (Descripción del algoritmo, Funciones utilizadas.).....	37
2.3. Conclusiones Parciales.....	42

Validación Experimental .....	43
3.1.    Herramienta de ensayos .....	44
3.2.    Análisis de los resultados .....	44
3.2.1.    Análisis de los resultados de acuerdo a la medida $F$ .....	44
3.2.2.    Análisis de los resultados de acuerdo a la Predictividad Negativa (Pn) .....	45
3.2.3.    Análisis de los resultados de acuerdo al Área bajo la curva ROC (A-ROC o AUC).....	46
3.3.    Análisis comparativo de los clasificadores .....	47
3.3.1.    Comparación con otros enfoques de la literatura estudiada .....	49
3.4.    Conclusiones Parciales.....	50
Conclusiones Generales .....	51
Recomendaciones .....	53
Referencias bibliográficas.....	54
Anexos .....	59

# Introducción

La prueba citológica, también llamada prueba de Papanicolaou o Pap, es un examen de pesquisa ginecológica que permite apreciar los cambios en la morfología de las células del cuello uterino. En ellas se toman muestras de células del interior y de los alrededores del cuello del útero. Las células se extienden en un portaobjetos o lámina de cristal, se fijan y se envían al laboratorio de citodiagnóstico para ser procesadas. La observación y el análisis de los tejidos permiten comprobar el estado de las células. Este examen posibilita alertar sobre anomalías, incluso asintomáticas, en el aparato genital femenino, contribuyendo a mejorar la calidad de vida de la mujer. Dicho estudio diagnostica infecciones, trastornos y cambios morfológicos en células del cuello y cuerpo del útero. Puede alertar sobre una patología tan frecuente en la mujer como el cáncer del cuello uterino, cuya detección temprana facilita el tratamiento y aumenta las oportunidades de curación.

La prueba de Papanicolaou, debido al carácter masivo de su aplicación en la población femenina, genera una considerable carga de trabajo para los laboratorios que analizan en el microscopio los frotis resultantes de esta técnica citológica. En un frotis típico se pueden encontrar hasta 300000 células, lo que limita la productividad a no más de unas 60-80 citologías por día de trabajo y observador [1]. A partir de la experiencia de la aplicación de esta prueba durante muchos años, se ha podido determinar que existen diferentes factores que afectan la calidad de los resultados. Estos son, principalmente, los errores en la toma de las muestras, en su procesamiento y en su lectura e interpretación. Sobre este último caso, la necesidad de analizar una gran cantidad de muestras con muy baja tasa de casos positivos tiende a sesgar el resultado de la evaluación, y además provoca errores debidos a la rutina y a la fatiga de los analistas.

Como consecuencia de los diferentes errores mencionados, aparece una cierta tasa de falsos negativos en el resultado final. Estos implican que una lesión precursora pueda progresar, o que un carcinoma in situ (curable) pueda devenir en un cáncer invasivo, sin que la paciente sea sometida a tratamiento alguno.

[1]



El análisis de imágenes de microscopía es cada vez más importante tanto para la generación de diagnósticos como para las investigaciones que se llevan a cabo en las Ciencias Informáticas y de la Salud. Por ello, se hace necesario el desarrollo de nuevas técnicas que efectúen un análisis práctico de las muestras, tales como el análisis celular a nivel del epitelio vaginal, lo cual proporciona información para la detección de deformaciones celulares; que a su vez, pueden llevar al diagnóstico de enfermedades o a la prevención de las mismas.

Los métodos tradicionales utilizados en la detección de células premalignas en el útero femenino se basan en análisis microscópicos, métodos moleculares o microscopia electrónica de transmisión y pruebas de reacción en cadena de la polimerasa (denominadas técnicas PCR) [2]. Ahora es más común utilizar sistemas de captura de datos, que poseen herramientas robustas o módulos de procesamiento de imágenes. Estas herramientas, permiten realizar el tratamiento y reconocimiento completo de las partículas o microorganismos en las imágenes muestreadas. [3]

Desde la década de los 90's se han desarrollado e implantado sistemas informáticos en los laboratorios, con el fin de clasificar y/o diagnosticar enfermedades a través de procesamiento digital de imágenes. El diagnóstico de enfermedades en el útero femenino se sigue realizando mediante los métodos tradicionales de análisis mediante el microscopio. En las investigaciones desarrolladas se utilizan técnicas de procesamiento para observar los cambios en las células. [3]

Por todo lo anteriormente planteado se tiene como problema científico: ¿Cómo reducir la tasa de falsos negativos, a través de los métodos que utilizan el procesamiento digital de imágenes en la prueba de Papanicolaou?

Este problema se desglosa en las siguientes preguntas de investigación:

1. ¿Qué métodos de clasificación pueden resultar efectivos en el análisis de las imágenes para determinar posibles anomalías en las células?
2. ¿Cómo programar un plugin especializado en la clasificación de células cervicales?
3. ¿Qué nivel de efectividad es posible lograr en una herramienta de clasificación automática aplicada a la prueba de Papanicolaou y cómo es posible cuantificar y expresar esta efectividad?

Para darle solución al problema científico se planteó el siguiente objetivo general de investigación que consiste en: desarrollar una herramienta de clasificación automática, haciendo uso de algoritmos de clasificación que contribuyan a reducir la tasa de falsos negativos en la prueba de Papanicolaou.

Este objetivo general fue desglosado en los siguientes objetivos específicos:

1. Caracterizar los procesos que se incluyen en la prueba de Papanicolaou.
2. Describir los algoritmos que se utilizan en el procesamiento de imágenes y visión computacional que se puedan utilizar para la detección de células anómalas en la prueba de Papanicolaou.
3. Desarrollar una herramienta de clasificación automática que permita reducir la tasa de falsos negativos en el análisis computacional de la prueba citológica haciendo uso de las técnicas de clasificación.
4. Evaluar el nivel de efectividad de la herramienta de clasificación desarrollada en base a las medidas de evaluación definidas y aplicadas a los clasificadores utilizados con el objetivo de reducir los falsos negativos en la prueba de Papanicolaou.

Después de haber evaluado el marco teórico se formuló la siguiente hipótesis de investigación como respuestas a las preguntas de investigación: Si se desarrolla una herramienta de clasificación automática, que haciendo uso de técnicas de clasificación analice las imágenes obtenidas de la prueba de Papanicolaou, se logrará disminuir los falsos negativos y realizar un control de calidad en las mismas.

Esta investigación permite obtener resultados cuya novedad científica radica en la obtención de una nueva herramienta de apoyo a la clasificación de células cervicales que, haciendo uso de algoritmos de clasificación, contribuye al trabajo de los especialistas que realizan la prueba de Papanicolaou. Podría aplicarse en consultas, hospitales y para las valoraciones de estudio de enfermedades en el campo de la citología y las enfermedades en el útero. Además la solución que se propone podría incrementar la productividad de los laboratorios y a su vez lograr reducir la tasa de falsos negativos en el resultado final de la prueba y propiciando un punto de inicio para posteriores investigaciones, desarrollo y avance de la propia solución propuesta en dicha investigación.

La investigación estará compuesta por tres capítulos que posibilitarán un entendimiento adecuado hasta llegar a la discusión de los resultados obtenidos. El capítulo 1 introduce aspectos sobre la detección de células anómalas en el cérvix y la Prueba de Papanicolaou, además se abordan temas como la clasificación de imágenes, aspecto fundamental de la investigación. El capítulo 2 aborda una breve descripción de la solución propuesta, así como la caracterización de los algoritmos de clasificación definidos y una comparación entre los parámetros internos para el funcionamiento de los mismos. Dicha comparación está basada en obtener el algoritmo con los parámetros más eficientes y eficaces para la clasificación de células cervicales que pueda ser comparado con otros algoritmos similares. Los resultados obtenidos en el capítulo 2, son reflejados en el capítulo 3 en la validación y evaluación estadística de cada clasificador con respecto

a los demás y a otras investigaciones reportadas en la literatura. En dicho capítulo concluye la investigación con la propuesta del clasificador que de acuerdo a la investigación y los resultados estadísticos obtenidos es el más adecuado para la clasificación de células cervicales en la prueba de Papanicolaou. Este documento culmina con las Conclusiones, Recomendaciones, Referencias Bibliográficas y Anexos.

# Capítulo 1

## Fundamento Teórico

En el presente capítulo son tratados los conceptos y definiciones fundamentales relacionadas con la prueba de Papanicolaou. Se describen brevemente los pasos que están presentes en el procesamiento de imágenes digitales, profundizando en la etapa de clasificación que es el objetivo fundamental de dicha investigación. Se abunda lo referente a las medidas o indicadores de evaluación a utilizar para evaluar los resultados de la clasificación de las células cervicales, así como las herramientas posibles a utilizar en dicha clasificación.

### 1.1. Detección de células anómalas en el cérvix

La mayoría de los cánceres de cuello uterino se originan en el revestimiento de las células de esta región. Estas células no se tornan en cáncer de repente, sino que las células normales del cuello uterino se transforman gradualmente en cambios precancerosos, los cuales se convierten en cáncer. En las ciencias médicas se usan varios términos para describir estos cambios precancerosos, incluyendo neoplasia intraepitelial cervical (CIN, por sus siglas en inglés), lesión intraepitelial escamosa (SIL, por sus siglas en inglés) y displasia. Estos cambios se pueden detectar mediante la prueba de Papanicolaou y se pueden tratar para prevenir el desarrollo del cáncer. [4]

Las células cervicales anómalas son células en el revestimiento del cuello que han cambiado de apariencia. Esto también se llama displasia cervical. Cuanto más severa es la anomalía cervical, mayor es la probabilidad de desarrollar cáncer de cuello uterino en el futuro. En la mayoría de los casos, el desarrollo de cáncer de cuello uterino a partir de cambios anómalos en las células cervicales puede tomar varios años. Sin embargo, en casos excepcionales se ha desarrollado en períodos de aproximadamente un año.

#### 1.1.1. Causa/Efecto de la presencia de células anómalas en el cérvix

El cuello uterino es la abertura del útero que se encuentra encima de la vagina. Está cubierto por una fina capa de tejido. Este tejido está compuesto por células. A medida que estas células se desarrollan, las que se encuentran en la capa inferior se trasladan lentamente a la superficie del cuello uterino. Durante este

proceso, algunas células pueden transformarse en células anómalas o alterarse. Las células alteradas crecen de modo diferente (Figura 1). En algunos casos, es necesario extraer estas células para evitar que se produzca una lesión cancerosa.

Las células anómalas muestran ciertas alteraciones precancerosas preliminares que se denominan displasia o neoplasia cervical intraepitelial (NCI). La displasia y la NCI se clasifican en grados: leve, moderado o grave. La displasia leve (NCI 1) generalmente tiende a desaparecer con el tiempo. La displasia moderada (NCI 2) y grave (NCI 3) indican alteraciones más peligrosas.

La prueba de Papanicolaou, también denominada frotis de Papanicolaou o evaluación de la citología cervical, detecta cambios anómalos en las células del cuello uterino y permite comenzar con el tratamiento a tiempo antes de que las células se vuelvan cancerosas.

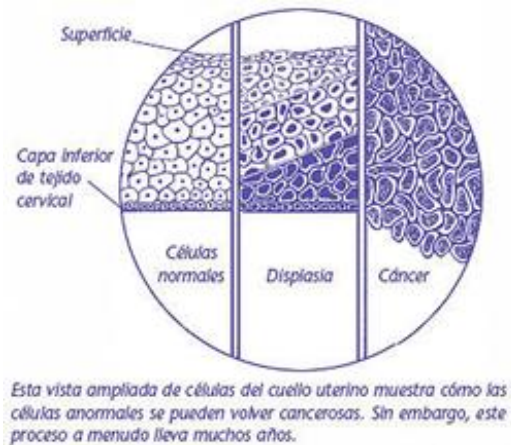


Figura 1: Vista ampliada de las células del cuellos uterino [5].

Las pruebas de detección ofrecen la mejor oportunidad para detectar el cáncer de cuello uterino en una etapa temprana cuando es probable que el tratamiento sea eficaz. Además, pueden prevenir la mayoría de los cánceres cervicales mediante la detección de los cambios anómalos de las células del cuello uterino (pre-cánceres) para que puedan ser tratados antes de que estos devengan en un proceso canceroso.

La importancia de la detección de infecciones cervicovaginales radica en que pueden evitarse complicaciones posteriores, como enfermedad pélvica inflamatoria, ruptura prematura de membranas, abortos, sepsis de un recién nacido, esterilidad. Por eso es importante la prevención mediante exámenes periódicos. Y aún más importante es que la calidad de esta prueba sea óptima.

La vaginosis bacteriana, las infecciones micóticas y el trichomonas vaginalis, son infecciones muy frecuentes, localizadas por la prueba. Esto demuestra que no solo previene el cáncer de cérvix, sino que detecta cualquier tipo de anomalía existente en esta región del organismo.

### Características de las células y los núcleos del cérvix

La mucosa del exocérnix consta de un estroma fibromuscular de superficie lisa, y de un revestimiento de epitelio pavimentoso estratificado no cornificado. Este posee un estrato basal, de células con núcleos alargados, perpendiculares a la membrana basal y escaso citoplasma; un estrato parabasal, de células poligonales y mayor cantidad de citoplasma; un estrato intermedio, con células de núcleos ovalados y abundante citoplasma y un estrato superficial, de células dispuestas horizontalmente y núcleos pequeños, hipercromáticos. Esta variación de los caracteres celulares al ascender hacia la superficie, corresponde a una maduración celular (Figura 2).

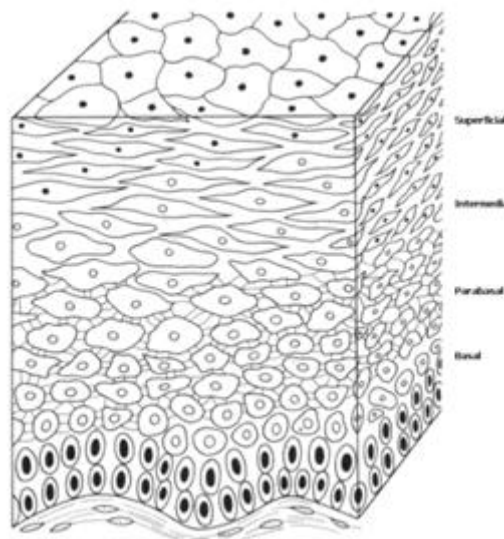


Figura 2: Diferenciación escamosa de células por capas.

Por regla general en un frotis normal se presentan células superficiales, intermedias y parabasales. La existencia de células de la capa más interna (capa basal) es síntoma inequívoco de anomalías; es decir, la aparición de este tipo de células informa la inmadurez o la no-diferenciación del tejido más externo del cérvix. Si la mujer es menopáusica, se suministran estrógenos y se presta mucha más atención a las demás características con el fin de obtener un diagnóstico más certero. Es importante mencionar que dos de los criterios que permiten diferenciar entre células basales, células parabasales, células intermedias y células superficiales son: la relación núcleo-citoplasma y la distribución de cromatina en el núcleo, sin detrimento de las diversas características morfológicas que poseen cada tipo de célula. Las figuras 3a, 3b, 3c y 3d, son representativas de los cuatro tipos de células.

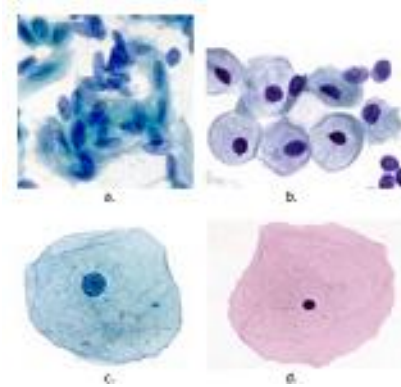


Figura 3: Ejemplo de células del epitelio escamoso del cérvix: a. Basal, b. Parabasal, c. Intermedia, d. Superficial.

**Tamaño del Núcleo:** La evaluación de la anomalía en el tamaño nuclear se basa en compararlo con el tamaño de una célula normal del mismo tipo de tejido, y no en el tamaño absoluto. El tamaño nuclear de las células somáticas normales es distinto de un tejido a otro y la relación nuclear-citoplasmática depende del grado de maduración de la célula. Las células basales del epitelio pavimentoso estratificado tienen una relación nuclear-citoplasmática mayor que las superficiales. Una relación nuclear-citoplasmática aumentada o disminuida se compara con las del resto del tejido originario normal.

**Aumento de la Relación Nuclear-Citoplasmática:** La relación entre el volumen nuclear y el citoplasmático se mantiene dentro de límites constantes en las células normales. El aumento del componente nuclear de esta relación [6], es uno de los rasgos más distintivos de malignidad. La relación nuclear-citoplasmática (relación N/C) no solo es valiosa para distinguir malignidad, sino que provee la pauta del grado de diferenciación de la célula cancerosa; cuanto menos diferenciadas son las células cancerosas, mayor es la relación N/C. Puede ocurrir agrandamiento nuclear en células no malignas por irradiación, administración de agentes alquilantes, inflamación, metaplasia y regeneración, pero en estos estados las células no malignas exhiben un aumento de tamaño general y conservan la relación N/C dentro de los límites benignos.

**Multinucleación y Multilobulación:** La multiplicación y pronunciada indentación o plegamiento de los núcleos se deben a mitosis anómalas. La diferenciación entre multinucleación maligna y benigna debe basarse en la alta relación nuclear-citoplasmática, en la mayor cantidad de material cromático y en la irregularidad de su distribución, todo lo cual es característico de malignidad. En la multinucleación benigna los núcleos son de tamaño y forma uniformes y de igual contenido de cromatina.

**Koilocitos (Virus del papiloma humano VPH):** Un aspecto de trascendental importancia hoy día es la identificación de mujeres afectadas por condiloma en relación con la infección por el virus del papiloma humano (VPH), agente del que cada vez hay más evidencia de su relación con el carcinoma de cérvix. Con

la toma citológica cervicovaginal se puede identificar un gran porcentaje de casos, ya que existe un cuadro citológico muy característico, tal como la presencia de koilocitos. Desgraciadamente, no siempre se observan dichas células, y además, disminuyen progresivamente en número conforme se incrementa la severidad de la displasia (aunque está ligada a dicha infección), siendo excepcionales en casos de cáncer invasivo. [7]

### 1.1.2. Métodos de diagnóstico

La prueba de Papanicolaou es la prueba de detección principal para el cáncer de cuello uterino y los cambios precancerosos. Aunque esta ha tenido más éxito que ninguna otra prueba de detección en la prevención del cáncer, no es perfecta. Una de sus limitaciones es que el análisis de las muestras de estas pruebas es realizado por los humanos, por lo que no siempre es posible un examen preciso de cientos de miles de células en cada muestra. Los investigadores están trabajando en conjunto para mejorar esta prueba. Existen además factores de riesgo que pueden afectar la calidad de las muestras tomadas [8] como son:

- Inflamación o infección vaginal.
- Embarazo.
- Manipulación física o irritación química como: tacto vaginal previo, desinfectantes en crema o líquidos, gel lubricante, medicación vaginal, duchas vaginales o geles espermicidas (menos de 24hs antes del estudio).
- Programar la cita médica durante el periodo menstrual. (El mejor momento es al menos cinco días después de que haya finalizado el período menstrual).
- Radioterapia.
- En el caso de embarazo o puerperio, la calidad del frotis disminuye debido a cambios reactivos inflamatorios; recomendándose esperar a 6-8 semanas postparto en mujeres que hayan realizado su examen dentro de los tres años previos y concurren al seguimiento.

Debido a la presencia real de los factores antes mencionados y la poca conciencia femenina de la importancia de realizarse la citología, es que se reconoce la necesidad de establecer estrategias para el mejoramiento de la prueba de Papanicolaou:

- Mejora de los programas de reclutamiento y seguimiento.
- Mejora de la calidad de la toma de la muestra citológica.
- Técnicas alternativas de preparación de la muestra citológica.
- Control de calidad de la observación.
- Desarrollo de métodos automatizados.



En este trabajo se utilizan imágenes de microscopía celular (Figura 4) tomadas en la prueba de Papanicolaou mediante el método convencional.

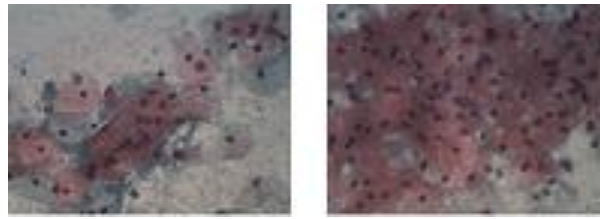


Figura 4: Imágenes de un campo, correspondiente a un frotis en la prueba de Papanicolaou convencional.

### **Equipos automatizados para el control de calidad de la prueba de Papanicolaou**

A pesar de la probada efectividad de los frotis de Papanicolaou en la reducción de la incidencia del cáncer cervical, muchos se han preocupado de la precisión de esta prueba diagnóstica, la cual depende de su especificidad y sensibilidad para detectar. Además con el transcurso de los años y la experiencia adquirida en la realización de las pruebas se han evidenciado e identificado factores de riesgo que ponen en peligro y afectan la calidad del resultado obtenido.

Los factores de riesgo son considerados como las diversas circunstancias, acciones, actividades, hábitos que contribuyen a producir un daño a nivel del tejido cervicouterino. Estos son, principalmente, los errores en la toma de las muestras, en su procesamiento y en su lectura e interpretación. Sobre este último caso, la necesidad de analizar una gran cantidad de muestras con muy baja tasa de casos positivos tiende a afectar el resultado de la evaluación, y además provoca errores debidos a la rutina y a la fatiga de los analistas. Como consecuencia de esto, aparece una cierta tasa de falsos negativos en el resultado final. Esto implica que una célula en estado pre-maligno pueda degenerarse a un estado maligno antes de una próxima prueba, sin que la paciente sea sometida a tratamiento alguno.

Debido a la preocupación de los médicos ante la aparición de falsos negativos se ha hecho necesaria la búsqueda de métodos que acierten de forma efectiva sobre las pruebas y la interpretación de las mismas. A esto se une la necesidad de incrementar la productividad de los laboratorios así como la de reducir la tasa de falsos negativos.

Los métodos que utilizan el procesamiento digital de imágenes y la visión computacional, han comenzado a jugar un papel importante en los últimos años. En [9] se exponen sistemas como *AutoPap Primary Screening System (Tri-Path Imaging)*, *BD FocalPoint™ Slide Profiler*, *ThinPrep Imaging System*.

***AutoPap Primary Screening System (Tri-Path Imaging)***: [9] presenta una valoración del estado de esta técnica y plantea que aglutina las tecnologías que le precedieron Autocyte SCREEN (Roche), Auto-Pap

(Neopath) y PapNet (Neuromedical Systems). Se emplea para el control de calidad del trabajo realizado por los analistas humanos, mediante un nuevo análisis de los casos negativos. Los casos que resultan sospechosos, son sometidos a un nuevo análisis. Está diseñado para tamizaje primario, para detectar carcinoma escamoso, adenocarcinoma y formas precursoras. Escanea láminas de citología convencionales preparadas con Pap y categoriza las muestras de acuerdo con el nivel de anomalía. Este sistema fue diseñado para buscar las anomalías lámina a lámina, y para excluir el 25% de las láminas con el riesgo más bajo. Estas láminas eran automáticamente excluidas de la lista que requería revisión microscópica manual y así reducir el trabajo de los examinadores un 25% [10]. Además realiza el análisis de las muestras en dos modalidades: el tamizaje primario y el secundario. En el tamizaje primario se persigue diagnosticar con alta confiabilidad un cierto porcentaje de casos como negativos, con la consiguiente reducción de la carga de trabajo de los analistas. Una evaluación de esta técnica [11] revela que aunque su valor predictivo negativo es muy alto, aún aparece una pequeña tasa de falsos negativos.

Durante los primeros años del siglo XXI, se ha adoptado como método para preparar las muestras la *citología de base líquida (CBL)* debido a que resuelve cinco problemas fundamentales [2]: 1) falta de captura de la totalidad de la muestra, 2) fijación deficiente, 3) distribución aleatoria de células anómalas, 4) existencia de elementos perturbadores y 5) mala calidad del frotis. Esta técnica incrementa la sensibilidad de lesiones intraepiteliales de alto grado y carcinomas, sin variación en la especificidad.

***BD FocalPoint™ Slide Profiler:*** Es una herramienta de evaluación automatizada de citologías cervicales que está actualmente disponible. Está diseñada para el uso en el tamizaje inicial de láminas de citología cervical. El FocalPoint identifica hasta un 25% de láminas exitosamente procesadas que no necesitan revisión adicional. El FocalPoint también identifica al menos 15% de las láminas exitosamente procesadas para una segunda revisión manual. Esta herramienta está diseñada para ser usada con preparaciones convencionales y láminas de citología cervical SurePath (anteriormente AutoCyte®PREP). En ambos métodos de preparación, esta herramienta detecta láminas con evidencia de carcinoma de células escamosas y adenocarcinoma y sus condiciones precursoras usuales. [10]

***ThinPrep Imaging System:*** Acrecienta la productividad del analista humano hasta más de 300 láminas por día, con un incremento significativo de la sensibilidad, pero al mismo tiempo con cierta reducción de la especificidad. Está diseñada para citología en base líquida y para ser teñidas con *Hologic Imager Stain*, escanea toda la preparación y selecciona 22 campos que son los que visualiza el lector que hace la revisión. Si se sospecha alguna anomalía en alguno de los 22 campos, se realiza un tamizaje completo de la preparación. Este sistema no asigna categorías a las preparaciones y es capaz de seleccionar las que no tienen anomalías y no requieren revisión, o seleccionar las que requieren control de calidad o revisión. [12]

Todos estos sistemas se encuentran instalados en varios laboratorios de EE. UU.

### 1.1.3. Características de las imágenes celulares

Las imágenes celulares al igual que cualquier otra imagen poseen características propias (rasgos) que permiten la catalogación o clasificación de las mismas de acuerdo a la información que brindan. Para la catalogación de dichas imágenes, es necesaria la realización de una serie de técnicas que apoyan a los clasificadores humanos o automáticos a poner en claro diagnósticos referentes a enfermedades o incertidumbres existentes de acuerdo a cada caso.

Idealmente las muestras son tomadas de varias áreas del cuello uterino. Las muestras a menudo contienen células del epitelio columnar y el epitelio escamoso. El epitelio columnar se encuentra en la parte superior del cuello del útero, y el epitelio escamoso en la parte inferior. Entre estos dos está el epitelio metaplásico, también conocido como región de transformación o de la unión escamo-columnar.

En el epitelio escamoso hay cuatro capas de células. Las células se forman en la capa basal y mientras se realiza la maduración se mueven hacia arriba a través de la capa parabasal, la capa intermedia, y finalmente la capa superficial. Las células en la capa basal se dividen y administran células a las capas superiores, mientras que las células maduran y se mueven a través de las capas, que cambian de forma, color y otras características. Cuando las células alcanzan la capa superficial son rechazadas y reemplazadas por las células que vienen de abajo. La capa basal tiene pequeñas células redondas con un núcleo relativamente grande y citoplasma pequeño. Cuando maduran, el núcleo se hace más pequeño y el citoplasma se hace más grande. La forma de las células es menos redondeada a medida que aumenta su maduración. [13]

El epitelio columnar solo contiene una única capa de células que posee células columnares y células de reserva. Las células de reserva se dividen en nuevas células de reserva y nuevas células columnares. En condiciones normales de células epiteliales columnares, el núcleo se encuentra en la parte inferior del citoplasma. Cuando se ve desde la parte superior, el núcleo parece más grande. Cuando se ve desde el lado, el citoplasma parece más grande.

El epitelio metaplásico está formado por células de reserva del epitelio columnar. Cuando las células han madurado completamente en el epitelio metaplásico, se parecen a las células que se encuentran en el epitelio escamoso (Figura 5).

En las células displásicas, la información genética es cambiada de algún modo, y la célula no se divide como debería. Esta es una célula precancerosa. Dependiendo al tipo de célula que se divide incorrectamente, se le da diagnósticos como la displasia y carcinoma in situ. El término "plasia" significa crecimiento medio,

y "displasia" significa crecimiento desordenado. Las células displásicas se dividen en displásicas leves, moderadas y graves. La clasificación se determina a partir de la probabilidad de que las células posteriormente se conviertan en células cancerosas malignas. Una cantidad elevada de las células displásicas leves desaparecerán sin llegar a ser malignas, mientras que las células displásicas graves probablemente se convertirán en células malignas. Las células escamosas displásicas generalmente tienen núcleos más grandes y más oscuros y tienden a unirse en grupos. En la displasia severa, los núcleos son grandes, con gránulos de color oscuro y deformado por lo general. El citoplasma es pequeño y oscuro en comparación con los núcleos. [13]

El carcinoma de cérvix es la neoplasia maligna más común del sistema genital femenino. La carcinogénesis es un proceso de larga duración, que comienza a partir de epitelio normal, que se convierte en displásico, evoluciona a carcinoma in situ, y luego con el cáncer. El intervalo de tiempo entre las etapas, permite la posibilidad de un diagnóstico precoz, con curación completa.

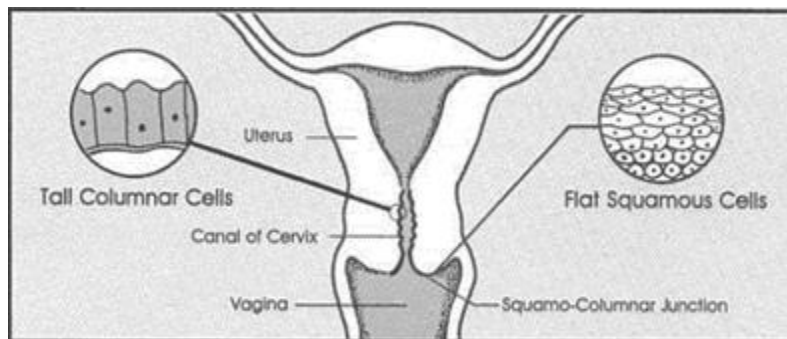


Figura 5: Ubicación de las células columnares y células escamosas. Las células escamosas se desarrollan de abajo hacia arriba de la basal, a la parabasales, la intermedia y la capa superficial. La displasia se desarrolla en la unión escamo-columnar.

#### 1.1.4. Rasgos de las imágenes celulares

Los métodos que se refieren a la clasificación de imágenes de Papanicolaou se basan en el cálculo de las características extraídas de las áreas del núcleo y el citoplasma [14] [15]. Estas características se basan generalmente en características de forma e intensidad de los objetos de interés. Sin embargo, las características calculadas no presentan la misma capacidad de discriminación [16]. Para la determinación del conjunto de características más eficaz que se utiliza como entrada en un clasificador, se han propuesto algunos esquemas de selección de características, y que se refieren a algoritmos genéticos [14] y optimización de nube de partículas [15].

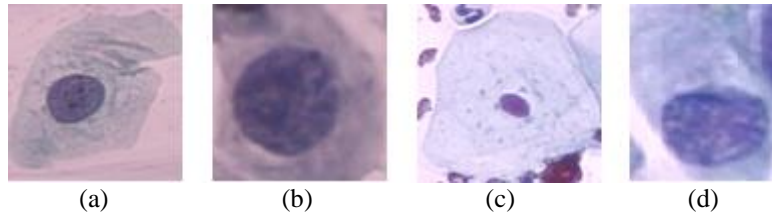


Figura 6: Ejemplo de células de la prueba de Papanicolaou incluidas en la base de datos Herlev: (a)-(b) Células Anómalas y (c)-(d) Células Normales.

En dicha investigación se evalúa la clasificación de las células del cuello uterino, sobre la base de las características extraídas del núcleo. Los experimentos a realizar se basarán en la base de datos Herlev de referencia presentada en [13]. La base de datos consta de 917 imágenes que contienen, cada una, una única célula (Figura 6), y las muestras se distribuyen irregularmente en siete clases. Tres de ellas son consideradas como normales, y las cuatro restantes como anómalas. La descripción detallada de la base de datos se representa en la tabla 1.

Tabla 1: Distribución de las células en la base de datos Herlev, de imágenes celulares de la prueba de Papanicolaou.

<b>NORMAL</b>	<b># DE CÉLULAS</b>
Epitelio Escamoso Superficial	74
Epitelio Escamoso Intermedio	70
Epitelio Columnar	98
<b>TOTAL</b>	<b>242</b>
<b>ANÓMALAS</b>	<b># DE CÉLULAS</b>
Carcinoma in situ	182
Displasia Ligera	146
Displasia Moderada	197
Displasia Severa	150
<b>TOTAL</b>	<b>675</b>

Las imágenes de la base de datos han sido manualmente segmentadas por expertos, y las áreas del núcleo y el citoplasma se definen con precisión. A partir de estas áreas, fueron determinadas veinte características relativas a la intensidad y las características de forma de la región específica tanto del núcleo como del citoplasma (Tabla 2). Nueve de estas veinte características se refieren a la región del núcleo, los que pueden ser calculados de forma independiente.

Tabla 2: Características extraídas de cada imagen en la base de datos [13].

RASGOS DEL CITOPLASMA	RASGOS DEL NÚCLEO
1. Área	1. Área
2. Brillo	2. Brillo
3. Diámetro Corto	3. Diámetro Corto
4. Diámetro más Largo	4. Diámetro más Largo
5. Elongación	5. Elongación
6. Redondez	6. Redondez
7. Perímetro	7. Perímetro
8. Máxima <sup>1</sup>	8. Máxima <sup>1</sup>
9. Mínima <sup>1</sup>	9. Mínima <sup>1</sup>
10. Posición del Núcleo	
11. Tamaño Núcleo/Citoplasma	

<sup>1</sup>El número de píxeles con el valor de intensidad máximo / mínimo en una zona de 3x3 de la región específica.

## 1.2. Procesamiento Digital de Imágenes: Clasificación de objetos en imágenes

El término procesamiento digital de imágenes versa sobre la manipulación y análisis de imágenes por computadora. El procesamiento de imágenes puede considerarse como un tipo especial del procesamiento digital en dos dimensiones, el cual se usa para revelar información sobre imágenes y que involucra hardware, software y soporte teórico.

Se podría decir que la visión artificial describe la deducción automática de la estructura y propiedades de un mundo tridimensional, posiblemente dinámico, o bien a partir de una o varias imágenes bidimensionales de ese mundo. Las imágenes pueden ser monocromáticas (de niveles de gris), a color o multibanda, y pueden provenir de una o varias cámaras e incluso cada cámara puede ser fija o móvil.

El procesamiento digital de imágenes es el conjunto de técnicas que se aplican a las imágenes digitales con el objetivo de mejorar la calidad o facilitar la búsqueda de información. González en [17], modela los pasos fundamentales que se siguen en el procesado de imágenes como se representa en la figura 7.

El primer paso del proceso es la *Adquisición de la imagen*, esto es, digitalizarla. Después de que la imagen digital ha sido obtenida, el siguiente paso es el *Pre-procesamiento*. Su función es mejorar la imagen de manera que se incremente la oportunidad de éxito de los siguientes procesos. El pre-procesamiento típicamente trata con técnicas para realzar el contraste y remover ruido. El siguiente paso es la *Segmentación*. Definida en forma amplia, la segmentación particiona una imagen de entrada en sus partes constituyentes u objetos. Generalmente, la segmentación automática es una de las tareas más difíciles en el

procesamiento de imágenes. En términos del reconocimiento sería, por ejemplo, separar el núcleo del citoplasma y del medio en una imagen celular.

Una vez segmentada la imagen, puede ser necesario extraer características de las distintas regiones de interés seleccionadas para determinar con mayor precisión el nivel de sospecha. El espacio de características puede ser muy grande y complejo, debido a la gran variabilidad que puede darse tanto en el tejido sano como en el canceroso. Sin embargo, no todas las características obtenidas son igualmente significativas. Usar un número excesivo de características puede incrementar la complejidad del clasificador y degradar el rendimiento del mismo. La extracción y selección de características es una etapa clave para la correcta detección de anomalías, ya que el rendimiento del sistema depende más de la optimización de la extracción y selección de características que del método de clasificación.

La descripción, más conocida como *Extracción de los rasgos*, no es más que la extracción de un conjunto de características o descriptores que identifican en forma de un vector de características, al objeto como tal, y que puedan ser utilizadas para su clasificación. Este conjunto de características normalmente ocupa menos espacio de almacenamiento que el objeto que describen, debido a que se almacenan de forma numérica para facilitar el procesamiento y análisis estadístico posterior. Además los descriptores de características aportan mayor información para agrupar y comparar diferentes objetos.

El último estado comprende al reconocimiento y la interpretación (*Clasificación*). El reconocimiento es el proceso que etiqueta, o asigna un nombre, a un objeto basándose en la información que proveen sus descriptores. La interpretación involucra la asignación de significado a un conjunto de objetos reconocido.

La *base de conocimientos* mantiene todo el conocimiento que se tiene acerca del problema tratado. Este conocimiento puede ser simple, teniendo solo las regiones de la imagen donde la información de interés se conoce que está localizada, con lo que se limita la búsqueda que conduce a esa información. La base de conocimientos también puede ser compleja, tal como una lista interrelacionada de todos los defectos mayores posibles en un problema de inspección de materiales o una base de datos de imágenes. Además de guiar la operación de cada módulo del proceso, la base de conocimientos también controla la interacción entre módulos (representado con flechas de doble punta en la figura). Con esto se indica que la comunicación entre módulos del proceso generalmente está basada sobre conocimiento previo de que resultado se podría esperar.

El problema de la clasificación es uno de los primeros que aparecen en la actividad científica y constituye un proceso consustancial con casi cualquier actividad humana. De tal manera, que en la resolución de problemas y en la toma de decisiones la primera parte de la tarea consiste precisamente en clasificar el

problema o la situación. Luego se aplica la metodología correspondiente que en buena medida dependerá de esa clasificación. La base principal de la solución a la problemática abordada en la presente investigación, consiste precisamente en la clasificación de las imágenes celulares en la prueba de Papanicolaou.

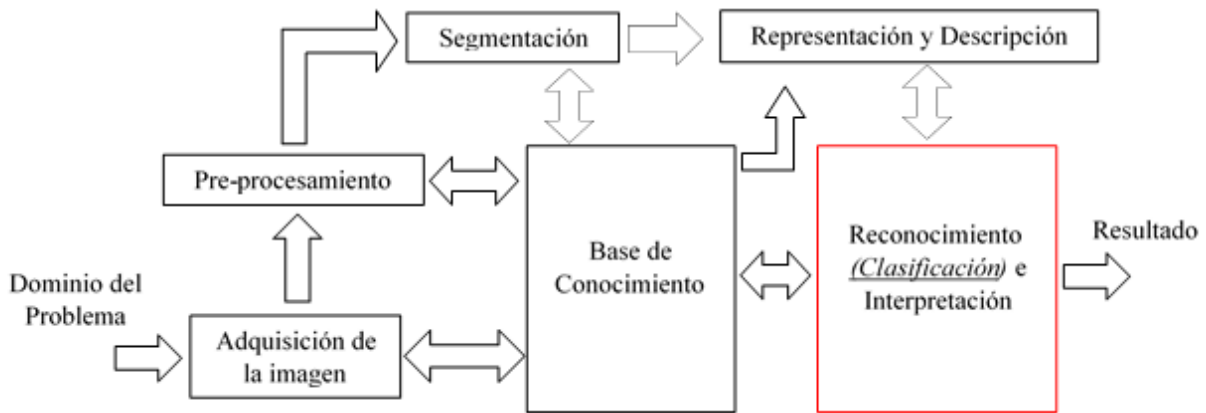


Figura 7: Pasos fundamentales para el procesamiento de imágenes.

La clasificación puede ser tratada como dos problemas por separado: la clasificación binaria y la multiclases. En la clasificación binaria solo están involucradas dos clases, mientras que en la multiclases un objeto puede ser asignado a una de varias clases.

De acuerdo a [18] la clasificación es una tarea que consiste en asignar una categoría o clase a cada elemento de un conjunto. Es decir, dado un conjunto de elementos  $E$  y un conjunto de clases  $C$ , se dice que un problema es de clasificación si consiste en encontrar la función  $M : E \rightarrow C$ , que asigne una clase  $c \in C$  para cada elemento  $e \in E$ .

Un tipo particular de problemas de clasificación es la clasificación binaria. En la clasificación binaria, solo existen dos categorías o clases que se pueden asignar a los elementos del conjunto  $E$ . La clasificación binaria es considerada el tipo más simple de clasificación. Muchos problemas de la vida real pueden ser modelados como problemas de clasificación binaria. Un ejemplo es la clasificación de las células en normales y anómalas en la prueba de Papanicolaou.

### 1.2.1. Componentes de un sistema de clasificación

En la resolución de problemas de toma de decisiones, la primera parte de la tarea consiste en clasificar el problema o la situación, para después aplicar la metodología correspondiente, que en buena medida dependerá de esa clasificación.



Cuando se diseña un sistema de clasificación es necesario definir cada uno de sus componentes. Los principales componentes de un sistema de clasificación general según [19] son:

- **Clase:** Conjunto de entidades que comparten alguna característica que las diferencia de otras.
- **Clase de rechazo:** Conjunto de entidades que no se pueden etiquetar como ninguna de las clases del problema.
- **Extractor de características:** Subsistema que extrae información relevante para la clasificación a partir de las entidades cuantificables.
- **Clasificador:** Subsistema que utiliza un vector de características de la entidad cuantificable y lo asigna a una de las  $M$  clases.
- **Evaluación de error de la clasificación:** Estimación del error que el clasificador comete en el conjunto de prueba.
- **Falso rechazo y Falsa aceptación:** También conocidos como falso negativo y falso positivo respectivamente, es el error producido cuando el sistema diagnostica una clase siendo la otra la correcta. Para problemas con dos clases, estas definiciones reflejan la importancia de una decisión contra la opuesta.

### 1.2.2. Clasificadores de datos, Técnicas de Clasificación y Reglas de clasificación

#### Clasificadores de Datos:

Cada elemento  $e$  de un conjunto de datos  $E$  puede ser representado como un vector  $\langle e_1; e_2; \dots; e_n \rangle$  de  $n$  atributos. Cada atributo  $e_i$  representa una característica específica del elemento  $e$  (Tabla 3). Estas características se representan mediante valores numéricos o categorizados. Cada vector de atributo  $e \in E$ , tiene una clase  $c \in C$  asociada. La clasificación consiste en la predicción del valor de la clase  $c \in C$ , dado un vector de atributos  $e \in E$ . Los problemas de clasificación pueden resolverse encontrando una función  $M(e) = c$ . [20]

Tabla 3: Ejemplo de relación entre los elementos con sus respectivos atributos y clases.

<i>Elemento</i>	<i>Atributo<sub>1</sub></i>	<i>Atributo<sub>2</sub></i>	<i>...</i>	<i>Atributo<sub>n</sub></i>	<i>Clase</i>
1	6	<i>Si</i>	<i>...</i>	360	<i>Clase<sub>1</sub></i>
2	4	<i>No</i>	<i>...</i>	240	<i>Clase<sub>2</sub></i>
3	7	<i>No</i>	<i>...</i>	670	<i>Clase<sub>1</sub></i>

### Técnicas de Clasificación:

Existen varias técnicas de clasificación. Cada una de ellas utiliza un conjunto de parámetros para caracterizar cada elemento; además, cuentan con un conjunto de clases a los que un elemento puede ser categorizado. Si el conjunto de parámetros es muy grande, existen tantas combinaciones de parámetros que encontrar la mejor solución mediante búsqueda exhaustiva se vuelve computacionalmente intratable. Por lo tanto, en la práctica se utilizan métodos de clasificación que utilizan heurísticas para encontrar soluciones suficientemente buenas. [21] [22]

Las técnicas de clasificación suelen dividirse en diferentes tipos de categorías como:

- *Aprendizaje supervisado*: funciona con ejemplos etiquetados previamente, por lo que a priori se conocen las clases existentes, luego, dado un conjunto de vectores de características intenta clasificarlos en una de esas clases.
- *Aprendizaje no supervisado*: consiste en el aprendizaje a partir de patrones de entrada sin valores o clases determinadas. El principal problema de esta técnica es cómo tomar una decisión entre los patrones dados. El sistema toma los objetos de entrada como un conjunto de variables aleatorias, construyendo de esa manera un modelo de densidad para el conjunto de datos. Esta técnica es conocida como clustering.
- *Aprendizaje semi-supervisado*: se basa en técnicas que combinan las dos anteriores, esto se debe a que en ocasiones resulta muy difícil clasificar todos los datos. La idea es combinar los datos etiquetados y no etiquetados para obtener un modelo de clasificación. Esta técnica toma importancia cuando el sistema diseñado tiene acceso a un número limitado de datos etiquetados.

En la presente investigación se prevé utilizar el aprendizaje supervisado, debido a que ya se poseen las clases establecidas (normal y anómala) y las imágenes con las que se entrenarán los clasificadores definidos ya están clasificadas en la base de datos a utilizar (Herlev).

Varios métodos han sido propuestos para la clasificación de las células en las imágenes de la prueba de Papanicolaou y que se refieren a las técnicas de aprendizaje supervisado tales como clasificadores bayesianos, redes neuronales artificiales (ANN), máquinas de soporte vectorial (SVM) y los clasificadores basados en los vecinos más cercanos (*kNN*). Debe tenerse en cuenta que la mayoría de estos métodos utilizan imágenes pre-segmentadas que contienen solo una célula, por lo que la segmentación correcta del núcleo y el citoplasma es factible. [16]

- **Clasificador Bayesiano:** es un clasificador de patrones basado en las teorías estadísticas de aprendizaje. Se calcula la probabilidad de cada hipótesis de los datos y se realiza la predicción a partir de esta, basándose en la teoría de decisión de Bayes. Se destacan las técnicas Nearest Neighbor y Naive Bayes. [23]
- **Redes Neuronales Artificiales (ANN, por sus siglas en inglés):** Según [24] una ANN es una estructura compuesta de un número de unidades interconectadas (neuronas artificiales). En las redes neuronales artificiales, la unidad análoga a la neurona biológica es el elemento procesador, (PE, process element). Un elemento procesador tiene varias entradas y las combina, normalmente con una suma básica. La suma de las entradas es modificada por una función de transferencia y el valor de la salida de esta función de transferencia se pasa directamente a la salida del elemento procesador. La salida del PE se puede conectar a las entradas de otras neuronas artificiales (PE) mediante conexiones ponderadas correspondientes a la eficacia de la sinapsis de las conexiones neuronales. Algunos ejemplos de las ANN son: el perceptrón simple, el perceptrón multicapa, los mapas auto-organizados, también conocidos como redes de Kohonen. [25]
- **Clasificadores de Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés):** SVM crea un hiperplano de separación óptima a partir de un conjunto de ejemplos de entrenamiento. Puede ser lineal o no lineal dependiendo de la función de optimización utilizada (kernel function), por lo que se puede decir que es un caso específico de un clasificador lineal, sin embargo su uso más común suele ser cuando el conjunto de datos no es linealmente separable. Es considerado más fácil de usar que las redes neuronales y su meta es crear un modelo a partir de los datos de entrenamiento. Entre los kernels más usados se encuentran el lineal, el polinomial, el sigmoideal y la función de base radial. [23]
- **Vecinos más cercanos (kNN por sus siglas en inglés):** está basado en un principio de distancia. La distancia está dada por la similitud de atributos entre elementos. En base a este principio de distancia, los elementos se encuentran próximos o apartados de otros elementos. Si todos los elementos están clasificados, entonces la clase de un nuevo elemento sin clasificar puede determinarse observando las clases de sus vecinos más próximos. [20]

### Reglas de Clasificación:

Las reglas de clasificación están compuestas por expresiones. Estas expresiones relacionan los atributos de cada elemento de un conjunto de datos. Estas reglas utilizan operadores lógicos, de comparación y aritméticos. El uso en conjunto de estos operadores hace posible la representación de un lenguaje más potente y expresivo que los obtenidos por las demás técnicas. Mediante el uso de gramáticas formales, se hace posible la definición de la estructura de estas reglas, como se muestra en el Ejemplo 1. [20]

$$Atributo_1, Atributo_2, Constante_3 \in \mathbb{R}$$

$$Atributo_5 \in \{\text{verdadero}, \text{falso}\}$$

Si  $\left( (Atributo_1 \div Atributo_2) < Constante_3 \right) \vee Atributo_5$  es verdadero entonces

$Clase_1$  si no  $Clase_2$

Ejemplo 1: Reglas de Clasificación.

### 1.2.3. Medidas de evaluación

La toma de decisiones clínicas es un proceso extremadamente complejo en el que deberá finalmente ser valorada la utilidad para el manejo del paciente de cualquier prueba diagnóstica. En este contexto, es imprescindible conocer detalladamente la exactitud de las distintas pruebas diagnósticas, es decir, su capacidad para clasificar correctamente a los pacientes en categorías o estados en relación con la enfermedad, típicamente dos, por ejemplo: estar o no estar enfermo o manifestar una respuesta positiva o negativa a la terapia. [26]

La tabla 4 refleja los posibles resultados de un proceso de clasificación binaria, ésta se analiza para comprender la importancia de la evaluación de la calidad en los clasificadores.

Tabla 4: Proceso de clasificación binaria.

		Grupo Verdadero	
		Anómalas $x_i = 1$	Normal $x_i = 0$
Detectadas como Y	Anómalas $y_i = 1 (+)$	<b>a</b>	<b>b</b>
	Normal $y_i = 0 (-)$	<b>c</b>	<b>d</b>

En la tabla mostrada se tienen las siguientes definiciones:

- a = cantidad de ocurrencias de  $y_i = 1 \mid x_i = 1$  (células anómalas detectadas como anómalas), conocidas como verdaderos positivos.
- b = cantidad de ocurrencias de  $y_i = 1 \mid x_i = 0$  (células normales detectadas como anómalas) conocidas como falsos positivos.
- c = cantidad de ocurrencias de  $y_i = 0 \mid x_i = 1$  (células anómalas detectadas como normales), conocidas como falsos negativos.
- d = cantidad de ocurrencias de  $y_i = 0 \mid x_i = 0$  (células normales detectadas como normales), conocidas como verdaderos negativos.

A partir de estas definiciones se pueden calcular diferentes indicadores de desempeño como la sensibilidad, la especificidad, la predictividad positiva, la predictividad negativa y la tasa de clasificación correcta.

La sensibilidad ( $S_e$ ) también conocida como *Recall* es la capacidad que tiene una prueba diagnóstica para detectar los que tienen la condición buscada en una población (verdaderos positivos) [26]. Se define como:

$$S_e = p(Y = 1|X = 1) = 1 - p(Y = 0|X = 1) = \frac{a}{a + c} = 1 - \frac{c}{a + c} \quad (1)$$

La especificidad ( $S_p$ ) es la capacidad que tiene una prueba para detectar los que no tienen la condición buscada en una población (verdaderos negativos) y se calcula como:

$$S_p = p(Y = 0|X = 0) = 1 - p(Y = 1|X = 0) = \frac{d}{b + d} = 1 - \frac{b}{b + d} \quad (2)$$

La predictividad positiva ( $P_p$ ) también conocida como *Precision* es la es la probabilidad de que cuando la prueba es positiva, corresponda a un verdadero positivo. Se define como:

$$P_p = p(X = 1|Y = 1) = 1 - p(X = 0|Y = 1) = \frac{a}{a + b} = 1 - \frac{b}{a + b} \quad (3)$$

La predictividad negativa ( $P_n$ ) o tasa bayesiana negativa es la probabilidad de que cuando la prueba es negativa, corresponda a un verdadero negativo. Se calcula como:

$$P_n = p(X = 0|Y = 0) = 1 - p(X = 1|Y = 0) = \frac{d}{c + d} = 1 - \frac{c}{c + d} \quad (4)$$

La tasa de clasificación correcta ( $R_c$ ) también conocida como *Accuracy* es la proporción del número total de casos que fueron clasificados correctamente.

$$R_c = D_{sm} = p(X = Y) = 1 - p(X \neq Y) = \frac{a + d}{a + b + c + d} = 1 - \frac{b + c}{a + b + c + d} \quad (5)$$

En resumen:

Si disminuye ‘c’ (falsos negativos), la sensibilidad y la predictividad negativa van a aumentar, por otro lado, si disminuye ‘b’ (falsos positivos), la especificidad y la predictividad positiva también aumentarían. Y, por consiguiente la clasificación general sería mejor.

La eficacia máxima se consigue cuando este es capaz de procesar un conjunto de datos sin producir ningún falso positivo o falso negativo. Esto se traduce en que las células cervicales detectadas son clasificadas correctamente en un 100% y no debe quedar ninguna anomalía sin detectar. En la literatura, además de las medidas de evaluación ya mencionadas, se abordan otras que permiten medir la eficacia de estos algoritmos [27].

La media armónica es una de ellas y es utilizada por los investigadores para la evaluación de los algoritmos de clasificación. En este caso están la medida  $F_1$  ( $F - measure$ ) que es media armónica entre  $P_p$  y  $S_e$  (predictividad positiva y sensibilidad), y la media  $H$  ( $H - mean$ ) que es la media armónica entre  $S_e$  y  $S_p$  (sensibilidad y especificidad): [28]

$$F - measure = \frac{2 \times (P_p \times S_e)}{P_p + S_e} \quad (6)$$

$$H - mean = \frac{2 \times (S_p \times S_e)}{S_p + S_e} \quad (7)$$

### 1.3. Herramientas y librerías de procesamiento de imágenes

#### 1.3.1. ImageJ

ImageJ es un programa de procesamiento de imagen digital de dominio público programado en Java desarrollado en el National Institutes of Health. ImageJ fue diseñado con una arquitectura abierta que proporciona extensibilidad vía plugins Java y macros (macroinstrucciones) grabables. Se pueden desarrollar plugins de escaneo personalizado, análisis y procesamiento usando el editor incluido en ImageJ y un compilador Java [29]. Los plugins escritos por usuarios hacen posible resolver muchos problemas de procesado y análisis de imágenes, desde imágenes en vivo de las células en tres dimensiones, procesado de imágenes radiológicas, comparaciones de múltiples datos de sistema de imagen, hasta sistemas automáticos de hematología.

ImageJ puede ejecutarse en cualquier sistema operativo, Linux, Windows, Mac OS X, como aplicación ejecutable, o en cualquier computadora con Máquina virtual Java 5 o superior. El código fuente de ImageJ está disponible gratuitamente. Puede mostrar, editar, analizar, procesar, guardar, e imprimir imágenes de 8 bits (256 colores), 16 bits (miles de colores) y 32 bits (millones de colores). Puede leer varios formatos de imagen incluyendo TIFF, PNG, GIF, JPEG, BMP, DICOM, FITS, así como formatos RAW. ImageJ soporta pilas o lotes, una serie de imágenes que comparten una sola ventana, y es multiproceso, de forma que las operaciones que requieren mucho tiempo se pueden realizar en paralelo en hardware multi-CPU. ImageJ puede calcular el área y las estadísticas de valor de píxel de selecciones definidas por el usuario y la intensidad de objetos umbralados (thresholded objects). Puede medir distancias y ángulos. Se puede crear histogramas de densidad y gráficos de línea de perfil. Es compatible con las funciones estándar de procesamiento de imágenes tales como operaciones lógicas y aritméticas entre imágenes, manipulación de contraste, convolución, análisis de Fourier, nitidez, suavizado, detección de bordes y filtrado de mediana. Hace transformaciones geométricas como ampliar, rotar y flips. [30]

### 1.3.2. Fiji

Es una plataforma que permite compartir aplicaciones para mejorar y avanzar en el procesamiento y análisis de imágenes biomédicas en código libre, se ha construido sobre los cimientos de una plataforma anterior (ImageJ). [31]

Fiji es un paquete de procesamiento de imágenes. Esto puede ser descrito como una distribución de ImageJ junto con Java, Java 3D y un conjunto de plugins organizados en una estructura de menú coherente y ofrece una amplia documentación. El objetivo principal de Fiji es ayudar a la investigación en ciencias de la vida.

Para los usuarios, Fiji es fácil de instalar y tiene una función de actualización automática y para los desarrolladores, Fiji es un proyecto de código abierto alojado en un repositorio de control de versiones Git, con acceso al código fuente de todas las partes internas, librerías y plugins, y facilita el desarrollo y la secuencia de comandos de plugins.

### 1.3.3. Matlab

Matlab es una herramienta de programación cuyo nombre proviene de la abreviación de “MATrix LABoratory”. Esta herramienta permite realizar cálculos numéricos con vectores y matrices. Como caso particular puede trabajar con números escalares reales o complejos, con cadenas de caracteres y con otras estructuras de información más complejas. Una de las capacidades más atractivas es la de realizar una amplia variedad de gráficos en dos y tres dimensiones. MATLAB es un programa de cálculo técnico y científico. Dispone de un código básico y de varias bibliotecas especializadas (toolboxes). [32]

El toolbox de procesamiento de imagen es una colección de funciones implementadas en Matlab. Comprenden una extensa colección de operaciones para el tratamiento digital de imágenes, éstas van desde la representación de imágenes hasta el filtrado, análisis o transformación de las mismas.

### 1.3.4. WEKA

El sitio oficial de Weka, describe la herramienta como una colección de algoritmos de aprendizaje automático para la tarea de extracción de datos. Los algoritmos pueden ser aplicados directamente a un conjunto de datos o ser llamados desde código java (usando el JDK de java y un IDE). Weka contiene herramientas para el pre-procesamiento, clasificación, regresión, clustering, reglas de asociación y visualización de datos. También es muy adecuado para el desarrollo de nuevos sistemas de aprendizaje de máquinas. Weka dispone de versiones developer, que permiten desde un IDE de Java hacer personalizaciones por código y en general gozar de las bondades típicas del software libre. [33]

Weka está compuesta por una serie de herramientas gráficas de visualización y diferentes algoritmos para el análisis de datos y modelado predictivo. Su interfaz gráfica de usuario facilita el acceso a sus múltiples funcionalidades. Esta potente herramienta de minería de datos se encuentra libremente disponible bajo la licencia pública general de GNU, además, al estar implementada en Java, puede ejecutarse prácticamente bajo cualquier entorno.

La interfaz gráfica de Weka cuenta con 4 formas de acceso a las diferentes funcionalidades de la aplicación. [34]

- Simple CLI (Simple command-line interface), permite el acceso a través de una consola de comandos a todas las opciones de Weka.
- Explorer, es la opción más intuitiva para el usuario, pues dispone de varios paneles que dan acceso a las principales características del programa.
- Experimenter, permite la comparación sistemática de una ejecución de los algoritmos predictivos de Weka sobre una colección de conjuntos de datos.
- Knowledge Flow, soporta esencialmente las mismas opciones que la interfaz.

### 1.4. Conclusiones Parciales

En el presente capítulo fueron expuestas las principales características de la prueba de Papanicolaou así como de su forma de realización tanto manual como a nivel de equipos automáticos. Se mencionan las características fundamentales de las células que se evalúan en este tipo de prueba, y de los sistemas diagnósticos utilizados (indicadores utilizados y forma de evaluación de la clasificación). Como conclusión se puede decir que uno de los pasos fundamentales para que la clasificación (la cual es el fin de dicha investigación) tenga buenos resultados es la extracción de los rasgos a utilizar. El empleo de métodos o algoritmos de clasificación para la realización de pruebas a células cervicales pudiera contribuir a la disminución de falsos negativos, además de la ayuda que brindarían los mismos a los especialistas que llevan a cabo este tipo de prueba.



# Capítulo 2

## Metodología de análisis, clasificación y diagnóstico

En el presente capítulo se realiza un análisis del proceso de clasificación así como una breve descripción de los algoritmos propuestos a utilizar. Se comparan los algoritmos de clasificación internamente, con el fin de encontrar el/los parámetros más eficientes en la clasificación de células cervicales, haciendo énfasis en los resultados obtenidos a través de la medida de evaluación predictividad negativa.

### 2.1. Descripción de la solución propuesta

La necesidad de agilizar y aumentar la calidad de las pruebas citológicas, sirve de motivación para la construcción de una herramienta de apoyo a la detección temprana de posibles enfermedades en el cérvix uterino. Mediante el uso de algoritmos de clasificación, y comparando una base de entrenamiento con una matriz de rasgos extraídos de las imágenes a evaluar, se puede obtener una clasificación de las células obtenidas a través de la prueba citológica.

Una muestra recogida en dichas pruebas contiene varias células en sí misma, las cuales se solapan unas a otras interfiriendo en la delimitación de la región del citoplasma perteneciente a una célula u otra, no siendo así en el caso del núcleo, que, como se muestra en las figuras 8(a) y 8(b) es posible identificarlo. De hecho, la región del núcleo celular resulta más factible de ser segmentada con precisión mediante algoritmos de procesamiento digital de imágenes, siendo este pre-procesamiento un tema no abordado dentro de esta tesis. Por tales motivos, a pesar de que la entrada de datos a la herramienta de clasificación a desarrollar no restringirá la cantidad de rasgos, y atendiendo a investigaciones realizadas por [16] [35] se tiene concebido que las pruebas y muestras a utilizar en la validación de la herramienta sean con los rasgos extraídos de la región del núcleo.

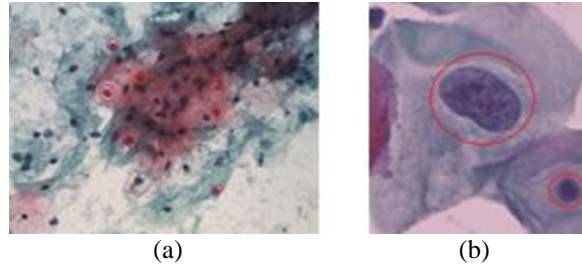


Figura 8: Células solapadas con identificación del núcleo.

### 2.1.1. Caracterización de las fases del proceso

En la figura 9 se muestra el diagrama de flujo de la solución propuesta. Dicha solución está compuesta por 2 fases en las cuales las matrices de rasgos (base de entrenamiento ( $m \times n$ )) y (prueba ( $m \times n$ )) son validadas (estructuralmente) para que puedan ser utilizadas por el clasificador, donde  $m$  son los casos y  $n$  los rasgos asociados a los casos. Luego de validada la estructura interna de dichas matrices de rasgos, comienza la fase de clasificación la cual brinda como resultado/salida una matriz de resultado ( $m \times 1$ ). Más adelante se caracterizarán los distintos algoritmos empleados en la etapa de clasificación.

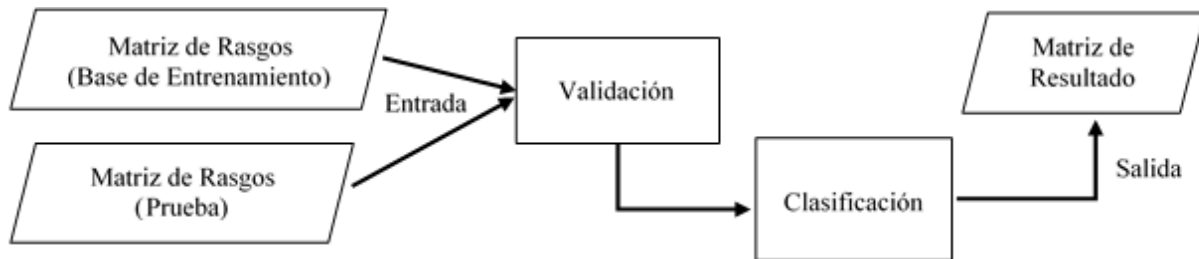


Figura 9: Diagrama de flujo de la solución propuesta.

En la fase de validación se propone como entrada que las matrices de rasgos utilizadas tanto como base de entrenamiento, así como los ficheros de prueba, posean una estructura interna de la siguiente manera (Figura 10):

```
1 1:1410 2:139.882251 3:38.49138889 4:46.8209946 5:0.905531895
1 1:1948 2:166.6101731 3:42.96818515 4:57.88582281 5:0.881852165
1 1:1487 2:143.882251 3:39.32615747 4:48.27557167 5:0.902622978
1 1:1623 2:148.1248917 3:42.35424047 4:48.86739614 5:0.929549009
-1 1:4524 2:269.8650071 3:58.63828838 4:98.82672345 5:0.78061935
-1 1:3078 2:275.6639969 3:58.97125237 4:68.24491694 5:0.509000703
-1 1:2267 2:206.7523087 3:43.91737894 4:66.93235594 5:0.666439375
-1 1:3499 2:231.8650071 3:60.70398481 4:74.32903026 5:0.81786832
```

Figura 10: Estructura de la matriz de rasgos (Base de Entrenamiento).

El primer valor representa la clasificación del caso (1 para los casos normales y  $-1$  para los casos anómalos), posteriormente se observa una estructura  $n:r$ , donde  $n$  representa el rasgo que puede variar desde *rasgo 1* ... *rasgo k* siempre de forma ascendente, los  $(:)$  actúan como separador y  $r$  que representa el valor del rasgo extraído de la célula. Así mismo se evidenciaría en el caso del fichero de prueba (Figura 11), con la diferencia de que el valor de la clasificación del caso siempre tomaría el valor 0.

```
0 1:7656.0 2:349.1614716 3:90.73562787 4:108.0589282 5:0.789149269
0 1:4450.0 2:253.2792206 3:66.72134228 4:85.16087607 5:0.87170744
0 1:3496.0 2:240.2081528 3:53.59108946 4:83.52040503 5:0.761387602
0 1:4979.0 2:268.6934342 3:67.66345652 4:93.94496702 5:0.866638308
0 1:2614.0 2:209.3797257 3:44.62966431 4:75.39764428 5:0.749283544
```

Figura 11: Estructura de la matriz de rasgos (Fichero de Prueba).

Para llevar a cabo la fase de clasificación es necesario haber obtenido con anterioridad las matrices de rasgos (base de conocimiento) y (fichero de prueba) a clasificar. Posteriormente se utiliza la funcionalidad de clasificar y la herramienta devolverá un fichero que contiene la clasificación (enferma o sana) de las células en el mismo orden en el que se encuentran en el fichero de prueba.

## 2.2. Descripción y metodología de uso de los algoritmos propuestos

Varios métodos han sido propuestos para la clasificación de las células en las imágenes de la prueba de Papanicolaou y que se refieren a las técnicas tales como clasificadores bayesianos [36], redes neuronales artificiales [37], máquinas de soporte vectorial (SVM) [38] y vecinos más cercanos [14]. Debe tenerse en cuenta que la mayoría de estos métodos utilizan imágenes pre-segmentadas que contienen solo una célula, por lo que la segmentación correcta del núcleo y del citoplasma es factible (Figura 12 (a)). En las imágenes que contienen grupos de células (Figura 12 (b)), la detección de la frontera del citoplasma es un problema difícil, y hasta ahora, no se ha encontrado ninguna técnica en la literatura estudiada que mencione algoritmos con buenos resultados en la delimitación de las fronteras de la región del citoplasma. Sin embargo, la detección y segmentación de los núcleos de las imágenes que contienen células superpuestas y agrupaciones de células ha sido abordado con éxito en varios estudios [39] [40].

Los métodos que se refieren a la clasificación de imágenes de Papanicolaou se basan en el cálculo de las características (Tabla 2) extraídas tanto de la región del núcleo como del citoplasma. En el caso de la presente investigación es importante este análisis, debido a que la entrada para el clasificador es una matriz  $[m(\text{casos}) \times n(\text{rasgos})]$  (donde los casos son cada una de las células) con cada uno de los rasgos extraídos de las imágenes. [41]

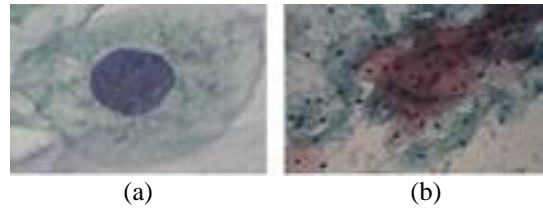


Figura 12: (a) Imagen de una célula simple, (b) imagen de células superpuestas.

En algunas investigaciones como es el caso de [16] se hace uso de los 9 rasgos del núcleo y técnicas como *spectral clustering* y *fuzzy C-means* con reducción de la dimensionalidad, a diferencia de la presente investigación que se dirige hacia la selección de un clasificador sin reducción donde solo se utilizan 5 rasgos de los 9 como son: *el área, perímetro, diámetro corto, diámetro más largo y la redondez*. Esta selección de rasgos persigue demostrar que, a partir de un conjunto simplificado de datos geométricos básicos, es posible realizar de forma efectiva la clasificación binaria de imágenes en la prueba de Papanicolaou [42].

La extracción de rasgos es uno de los pasos fundamentales en el procesamiento de imágenes debido a que mientras mejor sea la selección de los atributos, más acertada será la clasificación final de las células. Esto hace que la adecuada selección de los rasgos sea una de las limitantes en las investigaciones para la clasificación de imágenes, de no contar con los rasgos apropiados los resultados obtenidos por el clasificador no tendrán la calidad que se necesita. Para el caso de la presente investigación, las matrices de rasgos utilizadas fueron extraídas de las imágenes ya previamente segmentadas (Figura 13(b)) pertenecientes a la base de datos Herlev.

Para ello se hizo uso de la herramienta Matlab y de las funciones propias de la misma. De esta forma se desarrolló un algoritmo que toma aleatoriamente de la base de datos Herlev un 80% de las imágenes y extrae de ellas (Figura 13 (b)) los rasgos antes mencionados. A continuación el algoritmo crea una matriz con los rasgos extraídos de las imágenes, la que será utilizada en el entrenamiento del clasificador. Posteriormente el algoritmo extrae los rasgos del 20% de las imágenes restantes y conforma la matriz de rasgos que se utilizará para realizar las pruebas. Estas matrices tienen como característica fundamental que poseen vectores de rasgos diferentes lo que posibilita una evaluación correcta del funcionamiento del clasificador.

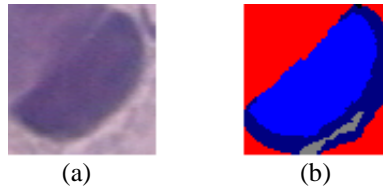


Figura 13: Células anómalas. (a) Imagen RGB, (b) Imagen Segmentada en la cual la región correspondiente al núcleo aparece resaltada en azul claro y a la cual se le realiza el cálculo de los rasgos propuestos.

Para dar cumplimiento al objetivo propuesto en la presente investigación se definió utilizar para la comparación y posterior selección, tres algoritmos de clasificación: las máquinas de soporte vectorial (SVM), la búsqueda por el vecino más cercano ( $kNN$ ) y las redes neurales artificiales (RNA). Con el fin de estimar la capacidad de discriminación de los parámetros internos para el funcionamiento de dichos algoritmos se realizaron una serie de experimentos donde se tuvieron en cuenta los kernels para SVM, las distancias en el caso de  $kNN$  y dos RNA (Perceptrón Multicapa, *RBF Network*). Para esto se utilizó un esquema experimental, validación cruzada de  $k$  particiones (*k-fold cross validation*). Se realizó un particionamiento de los datos de la forma antes descrita y se realizaron siete ejecuciones. Se comprobaron los mismos conjuntos de datos para cada clasificador, tanto en las comparaciones internas de estos algoritmos como en la posterior comparación entre los mismos (Anexos I y II). Para las comparaciones de los resultados en cada algoritmo, de los conjuntos de datos obtenidos en el particionamiento, se utilizaron las tres particiones que mejores resultados mostraron en cuanto a las medidas Pn, área bajo la curva ROC (AUC) y las medias armónicas  $H$  y  $F$ .

### 2.2.1. SVM (Descripción del algoritmo, Funciones utilizadas)

A continuación se dará una breve introducción sobre algunos de los conceptos básicos de la teoría de las Máquinas de Soporte Vectorial (SVM por su nombre en inglés Support Vector Machines). Para más detalles se puede consultar el trabajo de [43]. Como todo clasificador binario, las máquinas de soporte vectorial se basan en encontrar un hiper-plano discriminante que separe de manera óptima dos clases, a partir de un conjunto de elementos positivos y negativos.

La formulación matemática de las SVM varía dependiendo de la naturaleza de los datos; es decir puede ser lineal o no lineal, para esto se emplea una función kernel  $K(x, v)$ . En este caso los vectores de entrenamiento son mapeados en un espacio de mayores dimensiones donde los datos son linealmente separables (Figura 14). Los investigadores se han dedicado a proponer nuevos kernels para estos propósitos, sin embargo existen cuatro básicos que se pueden utilizar fácilmente: el lineal, el polinomial, el sigmoidal y la función de base radial (RBF) [44].

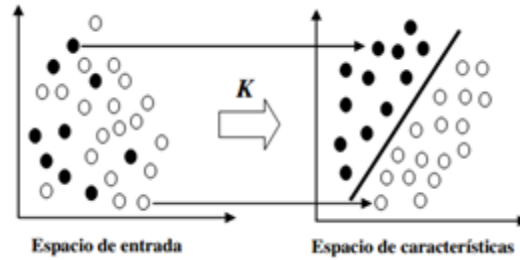


Figura 14: SVM no linealmente separable inducida por una función kernel ( $K$ ).

### Funciones principales utilizadas

Para la presente investigación se utilizó el kernel RBF, sobre la base en primer lugar de que es el más usado para casos que no son linealmente separables. Además se tuvo en cuenta que el número de descriptores o rasgos utilizados no es muy grande, de otro modo RBF no podría ser usado. Según los valores ofrecidos por los indicadores en las tablas 5, 6, 7, 8, 9 y 10, los kernels se mantienen bastante parejos en la clasificación de las imágenes, sin embargo el tiempo de ejecución del algoritmo con el kernel lineal es significativamente mayor en comparación con el RBF. Los resultados de la clasificación haciendo uso del kernel RBF por el clasificador SVM se comportan entre un 81-89% de predictividad negativa con respecto al kernel lineal que mantiene un comportamiento entre un 81-92% de  $P_n$ , de la misma manera que la predictividad positiva y el área bajo la curva ROC se mantienen entre rangos de valores que permiten validar la eficacia del clasificador empleado para cada uno de los conjuntos de datos. Los valores obtenidos por el kernel RBF de acuerdo a las medidas  $F$  y  $H$  se mantienen entre un 91-92% y 88-89% respectivamente, lo que muestra el nivel de efectividad del clasificador.

Tabla 5: Resultados de los índices de efectividad en la clasificación del conjunto de datos 3 para el clasificador SVM.

<i>Kernels</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>RBF</i>	0.894	0.923	0.890	0.923
<i>Lineal</i>	0.90	0.914	0.899	0.912

Tabla 6: Resultado del cálculo de los indicadores de desempeño.

<i>Valores de las variables</i>		
<i>Matriz de Confusión:</i>	<i>Lineal</i>	<i>RBF</i>
	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	125	129
<i>FP</i>	6	8
<i>FN</i>	10	6
<i>VN</i>	42	40
<i>Indicadores de desempeño</i>		
<i>Se</i>	0.93	0.96
<i>Sp</i>	0.88	0.83
<i>Pp</i>	0.95	0.94
<i>Pn</i>	0.81	0.87
<i>Rc</i>	0.91	0.92

Tabla 7: Resultados de los índices de efectividad en la clasificación del conjunto de datos 5 para el clasificador SVM.

<i>Kernels</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>RBF</i>	0.90	0.914	0.899	0.912
<i>Lineal</i>	0.894	0.923	0.890	0.923

Tabla 8: Resultado del cálculo de los indicadores de desempeño.

<i>Valores de las variables</i>		
<i>Detectadas como:</i>	<i>Lineal</i>	<i>RBF</i>
	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	129	125
<i>FP</i>	8	6
<i>FN</i>	6	10
<i>VN</i>	40	42
<i>Indicadores de desempeño</i>		
<i>Se</i>	0.96	0.93
<i>Sp</i>	0.83	0.88
<i>Pp</i>	0.94	0.95
<i>Pn</i>	0.87	0.81
<i>Rc</i>	0.92	0.91

Tabla 9: Resultados de los índices de efectividad en la clasificación del conjunto de datos 7 para el clasificador SVM.

<i>Kernels</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>RBF</i>	0.888	0.922	0.881	0.923
<i>Lineal</i>	0.843	0.903	0.821	0.907

Tabla 10: Resultado del cálculo de los indicadores de desempeño.

<i>Valores de las variables</i>		
<i>Detectadas como:</i>	<i>Lineal</i>	<i>RBF</i>
	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	132	130
<i>FP</i>	14	9
<i>FN</i>	3	5
<i>VN</i>	34	39
<i>Indicadores de desempeño</i>		
<i>Se</i>	0.98	0.96
<i>Sp</i>	0.71	0.81
<i>Pp</i>	0.90	0.94
<i>Pn</i>	0.92	0.89
<i>Rc</i>	0.91	0.92

A pesar de que las pruebas realizadas fueron con conjuntos de datos que contienen 5 rasgos extraídos de los núcleos de las imágenes, se desea que el clasificador no dependa de la cantidad de rasgos. Para ello se hace uso de un algoritmo *Grid Search* que determina los parámetros  $C$  y  $\gamma$  automáticamente utilizando la validación cruzada sobre la matriz de entrenamiento. Luego guarda el mejor valor de cada uno de los parámetros y lo utiliza en el momento de realizar la clasificación sobre las matrices de prueba. Esto posibilita que se puedan utilizar matrices de  $n$  rasgos para la clasificación de las imágenes. A partir de dicho entrenamiento, la herramienta quedó capacitada para llevar a cabo la clasificación de las muestras o conjuntos de datos a clasificar, dando la misma como salida una matriz de rasgos que muestra la clasificación de cada una de las imágenes en (1, -1) es decir (sana, enferma) respectivamente.

Se decidió utilizar como apoyo a la implementación la Librería de las Máquinas de Soporte Vectorial (LibSVM) en su versión 3.18, creada por Chih-Chung Chang and Chih-Jen Lin del Departamento de Ciencias de la Computación de la Universidad Nacional de Taiwan [45].

### 2.2.2. *kNN* (Descripción del algoritmo, Funciones utilizadas)

Las reglas de clasificación por vecindad conocidas como *kNN* (K-Nearest Neighbour) están basadas en un proceso de aprendizaje, el cual consiste en almacenar una tabla con los ejemplos disponibles, junto a la clase asociada a cada uno de ellos. Ante un nuevo ejemplo a clasificar, se calcula su distancia (son utilizadas algunas como: euclidiana, mahalanobis, cityblock o manhattan) que puede ser con respecto a los  $n$  ejemplos existentes en la tabla, y se consideran los  $k$  vecinos más cercanos. El nuevo ejemplo es clasificado según la clase mayoritaria de los  $k$  vecinos más cercanos (Figura 15). El caso más sencillo es cuando  $k = 1$ . Se recomienda que en el momento de indicar  $k$ , la misma sea un número impar para evitar un posible caso de empate.



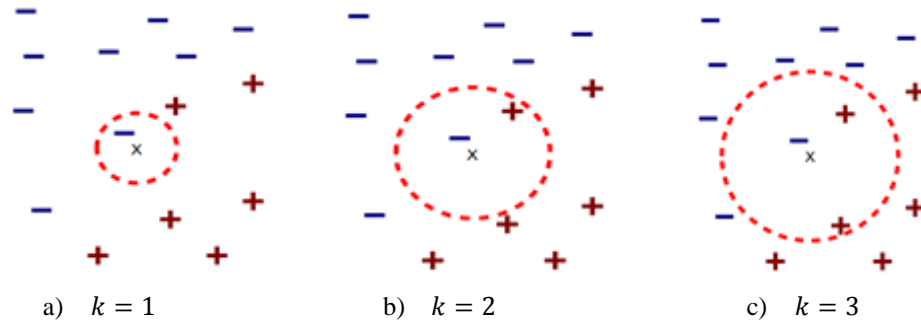


Figura 15: Ejemplo que demuestra la diversidad de posibilidades de clasificación para un vector que depende del tamaño de la  $k$ .

Este método supone que los vecinos más cercanos dan la mejor clasificación y esto se hace utilizando todos los atributos; el problema de dicha suposición es que es posible que se tengan muchos atributos irrelevantes que dominen sobre la clasificación: dos atributos relevantes perderían peso entre otros veinte irrelevantes. Es por esto que existen variantes para mejorar la clasificación haciendo que la misma no dependa únicamente de la cantidad de clases mayoritarias más cercanas a la muestra como el caso que a continuación se describe.

### Vecinos más cercanos con distancia ponderada

Se puede ponderar la contribución de cada vecino de acuerdo a la distancia entre él y el ejemplar a ser clasificado  $x_q$ , dando mayor peso a los vecinos más cercanos. Por ejemplo se puede ponderar el voto de cada vecino de acuerdo al inverso de sus distancias [46]

$$\hat{f}(x_q) \leftarrow \operatorname{argmax}_{v \in V} \sum_{i=1}^k w_i \delta(v, f(x_i)) \quad (8)$$

donde

$$w_i \equiv \frac{1}{d(x_q, x_i)} \quad (9)$$

De esta manera no hay riesgo al permitir que todos los ejemplos de entrenamiento contribuyan en la clasificación de  $x_q$ , ya que los que son muy distantes tienen muy bajo peso asociado. La desventaja de considerar todos los ejemplos sería su lenta respuesta (método global). Se quiere siempre tener un método local en el que solo los vecinos más cercanos son considerados.

Esta mejora es muy efectiva en muchos problemas prácticos. Es robusto ante los ruidos de datos y suficientemente efectivo en conjuntos de datos grandes. Se puede ver que al tomar promedios ponderados de los  $k$  vecinos más cercanos el algoritmo puede evitar el impacto de ejemplos con ruido aislados.

## **Funciones principales utilizadas**

A pesar de que para las pruebas realizadas se utilizaron conjuntos de datos que contienen 5 rasgos extraídos de los núcleos de las imágenes, se desea que el clasificador no dependa de la cantidad de rasgos por lo que se decidió utilizar el algoritmo IBK del Weka. Este permite hacer una validación cruzada interna en el algoritmo utilizando el proceso *hold-one-out* que recorre automáticamente el parámetro  $k$  comparando los valores desde  $k = 1, \dots, 21$  (dichos valores son definidos por los autores de la investigación) en busca del mejor  $k$  [33]. Dicha  $k$  es guardada internamente por el algoritmo para posteriormente utilizarla en las clasificaciones realizadas a las matrices de prueba, lo que posibilita que se puedan clasificar matrices con diferentes cantidades de rasgos. El algoritmo IBK no cuenta con la distancia de Mahalanobis por lo que se le incorpora el cálculo de dicha distancia para poder realizar las comparaciones pertinentes.

De acuerdo a los valores ofrecidos por los indicadores en las tablas 11, 12, 13, 14, 15 y 16, los resultados obtenidos por la distancia Manhattan en la clasificación se comportan entre un 85-93% de predictividad negativa con respecto a las distancias euclidiana y Mahalanobis que se comportaron entre un 87-93% y un 84-90% respectivamente, la predictividad positiva y el área bajo la curva ROC se mantienen entre rangos de valores que permiten validar la eficacia del clasificador con cada uno de los conjuntos de datos. Los valores obtenidos por el clasificador haciendo uso de la distancia Manhattan de acuerdo a las medidas  $F$  y  $H$  se mantienen entre un 91-93% y 87-91% respectivamente, lo que muestra el nivel de efectividad del clasificador. Al comparar los valores obtenidos por la distancia euclidiana y la distancia Manhattan, se observó que estos poseen valores muy similares entre sí y mejores con respecto a la distancia de Mahalanobis. Por otra parte, al comparar los tiempos de ejecución del algoritmo entre la distancia euclidiana y la distancia Manhattan, se observó que el de la primera es significativamente mayor, por lo que se decidió seleccionar el uso de la distancia Manhattan en el clasificador  $kNN$ .

Tabla 11: Resultados de los índices de efectividad en la clasificación del conjunto de datos 3 para el clasificador  $kNN$ .

<i>Distancias</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>Euclidiana</i>	0.95	0.917	0.878	0.918
<i>Mahalanobis</i>	0.908	0.91	0.846	0.913
<i>Manhattan</i>	0.952	0.918	0.887	0.918

## Metodología de análisis, clasificación y diagnóstico

Tabla 12: Resultado del cálculo de los indicadores de desempeño.

		<i>Valores de las variables</i>		
<i>Matriz de Confusión:</i>		<i>Euclidiana</i>	<i>Mahalanobis</i>	<i>Manhattan</i>
		<i># de Imágenes</i>	<i># de Imágenes</i>	<i># de Imágenes</i>
		<i>Reconocidas</i>	<i>Reconocidas</i>	<i>Reconocidas</i>
	<i>VP</i>	129	131	127
	<i>FP</i>	9	12	8
	<i>FN</i>	6	4	7
	<i>VN</i>	39	36	39
	<i>Indicadores de desempeño</i>			
	<i>Se</i>	0.96	0.97	0.95
	<i>Sp</i>	0.81	0.75	0.83
	<i>Pp</i>	0.93	0.92	0.94
	<i>Pn</i>	0.87	0.90	0.85
	<i>Rc</i>	0.92	0.91	0.92

Tabla 13: Resultados de los índices de efectividad en la clasificación del conjunto de datos 5 para el clasificador *kNN*.

<i>Distancias</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>Euclidiana</i>	0.969	0.934	0.913	0.934
<i>Mahalanobis</i>	0.961	0.894	0.837	0.896
<i>Manhattan</i>	0.971	0.934	0.913	0.934

Tabla 14: Resultado del cálculo de los indicadores de desempeño.

		<i>Valores de las variables</i>		
<i>Matriz de Confusión:</i>		<i>Euclidiana</i>	<i>Mahalanobis</i>	<i>Manhattan</i>
		<i># de Imágenes</i>	<i># de Imágenes</i>	<i># de Imágenes</i>
		<i>Reconocidas</i>	<i>Reconocidas</i>	<i>Reconocidas</i>
	<i>VP</i>	129	128	129
	<i>FP</i>	6	12	6
	<i>FN</i>	6	7	6
	<i>VN</i>	42	36	42
	<i>Indicadores de desempeño</i>			
	<i>Se</i>	0.96	0.95	0.96
	<i>Sp</i>	0.88	0.75	0.88
	<i>Pp</i>	0.96	0.91	0.96
	<i>Pn</i>	0.88	0.84	0.88
	<i>Rc</i>	0.93	0.90	0.93

Tabla 15: Resultados de los índices de efectividad en la clasificación del conjunto de datos 7 para el clasificador *kNN*.

<i>Distancias</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>Euclidiana</i>	0.952	0.921	0.862	0.923
<i>Mahalanobis</i>	0.85	0.851	0.697	0.863
<i>Manhattan</i>	0.951	0.927	0.874	0.928

Tabla 16: Resultado del cálculo de los indicadores de desempeño.

<i>Matriz de Confusión:</i>	<i>Valores de las variables</i>		
	<i>Euclidiana</i>	<i>Mahalanobis</i>	<i>Manhattan</i>
	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	132	132	132
<i>FP</i>	11	22	10
<i>FN</i>	3	3	3
<i>VN</i>	37	26	38
<i>Indicadores de desempeño</i>			
<i>Se</i>	0.98	0.98	0.98
<i>Sp</i>	0.77	0.54	0.79
<i>Pp</i>	0.92	0.86	0.93
<i>Pn</i>	0.93	0.90	0.93
<i>Rc</i>	0.92	0.86	0.93

### 2.2.3. Redes Neuronales Artificiales (Descripción del algoritmo, Funciones utilizadas.)

Según [25] las Redes Neuronales Artificiales (RNA por sus siglas en español o ANN (Artificial Neural Networks por sus siglas en inglés) (Figura 16) están inspiradas en las redes neuronales biológicas del cerebro humano. Están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes. Estos elementos están organizados de una forma parecida a la que presenta el cerebro humano. Plantean además que presentan una serie de características propias del cerebro como *Aprender, Generalizar y Abstraer*.

En las RNA, la unidad análoga a la neurona biológica es el elemento procesador, PE (process element). Un elemento procesador tiene varias entradas y las combina, normalmente con una suma básica. La suma de las entradas es modificada por una función de transferencia y el valor de la salida de esta función de transferencia se pasa directamente a la salida del elemento procesador. Una red neuronal consiste en un conjunto de unidades elementales PE conectadas de una forma concreta. El interés de las RNA no reside solamente en el modelo del elemento PE sino en las formas en que se conectan estos elementos procesadores. Generalmente los elementos PE están organizados en grupos llamados niveles o capas. Una red típica consiste en una secuencia de capas con conexiones entre capas adyacentes consecutivas.

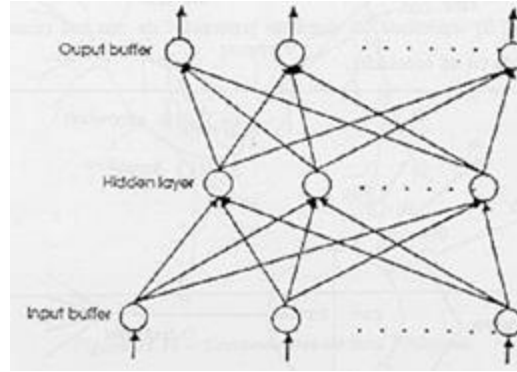


Figura 16: Arquitectura de una Red Neuronal Simple [25].

Para la presente investigación, se utilizaron las RNA: Perceptrón Multicapa y *RBF Network*. Fueron utilizados los algoritmos de la herramienta Weka, para ello se evaluaron los parámetros que más se adecuaban al tipo de clasificación en cuestión de acuerdo a las RNA utilizadas. Luego de realizada dicha evaluación se decidió adoptar en su mayoría los parámetros que por defecto propone la propia herramienta.

#### 2.2.3.1. Perceptrón Multicapa (MLP)

El Perceptrón Multicapa o Multilayer Perceptron (MLP, por sus siglas en inglés) es una de las redes más utilizadas en investigaciones analizadas referentes a la clasificación de células cervicales [47] [48]. Se basa en otra red más simple llamada perceptrón simple solo que el número de capas ocultas puede ser mayor o igual que una. Es una red de alimentación hacia adelante (feedforward) compuesta por una capa de unidades de entrada (sensores), otra capa de unidades de salida y un número determinado de capas intermedias de unidades de proceso, también llamadas capas ocultas porque no se ven las salidas de dichas neuronas y no tienen conexiones con el exterior. Cada sensor de entrada está conectado con las unidades de la segunda capa, y cada unidad de proceso de la segunda capa está conectada con las unidades de la primera capa y con las unidades de la tercera capa, así sucesivamente. Las unidades de salida están conectadas solamente con las unidades de la última capa oculta, como se muestra en la figura 17.

Con esta red se pretender establecer una correspondencia entre un conjunto de entrada y un conjunto de salidas deseadas, de manera que

$$(x_1, x_2, \dots, x_N) \in R^N \rightarrow (y_1, y_2, \dots, y_M) \in R^M \quad (10)$$

Se dispone de un conjunto de  $p$  patrones de entrenamiento, de manera que se sabe que al patrón de entrada  $(x_1^k, x_2^k, \dots, x_N^k)$  le corresponde la salida  $(y_1^k, y_2^k, \dots, y_M^k)$ ,  $k = 1, 2, \dots, p$ . Es decir, se conoce dicha correspondencia para  $p$  patrones. Así, el conjunto de entrenamiento será:

$$\{(x_1^k, x_2^k, \dots, x_N^k) \rightarrow (y_1^k, y_2^k, \dots, y_M^k): k = 1, 2, \dots, p\} \quad (11)$$

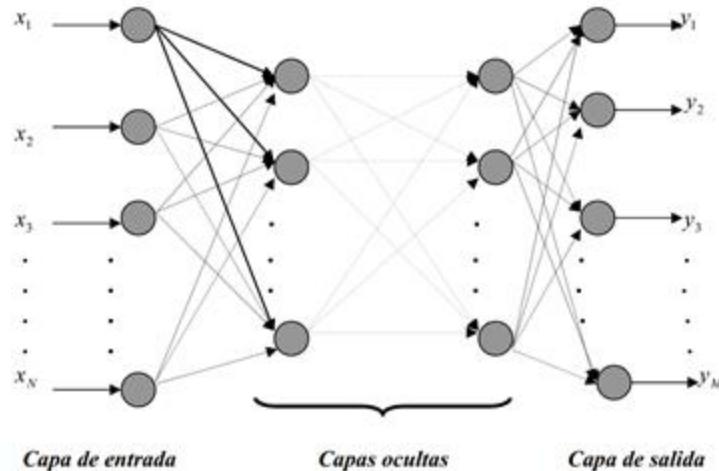


Figura 17: Topología de un Perceptrón Multicapa.

Para implementar dicha relación, la primera capa (sensores de entrada) tendrá tantos sensores como componentes tenga el patrón de entrada, es decir,  $N$ ; la capa de salida tendrá tantas unidades de proceso como componentes tengan las salidas deseadas, es decir,  $M$ , y el número de capas ocultas y su tamaño dependerán de la dificultad de la correspondencia a implementar.

Los valores de los parámetros que intervienen en la red MLP, que brindan mejores resultados y hacen que la red tenga un comportamiento estable por lo que logra la habilidad de clasificar fueron los siguientes:

- Capas ocultas: 1.
- Tasa de aprendizaje: 0,3 (Por defecto).
- *Momentum* (velocidad del proceso de convergencia): 0,2 (Por defecto).
- Tiempo de entrenamiento: 500 (Por defecto).

#### 2.2.3.2. RBF Network (RBF-N)

Las redes de neuronas de base radial han sido aplicadas a una variedad de problemas, aunque es necesario señalar que su aplicación no ha sido tan extendida como en el caso del MLP [49]. Sin embargo se han utilizado en algunas investigaciones [50] [37] para la clasificación de cáncer en las células cervicales.

A diferencia de la disposición que se tiene en las funciones de activación que permite construir modelos de entrenamiento mediante *backpropagation*, las redes basadas en RBF construyen sus modelos con funciones de activación que son diferentes tanto en la capa oculta como en la de salida. Están diseñadas con neuronas activadas en la capa oculta mediante funciones radiales de carácter no lineal con sus centros gravitacionales propios y en la capa de salida mediante funciones lineales.

A diferencia de las MLP, el modelo clásico de las redes RBF está construido con una arquitectura rígida de tres capas: la de entrada, la oculta y la de salida. [51]

En general, una red RBF tiene un mejor desempeño con un mayor volumen de datos de entrenamiento. La construcción de una red de este tipo requiere de una mayor cantidad de neuronas en los nodos ocultos que en las redes que usan *backpropagation*. Aunque las redes RBF no son comúnmente utilizadas en aplicaciones que impliquen un alto volumen de patrones de entrenamiento, se le reconoce como una red con una alta eficacia en la fase de entrenamiento [52].

Los valores de los parámetros que intervienen en la red RBF-N, que brindan mejores resultados y hacen que la red tenga un comportamiento estable por lo que logra la habilidad de clasificar fueron los siguientes:

- *ClusteringSeed*: 5.
- *maxIts*: -1 (Por defecto).
- *minStdDev*: 0,1 (Por Defecto).
- *numClusters*: 4.

Según los valores ofrecidos por los indicadores en las tablas 17, 18, 19, 20, 21 y 22, los resultados de la RNA – *RBF Network* en la clasificación se comportan entre un 86-91% de predictividad negativa con respecto a la MLP que se comportó entre un 85-88%, de la misma manera que la predictividad positiva y el área bajo la curva (ROC) se mantienen entre rangos de valores que permiten validar la eficacia del clasificador empleado para cada uno de los conjuntos de datos. Los valores obtenidos de acuerdo a las medidas *F* y *H* por el *RBF Network* se mantienen entre un 92-93% y 89-91% respectivamente, lo que muestra el nivel de efectividad del clasificador. Al comparar los valores obtenidos al emplear la red neuronal MLP y la RBF-N, se observó que estos poseen valores muy similares entre sí. A pesar de ello se denota en algunas pruebas que RBF-N muestra mejores resultados que MLP por lo que se decide utilizar la RBF-N como clasificador por parte de las RNA para comparar con los clasificadores antes expuestos.

Tabla 17: Resultados de los índices de efectividad en la clasificación del conjunto de datos 3 para el clasificador RNA.

<i>RNA</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>MLP</i>	0.94	0.918	0.887	0.918
<i>RBF-N</i>	0.94	0.928	0.893	0.928

Tabla 18: Resultado del cálculo de los indicadores de desempeño.

<i>Valores de las variables</i>		
<i>Matriz de Confusión:</i>	<i>MLP</i>	<i>RBF-N</i>
	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	128	130
<i>FP</i>	8	8
<i>FN</i>	7	5
<i>VN</i>	40	40
<i>Indicadores de desempeño</i>		
<i>Se</i>	0.95	0.96
<i>Sp</i>	0.83	0.83
<i>Pp</i>	0.94	0.94
<i>Pn</i>	0.85	0.89
<i>Rc</i>	0.92	0.93

Tabla 19: Resultados de los índices de efectividad en la clasificación del conjunto de datos 5 para el clasificador RNA.

<i>RNA</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>MLP</i>	0.969	0.929	0.910	0.928
<i>RBF-N</i>	0.971	0.929	0.910	0.928

Tabla 20: Resultado del cálculo de los indicadores de desempeño.

<i>Valores de las variables</i>		
<i>Matriz de Confusión:</i>	<i>MLP</i>	<i>RBF-N</i>
	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	128	128
<i>FP</i>	6	6
<i>FN</i>	7	7
<i>VN</i>	42	42
<i>Indicadores de desempeño</i>		
<i>Se</i>	0.95	0.95
<i>Sp</i>	0.88	0.88
<i>Pp</i>	0.96	0.96
<i>Pn</i>	0.86	0.86
<i>Rc</i>	0.93	0.93

Tabla 21: Resultados de los índices de efectividad en la clasificación del conjunto de datos 7 para el clasificador RNA.

<i>RNA</i>	<i>Área ROC</i>	<i>Medida F</i>	<i>Media H</i>	<i>% Clasificación</i>
<i>MLP</i>	0.949	0.917	0.868	0.918
<i>RBF-N</i>	0.939	0.933	0.896	0.934



Tabla 22: Resultado del cálculo de los indicadores de desempeño.

<i>Valores de las variables</i>		
	<i>MLP</i>	<i>RBF-N</i>
<i>Matriz de Confusión:</i>	<i># de Imágenes Reconocidas</i>	<i># de Imágenes Reconocidas</i>
<i>VP</i>	130	131
<i>FP</i>	10	8
<i>FN</i>	5	4
<i>VN</i>	38	40
<i>Indicadores de desempeño</i>		
<i>Se</i>	0.96	0.97
<i>Sp</i>	0.79	0.83
<i>Pp</i>	0.93	0.94
<i>Pn</i>	0.88	0.91
<i>Rc</i>	0.92	0.93

## 2.3. Conclusiones Parciales

En el presente capítulo fueron expuestos los parámetros internos que permiten un mejor funcionamiento de los algoritmos de clasificación seleccionados. Se realizó una breve descripción de los procesos que se llevan a cabo en la herramienta de clasificación desarrollada, además de las características estructurales que debe tener la matriz de rasgos a utilizar para la clasificación en la herramienta final. Como conclusión se puede decir que los algoritmos definidos para hacer las pruebas estadísticas en función de cumplir con el objetivo general de la presente investigación han arrojado resultados que demuestran la eficacia de cada uno en la clasificación del dominio que se trabaja y pudieran dar respuesta a la hipótesis planteada.

# Capítulo 3

## Validación Experimental

En el capítulo precedente se han presentado tres modelos de clasificación aplicables a dominios en los que el número de clases es finito e igual a dos y en los que los casos están descritos por un gran número de atributos. El objetivo del presente capítulo, es determinar desde un punto de vista práctico, cuál de las propuestas realizadas en los acápites (2.2.1, 2.2.2, 2.2.3) permite resolver más eficazmente el problema de clasificación de células cervicales. Para ello se presenta un estudio experimental que pretende analizar la viabilidad de los métodos propuestos al considerar los clasificadores seleccionados con igualdad en el número de clases, el número de atributos, el número de ejemplos de entrenamiento, la proporción de casos pertenecientes a cada clase.

Para analizar la calidad de los clasificadores propuestos, los resultados obtenidos en cada uno de los dominios han sido comparados con los derivados de aplicar algunos de los métodos de aprendizaje más usuales en la bibliografía. Dado que los clasificadores usados en este proceso experimental tienen características iguales, la metodología de evaluación, análisis y comparación será en función de los mismos parámetros de medida.

En los siguientes epígrafes se detalla el estudio experimental realizado y para facilitar la lectura, en la Tabla 23 se recogen las principales características asociadas a los clasificadores.

Tabla 23: Principales características de los clasificadores evaluados.

<i>Clasificadores</i>	<i>Casos entrenamiento</i>	<i>Casos de prueba</i>	<i>Clases</i>	<i>Rasgos</i>
SVM (RBF)	734	183	2	5
<i>kNN</i> (Manhattan)	734	183	2	5
RNA (RBF-N)	734	183	2	5

### 3.1. Herramienta de ensayos

Para realizar las pruebas prácticas con las técnicas propuestas se utilizó la herramienta de visión computacional desarrollada y el software WEKA v3.6.8 sobre Sistema Operativo GNU/Linux Ubuntu 14.10 con kernel 3.13.0-24. El hardware utilizado fue una notebook con procesador Intel Core i5 a 1.80 Ghz y 4 Gb de memoria RAM y java 7. Para la extracción de los rasgos de las imágenes de las células se utilizó el software MATLAB R2011a. La herramienta al ser implementada en java es multiplataforma.

### 3.2. Análisis de los resultados

A fin de determinar la calidad relativa de los resultados se realizó una comparación entre los clasificadores anteriormente mencionados. Se utilizó una programación propia escrita en lenguaje Java y se usó además las implementaciones disponibles en el paquete de Minería de Datos WEKA.

#### 3.2.1. Análisis de los resultados de acuerdo a la medida $F$

El objetivo de este estudio es realizar un análisis referente al comportamiento del índice de efectividad conocido como medida  $F$  o media armónica ( $F$ -measure, en la literatura en inglés). La media armónica determina hasta qué punto los resultados obtenidos se asemejan a los que se hubieran logrado con una categorización manual real por cada iteración de prueba. De acuerdo a los datos ofrecidos en la tabla 24 se muestra que el valor de la medida  $F$  se mantuvo por encima del 86% entre los tres clasificadores donde la media prevaleció por encima del 90%. En la corrida de las iteraciones de prueba el clasificador  $kNN$  obtuvo un 93.4% de medida  $F$  (mayor valor obtenido). Como puede observarse en la figura 18, se representa la distribución seguida por los clasificadores a través de las iteraciones de prueba y el por ciento de la media armónica alcanzado en cada uno de ellos.

Tabla 24: Comportamiento de la medida  $F$  en las iteraciones de pruebas realizadas por clasificador.

Clasificadores	<i>Particiones</i>						
	1	2	3	4	5	6	7
<i>SVM (RBF)</i>	0.863	0.866	0.923	<b>0.923</b>	0.914	0.895	0.922
<i>kNN (Manhattan)</i>	0.880	0.872	0.918	0.897	<b>0.934</b>	0.901	0.927
<i>RNA (RBF Network)</i>	<b>0.909</b>	<b>0.884</b>	<b>0.928</b>	0.918	0.929	<b>0.902</b>	<b>0.933</b>

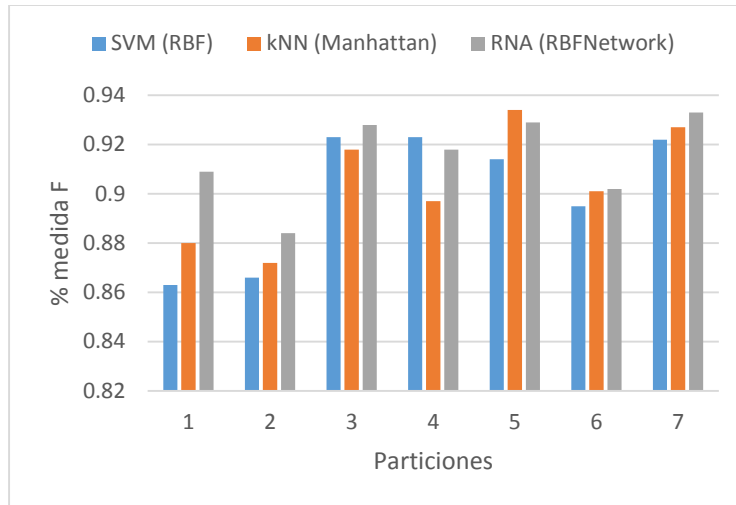


Figura 18: Comportamiento gráfico de la medida  $F$ .

### 3.2.2. Análisis de los resultados de acuerdo a la Predictividad Negativa ( $P_n$ )

El objetivo de este estudio es realizar un análisis referente al comportamiento de la medida de evaluación predictividad negativa ( $P_n$ ). La  $P_n$  es la probabilidad de que el clasificador detecte la célula como normal siendo en realidad el resultado de la prueba diagnóstica negativo. De acuerdo a los datos que se ofrecen en la tabla 25 se muestra que el valor de  $P_n$  se mantuvo por encima del 73% entre los tres clasificadores donde la media prevaleció por encima del 83%. En la corrida de las iteraciones de prueba el clasificador  $kNN$  obtuvo un 93% de  $P_n$  (mayor valor obtenido). En la figura 19, se representa la distribución seguida por los clasificadores a través de las iteraciones de prueba y el por ciento de la predictividad negativa alcanzado en cada uno de ellos.

Tabla 25: Comportamiento de la medida  $P_n$  en las iteraciones de pruebas realizadas por clasificador.

Clasificadores	<i>Particiones</i>						
	1	2	3	4	5	6	7
<i>SVM (RBF)</i>	0.74	0.79	0.87	0.85	0.81	0.82	0.89
<i>kNN (Manhattan)</i>	0.77	0.79	0.85	0.80	0.88	0.83	0.93
<i>RNA (RBF Network)</i>	0.73	0.80	0.89	0.84	0.86	0.81	0.91

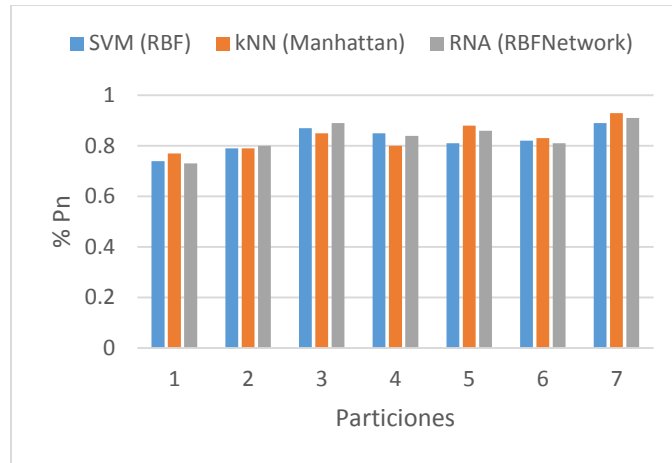


Figura 19: Comportamiento gráfico de la medida Pn.

### 3.2.3. Análisis de los resultados de acuerdo al Área bajo la curva ROC (A-ROC o AUC)

El objetivo de este estudio es realizar un análisis referente al comportamiento de la medida de evaluación área bajo la curva ROC (A-ROC o AUC). El AUC según [53] y [54], se define como la probabilidad de clasificar correctamente un par de individuos (sano y enfermo), seleccionados al azar de la población, mediante los resultados obtenidos al aplicar la prueba diagnóstica. Los datos ofrecidos en la tabla 26 muestran que el valor de la AUC se mantuvo por encima del 81% entre los tres clasificadores donde la media prevaleció por encima del 91%. En la corrida de las iteraciones de prueba el clasificador *kNN* y el *RBF Network* obtuvieron un 97.1% de AUC (mayor valor obtenido). En la tabla se muestra la estabilidad del clasificador *kNN* por encima de SVM y *RBF Network*, donde obtuvo los mayores valores del área bajo la curva ROC en todas las iteraciones de pruebas realizadas. De la misma manera en la figura 20, se representa la distribución seguida por los clasificadores a través de las iteraciones de prueba y el por ciento de la AUC alcanzado en cada uno de ellos.

Tabla 26: Comportamiento de la medida de evaluación AUC en las iteraciones de pruebas realizadas por clasificador.

Clasificadores	<i>Particiones</i>						
	1	2	3	4	5	6	7
<i>SVM (RBF)</i>	0.820	0.814	0.894	0.901	0.900	0.856	0.888
<i>kNN (Manhattan)</i>	<b>0.913</b>	<b>0.948</b>	<b>0.952</b>	<b>0.945</b>	<b>0.971</b>	<b>0.925</b>	<b>0.951</b>
<i>RNA (RBF Network)</i>	0.865	0.928	0.94	0.94	<b>0.971</b>	0.920	0.939

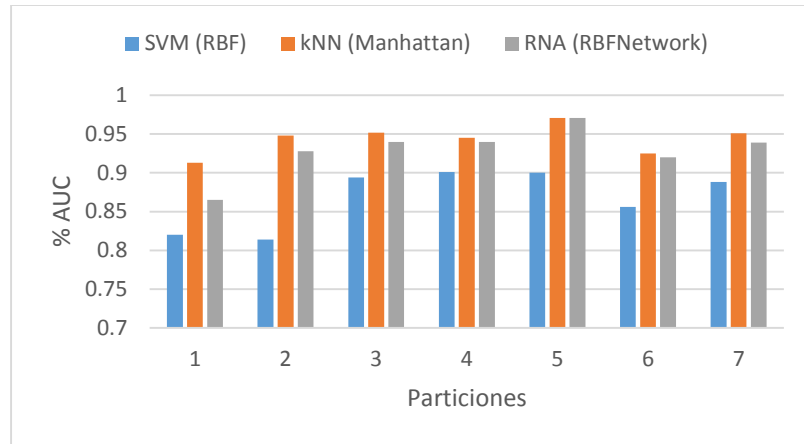


Figura 20: Comportamiento gráfico de la medida AUC.

### 3.3. Análisis comparativo de los clasificadores

Luego del análisis realizado de acuerdo a las medidas de evaluación aplicadas a los clasificadores es necesario evaluar si existe alguna diferencia significativa que demuestre la superioridad de un clasificador sobre los demás. Para ello se siguieron las recomendaciones realizadas por Demšar [55] y las extensiones presentadas por García y Herrera [57]. Se procedió a realizar un análisis estadístico a través de pruebas no paramétricas para  $k$  muestras no relacionadas mediante el Test de Friedman [55] con el objetivo de probar la hipótesis nula de que todos los clasificadores alcanzan los mismos resultados en promedio. Como pruebas post-hoc se aplicó primeramente el test Bonferroni-Dunn [55] para definir que existen diferencias significativas entre el clasificador mejor rankeado y el siguiente. Para contrastar los resultados se aplicó el test de Holm [55].

Tabla 27: Resultados obtenidos para AUC por los clasificadores.

<i>Particiones</i>	<i>SVM</i>	<i>kNN</i>	<i>RNA</i>
<i>1</i>	0.82	0.913	0.865
<i>2</i>	0.814	0.948	0.928
<i>3</i>	0.894	0.952	0.94
<i>4</i>	0.901	0.945	0.94
<i>5</i>	0.90	0.971	0.971
<i>6</i>	0.856	0.925	0.92
<i>7</i>	0.888	0.951	0.939
<i>Ranking</i>	2.99	1.07	1.92
<i>Posición</i>	3	1	2

Tabla 28: Resultados obtenidos para Pn por los clasificadores.

<i>Particiones</i>	<i>SVM</i>	<i>kNN</i>	<i>RNA</i>
1	0.74	0.77	0.73
2	0.79	0.79	0.80
3	0.87	0.85	0.89
4	0.85	0.8	0.84
5	0.81	0.88	0.86
6	0.82	0.83	0.81
7	0.89	0.93	0.91
<i>Ranking</i>	2.21	1.78	2.0
<i>Posición</i>	3	1	2

Tabla 29: Resultados obtenidos para medida  $F$  por los clasificadores.

<i>Particiones</i>	<i>SVM</i>	<i>kNN</i>	<i>RNA</i>
1	0.863	0.88	0.909
2	0.866	0.872	0.884
3	0.923	0.918	0.928
4	0.923	0.897	0.918
5	0.914	0.934	0.929
6	0.895	0.901	0.902
7	0.922	0.927	0.933
<i>Ranking</i>	2.57	2.14	1.28
<i>Posición</i>	3	2	1

Las tablas 27, 28 y 29 muestran los resultados experimentales de los tres clasificadores para cada una de las particiones realizadas del conjunto de datos. Las dos últimas filas muestran el ranking de cada uno de los clasificadores y su posición. Después de analizar estadísticamente los resultados para detectar si existen diferencias significativas entre los clasificadores para cada una de las medidas utilizadas, se comprobó que el test de Friedman rechazó la hipótesis nula para AUC con un valor de  $p = 0.01$  y para la medida  $F$  con un valor de  $p = 0.04$ , mientras que para Pn no se encontraron diferencias significativas. Por esta razón solo se aplicaron los test post-hoc para la medida AUC y medida  $F$ . Para ambas medidas los test Bonferroni-Dunn y de Holm rechazaron la hipótesis nula para valores de  $p \leq 0.025$  y  $p \leq 0.05$  respectivamente, con un valor de confianza  $\alpha = 0.05$  (Tabla 30) y (Tabla 31). Los resultados presentados en la tabla 30, demuestran que el clasificador  $kNN$  tiene un mejor desempeño que SVM y equivalente que RNA para el particionamiento realizado de los datos. En el caso de la tabla 31, los resultados presentados demuestran que el clasificador RNA tiene un mejor desempeño que SVM y equivalente que  $kNN$  para el particionamiento realizado de los datos.

Tabla 30: Orden de los clasificadores por el valor de  $p$  y ajustando el valor de  $\alpha$  para el procedimiento Holm con  $\alpha = 0.05$  para la medida AUC.

$i$	<i>Clasificador</i>	$z = (R_0 - R_i)/SE$	$p$	$\alpha/i$
2	SVM	3.60	$3.08 \times 10^{-4}$	0.025
1	RNA	1.60	0.10	0.05

Tabla 31: Orden de los clasificadores por el valor de  $p$  y ajustando el valor de  $\alpha$  para el procedimiento Holm con  $\alpha = 0.05$  para la medida  $F$ .

$i$	<i>Clasificador</i>	$z = (R_0 - R_i)/SE$	$p$	$\alpha/i$
2	SVM	2.40	0.01	0.025
1	<i>kNN</i>	1.60	0.10	0.05

De acuerdo al análisis realizado, se evidencia que existen diferencias significativas entre los clasificadores *kNN* y RNA con respecto a SVM de acuerdo a la medida AUC y a la medida  $F$ , no así entre *kNN* y RNA. Por tal motivo, es necesario realizar un análisis del comportamiento de los clasificadores *kNN* y RNA de acuerdo a los valores obtenidos en ambas medidas en las tablas 26 y 24. Se evidencia que el mayor valor obtenido en ambas medidas entre ambos clasificadores fue obtenido por *kNN*, con un valor de 93.4% en la medida  $F$  y del 97.1% para la medida AUC. Luego de evaluadas las medidas en cada clasificador, se evidencia que el clasificador con mejor desempeño en la clasificación de células cervicales, para el conjunto de datos estudiados, es *kNN* haciendo uso de la distancia Manhattan. Dicho clasificador presenta los mejores resultados para las medidas evaluadas, con valores promedios de medida  $F$  de 90.4%, Pn de 83% y AUC de 94.3%. Además este clasificador mostró un alto desempeño manteniendo los valores de AUC superiores al 91% y llegando hasta un 97.1% (Tabla 26). Por estas razones se recomienda utilizar el clasificador *kNN* con la distancia Manhattan en la herramienta de clasificación desarrollada.

### 3.3.1. Comparación con otros enfoques de la literatura estudiada

La tabla 32 muestra la comparación de la evaluación estadística obtenida durante la investigación con los resultados obtenidos en otras investigaciones analizadas. Dichas investigaciones utilizan al igual que la presente investigación la base de datos Herlev, aunque no especifican la forma en que se realizó la partición de los conjuntos de rasgos para las pruebas y el entrenamiento.

Se establece una comparación en función de la media armónica  $H$  debido a que es la medida de evaluación utilizada por dichas investigaciones. El resultado obtenido por la presente investigación de acuerdo a la media armónica muestra que se logra superar los resultados obtenidos en [16], mientras que son ligeramente inferiores a [35] aunque se debe tener en cuenta que este utiliza un mayor número de rasgos.



Tabla 32: Comparación de la clasificación en términos de la media armónica  $H$  ( $H$ -mean).

<i>Clasificadores utilizados por las investigaciones.</i>	<i># de Rasgos utilizados</i>	<i>% Media Armónica (H-mean)</i>
Fuzzy C-means [16]	7	90.58
Spectral Clustering [16]	9	88.77
SVM (RBF) [35]	9	96.89
$kNN$ (Manhattan)	5	91.35

### 3.4. Conclusiones Parciales

En el presente capítulo se realiza una comparación entre los clasificadores de acuerdo a los resultados obtenidos en cuanto a las medidas de evaluación predictividad negativa, medida  $F$ , media  $H$  y AUC. Se establece además una comparación entre el clasificador resultante de la validación experimental con resultados obtenidos por otras investigaciones, arrojando esta resultados que muestran la eficacia de la herramienta desarrollada. A raíz de esto se puede concluir que el clasificador  $kNN$  haciendo uso de la distancia Manhattan ha sido el más estable de los clasificadores evaluados en este trabajo, obteniendo durante las iteraciones de prueba los mayores valores en cuanto al área bajo la curva y la mejor probabilidad de detección correcta de los casos negativos.

# Conclusiones Generales

"Si tenemos en cuenta que la esperanza de vida promedio de la mujer en nuestro país es de 80 años, con cada fallecimiento se pierden 30 años de vida potenciales, por lo que es la principal causa de muerte prematura en la población femenina cubana. La carga para la sociedad es entonces muy alta, además de la discapacidad que genera en aquellas que no fallecen". [56]

Según [56], en el año 2011 se reportaron en Cuba 1334 casos nuevos de cáncer cervicouterino, y unas 455 mujeres fallecieron por esta causa. Atendiendo a tal situación, se plantea que los programas de pesquisa y diagnóstico precoz del cáncer cervicouterino y sus lesiones precursoras cobran cada vez mayor relevancia.

La presente investigación se basa fundamentalmente en la realización de una herramienta de clasificación automática que permita la reducción de los falsos negativos resultantes de la clasificación de las imágenes en la prueba de Papanicolaou. A medida que avanzó la investigación y en la realización de las pruebas se evidenció que las células sanas pertenecientes a la clasificación *Normal\_Columnar* de la base de datos Herlev, son células propensas a falsos positivos debido a que sus características son similares a las pertenecientes a las células enfermas.

En la presente investigación se analizó la aplicación de tres algoritmos en la clasificación de células cervicales, y se comparó su desempeño en la solución de este problema con el propósito de dar respuesta a la hipótesis planteada y cumplir con el objetivo general del presente trabajo. A continuación se enuncian algunos de los resultados arrojados:

1. El empleo de técnicas de clasificación en la realización de pruebas a células cervicales permite obtener una clasificación con baja tasa de falsos negativos.
2. A través de una comparación fueron obtenidos los parámetros propios de funcionamiento (kernel, distancias, números de clusters) con los cuales los algoritmos muestran los mejores resultados en la clasificación de este tipo de células. El algoritmo SVM mostró ser más apropiado cuando hace uso del kernel RBF, debido a su comportamiento con respecto al kernel lineal, y a su bajo costo computacional. En el caso de las redes neuronales comparadas se evidencia que la *RBF Network* es más eficaz en el momento de la clasificación por los resultados obtenidos, así como en el caso

del algoritmo *kNN*, la distancia Manhattan evidenció obtener mejores resultados además de un bajo costo computacional. Dichos algoritmos arrojaron resultados que demuestran su eficacia en la clasificación de las células cervicales.

3. Se demostró que a partir de un conjunto simplificado de datos geométricos básicos, es posible realizar de forma efectiva la clasificación binaria de imágenes en la prueba de Papanicolaou.
4. Los procedimientos estadísticos llevados a cabo por cada índice de efectividad evaluado, determinaron que con respecto a la medida AUC y a la medida *F* los clasificadores RNA y *kNN* poseían diferencias significativas con respecto a SVM, no siendo así entre *kNN* y RNA para ambas medidas. Luego del análisis estadístico *kNN* mostró un alto desempeño manteniendo los valores de AUC superiores al 91% y llegando hasta un 97.1%.

La investigación concluyó con dos herramientas, una para el apoyo y la continuidad de las investigaciones que se llevan a cabo en el Centro de Estudios de Eléctrica y Tecnologías de la Información (CEETI) de la UCLV, y la otra para el apoyo de los especialistas en el momento de la realización de la prueba de Papanicolaou en los centros médicos del país.

La herramienta de clasificación desarrollada como resultado de la presente investigación, si bien no es una herramienta de diagnóstico, sirve como ayuda en el proceso. La misma es capaz de clasificar entre células anómalas y normales obteniendo resultados alentadores. Las iteraciones de prueba realizadas y la comparación de los resultados obtenidos en otras investigaciones relacionadas con el dominio abordado demostraron la reducción de falsos negativos que es uno de los propósitos principales de la presente investigación.

# Recomendaciones

Una vez cumplido los objetivos del presente trabajo y en correspondencia con los resultados obtenidos, los cuales se expusieron en el propio documento, se recomienda como trabajos futuros:

- Aplicar la herramienta de clasificación propuesta, en las investigaciones realizadas referentes a la reducción de dimensionalidad de rasgos llevadas a cabo en el CEETI para comprobar su eficacia en cuanto a la clasificación de células.
- Realizar un estudio para la definición del mayor conjunto posible de rasgos relativos al núcleo celular en estos casos y la selección rigurosa de los rasgos, así como el ensayo de otros métodos de reducción de la dimensionalidad pudiendo aplicar los relacionados en [35].
- Ampliar la prueba con el clasificador a imágenes fuera de la base Herlev y determinar cuán efectivo puede ser el algoritmo *kNN* usando la base Herlev como “entrenamiento” e imágenes no pertenecientes a esa base de datos como prueba.

# Referencias bibliográficas

- [1] Lorenzo-Ginori JV, Rodríguez-Santos I. Aplicación de técnicas de visión computacional en la prueba de Papanicolaou. *Medicentro Electrónica*. 2012;16(3):196-8.
- [2] Campo P, Bonilla LJ, Calderon A. Cáncer cervical: Citología en base líquida, convencional y otras pruebas de tamizaje. *Revista Repertorio de Medicina y Cirugía*. 2012;21(3).
- [3] Ramos CMÁ. Sistema de reconocimiento y clasificación de agentes patógenos de neoplasias [Máster en Investigación en Informática]. Madrid, España: Universidad Complutense de Madrid; 2010.
- [4] Cáncer de cuello uterino: detección temprana y prevención. American Cancer Society; 2014; Available from: <http://www.cancer.org/acs/groups/cid/documents/webcontent/002580-pdf.pdf>.
- [5] Pruebas de detección de cáncer de cuello uterino. Washington, DC: The American College of Obstetricians and Gynecologists; 2013; Available from: <http://www.acog.org/Patients/Search-Patient-Education-Pamphlets-Spanish/Files/Pruebas-de-deteccion-de-cancer-de-cuello-uterino>.
- [6] Takahashi M. Atlas Color Citología del Cáncer. Citología del cáncer Argentina 2da Edición Editorial Médica Panamericana. 1982;50.
- [7] Maisanava JML, Soriano JÁM, Murcia AP. Citología Exfoliativa Cervicovaginal (Método de Papanicolaou). Servicio de Anatomía Patológica del Hospital Obispo Polanco de Truel. *Boletín Oncológico*. 1998;8:46-54.
- [8] Arbyn M, Anttila A, Jordan J, Ronco G, Schenck U, Segnan N, et al. European Guidelines for Quality Assurance in Cervical Cancer Screening. *Annals of Oncology*. 2010;21(3):448-58.
- [9] Mas JAG, Moncasi PS, Torres JA, Hörndler C, Marcilla EU. Evaluación de dispositivos automatizados para diagnóstico citológico en la prevención del cáncer de cérvix. *Revista Española de Patología*. 2002;35(3):301-14.
- [10] Patten SF, Lee JS, Wilbur DC, Bonfiglio TA, Colgan TJ, Richart RM, et al. The AutoPap 300 QC system multicenter clinical trials for use in quality control rescreening of cervical smears. *Cancer Cytopathology*. 1997;81(6):337-42.
- [11] Troni GM, Ciatto S, Cariaggi MP, Bulgaresi P. Quality control of the autopap screening system employed as a primary screening device: rapid review of smears coded as no further review. *Tumori - a Journal of Experimental and Clinical Oncology* - 2006;92(4):276.

- [12] Soler Font I, Romero Martos E, Pijuan Andújar L, Lloveras Rubio B, Carreras Collado R, Serrano Figueras S, et al. Aplicación de la lectura automatizada de citología ginecológica. El punto de vista de los citotécnicos. *Revista Española de Patología*. 2010;43(2):69-72.
- [13] Jantzen J, Norup J, Dounias G, Bjerregaard B. Pap-smear Benchmark Data For Pattern Classification. *Proc NiSIS 2005*. Albufeira, Portugal: Nature inspired Smart Information Systems (NiSIS); 2005. p. 1-9.
- [14] Marinakis Y, Dounias G, Jantzen J. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. *Computers in Biology and Medicine*. 2009;39(1):69-78.
- [15] Marinakis Y, Marinaki M, Dounias G. Particle swarm optimization for pap-smear diagnosis. *Expert Systems with Applications*. 2008;35(4):1645-56.
- [16] Plissiti M, Nikou C. Cervical Cell Classification Based Exclusively on Nucleus Features. In: Campilho A, Kamel M, editors. *Image Analysis and Recognition: Springer Berlin Heidelberg*; 2012. p. 483-90.
- [17] Gonzalez RC, Woods RE. *Digital Image Processing: Addison-Wesley Longman Publishing Co., Inc.*; 2001.
- [18] Clancey WJ. *Classification Problem Solving*. Department of Computer Science, Stanford University; 1984.
- [19] Padilla FV. Implementación eficiente de clasificadores PRIOR-SVM para Matlab [Proyecto Fin de Carrera]. Madrid, España: Universidad Carlos III de Madrid; 2010.
- [20] Abente M, Martínez J, Barán B. Evolución de Reglas de Clasificación Binaria utilizando Programación Genética Lineal. Una Aplicación al Descarte de Ganado. XXXVI Conferencia Latinoamericana de Informática, CLEI'2010; Asunción, Paraguay 2010.
- [21] Hand DJ, Mannila H, Smyth P. *Principles of data mining*. Cambridge, Massachusetts: The MIT Press; 2001.
- [22] Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques. *Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, . 1566773: IOS Press; 2007. p. 3-24.
- [23] García MRH. Detección de anuncios en emisiones de televisión basada en la caracterización del audio [Máster en Informática Aplicada]. Ciudad de la Habana, Cuba: Universidad de las Ciencias Informáticas; 2011.
- [24] Schalkoff RJ. *Artificial Neural Networks: McGraw-Hill Higher Education*; 1997.

- [25] Olabe Xabier B. *Redes Neuronales Artificiales y sus Aplicaciones*. Bilbao, España: Publicaciones de la Escuela de Ingenieros; 1998.
- [26] de Ullibarri Galparsoro L, Fernández P. Curvas ROC. *Atención Primaria en la Red*. 1998;5(4):229-35.
- [27] Olson DL, Delen D. *Advanced Data Mining Techniques*: Springer Publishing Company, Incorporated; 2008.
- [28] Menon A, Narasimhan H, Agarwal S, Chawla S. On the Statistical Consistency of Algorithms for Binary Classification under Class Imbalance. *Proceedings of The 30th International Conference on Machine Learning*, 2013. p. 603-11.
- [29] Montes MAP, Llorente DÁ, Pérez EP, González BF. Implementación de un paquete de plugins en ImageJ para el control de calidad de gammacámaras. *Revista de Física Médica*. 2008;9(2):88-90.
- [30] Pérez JMM, Pascau J. *Image Processing with ImageJ*: Packt Publishing; 2013.
- [31] Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*. 2012;9(7):676-82.
- [32] Gonzalez RC, Woods RE, Eddins SL. *Digital Image Processing Using MATLAB: AND " Mathworks, MATLAB Sim SV 07 ": Prentice Hall Press; 2007*.
- [33] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Exploration Newsletter*. 2009;11(1):10-8.
- [34] Corso CL, Alfaro SL. Alternativa de herramienta libre para la implementación de aprendizaje automático. *Congreso Ingeniería en Sistemas (CNEISI 2009)*; Congreso en San Francisco 2009.
- [35] Plissiti ME, Nikou C. On the importance of nucleus features in the classification of cervical cells in Pap smear images. *University of Ioannina [serial on the Internet]*. 2012.
- [36] Riana D, Murni A. Performance evaluation of Pap smear cell image classification using quantitative and qualitative features based on multiple classifiers. *International Conference on Advanced Computer Science and Information Systems, ACSIS2009*.
- [37] Mat-Isa NA, Mashor MY, Othman NH. An automated cervical pre-cancerous diagnostic system. *Artificial Intelligence in Medicine*. 2008;42(1):1-11.
- [38] Huang P-C, Chan Y-K, Chan P-C, Chen Y-F, Chen R-C, Huang Y-R. Quantitative Assessment of Pap Smear Cells by PC-Based Cytopathologic Image Analysis System and Support Vector Machine. In: Zhang D, editor. *Medical Biometrics: Springer Berlin Heidelberg*; 2007. p. 192-9.
- [39] Plissiti ME, Nikou C, Charchanti A. Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images. *Pattern Recognition Letters*. 2011;32(6):838-53.

- [40] Plissiti ME, Nikou C, Member, IEEE, Charchanti A. Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering. *IEEE Transactions on information technology in biomedicine*. 2011;15(2):233–41.
- [41] Lorenzo-Ginori JV, Curbelo-Jardines W, López-Cabrera JD, Huergo-Suárez SB. Cervical Cell Classification Using Features Related to Morphometry and Texture of Nuclei. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: Springer*; 2013. p. 222-9.
- [42] Velezmoro GAB, Villafuerte DF. Factores de riesgo que pronóstican el hallazgo de citologías cervicales anormales en dos poblaciones: mujeres de obreros de construcción civil vs. mujeres control en la posta médica "Construcción Civil" ESSALUD, de junio a septiembre del 2000 [Licenciatura]. Lima, Perú: Universidad Nacional Mayor de San Marcos; 2001.
- [43] Vapnik VN. *The Nature of Statistical Learning Theory*: Springer; 2000.
- [44] Betancourt GA. Las máquinas de soporte vectorial (SVMs). *Scientia et Technica*. 2005;1(27).
- [45] Fan R-E, Chen P-H, Lin C-J. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*. 2005;6:1889-918.
- [46] Silverman BW, Jones MC. E. Fix and JL Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). *International Statistical Review/Revue Internationale de Statistique*. 1989:233-8.
- [47] Kruse R, Borgelt C, Klawonn F, Moewes C, Steinbrecher M, Held P. *Multi-Layer Perceptrons. Computational Intelligence*: Springer; 2013. p. 47-81.
- [48] Jusman Y, Sulaiman SN, Isa NAM, Yusoff IA, Adnan R, Othman NH, et al. Capability of new features from FTIR spectral of cervical cells for cervical precancerous diagnostic system using MLP networks. *TENCON 2009 - 2009 IEEE Region 10 Conference*; 23-26 Jan. 2009/2009. p. 1-6.
- [49] Silva EL. *Método de reducción de Redes Neuronales RBF usando la descomposición QLP para aplicaciones de ajuste de datos y clasificación [Tesis Doctoral]*. Granada, España: Universidad de Granada; 2007.
- [50] Kim K-B, Song D, Woo Y. Nucleus Segmentation and Recognition of Uterine Cervical Pap-Smears. In: An A, Stefanowski J, Ramanna S, Butz C, Pedrycz W, Wang G, editors. *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*: Springer Berlin Heidelberg; 2007. p. 153-60.
- [51] Schwenkar F, Kestler HA, Palm G. Radial basis function networks: learning and applications. *Fouth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, 2000. p. 33-43.
- [52] Hamed M, Salleh S-H, Astaraki M, Noor A, Harris A. Comparison of Multilayer Perceptron and Radial Basis Function Neural Networks for EMG-Based Facial Gesture Recognition. In: *Mat Sakim*



- HA, Mustaffa MT, editors. The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications: Springer Singapore; 2014. p. 285-94.
- [53] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975;12(4):387-415.
- [54] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
- [55] Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*. 2006;7:1-30.
- [56] Acosta LF. El diagnóstico temprano es garantía de vida. *Periódico Granma*. 2012.
- [57] García S, Herrera F. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research*. 2008;9:2677-94.

# Anexos

## I. Comportamiento de las matrices de confusión y de las medidas de evaluación referente a los algoritmos $kNN$ , SVM, RNA

### $kNN$ – Manhattan

Tabla 33: Matriz de confusión referente al comportamiento del algoritmo  $kNN$  con la distancia Manhattan en las iteraciones de prueba.

<i>Variables</i>		<i>Iteraciones de Pruebas</i>						
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
VP	(a)	123	125	127	124	128	126	131
FP	(b)	10	13	8	8	5	9	10
FN	(c)	11	9	7	10	6	8	3
VP	(d)	37	34	39	39	42	38	37

Tabla 34: Comportamiento de las medidas de evaluación en cuanto al algoritmo  $kNN$  con la distancia Manhattan en las iteraciones de prueba.

<i>Medidas de evaluación</i>	<i>Iteraciones de Pruebas</i>						
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
Sensibilidad ( $S_e$ )	0.92	0.93	0.95	0.93	0.96	0.94	0.98
Especificidad ( $S_p$ )	0.79	0.72	0.83	0.83	0.89	0.81	0.79
Predictividad Positiva ( $P_p$ )	0.92	0.91	0.94	0.94	0.96	0.93	0.93
Predictividad Negativa ( $P_n$ )	0.77	0.79	0.85	0.80	0.88	0.83	0.93
Tasa de Clasificación ( $R_c$ )	0.88	0.88	0.92	0.90	0.94	0.91	0.93
$H - mean$	0.8475	0.8148	0.8848	0.8749	0.9233	0.8694	0.8721

### SVM – RBF

Tabla 35: Matriz de confusión referente al comportamiento del algoritmo SVM haciendo uso del kernel RBF en las iteraciones de prueba.

<i>Variables</i>		<i>Iteraciones de Pruebas</i>						
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
VP	(a)	123	126	128	124	122	125	129
FP	(b)	10	13	8	6	5	10	9
FN	(c)	11	8	6	10	12	9	5
VP	(d)	37	34	39	41	42	37	38

Tabla 36: Comportamiento de las medidas de evaluación en cuanto al algoritmo SVM haciendo uso del kernel RBF en las iteraciones de prueba.

<i>Medidas de evaluación</i>	<i>Iteraciones de Pruebas</i>						
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
Sensibilidad ( $S_e$ )	0.92	0.94	0.96	0.93	0.91	0.93	0.96
Especificidad ( $S_p$ )	0.79	0.72	0.83	0.87	0.89	0.79	0.81
Predictividad Positiva ( $P_p$ )	0.92	0.91	0.94	0.95	0.96	0.93	0.93
Predictividad Negativa ( $P_n$ )	0.77	0.81	0.87	0.80	0.78	0.80	0.88
Tasa de Clasificación ( $R_c$ )	0.88	0.88	0.92	0.91	0.91	0.90	0.92
<i>H – mean</i>	0.8475	0.8177	0.8880	0.8980	0.9019	0.8538	0.8788

### RNA – RBF Network

Tabla 37: Matriz de confusión referente al comportamiento de la RNA-RBF Network en las iteraciones de prueba.

<i>Variables</i>		<i>Iteraciones de Pruebas</i>						
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
VP	(a)	121	125	128	127	127	125	130
FP	(b)	10	12	8	6	5	8	8
FN	(c)	13	9	6	7	7	9	4
VP	(d)	37	35	39	41	42	39	39

Tabla 38: Comportamiento de las medidas de evaluación en cuanto a la RNA-RBF Network en las iteraciones de prueba.

<i>Medidas de evaluación</i>	<i>Iteraciones de Pruebas</i>						
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
Sensibilidad ( $S_e$ )	0.90	0.93	0.96	0.95	0.95	0.93	0.97
Especificidad ( $S_p$ )	0.79	0.74	0.83	0.87	0.89	0.83	0.83
Predictividad Positiva ( $P_p$ )	0.92	0.91	0.94	0.95	0.96	0.94	0.94
Predictividad Negativa ( $P_n$ )	0.74	0.80	0.87	0.85	0.86	0.81	0.91
Tasa de Clasificación ( $R_c$ )	0.87	0.88	0.92	0.93	0.93	0.91	0.93
<i>H – mean</i>	0.8411	0.8282	0.8880	0.9084	0.9198	0.8782	0.8944

## II. Comparación gráfica del comportamiento entre los resultados de las matrices de confusión y las medidas de evaluación referente a los algoritmos *kNN*, SVM y RNA

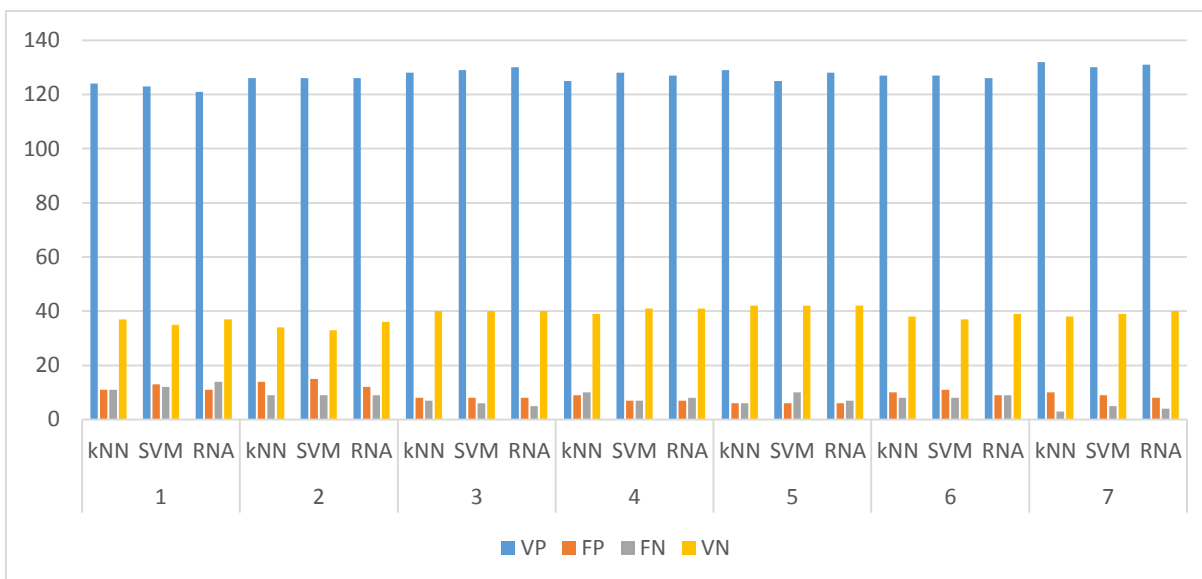


Figura 21: Comportamiento de los algoritmos referente a los valores de la matriz de confusión durante las iteraciones de prueba.

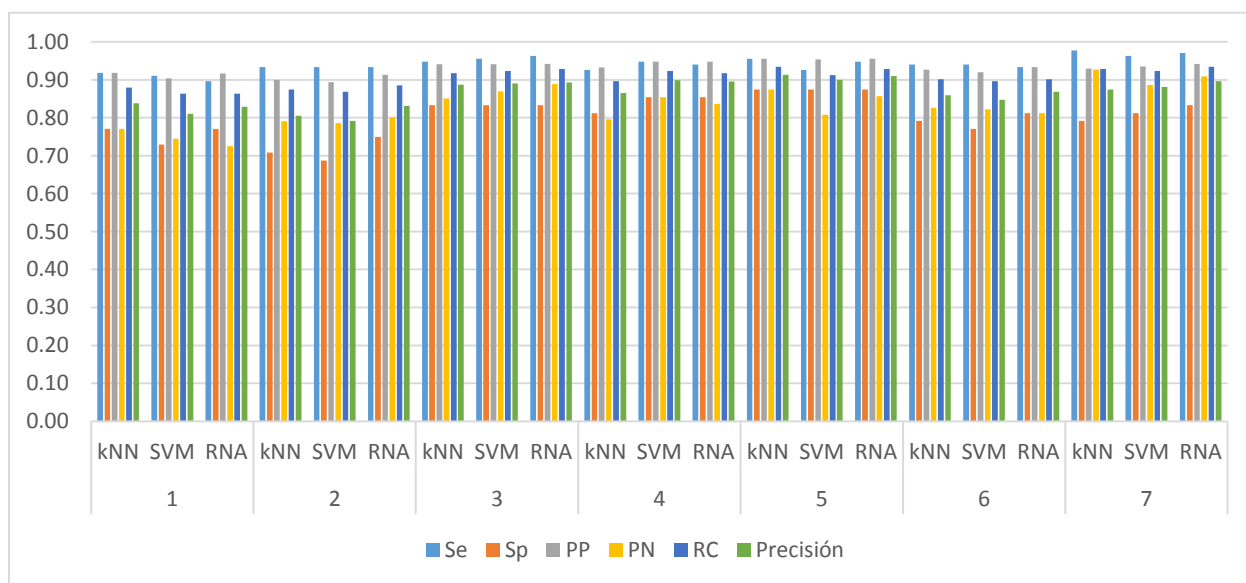


Figura 22: Comportamiento de los algoritmos referente a las medidas de evaluación durante las iteraciones de pruebas.