Tejera, Y. & Hernández, Y.
Pág. 58-68
Recibido: 15/09/2019
Aceptado: 18/11/2019
Publicado: 13/01/2020

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

EVALUACIÓN Y CLASIFICACIÓN DE LAS IMÁGENES DE CÉLULAS DEL CÉRVIX USANDO RASGOS DE TEXTURAS Y MORFOLÓGICOS

EVALUATION AND CLASSIFICATION OF THE IMAGES OF CELLS OF THE CERVIX USING TEXTURAL AND MORPHOLOGICAL FEATURES

Autores: Yosmany Tejera Marante
Yoelkis Hernández Victor
Institución: Universidad de Ciego de Ávila Máximo Gómez Báez, Cuba
Correo electrónico: yosmany@unica.cu

RESUMEN

El análisis de imágenes de células del *cervix* extraídas a través de la citología cervical es utilizado para la detección de anormalidades en las células, así como el estadio en que se encuentra el carcinoma detectado. En este articulo se presenta una evaluación experimental para la clasificación de las células cervicales en las condiciones normal y anómala, se basa solamente en las características extraídas de la región ocupada por el núcleo, sin hacer uso de las características del citoplasma y se utilizan como atributos de entrada los rasgos morfológicos y de texturas. Las técnicas de clasificación fueron realizadas por los algoritmos, Máquinas de Soporte Vectorial (MSV), los clasificadores de k vecinos más cercanos (kNN), árboles de decisión generados mediante C45, las técnicas estadísticas de clasificación como Regresión Logística y el perceptrón multicapa o MLP (*Multi-Layer Perceptron*). Este trabajo tiene como objetivo comparar los resultados de los algoritmos de clasificación de células cervicales en imágenes digitales usando como atributos de entrada rasgos morfológicos y rasgos de textura.

Palabras clave: Células, Cérvix, Clasificación, Morfológicos, Textura.

ABSTRACT

The analysis of images of cervix cells extracted through cervical cytology is used to detect abnormalities in the cells, as well as the stage in which the detected carcinoma is found. This article presents an experimental evaluation for the classification of cervical cells in normal and abnormal conditions, based only on the characteristics extracted from the region occupied by the nucleus, without using the characteristics of the cytoplasm and using as input attributes the morphological features and textures. The algorithms performed the classification techniques, Vector

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Support Machines (MSV), the closest k-k classifiers (kNN), decision trees generated by C45, the statistical classification techniques such as Logistic Regression and the multilayer perceptron or MLP (Multi-Layer Perceptron). This work aims to compare the results of the cervical cell classification algorithms in digital images using morphological features and texture features as input attributes.

Keywords: Cells, Cervix, Classification, Morphological, Texture.

INTRODUCCIÓN

El cáncer cervical uterino (CCU) es el segundo tipo de cáncer más común en las mujeres. La edad media de aparición es a los 45 años. Es uno de los problemas de salud pública más frecuentes en países en vías de desarrollo, diagnosticándose más de 400.000 casos nuevos cada año. (Arzuaga-Salazar, Souza, Lima, 2012).

Con el procesamiento digital de imágenes, es posible hacer mejoras en las imágenes, aumentando o disminuyendo la luminosidad y/o la cromaticidad para percibir mejor los detalles. El procesamiento digital de imágenes tiene una serie de pasos para lograr la extracción de objetos de una imagen y obtener de estos, datos que permitan diferenciarlos y clasificarlos entre varios objetos.

Según datos tomados del Anuario Estadístico de Salud en Cuba, en 1965 la tasa de mortalidad por CCU en Cuba era de veinte por cada 100 mil mujeres. Por lo que para el año 1968 fue creado el Programa Nacional para el Diagnóstico Precoz del Cáncer de Cuello Uterino. El Programa se enfocó en la masificación de la prueba de Papanicolaou, conocido en el país como Prueba Citológica, con cobertura para casi el 70 % de la población en riesgo. No es una prueba muy sensible, pero ha logrado salvar la vida a millones de mujeres cubanas. Sin embargo, desde 1970 hasta la fecha, la situación del CCU en Cuba no ha hecho más que empeorar; la tasa de mortalidad es hoy el doble (Martínez, PimentelI, 2015).

El crecimiento de esta tasa es resultado de una mayor exposición a factores de riesgo tales como precocidad y promiscuidad sexual, embarazos tempranos y abundantes, tabaquismo, no asistir a exámenes citológicos, edad avanzada, nivel socioeconómico bajo e infección con cepas endógenas del Virus de Papiloma Humano, entre otros (Salvá, García, 2001).

El cuello uterino es la parte más baja del útero (matriz) que desemboca en la parte superior de la vagina. La mayoría de los cánceres del cuello uterino se pueden

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

detectar a tiempo si una mujer se hace pruebas de Papanicolaou de manera rutinaria, las cuales deberían empezar a la edad de 21 años.

Para detectar posibles lesiones en el útero de la mujer, los especialistas deben revisar las muestras usando equipamientos como microscopios o cámaras especiales. El proceso de interpretación manual de estas imágenes puede resultar muy tedioso por la gran cantidad de imágenes a analizar en un día o puede no existir un especialista con la suficiente experiencia para emitir un diagnóstico. Estos problemas estimulan el desarrollo de algoritmos utilizados en aplicaciones informáticas que faciliten la interpretación de imágenes médicas. Estas aplicaciones incluyen la adquisición de las imágenes, pre procesamiento, segmentación, extracción de características y clasificación en la etapa final. En todos los casos, sus resultados deben servir de complemento a los diagnósticos de los especialistas médicos (Martínez Pinillo… et al., 2010).

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, redes neuronales artificiales, y máquinas de vectores soporte (SVM) (Huang …et al., 2008).

La citometría puede medir la textura y localización de proteínas con la misma precisión que se mide el tamaño y la forma de las células. Mientras que el ser humano hace un análisis cualitativo de las imágenes en la mayoría de los casos, la citometría hace un análisis cuantitativo para cada una de las células. La medición de una gran cantidad de características, incluso las que no son observables por los seres humanos, es muy útil para la detección de enfermedades, y para la medición del comportamiento celular.

El análisis cuantitativo permite detectar algunas características que no son fáciles de descubrir por los especialistas como, por ejemplo, el aumento en un 10% del tamaño del núcleo de una célula. Por otro lado, el suavizado de la textura de proteínas o de una mancha de ADN puede ser observado por humanos, pero no cuantificado. Grandes o pequeños cambios en la textura de las proteínas o del ADN puede traer cambios fisiológicos en las células por lo que es utilizado en el diagnóstico de enfermedades, y estos cambios no necesariamente deben de ser visibles por los especialistas.

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Para la clasificación de este tipo de imágenes es común el uso de rasgos morfológicos de las células (área, perímetro, redondez, excentricidad, diámetro, etcétera). Y es menos común el uso de rasgos de texturas en dicho proceso de clasificación (energía, entropía, segundo momento angular, disimilitud, etcétera). Sin embargo, no es frecuente la combinación de dichos rasgos para la clasificación y tampoco la realización de experimentos que corroboren el uso de unos u otros, o la combinación de ambos (morfológicos y texturas). Lo antes expuesto hace que se plantee como problema de investigación las insuficiencias en la clasificación de células cervicales en imágenes médicas digitales. Definiéndose como objeto de estudio el proceso de clasificación de células cervicales en imágenes digitales.

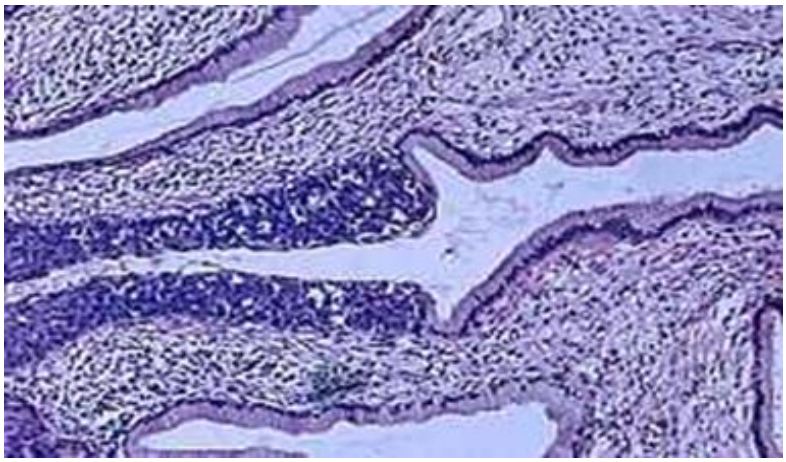

Figura 1 Imagen histológica de tejido de cáncer cervical en el útero.

El objetivo de este trabajo es comparar los resultados de los algoritmos de clasificación de células cervicales en imágenes digitales usando como atributos de entrada rasgos morfológicos y rasgos de textura.

Dado un grupo de láminas de cristal, una base de datos de imágenes individuales de células cervicales se recopila en el Hospital Universitario de Herlev. Los técnicos en citología calificados utilizan un microscopio con resolución de 0,201 m/pixel para capturar las imágenes digitales de las células individuales. Luego cada imagen de célula se clasifica de forma manual en siete tipos diferentes, los cuales se describen en la tabla 1.

Para la validación, dos técnicos diferentes realizan la clasificación. Si la validación resulta negativa, la imagen es descartada. Los siete diagnósticos son idénticos a

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

aquellos que se utilizan para la clasificación manual estándar. A continuación, se muestra la serie de datos.

Tabla 1 Los 7 tipos diferentes de citología cervical.

Normal - 242 células
1. epitelial escamoso superficial, 74 células.
2. epitelial escamoso intermedio, 70 células.
3. epitelial columnar, 98 células.
Anormal - 675 células
4. displasia escamosa ligera no queratinizada, 182 células.
5. displasia escamosa moderada no queratinizada, 146 células.
6. displasia escamosa severa no queratinizada, 197 células.
7. carcinoma de célula escamosa en fase intermedia, 150 células.

MATERIALES Y MÉTODOS

La mayoría de los cánceres de cérvix se originan en el revestimiento de las células del cuello uterino. Estas células no se tornan en cáncer de repente, sino que las células normales del cuello uterino se transforman gradualmente en precancerosas, y pueden con el tiempo convertirse en células malignas. 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 de cáncer.

Las células tienen ciertos rasgos tales como tamaño, área, forma y brillo tanto en el núcleo como en el citoplasma.

Estos rasgos – entre otros- pueden ser extraídos de una combinación de imágenes de células segmentadas y no segmentadas. Algunas medidas solo usan la imagen

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

segmentada y otras utilizan varias. Algunos rasgos se seleccionan con anterioridad por conocimiento de expertos. Los rasgos restantes se seleccionan para obtener una descripción completa de la célula.

Tabla 2 Rasgos extraídos para los datos de la citología cervical. N y C son las abreviaturas de Núcleo y Citoplasma

1. Área N	8. Elongación N	15. Perímetro C
2. Área C	9. Redondez N	16. Posición relativa N
3. Razón N/C	10. Diámetro más corto C	17. Máxima en N
4. Brillo N	11. Diámetro más largo C	18. Mínima en N
5. Brillo C	12. Elongación C	19. Máxima en C
6. Diámetro más corto N	13. Redondez C	20. Mínima en C
7. Diámetro más largo N	14. Perímetro N	

A pesar de que se recopile una base de datos de imagen que contiene imágenes clasificadas de células individualmente segmentadas, se dificulta hacerla funcionar mediante un algoritmo de clasificación. Las imágenes de células que van a ser clasificadas no son casos estereotipados de los 7 tipos de células, sino imágenes de orientación y tamaño bastante diferentes.

En la figura 2 se muestra una sola imagen de una célula normal. Tanto el núcleo como el citoplasma circundante son posibles de identificar por el ojo humano, incluso cuando los bordes son ambiguos y una segunda célula forma una intersección.

Para lograr una imagen sólida, todas las imágenes se segmentan en tres partes: fondo, citoplasma y núcleo. Esta segmentación se realiza al unísono utilizando el *software* CHAMP y de la misma se encargan los técnicos en citología cervical del Hospital Universitario de Herlev. El **software** CHAMP es un sistema de análisis de imagen médica basado en un algoritmo de reconocimiento de objeto a color. Una imagen segmentada se ilustra en la fig. 3. Al comparar la célula segmentada y no segmentada (figuras 2 y 3) se muestra que la célula individual se encuentra aislada y el punto de intersección ha cambiado. Además, los técnicos en citología validan el proceso de segmentación de forma manual mirando a través de todas las células segmentadas para cambiar las imágenes fallidas.

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

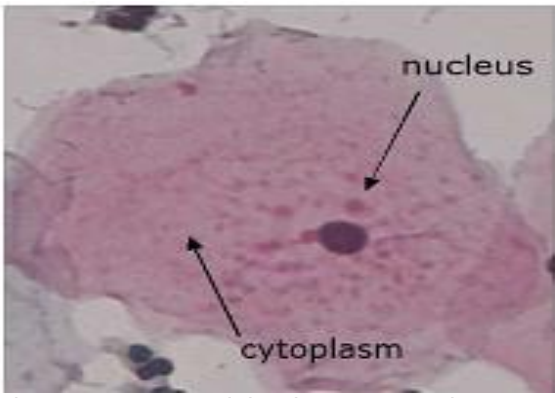

Figura 2 Imagen celular de una normal (Célula superficial normal tipo 1).

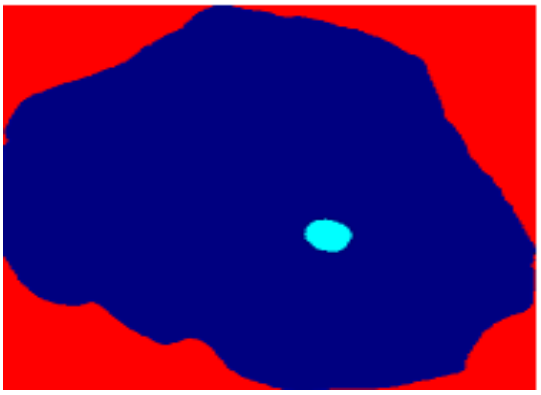

Figura 2 Célula superficial segmentada clase 1.

Para la realización del experimento se utilizaron las 917 imágenes de la base de datos de *Herlev*. Para ello se desarrolló una herramienta en Matlab Versión R2015a (8.5.0.197613), de esta manera se calcularon 27 rasgos de texturas basados en la matriz de co-ocurrencia calculada con distancia 1 y en los cuatro ángulos (00, 450, 900, 1800) y 20 rasgos morfológicos. Estos cálculos se realizaron sobre los canales R, G y B del espacio de color RGB; los canales H, S y V del espacio de color HSV y a las imágenes convertidas a escalas de grises.

Se proponen, además, dos bases de datos con las características de texturas extraídas de las imágenes antes mencionadas usando la herramienta Matlab. La primera solamente con rasgos de texturas y la otra con la combinación de rasgos de texturas y morfológicas. Estas bases de datos se encuentran en ficheros de texto plano con extensión *arff*, *csv*, *data* y *name*. De esta manera puede ser analizada dicha información por herramientas informáticas tales como *Keel* (García, Luengo , y Herrera, 2015), *Weka*, *Knime* (Cleophas, Zwinderman, 2017), entre otras.

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

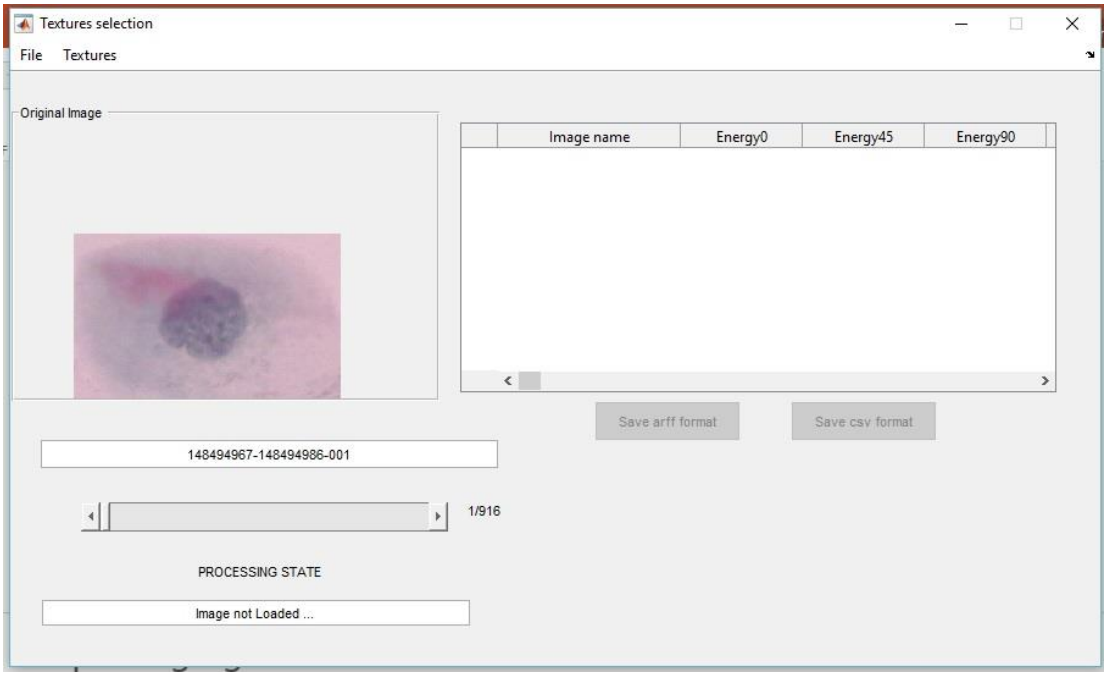

Figura 4 Muestra de clasificador utilizando la herramienta *Matlab.*

El criterio de evaluación es un factor clave a la hora de medir el desempeño de un clasificador supervisado. En un problema de dos clases como el presentado, la matriz de confusión (ver Tabla 3) registra los resultados de los objetos clasificados (correctamente e incorrectamente) en cada clase (López …et al., 2013).

Tabla 3 Matriz de confusión para problemas de dos clases

	Positiva Real	Negativa Real
Positiva Predicha	Verdaderos Positivos(TP)	Falsos Positivos(FP)
Negativa Predicha	Falsos Negativos(FN)	Verdaderos Negativos(TN)

En concreto, podemos obtener cuatro métricas para medir el rendimiento de clasificación para cada una de las clases. Donde:

$TP_{rate} = \frac{TP}{TP+FN}$ es la fracción de objetos bien clasificados en la clase positiva.

$TN_{rate} = \frac{TN}{FP+TN}$ es la fracción de objetos bien clasificados en la clase negativa.

$FP_{rate} = \frac{FP}{FP+TN}$ es la fracción de objetos mal clasificados en la clase positiva.

$FN_{rate} = \frac{FN}{TP+FN}$ es la fracción de objetos mal clasificados en la clase negativa.

La medida que fue utilizada para evaluar el desempeño de los clasificadores supervisados en problemas con clases desbalanceadas fue la gráfica Receiver *Operating Characteristic* (ROC). En esta gráfica se puede visualizar el equilibrio

Tejera, Y. & Hernández, Y.
Pág. 58-68
Recibido: 15/09/2019
Aceptado: 18/11/2019
Publicado: 13/01/2020

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

costo-beneficio; mostrando que cualquier clasificador no puede incrementar el número de TP sin aumentar los FP. Calcular el área bajo la curva ROC (AUC) es una de las medidas de evaluación más utilizadas para medir el desempeño los clasificadores supervisados en problemas con clases desbalanceadas. Esta se define como:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

En dicha clasificación se emitió un diagnóstico de forma genérica ¨normal¨ o ¨anormal¨.

RESULTADOS Y DISCUSIÓN

En la evaluación experimental de las diferentes bases de datos con los 4 algoritmos definidos al comienzo del artículo, se puede decir:

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Area_under_ROC
Datasets:   3
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       23/05/17 14:32

Dataset                 (1)      C45 '-C' |    (2) knn       (3) MSV          (4)logistic
-----------------------------------------------------------------------------------------
unión                   (100)    0.87(0.06) |  0.84(0.04)    0.50(0.00) *     0.93(0.03) v
morfológicos            (100)    0.89(0.06) |  0.88(0.04)    0.50(0.00) *     0.98(0.01) v
texturas                (100)    0.74(0.06) |  0.79(0.04) v  0.50(0.00) *     0.83(0.06) v
-----------------------------------------------------------------------------------------
                                 (v/ /*) |      (1/2/0)      (0/0/3)          (3/0/0)
```

Figura 5 Resultados de la clasificación con cuatro algoritmos y analizado con el campo de comparación Área bajo la Curva ROC.

Como se puede observar en la figura 5 se observa una tabla de análisis emitida por el Weka (Versión 3.7.2) donde se clasifican las tres bases de datos (unión, morfológicos y texturas) que fueron clasificados por los algoritmos Árboles de decisión generados mediante C45, Técnicas estadísticas de clasificación-Regresión

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Logística, Clasificador de k vecinos más cercanos y Máquina de Soporte Vectorial (MSV).

Todos los algoritmos fueron comparados con los Árboles de decisión generados mediante C45 donde el algoritmo que mejor clasificó fue la Técnica estadística de clasificación-Regresión Logística porque es donde los valores se acercan más a 1. Aunque no existe mucha diferencia entre los valores resultantes, los resultados de la clasificación de los 4 algoritmos con respecto a las bases de datos son favorables para la colección de datos de rasgos morfológicos, tal y como se muestra, en el 75% de los casos, exceptuando aquellos referentes al algoritmo Máquina de Soporte Vectorial (MSV), que muestran datos iguales para las tres bases de datos. Además, se muestra en la tabla la Desviación Estándar de cada valor de la clasificación.

CONCLUSIONES

En el artículo se comparan tres bases de datos utilizando como medida de evaluación del área bajo la curva. Según el proceso el realizado se evidenció que los clasificadores estudiados reflejan mejores resultados usando como atributos de entrada sólo los rasgos morfológicos, aunque usando la unión de los rasgos morfológicos y de texturas como atributos de entrada los resultados obtenidos se consideran aceptables, no así el uso de los rasgos de texturas solamente.

Se propone usar otros tipos de algoritmos de clasificación donde sean usados como atributos de entradas rasgos de texturas, morfológicos y la unión de ambos, además, usar algoritmos de reducción de atributos para determinar cuáles rasgos son los más representativos en la clasificación de células de cérvix.

REFERENCIAS BIBLIOGRÁFICAS

ARZUAGA-SALAZAR, M. A., SOUZA, M. de L. de y AZEVEDO LIMA, V. L. de. (2012). El cáncer de cuello de útero: un problema social mundial. *Rev. Cuba. Enferm.*, vol. 28, n.º 1, pp. 63-73.

BETANCOURT, G. A. (2005). Las máquinas de soporte vectorial (svms). *Sci. Tech.*, vol. 1, n.º 27.

CLEOPHAS, T. J. y ZWINDERMAN, A. H. (2017). Data Mining for Visualization of Health Processes (150 Patients with Pneumonia). En *Machine Learning in Medicine-Cookbook Three*, Springer, pp. 3-14.

GARCÍA C. y GÓMEZ, I.(2012). Algoritmos de aprendizaje*: KNN & Kmeans*. Documento.

Universidad&Ciencia
Vol. 9, No. 1, enero-abril (2020)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

GARCÍA, S., LUENGO, J. y HERRERA, F. (2015). A Data Mining Software Package Including Data Preparation and Reduction: KEEL. En *Data Preprocessing in Data Mining*, Springer, pp. 285-313.

MARTÍNEZ, J. C. y Pimenteli, M. G. (2015). Citologías alteradas y diferentes factores de riesgo para el cáncer cervicouterino Altered cytology and different risk factors of cervical uterine cancer. *Rev. Cienc. MÉDICAS HABANA*, vol. 21, n.º 2.

HUANG, P.-C…[et al.](2008). Quantitative assessment of Pap smear cells by PC-based cytopathologic image analysis system and support vector machine. En *International Conference on Medical Biometrics*, 2008, pp. 192-199.

LÓPEZ, V…. [et al.] (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics., *Inf. Sci.*, vol. 250, pp. 113-141.

MARTÍNEZ PINILLO, A. …et al. (2010). Análisis de los principales factores de riesgo relacionados con el cáncer cérvico uterino en mujeres menores de 30 años. *Rev. Cuba. Obstet. Ginecol.*, vol. 36, n.º 1, pp. 52-65.

PALOMINO, N. L. S. y CONCHA, U. N. R. (2009). Técnicas de segmentación en Procesamiento digital de imágenes. *Rev. Investig. Sist. E Informática*, vol. 6, n.º 2, pp. 9-16.

SALVÁ, A. R y GARCÍA, A. M. (2001). El registro nacional de cáncer de Cuba. Procedimientos y resultados. *Rev. Bras. Cancerol.*, vol. 47, n.º 2, pp. 171-77.

SOTOLONGO, D. (2010). Combinación de forma y textura para el análisis del cáncer de mama. Máximo Gómez Báez, Ciego de Ávila.