
Universidad de Costa Rica

Maestría Académica en Bioinformática

PF-5028 Minería de Datos

---

# Generación de clasificadores que predigan la necesidad de biopsia de cuello uterino en pacientes con sospecha de Cáncer Cervical

---

*Autora:* Paola Calderón

*Autor:* Carlos Solano

26 de noviembre de 2017

**Resumen**

El cáncer cervical produce la muerte de aproximadamente 300,000 mujeres alrededor de mundo, aún y cuando es un padecimiento facilmente prevenible. Sin embargo, las poblaciones con mayor probabilidad de desarrollar este padecimiento son las de menor acceso a recursos económicos y las infectadas con el virus del Papiloma Humano. El Papanicolau (Pap) o citología de cuello uterino permite la detección temprana de lesiones que puedan hacer sospechar cáncer incipiente, sin embargo, las poblaciones antes mencionadas no poseen un acceso directo a estos procedimientos de detección. Los modelos predictivos de mineria de datos para la detección temprana de un diagnóstico podrian disminuir la morbilidad y mortalidad de este padecimiento, aún cuando el paciente provenga de un ambiente de bajos recursos . El uso de un conjunto de datos proveniente de pacientes reales y la implementación de librerias de `Python` para análisis de minería de datos facilita la generación de resultados que impacten a largo plazo la detección temprana. Durante este estudio se procede a utilizar un conjunto de datos proveniente de 856

pacientes provenientes de Venezuela, con atributos como historial medico, información demográfica, y hábitos diarios, permitiendo ajustar un modelo que se aplique a la realidad de otro conjunto de personas alrededor del mundo.

# Índice general

# Capítulo 1

# Objetivos

## 1.1. Objetivo General

Generar clasficadores que predigan la necesidad de tomar una biopsia de cuello uterino en pacientes vulnerables a desarrollar cáncer de cuello uterino.

## 1.2. Objetivos Específicos

- Elegir los features predominantes y el número óptimo de features a partir del set de datos recopilado del Hospital Universitario de Caracas.
- Evaluar diferentes algoritmos de clasificación para predecir si para una

determinada paciente, dadas ciertas caracterísiticas, se requiere tomar una biopsia de cuello uterino.

- Implementar métodos de clasificación no supervisada para evaluar potenciales interacciones entre los features que hagan sospechar de la presencia de cáncer de cuello uterino.

# Capítulo 2

# Marco Teórico

El cáncer cervical es tipo de cáncer con mayor posibilidad de prevención. A diferencia de muchos tipos de cáncer, la fase premaligna es muchas veces larga y fácilmente detectable. El Papanicolau (Pap) o citología de cuello uterino permite la detección temprana de lesiones que puedan hacer sospechar cáncer incipiente. Sin embargo, este padecimiento afecta a las poblaciones con menor acceso a recursos económicos desproporcionadamente. En el mundo desarrollado se da el 20 % de todos los casos nuevos de cáncer de cervix, mientras que en países pobres se cuenta el 80 % de los casos. La mortalidad a causa de este cáncer se ve afectada por la detección temprana, tratamiento oportuno y la presencia o ausencia de infección por Virus de

Papiloma Humano, que es lastimosamente, muy prevalente en la población general. Sin embargo, un puñado de serotipos son los implicados en aparición de cáncer de cervix, siendo 16 y 18 los que presentan una correlación más fuerte. (Téguété y cols., 2017)

En caso de que el Pap muestre células anómalas, se procede a realizar una biopsia.(Frega y cols., 2016) Algunos tipos de biopsia cervical son:

1. Biopsia colposcópica: Se utilizan unas pinzas de biopsia para extirpar una pequeña sección (de aproximadamente 1/8 de pulgada) del área anormal en la superficie del cuello uterino.

2. Curetaje endocervical (raspado endocervical): con una cureta, se raspa el canal endocervical y se extrae tejido que posteriormente se envía al laboratorio de patología. Este tipo de biopsia es necesaria en caso de que no se observen alteraciones en la colposcopia.

3. Biopsia de cono (conización): Se extrae una porción de tejido en forma de cono, donde la base del cono es el exocervix y el ápice está conformado por endocervix. Con este tipo de muestreo se puede evidenciar la zona de transformación, que es el límite entre endo y exocervix, donde es más probable encontrar precáncer. Al remover una cantidad consi-

derable de tejido, este procedimiento puede ser utilizado como terapia para extirpar tejido comprometido.

Aunque se dice que las biopsias de cervix son procedimientos poco riesgosos, pueden darse complicaciones como sangrado profuso, infección secundaria, dolor pélvico y dolor abdominal severo. Más aún, en zonas donde no existe acceso a servicios de salud, que casualmente se concentran en países no desarrollados, la posibilidad de realizar estos procedimientos es casi nula. Esto pone en riesgo de desarrollar cáncer mortal altamente prevenible.(Samaniego y Mariela, 2017)

Shobha y Davuluri evaluaron la eficacia de los análisis citopatológicos, como biopsias tomadas con colposcopía, para detectar lesiones precancerosas y cancerorsas. La eficiencia del Papanicolau es baja, sin embargo esto se solventa con monitoreos continuos y regulares y con seguimiento adecuado de los hallazgos patológicos. El uso extendido del Papanicolau como prueba de elección para monitorear cambios en el cuello uterino se debe a su facilidad y rapidez. (Shobha y Davuluri, 2016) Sin embargo, existen otras pruebas sencillas que se realizan durante la exploración colposcópica.

La colposcoipia endocervical con cureta busca la detección de lesiones que no son directamente observables. Liu y cols. evaluaron la detección de neo-

plasia intraepitelial cervical de alto grado por medio de este tipo de biopsias. Se recomienda que este muestreo, por ser más invasivo, se realice a mujeres de más de 45 años o más, que estén infectadas por el HPV 16 y mujeres mayores de 30 años con lesiones de alto grado. (Liu y cols., 2017)

La prueba de Schiller, desarrollada por Dr. Walter Schiller que era contemporáneo del Dr. Hans Hinselmann que desarrolló la colposcopia, se basa en la excreción normal de glucosa por la mucosa cervical. Al aplicar una solución de yodo directamente al cuello uterino se produce una coloración café en caso de que la mucosa sea normal. Las zonas alteradas no excretan glucosa, por lo que estas zonas son candidatas a que se tome una biopsia para investigar más a fondo el origen de la lesión. (Aba y cols., s.f.) La prueba de Schiller complementa a la colposcopia y es barata, de rápida ejecución y mínimamente invasiva.

Dada la alta mortalidad del cáncer de cervix y su difícil detección en etapas tempranas, y la importancia del diagnóstico temprano en la disminución de morbilidad y mortalidad, enfoque de minería de datos se han utilizado para generar modelos predictivos más sensibles que las guías diagnósticas actualmente utilizadas. (Wu y Zhou, 2017) utilizaron un enfoque de máquina de soporte de vectores (SVM) para 32 factores de riesgo y 4 clases: Hinsel-

mann o colposcopia, la prueba de Schiller, el resultado del Papanicolau y si se realizó biopsia a la paciente. En su estudio determinaron que SVM acoplado a análisis de componentes principales presentan los mejor resultados de clasificación. (Wu y Zhou, 2017)

Los enfoques de aprendizaje de máquina para el diagnóstico de cáncer cervical permiten identificar de forma más temprana y más precisa potenciales nuevos casos a partir ya sea de metadata obtenida de entrevistas con las pacientes o hallazgos de imágenes de las lesiones. (Selvathi, Sharmila, y Sankari, 2018)

# Capítulo 3

# Área de aplicación

Un enfoque de minería de datos permite encontrar patrones o tendencias estadísticas emergentes que permiten clasificar la data de maneras que no son evidentes previo al análisis.(Kurniawati, Permanasari, y Fauziati, 2016)

La detección de cáncer cervical puede ser un reto complicado en algunos casos, ya que se desarrolla de forma asintomática y la diferenciación de células anormales en la citología de cuello uterino requiere de analistas experimentados. (Devi, Ravi, Vaishnavi, y Punitha, 2016)

La generación de modelos que determinen la necesidad de realizar un biopsia, que es el exámen de elección para descarta o confirmar la presencia de precáncer o cáncer de cervix, a partir de metadata generada en los centros

de salud promete ser un gran soporte diagnóstico para el médico tratante. Con este enfoque es posible detectar posibles cuadros de cáncer asintomáticos y llevar la atención de los servicios de salud hacia las pacientes vulnerables.

# Capítulo 4

# Metodología: Desarrollo del Sistema

## 4.1. Selección del dataset

Posterior a la exploración de bases de datos como 1000 Genomes, Gene Expression Omnibus (GEO), Human Microbiome Project (HMP), ICOS PSP Benchmark, MIT Cancer Genomics Data, NIH, Microarray data (FTP), Protein Data Bank, PubChem Project y Kaggle se eligió el dataset **Cervical Cancer Risk Classification** encontrado en `https://www.kaggle.com/loveall/cervical-cancer-risk-classification/data`. Este dataset in-

cluye una lista de factores de riesgo para Cáncer Cervical que indican la toma de una biopsia de cuello uterino.

## 4.2. Selección de plataforma de análisis

Se eligió la plataforma `Jupyter` para realizar los análsis debido a que permite tanto correr código de `Python` y sus librerías `pandas` y `scikit-learn` como hacer anotaciones *human-readable* que permiten la explicación detallada de los pasos del análisis.

La librería de `Python` `scikit-learn` ofrece métodos para hacer análisis de minería de datos fácilmente implementables. Diferentes tipos de análisis que se pueden realizar son: preprocesamiento, selección de modelos, reducción de dimensionalidad, agrupamiento, regresión y clasificación.

## 4.3. Exploración y preprocesamiento del dataset

El dataset presenta 35 atributos y una clase que representa si a la paciente se le realizó biopsia de cuello uterino. Tiene además 858 instancias que son el

número de pacientes cuya información fue recopilada para generar el dataset.

Una primera visualización de los datos demostró que algunos atributos presentaban datos faltantes en la mayoría de sus instancias. Por esta razón, dos atributos, **STDs: Time since first diagnosis** y **STDs: Time since last diagnosis** fueron eliminadas del dataset. La mayoría de los atributos corresponden a variables binarias, respuesta de sí o no. Se eligió un enfoque de **clasificación supervisada** debido a las características del dataset y la naturaleza de la pregunta de investigación.

## 4.4. Selección de Atributos

Por medio de una implementación del algoritmo `ExtraTreesClassifier` se calculó el peso de cada atributo en la clasificación y con el método `RFECV` se seleccionó el número óptimo de features.

## 4.5. Clasificación Supervisada

El conjunto de datos se dividió en 70 % entrenamiento y 30 % pruebas usando `train_test_split`. El primer clasificador utilizado fue `DecisionTreeClassifier` que es un clasificador basado en árboles de decisión que funciona bastante

bien con datos binarios.

Se optimizó el parámetro `max_depth` por medio de una implementación de un método iterativo que representa de forma gráfica los valores de precisión de la clasficación en un rango de valores del parámetro optimizado. Se optimizó tambien `min_samples_split` para validar el comportamiento del modelo.

El segundo clasificador utilizado fue `MLPClassifier` que es una implementación de algoritmos de redes neuronales básicas en `scikit-learn`. Se aplicó el anterior método de optimización de parámetros a `hidden_layer_sizes` que controla el número de capas ocultas dentro del modelo de predicción.

## 4.6. Validación Cruzada

Para la validación cruzada del modelo se utilizaron `ShuffleSplit` y `LeaveOneOut`. `ShuffleSplit` es un método para validación cruzada que realiza permutaciones aleatorias de las separaciones de los datos en entrenamiento y prueba. `LeaveOneOut` divide el set de datos en entrenamiento y prueba de tal forma que cada muestra es usada una vez como el set de prueba mientras que el resto del set de datos conforma el entrenamiento, lo que lo hace computacionalmente costoso y su utilidad se reduce de forma proporcional

al tamaño del set de datos.

## 4.7. Clasificación No Supervisada

Adicionalmente se implementó el algoritmo `KMeans` de clasificación no supervisada para explorar propiedades emergentes dentro del set de datos que potencialmente pueden relacionarse con el diagnóstico de cáncer cervical.

# Capítulo 5

# Resultados

La aplicación de los clasificadores mencionados anteriormente, generaron resultados interesantes:

| Clasificador | Score Entrenamiento | Score Pruebas |
|---|---|---|
| Decision Tree | 97.5 % | 94.6 % |
| Multi-layer Perceptron | 98.3 % | 93 % |

Se identifica un *overfitting* notable en los resultados de ambos clasificadores, sugiriendo la necesidad de una mayor precisión de los datos o mayor cantidad de atributos para un mejor ajuste. Lo confirma la ejecución de la validación cruzada con un resultado de `96 %` en ambos métodos.

Sin embargo, ambos resultados muestran ajustes altos, lo cual es consi-

derado alentador para el conjunto de datos utilizados. Como recomendación para implementaciones futuras, es importante recurrir a otras fuentes de información para la obtención de datos confiables y precisos, como expedientes médicos y resultados directos de examenes médicos.

Los principales atributos que contribuyen al modelo son:

- Schiller
- Hinselmann
- Edad
- Edad en la primera relación sexual
- Años con tratamiento anticonceptivo hormonal
- Número de parejas sexuales
- Resultado de la citología de cuello uterino

A pesar de su asociación con cáncer cervical, el diagnóstico del virus de Papiloma Humano (HPV) es el número 19 en el ranking de atributos. El ranking de los atributos concuerda con lo que se espera en la realidad. La prueba de Schiller en conjunto con la prueba de Hinselmann, conocida como colposcopia, son los mejores predictores para determinar si una mujer requiere biopsia

de cuello uterino. Estas pruebas evidencian directamente la presencia de lesiones en la mucosa del cuello uterino. Los demás atributos tienen que ver con factores de riesgo asociados a la infección con el virus de Papiloma Humano. Conforme aumenta la edad, aumentan el número de encuentros sexuales y por tanto la probabilidad de infectarse con alguna cepa del HPV carcinogénica. La edad a la cuál ocurrió el primer encuentro sexual indica la extensión del período de actividad sexual de la mujer, que conforme aumenta, aumenta la probabilidad de infección. El número de parejas sexuales sigue una lógica similar a los atributos anteriores. El uso de anticonceptivos orales se puede asociar al no uso de preservativos que ofrecen una barrera física ante la infección con el HPV. Se puede atribuir el bajo efecto del diagnóstico de HPV al modelo al subdiagnóstico de la infección, ya que los métodos diagnósticos suelen ser pruebas moleculares sofisticadas que no están al alcance de países pobres como lo es Venezuela, perpetuando la situación de riesgo que tienen las mujeres que habitan estos países pobres ante la aparición de cáncer de cuello uterino.

En cuanto a la implementación de una metodología no supervisada, se abre la posibilidad de obtener predicciones instantáneas que determinen si al paciente se le identifica o no un padecimiento cervical. Si esta implementación

se realiza de forma exitosa, la ejecución en tiempo real de este algoritmo bajo condiciones de investigación, puede beneficiar las poblaciones con mas riesgo de padecer esta enfermedad, sin la necesidad de aplicar gran cantidad de examenes físicos dolorosos e invasivos.

# Capítulo 6

# Conclusiones

- Los principales atributos que aportan poder de prediccón al modelo son: Schiller, Hinselmann, Edad, Edad en la primera relación sexual, Años con tratamiento anticonceptivo hormonal, Número de parejas sexuales y el Resultado de la citología de cuello uterino.

- El modelo predictivo permite predecir si la paciente requiere una biopsia de cuello uterino basado en datos epidemiológicos y no exclusivamente de métodos analíticos.

- La representación de los clusters en el aprendizaje supervisado permite identificar los comportamientos existentes en el set de datos, con el fin de estudiarlos a más detalle e identificar a qué patología puede

corresponder cada grupo.

- Se comprueba que el flujo de trabajo de un conjunto de datos va a depender del comportamiento que se obtenga de los diferentes clasificadores. No existe una sola forma de trabajar los datos, y el siguiente paso va a depender solamente del resultado del paso actual.

# Referencias

Aba, Y. A., ŞIK, B. A., ŞENTÜRK, M., KUMBASAR, S., SÜMER, E., ABA, Y. A., y DÜLGER, Ö. (s.f.). Cervical cytologic and colposcopic changes in cases using iuds for a long time. *International Journal of Health Services Research and Policy*, *2*(1), 1–9.

Devi, M. A., Ravi, S., Vaishnavi, J., y Punitha, S. (2016). Classification of cervical cancer using artificial neural networks. *Procedia Computer Science*, *89*(Supplement C), 465 - 472. Descargado de `http://www.sciencedirect.com/science/article/pii/S187705091631170X` (Twelfth International Conference on Communication Networks, ICCN 2016, August 19– 21, 2016, Bangalore, India Twelfth International Conference on Data Mining and Warehousing, ICDMW 2016, August 19-21, 2016, Bangalore, India Twelfth International Conference on Image and Signal Processing, ICISP 2016,

August 19-21, 2016, Bangalore, India) doi: https://doi.org/10.1016/j.procs.2016.06.105

Frega, A., Boselli, F., Buttignol, M., Cervico-Vaginal, P. S., Pieralli, A., Ciavattini, A., ... others (2016). High-grade vaginal intraepithelial neoplasia and risk of progression to vaginal cancer: a multicentre study of the italian society of colposcopy and cervico-vaginal pathology (sicpcv). *European review for medical and pharmacological sciences*, *5*(20), 818–824.

Kurniawati, Y. E., Permanasari, A. E., y Fauziati, S. (2016, Oct). Comparative study on data mining classification methods for cervical cancer prediction using pap smear results. En *2016 1st international conference on biomedical engineering (ibiomed)* (p. 1-5). doi: 10.1109/IBIOMED.2016.7869827

Liu, A. H.-C., Walker, J., Gage, J. C., Gold, M. A., Zuna, R., Dunn, S. T., ... Wentzensen, N. (2017). Diagnosis of cervical precancers by endocervical curettage at colposcopy of women with abnormal cervical cytology. *Obstetrics & Gynecology*.

Samaniego, P., y Mariela, J. (2017). *Prevalencia y factores asociados a las lesiones intraepiteliales del cuello uterino, en mujeres de 15 a 65 años*

*de edad que, acudieron al servicio de colposcopia del hospital vicente corral moscoso de la ciudad de cuenca, periodo 2014–2015.* (B.S. thesis). Universidad Católica de Cuenca.

Selvathi, D., Sharmila, W. R., y Sankari, P. S. (2018). Advanced computational intelligence techniques based computer aided diagnosis system for cervical cancer detection using pap smear images. En *Classification in bioapps* (pp. 295–322). Springer.

Shobha, T., y Davuluri, S. (2016). Cervicalcytopathology: evaluation of its efficacy in detecting cervical precancerous and cancerous lesions, as evidenced by colposcopic biopsy. *Int J Sci Res*, *5*(5).

Téguété, I., Dolo, A., Sangare, K., Sissoko, A., Rochas, M., Beseme, S., ... Koita, O. A. (2017). Prevalence of hpv 16 and 18 and attitudes toward hpv vaccination trials in patients with cervical cancer in mali. *PloS one*, *12*(2), e0172661.

Wu, W., y Zhou, H. (2017). Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access*.