INSTITUTO NACIONAL DE SALUD PÚBLICA

ESCUELA DE SALUD PÚBLICA DE MÉXICO

PROYECTO DE TITULACIÓN

“PREDICCIÓN DE PERSISTENCIA DE INFECCIÓN POR VIRUS DEL PAPILOMA HUMANO EN CÉRVIX: BASADA EN MODELO DE APRENDIZAJE DE MÁQUINAS”

Maestría en Salud Pública área de concentración en Enfermedades Infecciosas

ALUMNA: JOCELYN ISABEL RODRÍGUEZ ESQUIVEL

Correo: joce.rdz.esq@gmail.com

Teléfono: (33) 3157 4902

Generación 2020-2022

Directora:

Dra. Kirvis Torres Poveda

Asesores:

MCCC. Blanca Hilda Vázquez Gómez

Dr. Vicente Madrid Marina

Cuernavaca, Morelos

Agosto, 2022

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE SIGLAS Y ACRÓNIMOS

- ESPM: Escuela de Salud de Pública de México
- INSP: Instituto Nacional de Salud Pública
- VPH: Virus del papiloma humano
- VPH-AR: Virus del papiloma humano de alto riesgo
- CaCU: Cáncer cervicouterino
- CAPASAM: Centro de Atención para la Salud de la Mujer
- NIC: Neoplasia intraepitelial
- E: Región temprana
- L: Región tardía
- LCR: Región reguladora
- VLP: Partícula similar a un virus
- LEI-BG: Lesiones escamosas intraepiteliales de bajo grado
- LEI-AR: Lesiones escamosas intraepiteliales de alto grado
- ML: Modelo de aprendizaje de máquinas
- SVM: Máquinas de soporte vectorial
- RF: Árboles de decisión
- XGBoost: Máquinas de aumento de gradiente extremo
- LR: Regresión logística
- LightGBoost: Máquinas ligeras de gradiente extremo
- SMOTE: Técnica de sobremuestreo de minorías sintéticas
- AUC: Área bajo la curva ROC
- ADASYN: Enfoque de muestreo sintético adaptativo para el aprendizaje desequilibrado
- SHAP: Explicaciones del aditivo SHapley
- EM: Expectativa de Maximización regularizada
- MI: Imputación múltiple
- kNNI: Imputación K-vecinos
- FS: Selección de características
- ANN: Red neuronal artificial
- FIGO: Federación Internacional de Ginecología y Obstetricia
- MLP: Perceptrón multicapa
- RBF: Función de base radial
- DNN: Red neuronal profunda
- ITS: Infecciones de transmisión sexual
- NOM: Norma oficial mexicana

- EHR: Historia clínica electrónica
- LES: Lupus eritematoso sistémico
- AR: Artritis reumatoide
- OTB: Oclusión Tubárica Bilateral
- DIU: Dispositivo Intrauterino
- AVE: Evaluación visual automatizada
- DL: Aprendizaje profundo

RESUMEN

El 99% de los casos de cáncer cervicouterino (CaCU) están relacionados con una infección genital por el Virus del Papiloma Humano (VPH). La persistencia del virus es esencial para el desarrollo de neoplasias malignas de alto grado a nivel del cérvix o CaCU. Sin embargo, en una minoría de casos se puede detectar la persistencia una vez transcurridos 12 meses, lo que aumenta el riesgo a desarrollar CaCU o una lesión precancerosa.

El objetivo del presente estudio fue analizar los marcadores de riesgo de persistencia de infección por VPH en cérvix mediante modelos de aprendizaje de máquinas en mujeres que recibieron atención en el Centro de Atención para la salud de la Mujer (CAPASAM) de los Servicios de Salud del Estado de Morelos en el período de 2015-2017.

Entre los resultados del modelo de predicción para persistencia viral anual se identificó que el modelo que obtuvo el mejor rendimiento fue XGBoost. Los rendimientos obtenidos fueron del 100% de especificidad, 90% de sensibilidad y 95% de precisión. Usando un enfoque de interpretabilidad se identificó que la variable más relevante para la persistencia viral a la co-infección de VPH genera, seguido de las variables de número de parejas sexuales y el antecedente de tabaquismo

En relación a la predicción de predicción para persistencia al segundo año, se usó el modelo de XGBoost logrando rendimientos del 100% de especificidad, sensibilidad del 98% y precisión de 99%. En contraste con el primer año, las variables más relevantes fueron la edad de debut sexual, seguido del número de parejas sexuales y en tercer lugar se encontró a la co-infección por genotipos de VPH general.

Los modelos desarrollados en el presente trabajo presentaron buenos resultados para la predicción de persistencia de Virus del papiloma humano de alto riesgo (VPH-AR). No obstante, es necesario la implementación de estudios más amplios para poder tener una mayor representatividad de la población mexicana. Por lo que, este estudio puede ser considerado como una guía para próximos estudios que ayuden a centrar las intervenciones en los factores de pronóstico específicos y lograr la optimización en el diagnóstico precoz de persistencia de infección por VPH a nivel de cérvix y el tratamiento oportuno para cada paciente en caso de lesiones premalignas.

INTRODUCCIÓN

La infección por VPH es considerada la infección de transmisión sexual (ITS) más prevalente. En los últimos años, la presencia del VPH ha tenido gran relevancia entre la población femenina, pues la infección por genotipos de VPH-AR es un factor de riesgo para presentar persistencia viral y con ello contribuir en el desarrollo de CaCU (1-5).

La infección por VPH representa un problema de salud importante en México, se presenta una prevalencia del VPH en mujeres en edad reproductiva de 11-13% (3,4). Aproximadamente el 25% de las mujeres en México han presentado al menos una prueba de triage positiva y se espera que en el transcurso de pocos años, aquellas mujeres positivas a VPH, desarrollen alguna neoplasia intraepitelial (NIC) -2 (5).

Por su parte el CaCU ocupa el cuarto lugar como causa de muerte por cáncer entre las mujeres en todo el mundo, con una tasa de mortalidad global estimada de 7.3 por cada 100,000 mujeres (1). Además, en México el CaCU fue la segunda causa de muerte por cáncer en mujeres en el año 2020 (5.7 muertes por cada 100,000 mujeres) (2).

Como ya se mencionó anteriormente la persistencia del virus es esencial para el desarrollo de neoplasias malignas de alto grado a nivel del cérvix o CaCU. Sin embargo, en una minoría de casos se puede detectar la persistencia una vez transcurridos 12 meses, lo que aumenta el riesgo a desarrollar CaCU o una lesión precancerosa (3).

En diversos estudios se han correlacionado algunos factores con la presencia de tasas de persistencia más altas entre ellos se incluyen a la edad, la inmunodeficiencia, el tabaquismo, el uso prolongado de anticonceptivos orales, la infección por *Chlamydia trachomatis*, los cambios en el microbioma cervicovaginal y la multiparidad (6,7).

Es por esto que, desde hace algunos años, se ha prestado atención en la identificación de marcadores predictores de riesgo de persistencia de VPH en cérvix en población mexicana con el objetivo de detectar tempranamente aquellas mujeres que tendrían un riesgo mayor de desarrollo de lesiones premalignas y CaCU. Para la identificación de los marcadores de riesgo se han empleado múltiples estudios, tanto de control clínico, como técnicas computacionales con modelos de aprendizaje de máquinas (5, 8).

Los modelos de aprendizaje de máquinas son herramientas auxiliares que han sido enfocadas en la predicción de mortalidad, anticipación de eventos importantes, identificación de factores de riesgo, entre otros, con el objetivo de contribuir en la mejora del diagnóstico y en los resultados de los pacientes (8).

Por ello el empleo de métodos de aprendizaje de máquinas puede ayudar en el reforzamiento de la metodología científica y brindar una alternativa más confiable, eficiente y precisa para apoyar el diagnóstico de persistencia viral y detectar factores de riesgo relacionados en su presentación y así reducir la mortalidad por CaCU.

1. ANTECEDENTES

1.1 Generalidades del Virus del Papiloma Humano (VPH)

El VPH pertenece a la familia *Papillomaviridae* (9). Es un virus pequeño de ADN de 60 nm de diámetro que consta de una sola molécula de ADN circular bicatenaria de aproximadamente 8.000 pares de bases. Este virus no cuenta con envoltura y se divide en tres regiones: la región temprana (también conocida como *Early o E* por sus siglas en inglés), la región tardía (*late* o L por sus siglas en inglés) y la región reguladora (LCR, por sus siglas en inglés *Locus Control Region*) (1). La región E, codifica las proteínas E1, E2, E4, E5, E6 y E7, las cuales son usadas durante la replicación del ADN viral, la regulación de la trascripción y la transformación e inmortalización celular (1). La región L, en la cual se producen proteínas estructurales L1 y L2, que son necesarias para encapsular el genoma viral, mientras que la región LCR en la que se encuentra la secuencia de ADN es dónde se realiza la replicación y de la expresión del genoma viral (10).

Hasta este momento se han identificado alrededor de 200 genotipos del VPH, éstos pueden infectar las células epiteliales basales de la piel o el revestimiento interno de los tejidos y se clasifican en tipos cutáneos o tipos mucosos (11,12). También de acuerdo a su asociación con el CaCU y las lesiones precursoras, se pueden agrupar en tipos de VPH de alto y bajo riesgo (13). El mecanismo de acción de los VPH-AR en el desarrollo de la neoplasia en cérvix, se explica principalmente por la acción de dos de sus proteínas virales: E6 y E7. Estas tienen la capacidad de inmortalizar y transformar queratinocitos, confiriéndoles un alto grado de inestabilidad cromosómica (11).

La actividad oncogénica de la proteína E6 de los VPH de alto riesgo se ha correlacionado con su capacidad para interactuar con la proteína p53 celular e inactivarla, la cual tiene como una de sus principales funciones ser supresora de tumores (10). La oncoproteína E7 se ha demostrado que es capaz de formar complejos con el gen supresor de tumores de retinoblastoma (p105-RB), las deleciones o mutaciones de éste, están implicados en la progresión del ciclo celular y son características comunes de muchos tumores (12).

1.2 Ciclo viral de VPH

Las micro abrasiones en el epitelio del cuello uterino son la entrada del VPH a la célula del huésped. Luego de la infección pueden presentarse cuatro situaciones clínicas, como se muestra en la figura 1:

1. Infección activa.
2. Regresión inmune.
3. Estado de latencia viral, sin signos ni síntomas presentes.
4. Estado de latencia con reactivación posterior.

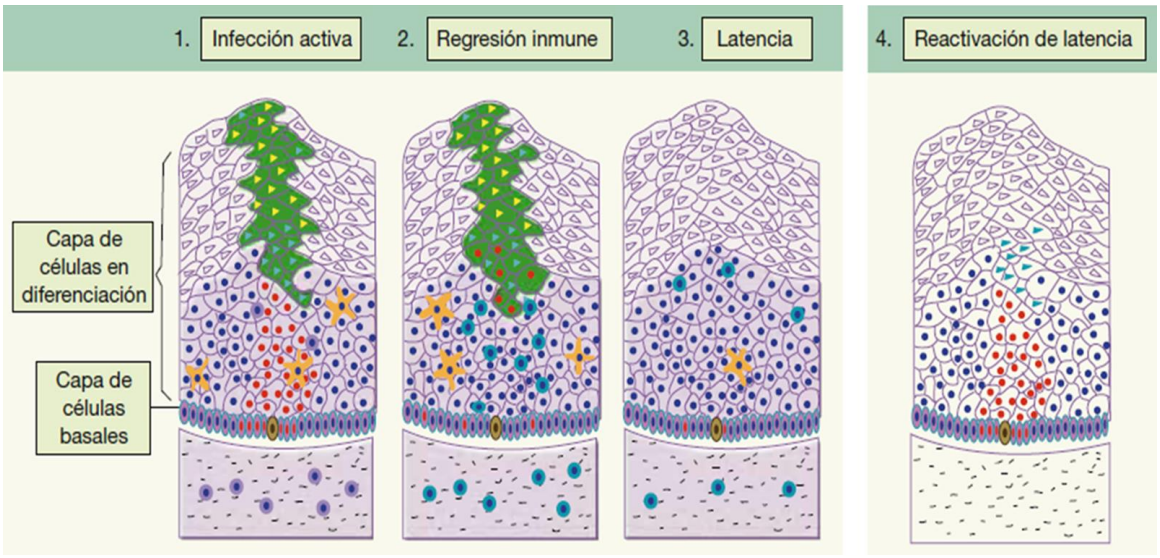


Figura 1. Esquema del ciclo viral de VPH
Imagen modificada de Meglennon G. et al., 2012 (14)

A continuación, se presenta un breve resumen descriptivo de cada una de estas situaciones (14).

1.2.1 Infección activa

La infección activa inicia con la expresión regulada de proteínas virales, aquellas células infectadas se sitúan en la superficie epitelial. Posteriormente se inicia la amplificación y la expresión del genoma viral y por medio de la proteína L1 los virones son liberados en el epitelio. Las células portadoras de VPH mantienen una expresión viral baja y de esta manera logran

poder evadir a las células de Langerhans (células presentadoras de antígenos) y las células T que se encuentran en capas inferiores (14).

1.2.2 Infección activa

Para que se inicie la regresión inmune es necesario que los antígenos virales sean presentados a las células de Langerhans, éstas a su vez activan a las células TCD4+ y CD8+ que se encargan de la eliminación de las células infectadas, penetrando y rodeando la zona lesionada (14).

1.2.3 Latencia

En esta etapa las células basales infectadas proliferan; sin embargo, no expresan el genoma viral. También se relaciona con los cambios que ocurren en los linfocitos que modifican la activación de las citocinas. Se cree que el genoma puede persistir en dichas células por un periodo largo con un ciclo lento de replicación (14).

1.2.4 Reactivación de latencia

Un período de inmunosupresión puede favorecer la replicación de copias virales, al disminuir la presencia de células T de memoria que son las encargadas de regular que no existe esta replicación, y con esto presentarse la reactivación de la latencia, por lo tanto, tener nuevamente manifestaciones clínicas (14).

1.3 Historia natural de la infección por VPH

La infección por VPH se considera una enfermedad de transmisión sexual. Para la transmisión es necesario el contacto de una mucosa infectada con otra, por lo que se encuentran expuestos tanto los hombres como las mujeres (11). Es por esto que se realizan medidas de prevención primaria, basada en educación y salud sexual para el uso correcto de métodos de barrera. Aunado a esto, se aplica la vacunación con partículas similares a virus (VLP, por las siglas en inglés de *Virus-like particle*). En México, el esquema de vacunación contra VPH incluye 2 dosis en niñas de 5° grado de primaria o de 11 años de edad, no escolarizadas y más recientemente el esquema en una sola dosis (12).

Una vez presente la infección por VPH, en el 90% de los casos en las células supra basales del epitelio cervical es el sitio donde se lleva a cabo la amplificación inicial del genoma, esto por medio de la expresión de los genes L1 y L2. Este virus por sí solo no puede llegar a los órganos linfoides regionales, debido a que afecta principalmente a los queratinocitos, lo que nos indica que la propia respuesta inmune (que es mediada por las células de Langerhans en el epitelio) puede inducir una respuesta inmunológica eficaz contra el virus (10–12). La eliminación viral ocurre al transcurso de uno o dos años, que es cuando la expresión del genotipo viral se vuelve indetectable (15).

En una minoría de casos se puede detectar la persistencia una vez transcurridos 12 meses, lo que aumenta el riesgo a desarrollar CaCU o una lesión precancerosa (lesiones intraepiteliales escamosas de alto grado o neoplasia intraepitelial 2° y 3°) (15). En este nivel es importante la realización de medidas de prevención secundaria, enfocadas en el diagnóstico oportuno, detección en estadios tempranos para recepción inmediata de tratamiento (12). Se sabe que cada infección por VPH es distinta, no excluyente entre sí a lo largo de la vida de cada mujer. Ejemplo de esto se puede explicar a la nueva detención del virus, que puede ser por una infección reciente o por una reactivación de enfermedad controlada o latente, o incluso por la autoinoculación en otros sitios epiteliales (15).

Finalmente, debido a la función de las proteínas E6 y E7, el virus del VPH causa que las células epiteliales no realicen la apoptosis e inhiban la actividad de la proteína quimiotáctica, permitiendo que el virus se siga replicando. Además, se sabe que específicamente el gen E6 inhibe la interacción de la célula epitelial con la célula dendrítica e inactiva el gen supresor de tumores p53, facilitando con esto la progresión al cáncer. Mientras que la proteína E7 tiene la capacidad de inhibir la función del supresor de tumores de retinoblastoma pRB, CDK2, CDKN1A y CDKN1B (10–12).

La persistencia del virus es esencial para el desarrollo de neoplasia intraepitelial del cérvix (NIC) de alto grado y para el desarrollo de CaCU. Los factores que se han correlacionado con tasas de persistencia más altas incluyen edad, inmunodeficiencia, tabaquismo, uso prolongado de anticonceptivos orales, infección por *Chlamydia trachomatis*, el microbioma cervicovaginal y múltiples nacidos vivos (6,7). Se cree que las respuestas inmunitarias sistémicas y locales son importantes para la persistencia frente a eliminación viral (5). Una vez detectado el CaCU, sin importar la etapa de éste, es importante seguir realizando intervenciones de prevención

terciaria para limitar el daño y asegurar la asistencia de salud en el transcurso de la enfermedad (12).

1.4 Epidemiología de la infección por VPH

El 99% de los casos de CaCU están relacionados con una infección genital por el VPH (15). En el mundo, anualmente se presentan aproximadamente 630,000 casos nuevos de cáncer causados por una infección por VPH (16). El CaCU ocupa el cuarto lugar como causa de muerte por cáncer entre las mujeres en todo el mundo, con una tasa de mortalidad global estimada de 7.3 mujeres por cada 100,000 mujeres (1,2). En el 2020, el CaCU fue la segunda causa de muerte por cáncer en mujeres mexicanas (5.7 por 100,000 mujeres) (1,2).

La prevalencia mundial de la infección por VPH en mujeres sin anomalías del cuello uterino es del 11 al 12%, y las tasas más altas se encuentran en África subsahariana (24%), Europa del Este (21%) y América Latina (16%) (4). Desde el punto de vista de la patología cervical, el VPH 16 y el 18 representan la gran mayoría de los casos de CaCU. En todo el mundo, el VPH 16 por sí solo representa casi el 60% de los casos (más frecuente) y el VPH 18 el 10% de los casos (3).

La prevalencia general del VPH es alta alrededor de la edad del debut sexual. Como resultado de esta transmisión elevada, las infecciones por VPH son muy comunes en mujeres jóvenes entre los 20 y los 25 años de edad. Posterior a esta edad de debut sexual, se produce un descenso abrupto como resultado de la eliminación natural de la infección y a una menor exposición a nuevas parejas sexuales. En la figura 2, se observan las variaciones a la izquierda/derecha de la edad pico (edad de mayor prevalencia de VPH) que se relacionarán con la edad promedio de la iniciación sexual en una población dada. El número promedio de intercambio de parejas sexuales en hombres y mujeres modulará la intensidad del pico de mayor prevalencia de VPH. La forma de la curva también podría verse afectada por las prácticas de tamizaje (17).

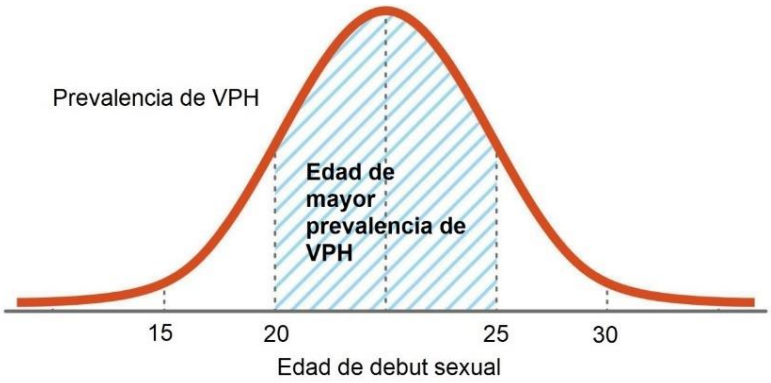


Figura 2. Curva descriptiva de prevalencia de VPH según edad de debut sexual
Basado en el texto de Gravitt et al (17), elaboración propia.

Particularmente en población mexicana, en una revisión sistemática elaborada por Peralta et al. (18) se reportaron datos de prevalencia y distribución de genotipos de VPH en mujeres mexicanas con CaCU, lesiones escamosas intraepiteliales de bajo grado (LEI-BG), LEI-alto grado (AG) y en citologías normales. De un total de 8,706 muestras de tejidos de mujeres mexicanas estratificados de acuerdo al diagnóstico (499 para CaCU, 364 para LEI-AG, 1,425 para LEI-BG y 6,418 para citología normal), los genotipos se resumen en la tabla 1. De tal manera, que los genotipos 58 y 31, suman un 10% de los genotipos más prevalentes en CaCU.

Tabla 1. Prevalencia y distribución de genotipos de VPH en mujeres mexicanas con CaCU,LEI-BG, LEI AG y citologías normales (18)

CACU	
Genotipos	Porcentaje
VPH 16	63.10%
VPH 18	8.60%
VPH 58	5%
VPH 31	5%
LEI-AG	
VPH 16	28.30%
VPH 58	12.60%
VPH 18	7.40%
VPH 33	6.50%
LEI-BG	
VPH 16	13.10%
VPH 33	7.40%
VPH 18	4.20%
VPH 58	2.60%
Citologías normales	

VPH 33	2.10%
VPH 18	1.20%
VPH 58	1.20%

En otro estudio realizado por Parada et al. (19) en parejas heterosexuales, se reportó infección por VPH en 13,7% de las mujeres. Los genotipos de alto riesgo más frecuentemente detectados fueron VPH 59, 16, 31, 52 y 58. Los genotipos de bajo riesgo más frecuentes fueron los VPH 62, 71, 81 y 54.

Finalmente es importante recordar que la incidencia en esta enfermedad está relacionada principalmente al comportamiento sexual, aunque se sabe que este virus es resistente al calor y la desecación, así como también que puede haber transmisión no sexual a través de fómites o el usar ropa interior contaminada, sin embargo, es la exposición a múltiples parejas sexuales, la falta de uso de métodos de barrera y el inicio de vida sexual a temprana edad los factores son los que aumentan el riesgo de contraerla (3,4,10,11).

1.5 Diagnóstico

Para realizar el diagnóstico de VPH, existen métodos de detección por medio de ADN viral o por ARNm de oncogenes E6/E7, también varían en tanto a la capacidad de genotipos virales que pueden detectar y a la sensibilidad de las pruebas, en la tabla 2 se muestra un resumen de las diferentes técnicas de detección disponibles comercialmente (20).

Tabla 2. Tabla resumen de técnicas de detección de VPH disponibles comercialmente

Pruebas de detección de ADN del VPH de alto riesgo							
Prueba	**Fabricante**	**Genotipos de VPH detectados**	**Molécula diana**	**S**	**E**	**Resultado detectado**	**Calidad de la evidencia (GRADE)**
Captura híbrida® 2 (HC2) *	QIAGEN	Genoma viral	13 genotipos cancerígenos	97%	85%	NIC2+	Alto
2 COBAS ® 4800 *	Roche	L1	13 genotipos cancerígenos y VPH 66; genotipado para HPV 16 y 18	90%	94%	NIC2+	Alto
BD ONCLARIDAD VPH	BD	E6/E7	14 genotipos. Tiene los genotipos 16, 18 y 45	95%	87,7%	NIC2	Alto

1 CareHPV™^	QIAGEN	genoma viral	13 genotipos cancerígenos y VPH 66	90%	84%	NIC2+	Alto
GP5 * /GP6 + PCR-EIA		L1	13 genotipos cancerígenos y VPH 66	94%	90%	NIC2+	Alto
CUÁDRUPLE INFINITI® HPV-HR	Autogenómica	E1	13 genotipos cancerígenos y VPH 66	97,30%	90%	NIC2	Moderado
Anyplex II VPH HR	SeaGen INC		14 genotipos	94-92%	81%	HSIL, CIN2+	Moderado
Cervista ™ HPV HR	Hologic		14 genotipos	89%	91%	NIC2+	Moderado
Alto riesgo en tiempo real	Abad	L1	13 genotipos cancerígenos y VPH 66; genotipado para HPV 16 y 18	95%	92%	NIC2+	Moderado
Pruebas de detección de ARN del VPH de alto riesgo							
APTIMA® *	GenProbe	ARNm de E6/E7	13 genotipos cancerígenos y VPH 66	94,2 % 87-98% 98% 90-100 %	94,5% 63-90% 55-60%	CIN2+ ASCUS CIN2/CIN3 + LSIL CIN2+/CIN 3+	Alto
Proofer PreTect	NorChip	ARNm de E6/E7	VPH 16, 18, 31, 33 y 45	75-79%	100%	NIC2+	Moderado

Nota: S: Sensibilidad; E: Especificidad.
1 Recomendado por la OMS para países de bajos y medianos ingresos.
2 Aprobado para detección primaria.
* Pruebas aprobadas por la FDA.
^ Prueba de bajo costo validada en el medio rural.
Modificado de Torres-Poveda, K et al. 2020 (20).

En México, las estrategias de prevención secundaria del CaCU incluyen la detección temprana de lesiones cervicales premalignas en mujeres de 25 a 34 años, ya sea mediante una prueba de Papanicolaou (Pap) o visualización directa con ácido acético cuando no se dispone de una prueba de Papanicolaou, y pruebas biomoleculares para la detección del VPH en mujeres de 35 a 64 años. Estas pruebas deben realizarse de forma gratuita a todas las mujeres solicitantes en los establecimientos de salud del sector público. De acuerdo con las políticas institucionales, el personal de atención primaria de salud puede ofrecer pruebas de captura híbrida (HC) o de reacción en cadena de la polimerasa (PCR) (21).

1.6 Tratamiento para infección por VPH

Según las manifestaciones clínicas presentadas por el huésped se presentan distintos esquemas de tratamiento. En la presencia de verrugas genitales las opciones de tratamiento incluyen ablación, escisión o agentes tópicos como podofilina al 0,5% (Podocon) o imiquimod al 5,0% (Aldara). En las lesiones neoplásicas el tratamiento dependerá en el estadio en que se encuentren (7). Mientras que el tratamiento para las lesiones anogenitales asociadas al VPH se ha basado principalmente en la escisión quirúrgica (10).

2. MARCO NORMATIVO

El marco normativo que reglamenta la vigilancia epidemiológica de la infección por el VPH asociada a CaCU y como ITS incluye dos normas oficiales, la NOM-014-SSA2–1994 para la prevención, detección, diagnóstico, tratamiento, control y vigilancia epidemiológica del CaCU. Ésta establece que el diagnóstico presuncional de CaCU se puede establecer por examen clínico, citología de cuello uterino y/o colposcopía. El diagnóstico definitivo se establece únicamente con el examen histopatológico (21).

Así como la NOM-039-SSA2-2014 para la prevención y control de las infecciones de transmisión sexual, en la cual se recomienda realizar programas educativos para capacitar a los profesionales, técnicos y auxiliares de la salud en ITS que incluyan: la prevención, consejería, mecanismos de transmisión, diagnóstico y tratamiento y realizar programas educativos para informar sobre las ITS, sus mecanismos de transmisión, diagnóstico y tratamiento a la población general. También se menciona que el esquema de vacunación en niñas de nueve años se realizará conforme a lo establecido en los Lineamientos Generales del Programa de Vacunación Universal vigente (22).

Asimismo, se cuenta con la Guía de Práctica Clínica “Prevención, detección, diagnóstico y tratamiento de lesiones precursoras del CaCU en primer y segundo nivel de atención” fecha de actualización: Junio, 2018. La cuál pretende unificar el manejo y seguimiento de la mujer en riesgo de padecer CaCU a partir de su contacto con el personal de salud del primer nivel de atención (23).

3. MARCO TEÓRICO

3.1 Aprendizaje de máquinas

De acuerdo con Tom M. Mitchell (24) describe que el aprendizaje de máquinas es un área que estudia el desarrollo de programas de computadoras que optimicen su desempeño en alguna tarea específica por medio de la experiencia. Según otros autores refieren a los modelos de aprendizaje de máquinas (Machine learning o ML por sus siglas en inglés) como *"adquirir conocimientos, comprensión o destreza mediante el estudio, la instrucción o la experiencia"*, y la *"modificación de una tendencia conductual mediante la experiencia"* (24,25).

En palabras más simples el aprendizaje de máquinas es un conjunto de métodos que usan las computadoras para hacer y mejorar predicciones o comportamientos basados en datos, según lo refiere Christoph Molnar (26). En la figura 3 se muestra el proceso de construcción de los métodos de aprendizaje de máquinas, el cual consiste en la extracción de las características necesarias para que los algoritmos puedan detectar automáticamente patrones en esos datos, estas características son la entrada a los algoritmos de aprendizaje de máquinas, donde son interpretadas y usadas para desarrollar una salida que sería nuestra predicción (27). Por ejemplo, es lo mismo que ocurre cuando recibes correos electrónicos, en un inicio aquellos correos son recibidos en la bandeja de entrada, pero aquellos que nosotros no queremos, los enviamos a la carpeta de spam, así poco a poco le vamos enseñando a nuestra máquina que correos no son bien recibidos, ella aprende los patrones que se encuentran en ellos y posteriormente de manera automática los procesa y envía a la carpeta de spam, y no a nuestra bandeja de entrada.

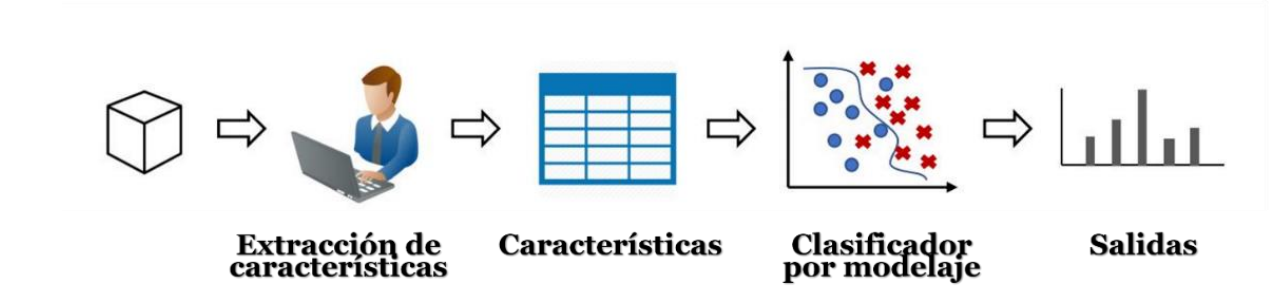


Figura 3. Proceso general para la generación de modelos de aprendizaje de máquinas
Imagen modificada de Megan D. et al., 2020 (29).

Este método se apoya principalmente con otras áreas del conocimiento como la teoría de la información, la estadística, la probabilidad, la inteligencia artificial, la psicología, la complejidad computacional y la teoría de control (24).

Los algoritmos de aprendizaje se clasifican en supervisados y no supervisados. Los algoritmos supervisados, son aquellos en donde se conoce el resultado final de la función que se desea realizar (etiqueta), es decir el modelo se entrena para predecir dicho resultado. Algunos ejemplos de algoritmos de aprendizaje de máquinas supervisado son árboles de decisión, máxima entropía, clasificador bayesiano, máquinas de soporte vectorial, etcétera (24,27).

En las técnicas de aprendizaje no supervisado, el objetivo es encontrar patrones similares o interesantes entre los datos que le son presentados. Ejemplos de las técnicas de aprendizaje no supervisado son las técnicas de *clustering*. Para la construcción de un modelo de aprendizaje de máquinas, en la práctica, regularmente se prueban distintos algoritmos de aprendizajes de máquinas y se comparan entre sí, esto con la finalidad de reducir sus fragilidades y aprovechar mejor las ventajas de cada uno. (24,25,27).

3.2 Modelos de aprendizaje de máquinas

3.2.1 Máquinas de soporte vectorial

Los algoritmos de máquinas de soporte vectorial (SVM, por sus siglas en inglés) forman parte de un ligado de algoritmos de aprendizaje estadístico supervisado desarrollados por Vladimir Vapnik en 1995 (28). La teoría de la SVM se basa en la idea de minimización de riego estructural. Constantemente se utilizan como herramientas para resolver problemas de clasificación, regresión y detección de valores atípicos (28,30).

En cuanto al funcionamiento de las SVM, como se ejemplifica en figura 4, en primer lugar, se realiza un mapeo y separación de los dos puntos de entrada (clase 1 y clase 2) y las plasma en un espacio de características de una dimensión mayor, es decir un hiperplano. De tal forma que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano, denominado margen, sea máxima, con esto se intenta forzar el aprendizaje durante el entrenamiento. La creación del hiperplano óptimo es la combinación de los puntos de entrada y es a lo que se le denomina vectores de soporte (28–31).

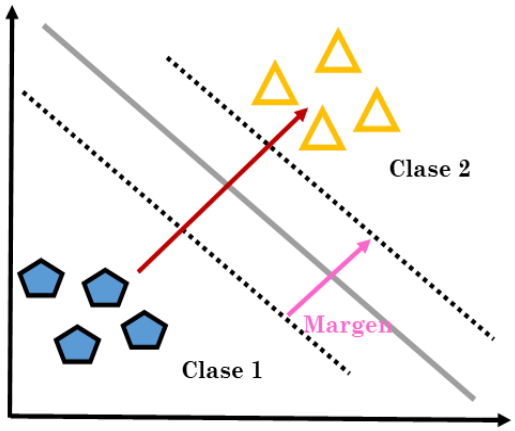


Figura 4. Algoritmo basado en máquinas de soporte vectorial
Imagen modificada de Betancourt et al., 2019 (30).

3.2.2 Árboles de decisión

Los algoritmos de árboles de decisión (RF, por sus siglas en inglés) es una técnica de aprendizaje supervisado que genera múltiples variables sobre un conjunto de datos de entrenamientos, con el fin de generar una aproximación a las funciones objetivo se combinan los resultados para obtener un modelo único más robusto (31–35). Como se muestra en la figura 5, inicialmente se tiene un conjunto de datos y posteriormente son generados cada árbol, los cuáles contienen distintas características aleatorias, que son elegidas por una técnica estadística denominada *bootstrap,* la cual proporciona estimaciones del error estadístico y selecciona el valor estadístico de interés. Cada árbol que se crea contiene un subconjunto aleatorio de variables, que serían los predictores (m), así mismo cada árbol crece hasta su máxima combinación posible (∑), para finalmente elegir el que presenta mejores resultados (y) (32).

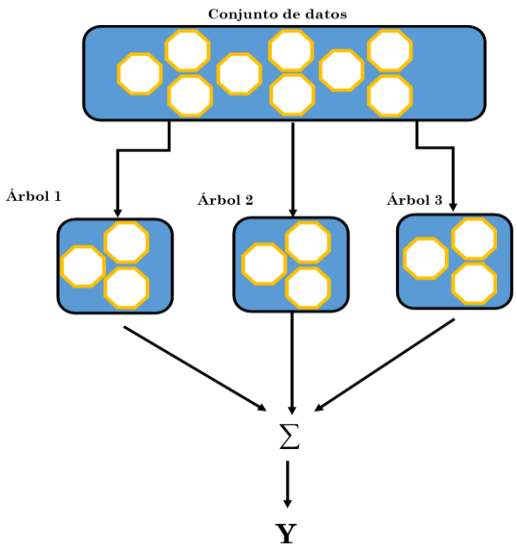


Figura 5. Algoritmo basado en árboles de decisión

Imagen modificada de Espinoza et al., 2020 (32).

3.2.3 XGBoost

Los algoritmos de máquinas de XGBoost (por sus siglas en inglés de Extreme Gradient Boosting), es un algoritmo de aprendizaje supervisado que utiliza los árboles de decisión para aumentar el gradiente de predicción del aprendizaje. Es conocido por su rápida ejecución y escalabilidad (33,36,37).

Sus principales características es el ensamblado secuencial de árboles de decisión, como se muestra en la figura 6, que van aprendiendo de los árboles previos y corrigiendo errores al momento de ser generados, lo que se le denomina “gradiente descendente”, para finalmente escoger al mejor árbol (y). Además de que optimiza los datos para evitar el sobreajuste o sesgo de dicho modelo (33).

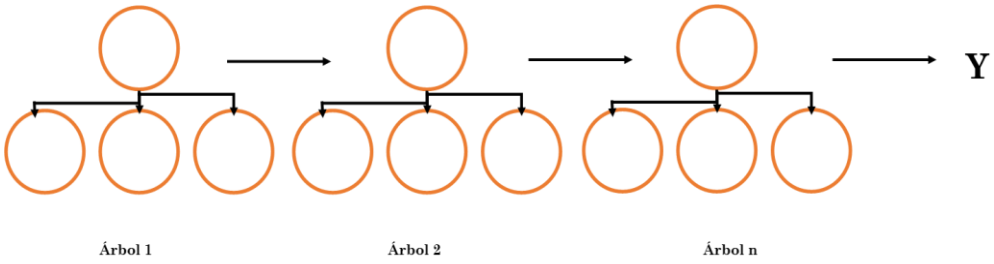


Figura 6. Algoritmo basado en XGBoost

Imagen modificada de Espinoza et al., 2020 (33).

3.2.4 Regresión logística

Los algoritmos de máquinas de regresión logística (LR, por sus siglas en inglés) es un algoritmo usado para la clasificación de variables dicotómicas, es de los más simples y utilizados en el ML. Se basa en la probabilidad de que una variable dependiente esté relaciona con algunas variables independientes (38–39).

Su fórmula se muestra en la figura 7, en la dónde p es la probabilidad de la salida de interés, ß0 es un término de intercepción, ß1,…,ßi son los coeficientes ß asociados con cada variable y X1,…,Xi son las variables independientes (40).

$$\log [\, p/1 - p \,] = ß0 + ß1\, X1 + . . . + ßi Xi$$

Figura 7. Fórmula de regresión logística

3.2.5 LightGBoost

LightGBoost (por sus siglas en inglés de Light Gradient Boosted Machine) es una biblioteca de código abierto desarrollada en Microsoft. LightGBoost es un marco de mejora de gradiente y está basado en árboles de decisión, proporciona una implementación eficiente y más ligera del algoritmo de aumento de gradiente. El beneficio principal de éste son los cambios en el algoritmo de entrenamiento que hacen que el proceso sea mucho más rápido y, en muchos casos, dan como resultado un modelo más efectivo. LightGBoost además permite un entrenamiento altamente eficiente sobre conjuntos de datos a gran escala con un bajo costo de memoria (41,42).

A diferencia de otros algoritmos de gradiante que hacen crecer el árbol por niveles, como se muestra en la figura 8, LightGBoost divide el árbol por hojas. Elige la hoja con máxima pérdida delta para crecer y de esta forma va aumentando su clasificación de forma específica. Dado que la hoja es fija, el algoritmo de hoja tiene una pérdida menor en comparación con el algoritmo de nivel (43).

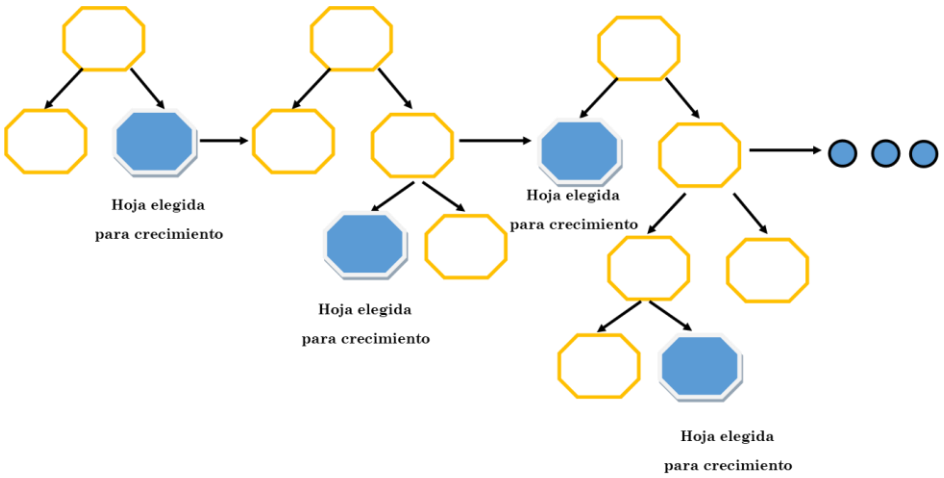


Figura 8. Algoritmo basado en LightGBoost
Imagen modificada de GeekforGeeks 2021 (43)

3.3 Interpretabilidad de los modelos predictivos

La interpretabilidad es un término usado en el contexto de modelos de aprendizaje de máquinas que se refiere a la explicación de la relación causa-efecto del algoritmo que busca comprender como varía la salida de cada modelo ante los cambios de las entradas o en los parámetros. Este proceso es empleado para el análisis de los modelos, con esto se puede determinar si son confiables, apoyados en la forma en que hacen las predicciones y no ciegamente basándose solo en la precisión de cada modelo (44).

3.3.1 Algoritmo SHAP

SHAP (por sus siglas en inglés de SHapley Additive exPlanations) es un enfoque de basado en la teoría de juegos para explicar el resultado de cualquier modelo de aprendizaje automático (45).

El objetivo de SHAP es explicar la predicción de una instancia *x* calculando la contribución de cada característica a la predicción. El método de SHAP calcula los valores de Shapley, éstos nos dicen cómo distribuir equitativamente el “pago” entre las características a partir de la teoría de juegos de coalición (45). Los valores de las características de una instancia de datos actúan como jugadores en una coalición. Los valores de SHAP nos dicen cómo distribuir equitativamente la predicción entre las características. En la figura 9, podemos observar la fórmula para el desarrollo del método SHAP, dónde *g(z’)* es la explicación del modelo, *M* es el

tamaño máximo de la coalición, $Ø_0$ es el vector de coalición, $Ø_j z'_j$ es la atribución a una característica *j* (46).

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

Figura 9. Fórmula del método SHAP
Imagen tomada de Christoph Molnar,2021 (49)

Las ventajas de los valores de SHAP es que la predicción está bastante distribuida entre los valores de las características. Obtenemos explicaciones contrastantes que comparan la predicción con la predicción promedio. Sin embargo, las desventajas es que los valores de SHAP pueden malinterpretarse y se necesita acceso a los datos para calcularlos nuevamente (46).

3.4 Ciclo para entrenamiento de modelos de aprendizaje de máquinas

Para la construcción de un modelo, se deben realizar distintos pasos, en la figura 10 se observan los pasos para la construcción de un modelo de aprendizaje de máquinas. Los cuáles son: obtención de datos, análisis de datos, pre-procesamiento, entrenamiento del modelo, evaluación del modelo, monitoreo y optimización.

El primer paso consiste en obtener la información que se desea analizar y tener claro que tarea queremos que aprenda nuestro modelo. Posteriormente se inicia el análisis de dicha información, en este paso se explora, comprende y se selecciona el subconjunto de todos los datos con los que se trabajará. Se debe considerar qué datos se necesitan realmente para abordar el problema en el que se trabajará (26,27,47).

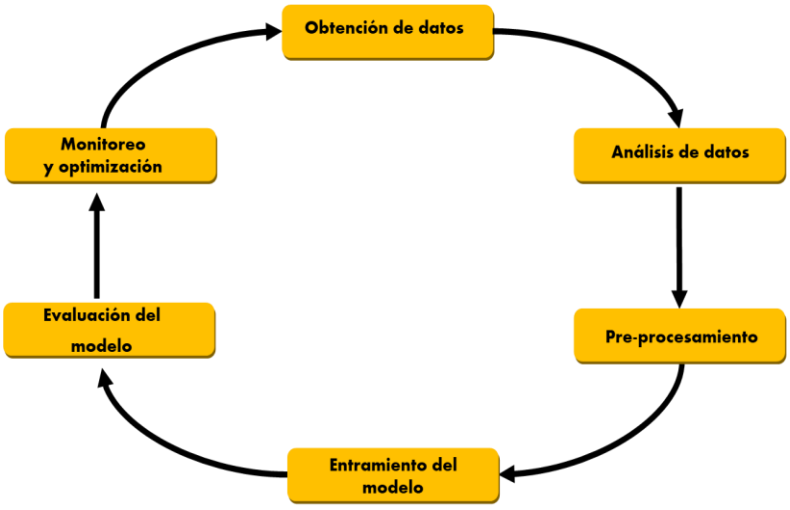


Figura 10. Pasos para el desarrollo de un modelo de aprendizaje de máquinas

El siguiente paso consiste en el pre-procesamiento, en este ocurre la división de la información, frecuentemente los datos se dividen en un 80% de la información, para utilizarla en el entrenamiento del modelo, y 20% restante para la evaluación del mismo (26,27). Además, los datos seleccionados son sometidos a varios procesos comunes que son el formateo, imputación de datos, selección de variables representativas, la codificación, la normalización y el tratamiento de desbalance de clases (47):

- Formateo: Los datos que se seleccionaron pueden encontrarse en un formato inadecuado para trabajarlos. Por lo que se debe modificar en el formato en el que se desea trabajar, por ejemplo, los datos pueden estar en una base de datos relacional y se necesitan en un archivo de texto o viceversa.
- Imputación: Consiste en la eliminación o reparación de datos faltantes. Un valor faltante se define como “un atributo que tiene no ha sido muestreado en el conjunto de datos, o que nunca fue registrado” (48). Es crucial encontrar el método que permita la mejor imputación para analizar la información con una mayor calidad (49,50). Las imputaciones más comunes son las siguientes (48):
    - Maximización de expectativa regularizada (EM, por sus siglas en inglés de Regularized Expectation-Maximization): es un meta-algoritmo aplicado para optimizar la máxima similitud de los datos repitiendo dos pasos hasta lograrlo: primero utiliza otras variables para imputar un valor (paso de

expectativa) y luego verifica si ese es el valor más probable (paso de maximización).

- Múltiple imputación (MI, por sus siglas en inglés de Multiple Imputation): es un algoritmo estadístico para manejar conjuntos de datos incompletos, MI crea M > 1, sin embargo, normalmente M ≤ 10 que los datos originales, cada conjunto de datos es analizado por separado y luego se combina para producir un conjunto de resultados generales.
- Imputación kNN (kNNI, por sus siglas en inglés de k-Nearest Neighbor Imputation): define cada muestra o individuo con su más cercano *k* vecinos en un espacio multidimensional y luego imputa los datos faltantes con una variable dada promediando los valores no faltantes de estos k vecinos.
- Imputación por media: es un método en el que el valor perdido se imputa por la media de los valores disponibles.

- Selección de variables representativas: Conocido como FS (por sus siglas en inglés de feature selection), en un inicio puede haber muchos más datos seleccionados disponibles de los que necesita para trabajar. Con esto se busca reducir la dimensionalidad y los grandes volúmenes de datos, así como también rescatar las variables más importantes del estudio (48, 51).
- Codificación: Todos los algoritmos necesitan que los datos se encuentren en valores numéricos para su procesamiento, es por esto que realizan distintos métodos para convertir las características categóricas en numéricas, el más utilizado es el *one-hot,* en éste, los valores pasan a ser 0 o 1 para cada etiqueta en las características categóricas; también se puede codificar por medio de la frecuencia, en este método se realiza un cálculo de porcentaje de cada característica categórica y se le asigna la misma etiqueta. Finalmente se puede emplear la codificación ordinal, en la cual se asigna valores distintos a cada característica categórica que puede tomar un valor de 0-1 (52).
- Normalización: consiste en cambiar los valores de las columnas numéricas en una escala común, sin cambiar las diferencias entre cada uno de ellos ni extraviar la información. Este paso consiste en tomar un atributo del conjunto de datos y reduciéndolo a uno de rango más pequeño (51,53). Las normalizaciones más comunes son las siguientes:

- Normalización Min Max: En este método el valor mínimo y máximo de la base de datos se desplaza a los valores 0 y 1, respectivamente, y todos los demás datos se transforman en el rango {0,1}.
- Normalización Z-Score: Este método transforma los datos a una distribución con media 0 y desviación estándar 1.
- Normalización por escala decimal: Este método normaliza los datos en un valor entre -1 y 1, sin incluirlos (52).

- Tratamiento de desbalance de clases: el desbalance de clases ocurre cuando se tiene una muestra de datos con un gran número de valores de una clase (o una menor cantidad) que los valores de otras clases, es decir nuestros datos se encuentran desequilibrados. Esto puede dificultar la construcción de un modelo y obtener resultados ineficaces que nos conducirán a falsos negativos. Para corregir este tipo de desbalanceo se pueden utilizar 2 técnicas: el sobremuestreo que consiste en duplicar las clases con menor número de datos y el submuestreo, que es la eliminación de la clase mayoritaria (54). Las técnicas básicas para tratar este problema son:
    - La técnica de Synthetic Minority Oversampling TEchnique (SMOTE, por sus siglas en inglés), fue descrita por Nitesh et al (55), consiste en un enfoque para la construcción de clasificadores a partir de conjuntos de datos desequilibrados, este método implica la creación de ejemplos sintéticos de clases minoritarias, el método se evalúa por medio del área bajo la curva (AUC) y la curva ROC. SMOTE primero elige un ejemplo aleatorio de la clase menor cantidad de datos. Posteriormente se encuentran (*k)* vecinos más cercanos *a* y *b*. A continuación, se elige al azar un vecino y se crea un ejemplo “sintético”, este se genera en el espacio entre dos datos (*a* y *b*) de las características minoritarias. El enfoque es efectivo porque se crean nuevos ejemplos sintéticos de la clase minoritaria que son plausibles, es decir, están relativamente cerca en el espacio de características de los datos existentes de la clase minoritaria. Su principal desventaja es que los datos sintéticos se crean sin considerar a la clase con mayores datos, lo puede crear datos ambiguos, en el caso de que hubiera una superposición de las clases (55).

- El algoritmo de enfoque de muestreo sintético adaptativo llamado *Adaptive Synthetic Sampling Approach for Imbalanced Learning* (ADASYN, por sus siglas en inglés), es una extensión del método de SMOTE. El método ADASYN crea datos sintéticos a través de la interpolación lineal, entre las características de la clase minoritaria y la clase mayoritaria. Los datos generados cambian adaptivamente el límite de decisión de clasificación, utilizan la distribución de densidad para crear diferentes ejemplos de clases minoritarias según su nivel de dificultad en el aprendizaje. ADASYN busca darle mayor peso a los datos de la clase minoritaria que son difíciles de aprender (56).

El análisis y el pre-procesamiento son los pasos que pueden tomar la mayor cantidad de tiempo, debido a su complejidad. Una vez finalizado el pre-procesamiento se inicia el entrenamiento del modelo (con el 80% de los datos), en el cual se emplean distintos algoritmos, como los descritos anteriormente, esto con la búsqueda de aquel que nos genere un mejor rendimiento, la medida de rendimiento es la forma en que se evalúa una solución al problema estudiado (27,28,50).

Es importante recalcar que el modelo entrenado no está expuesto al conjunto de datos de prueba (20% de datos) durante el entrenamiento y cualquier predicción realizada en ese conjunto de datos está diseñada para ser indicativa del rendimiento del modelo en general. Con los datos de prueba se realiza la evaluación del modelo, completando así el penúltimo paso (50).

Finalmente, en el monitoreo y la optimización se busca que el modelo siga dando resultados más exactos. Se puede optimizar el modelo ajustando los parámetros de un algoritmo, con la ampliación de los datos o usando un modelo con nuevos datos y correlacionando los resultados con evidencia científica (27,28,50).

3.5 Estado del arte de trabajos relacionados en la predicción de CaCU y en la identificación de factores de riesgo basados en modelos de aprendizaje de máquinas

Actualmente no existen reportes de estudios de aplicación de aprendizaje de máquinas en predicción de persistencia de VPH. Lo más cercano usando esta metodología son los estudios de predicción de riesgo de CaCU y su supervivencia y aquellos implementados para la

identificación de factores de riesgo para CaCU, por ello se describen a continuación como estado del arte.

Los trabajos en esta sección fueron agrupados en dos categorías. Los primeros trabajos se caracterizan por utilizar métodos de aprendizaje de máquinas para predecir el riesgo de desarrollar CaCU y su supervivencia. El segundo grupo de trabajo aborda la identificación de factores de riesgo para el desarrollo de CaCU basados también en aprendizaje de máquinas.

3.5.1 Trabajos relacionados en la predicción de CaCU y supervivencia

Según lo reportado, el primer estudio donde se usó por primera vez el aprendizaje de máquinas para predecir la supervivencia general de 134 pacientes con CaCU fue el realizado por Ochi et al. en 2002 (57), se utilizó un modelo de red neuronal artificial (ANN) que incluyó 11 factores pronósticos (edad, estado funcional, estadio el estadio de la Federación Internacional de Ginecología y Obstetricia (FIGO), entre otras), en pacientes tratadas con radioterapia intracavitaria combinada externa y de carga diferida remota de alta tasa de dosis entre 1978 y 1993. El modelo se entrenó con los datos de 67 pacientes seleccionadas al azar. Usando las ANN entrenadas, se predijo la supervivencia a cinco años en las 67 pacientes restantes y se comparó con la supervivencia a cinco años conocida. El resultado de supervivencia pronosticado fue capaz de lograr un AUC de 0.7782. Se concluyó que, usando ANN, la combinación de la clasificación histológica del efecto de la radiación determinada por el examen periódico de biopsias, además de los factores fundamentales, es la más efectiva para predecir la supervivencia en pacientes con CaCU comparado con métodos convencionales.

En la investigación hecha por Chunnv Yuan, et al (58), se hizo la comparación de la precisión diagnóstica entre colposcopistas y modelos de aprendizaje por reconocimiento en imágenes, para ello se seleccionaron 22,330 casos para capacitación y evaluación modelo, incluidos 10, 365 casos normales, 6,357 casos lesiones escamosas intraepiteliales de bajo grado y 5,608 casos lesiones escamosas intraepiteliales de alto grado. Los resultados del mejor rendimiento del modelo en cuanto a sensibilidad, especificidad y precisión para diferenciar los casos negativos de los positivos fue del 85.38%, 82.62% y 84.10% respectivamente, con un AUC de 0.93. A modo de comparación, se calcularon la sensibilidad, la especificidad y la precisión de cinco expertos en colposcopia del hospital de mujeres de la facultad de medicina de la Universidad de Zhejiang para diferenciar los casos positivos de los negativos, con un promedio

en los expertos de 70%, 72.92% y 71.83% respectivamente y con un AUC de 0.715. Se concluyó que, en comparación con los colposcopistas, el diagnóstico fue mejor en el modelo de reconocimiento de imágenes que en el de colposcopia ordinaria.

3.5.2 Trabajos relacionados en la identificación de los factores de riesgo de CaCU

Gupta S et al. (59) propuso un modelo de aprendizaje de máquinas para la identificación de los factores de riesgo de desarrollar CaCU de una manera más eficiente, para esto utilizó una base de datos de la Universidad de California en Irvine con 858 participantes, se implementaron diferentes enfoques de RF, percentrón multicapa (MLP, por sus siglas en inglés), aumento de gradiente (*gradient boosting*), y *N-Nearest Neighbors.* El modelo con mayor rendimiento fue RF con precisión de 0.985, sensibilidad de 1.0, especificidad de 0.927 y un AUC de 0.985. Se concluyó que el modelo ofrece resultados de predicción fiables, debido a que el estudio enfatizó los factores de riesgo clínicos para el desarrollo de CaCu, entre ellos la edad, debut sexual, número de parejas sexuales y el uso de anticonceptivos orales, principalmente. La principal desventaja del modelo fue el desequilibrio de los datos y que tuvieron que utilizar técnicas de sobre muestreo, lo que demoró más tiempo su construcción en comparación con otros modelos que tienen funciones reducidas.

Por otra parte, Asadi F et al. (60) publicaron un estudio transversal, que contenía los datos de 145 pacientes, referentes del Hospital Shohada de Teherán, Irán en el periodo 2017–2018, dichos datos fueron analizados por medio de algoritmos de clasificación de aprendizaje de máquinas que incluían enfoques de SVM, QUEST, C&R tree, MLP y función de base radial (RBF). Sus resultados más altos fueron con C&R tree y arrojaron una precisión de 95.55, sensibilidad 90.48, especificidad 100 y AUC de 95.20. Se pudieron identificar factores de riesgo relacionados al nivel de salud personal, el estado civil, estatus social, la dosis de anticonceptivos utilizados, el nivel de educación y el número de partos por cesárea para el desarrollo de CaCu.

En contraste con el elaborado por Komala Rayavarapu (61) el cual tenía como objetivo crear una herramienta médica para valorar el crecimiento de CaCU y con ello mejorar la tasa de supervivencia de los pacientes. Se analizaron 858 pacientes y 36 variables, se emplearon los modelos de *Voting Classifier* y *Deep Neural Network* (DNN, por sus siglas en inglés) *Classifier*,

sus mejores resultados fueron los obtenidos en modelo de Voting con una precisión de 0.99, sensibilidad de 1, especificidad de 0.96 y un AUC de 0.68. Los factores de riesgo identificados para el crecimiento de CaCu fueron la edad de inicio a la vida sexual, número de parejas sexuales, embarazos, uso de anticonceptivos orales, tabaquismo, entre otros.

Existen otros estudios como es el propuesto por Laboni Akter et al. (62), para identificar los factores de riesgo para el desarrollo de CaCU que tienen un enfoque más relacionado con el comportamiento individual y sexual. En este estudio se implementaron modelos de aprendizaje de máquinas como Decision Tree, RF, XGBoost y se analizaron a 72 pacientes. Se encontró que de los principales factores de riesgo es el nivel de empoderamiento de las decisiones, la autopercepción de riesgo y la agregación de intenciones. Se obtuvo una mayor precisión en éstos del 93.3%, una sensibilidad de 92% y una especificidad de 100%.

En la tabla 1 se muestra un cuadro comparativo de los trabajos analizados. En la columna 1 se observan los autores del trabajo, en la columna 2 se describe el tamaño de la población, en la columna 3 el tipo de variables analizadas, también se pueden observar las métricas obtenidas y los factores de riesgo identificados.

3.5.3 Áreas de oportunidad

De manera general es importante hacer ciertas menciones de los trabajos antes analizados, la primera es que la revisión propuesta por Chunnv Yuan et al (58) destacó que el aprendizaje de máquinas superó a los modelos estadísticos en la predicción de CaCU, pues éstos son capaces de resolver las relaciones complejas y no lineales en los datos a gran escala, además de poder aprender las representaciones de sus características y así descubrir nuevos factores pronósticos.

La segunda mención es referente a los estudios elaborados por Gupta S et al. (59) y Komala Rayavarapu (61) es importante resaltar que, aunque ambos usaron diferentes enfoques de aprendizaje de máquinas los factores de riesgo identificados en los dos trabajos son muy similares. Sin embargo, una de sus principales desventajas es que en ambos trabajos utilizaron la misma base de datos que es de acceso libre de la Universidad de California en Irvine con

858 participantes y 36 variables, lo que pudo haber causado que sus resultados fueran casi idénticos.

La tercera mención, es sobre el estudio de Laboni Akter et al. (62) que, al ser un trabajo más enfocado al comportamiento social, es un poco distinto a los revisados anteriormente, motivo por el cual es difícil compararlo con el resto, pero a su vez destaca que los factores de riesgo para el desarrollo de CaCu son multifactoriales, en la tabla 3 se muestra un resumen de los trabajos antes mencionados.

Aunque en los trabajos hubo algunas desventajas, fueron más relevantes sus aportaciones, demostraron en general un buen rendimiento en sus predicciones y coincidieron en su mayoría en los factores de riesgo identificados para el desarrollo de CaCU. Además, la metodología empleada en estos estudios puede servir como una base para futuros trabajos que se centren en el desarrollo de la persistencia viral.

Tabla 3. Resumen del estado del arte de trabajos que describen métodos basados en técnicas de aprendizaje de máquinas para la predicción e identificación de factores de riesgo de CaCU

Trabajo	P	Modelos usados	Tipo de información analizada	Métricas obtenidas				Factores de riesgo identificados
				AUC	Pre	S	E	
Gupta S. et al. [59]	858	-MLP -RF -N-Nearest Neighbors -Gradient Boosting	Laboratoriales, resultados de biopsia, citología y colposcopía.	0.98	0.98	1	0.92	Edad, debut sexual, número de parejas sexuales, uso de anticonceptivos orales, tabaquismo, entre otros.
Asadi F. et al. [60]	145	-SVM -QUEST -C&R tree -MLP -RBF	Variables cualitativas relacionadas al comportamiento sexual.	0.95	0.95	0.90	1	Nivel de salud personal, el estado civil, estatus social, la dosis de anticonceptivos utilizados, el nivel de educación y el número de partos por cesárea.
Laboni Akter et al. [62]	72	-Decision Tree -RF -XGBoost	Variables cualitativas relacionadas al comportamiento sexual.	*	0.93	0.92	1	Nivel de empoderamiento de las decisiones, percepción de riesgo, agregación de intenciones

Komala Rayavarapu [61]	858	-Voting Classifier -DNN	Laboratoriales, resultados de biopsia, citología y colposcopía.	0.68	0.99	1	0.96	Edad, inicio de vida sexual, número de parejas sexuales, embarazos, uso de anticonceptivos orales, tabaquismo, entre otros.
Chunnv Yuan, et al. [58]	22,330	- ResNet multimodal - Segmentación U-Net 42 -Detección Mask R-CNN	Imágenes de colposcopia	0.93	0.84	0.85	0.82	*

Nota: P: población; AUC: área bajo la curva; Pre: precisión; S: sensibilidad; E: especificidad.
*: No se menciona en el estudio.
MLP: multilayer perceptron; RF: random forest; SVM: máquinas de soporte vectorial; RBF: función de base radial y DNN: deep neural network classifier.

4. MARCO CONCEPTUAL

En la figura 11 se presenta el diagrama del marco teórico-conceptual del presente proyecto, en el cual se resume el proceso de desarrollo de aprendizaje de máquinas y como por medio de éstos se pueden generar predicciones de persistencia de VPH y con ello implementar mejores intervenciones que logren reducir la prevalencia de lesiones premalignas en cérvix y la mortalidad por CaCU.

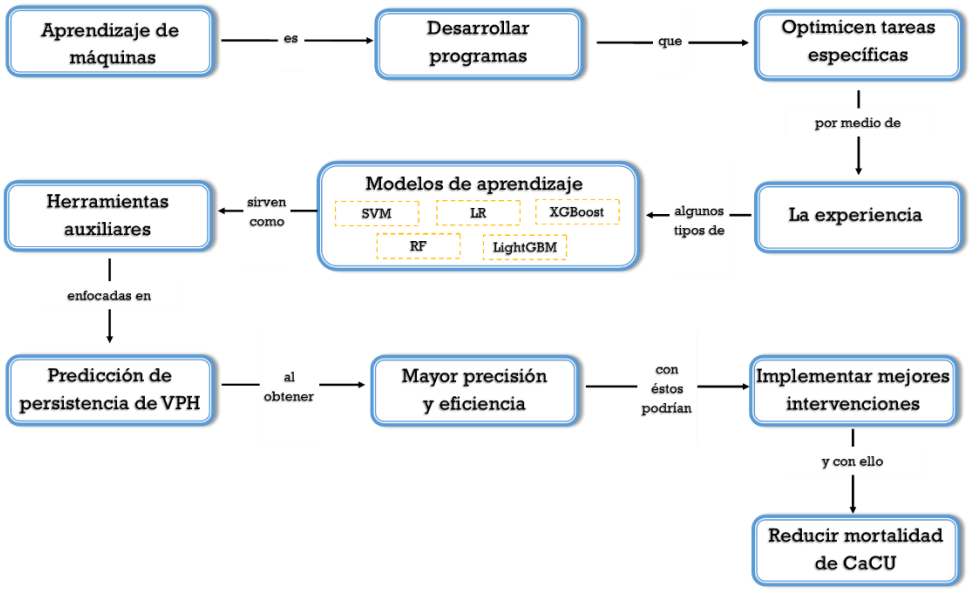


Figura 11. Mapa conceptual resumen de marco teórico

5. PLANTEAMIENTO DEL PROBLEMA

La infección por VPH es una de las principales infecciones transmitidas por contacto sexual, su prevalencia es alta alrededor de la edad del debut sexual, siendo muy común en mujeres jóvenes entre los 20 y los 25 años de edad (17). En población mexicana una revisión sistemática en 8.706 mujeres reportó que los genotipos más frecuentes a nivel del cérvix fueron el VPH 16 (63,1%), VPH 18 (8,6%), VPH 58 y VPH 31 (5%) que se encuentran asociados a lesiones precancerosas (18).

Se sabe que la persistencia viral es esencial para el desarrollo de una lesión precursora o para el crecimiento de CaCU, y también se ha comprobado que en una minoría de casos se puede detectar la persistencia viral una vez transcurridos 12 meses (6,7), además que en México no se realizan acciones específicas para la búsqueda de casos con persistencia viral.

Por otro lado, las respuestas inmunitarias sistémicas y locales juegan un papel muy importante frente a la eliminación viral (5), además se asocian otros factores como lo son el comportamiento sexual, la edad, tabaquismo, uso prolongado de anticonceptivos orales, infección por *Chlamydia trachomatis*, cambios en el microbioma cervicovaginal y múltiples nacidos vivos. (6,7,10,11). Es por estas razones que la detección de la persistencia del virus es tan compleja, y que, a pesar de las estrategias de prevención primaria y secundaria, sigue siendo un reto, dado que las características intrínsecas del virus dificultan su detección temprana, seguido del comportamiento sexual de la población, y de la respuesta inmunitaria de cada huésped.

El mundo cambia constantemente, por dicha razón es necesario ir implementando el uso de nuevas tecnologías con el fin de apoyar al personal de salud en la detección de aquellos casos persistentes de VPH que son más propensos para el desarrollo de CaCU. El uso de métodos de aprendizaje de máquinas ayudaría al reforzamiento en la metodología científica y brinda una alternativa más confiable, eficiente y precisa para apoyar al diagnóstico de persistencia viral y con ello reducir la mortalidad por CaCU. La predicción del riesgo de desarrollo futuro de CaCU, será una guía para centrar las intervenciones en los factores de pronóstico específicos y optimizar tanto el diagnóstico como el tratamiento para cada paciente y mejorar sus resultados clínicos.

Finalmente, la mayoría de los estudios elaborados en aprendizaje de máquinas han sido en la predicción diagnóstica de CaCU y en el contexto de otras poblaciones y demostraron tener un buen rendimiento en comparación con otras herramientas, lo que hace pensar que su aplicación en México y para la predicción de infección persistente de infección por VPH a nivel del cérvix antes del desarrollo de CaCU, podría ser una ventana de oportunidad.

Todo lo anteriormente mencionado lleva a formular la siguiente pregunta de investigación: ¿cuáles son los marcadores de riesgo de persistencia de infección por VPH en cérvix que pueden predecirse mediante modelos de aprendizaje de máquinas en población mexicana?

6. JUSTIFICACIÓN

En México, la mayoría de casos de lesiones premalignas de alto grado y CaCU, se detecta tardíamente, por lo que la identificación de marcadores predictores de persistencia de VPH en cérvix, permitirían mejorar el triage de mujeres VPH positivas que realmente desarrollan persistencia viral y así reducir los costos que para el sistema de salud mexicano representa el manejo de los casos de lesiones premalignas de alto grado y CaCU.

En los últimos años, el análisis de los registros clínicos electrónicos (EHR, por sus siglas en inglés) ofrecen nuevas oportunidades en la creación de herramientas basadas en evidencia para el soporte de toma de decisiones clínica (63). De manera general, los EHR describen la historia clínica de los pacientes, tales como, datos demográficos, tratamientos, resultados de laboratorio, entre otros. Los métodos de aprendizaje de máquinas, aprovechando la disponibilidad de los EHR, han sido propuestos para el desarrollo de herramientas auxiliares enfocadas en la predicción de mortalidad, anticipación de eventos importantes, identificación de factores de riesgo, entre otros, con el objetivo de contribuir en la mejora del diagnóstico y en los resultados de los pacientes (64). Particularmente, algunos trabajos se han enfocado en la predicción de CaCU y en la identificación de sus respectivos marcadores de riesgo basado en modelos de aprendizaje de máquinas. Aunque algunos trabajos han abordado el uso de aprendizaje de máquinas en la predicción de CaCU (57,58,59), la identificación de marcadores de persistencia de VPH a nivel de cérvix en población mexicana basado en aprendizaje de máquinas ha sido muy poco explorada y lleva al planteamiento de los siguientes objetivos.

7. OBJETIVOS

7.1 General

Analizar marcadores de riesgo de persistencia de infección por Virus del Papiloma Humano en cérvix mediante modelos de aprendizaje de máquinas en las mujeres que recibieron atención en CAPASAM en el período de 2015-2017.

7.2 Específicos

- Examinar modelos de aprendizaje de máquinas que permitan predecir eventos clínicos.
- Desarrollar un modelo de predicción de persistencia viral de VPH a nivel del cérvix.
- Identificar marcadores de riesgo de persistencia de VPH a nivel del cérvix utilizando un enfoque de interpretabilidad.
- Evaluar los marcadores identificados como predictores de persistencia de VPH a nivel del cérvix.

8. MÉTODOS Y MATERIALES

8.1 Tipo y diseño de estudio

El presente estudio es de diseño transversal, el cual está anidado al estudio de cohorte dinámica “Las citocinas inmunosupresoras Th2 y Th3 como predictores de persistencia viral y progresión a lesión pre-maligna en cérvix en mujeres infectadas con VPH-AR: estudio de cohorte” (65).

8.2 Definición de la población

La base de datos con la que se trabajó corresponde a información recabada de mujeres que recibieron atención en Centro de Atención para la Salud de la Mujer (CAPASAM) de los Servicios de Salud del Estado de Morelos, diagnosticadas con infección por el VPH y que entraron en el estudio cohorte antes mencionado, durante septiembre de 2015 y se realizó seguimiento posterior al primer y segundo año.

8.3 Tamaño de la muestra

Se cuenta con el acceso a una base de datos que contiene el registro de variables sociodemográficas, de historia sexual y reproductiva, clínicas y de laboratorio de 267 mujeres que ingresaron al estudio basal de la cohorte de acuerdo a los criterios de inclusión, exclusión y eliminación que se describen en Torres-Poveda et al 2018 (66) y se muestran en la tabla 4.

Tabla 4. Criterios de inclusión, exclusión y eliminación estudio de cohorte dinámica Torres-Povea et. Al, 2018 (66)

Criterios de inclusión
M Mujeres mayores de 15 años.
A Aceptar consentimiento informado.
R Residencia en Morelos mayor de 5 años.
P Prueba de VPH.
Criterios de eliminación
Diagnóstico de inflamación crónica o enfermedad autoinmune al inicio del estudio: ● Lupus eritematoso sistémico (LES) ● Artritis reumatoide (AR)
I Infección activa de transmisión sexual.
Criterios de exclusión
V Vacunación contra VPH.

La operacionalización de las variables a considerar en el desarrollo del presente estudio se describe en la tabla 5.

- Como variable dependiente o desenlace se tomará: persistencia viral.
- Como variables independientes: variables sociodemográficas (Edad, nivel socioeconómico); variables de historia sexual y reproductiva (edad menarca, edad inicio vida sexual, número de parejas sexuales, paridad), variables de estilo de vida (tabaquismo), variables intrínsecas del virus (genotipo y carga viral por genotipo).

Tabla 5. Operacionalización de las variables

Variable	Definición conceptual	Definición operativa	Tipo	Unidad de Medición	Escala de medición
Persistencia viral	Presencia del mismo genotipo viral en una infección para un periodo de seguimiento.	Detección del mismo genotipo viral en dos o más intervalos consecutivos.	Cualitativa	Genotipo de VPH	Nominal dicotómica

Infección incidente por VPH-tipo específico	Primera infección a nivel de cérvix por un genotipo de VPH diferente al detectado en el estudio basal.	Primer resultado positivo de VPH tipo específica después de un resultado negativo para el mismo tipo de VPH en el estudio basal.	Cualitativa	Genotipo de VPH	Nominal dicotómica
Edad	Tiempo que ha vivido una persona.	Mujeres mayores de 30 años	Cuantitativa Continua	Años	De razón
Número de parejas sexuales	Número de personas con las que se ha tenido actividad sexual.	Número de parejas sexuales que ha tenido la paciente.	Cuantitativa discreta	Número de parejas	De razón
Edad menarca	Edad de primera menstruación en una mujer	Año de vida en la que se presentó la primera menstruación en las pacientes del estudio.	Cuantitativa Continua	Años	De razón
Edad de inicio vida sexual	Edad en la que se tiene sexualmente activa	Año de la vida en la que se tiene primera relación sexual.	Cuantitativa Continua	Años	De razón
Nivel socioeconómico	Medida total económica y sociológica combinada con su preparación laboral, y la posición económica y social individual o familiar en relación a otras personas, basada en sus ingresos, educación, y empleo.	Estrato socioeconómico al que pertenecen las pacientes.	Cualitativa ordinal	Alto, medio y bajo	Ordinal
Paridad	Clasificación de una mujer por el número de niños nacidos vivos y de nacidos muertos con más de 28 semanas de gestación.	Número total de embarazos de las pacientes del estudio.	Cuantitativa discreta	Número de hijos	De razón
Tabaquismo	Adicción al tabaco, provocada principalmente por nicotina.	Antecedentes de tabaquismo: evaluado en las pacientes del estudio	Cualitativa Dicotómica	Ausente o presente	Nominal

		considerando años previos de consumo de tabaco y número de cigarrillos al día.	Cuantitativa discreta	Años de consumo	Ordinal
		Tabaquismo actual: Hábito de fumar actual evaluado en las pacientes del estudio considerando número de cigarrillos al día.	Cuantitativa discreta	Número de cigarros al día	Ordinal
			Cualitativa Dicotómica	Ausente o presente	Nominal
Anticoncepción hormonal	Los métodos hormonales o anticonceptivos hormonales son varios métodos que liberan hormonas habitualmente estrógeno y progesterona impidiendo que se produzca el embarazo.	Uso o no de tratamiento hormonal en las pacientes que participan en el estudio.	Cualitativa Politómica	Tipo de método	Nominal
Genotipo viral	Información genética que diferencia a los tipos virales de la familia de VPH.	Tipo de VPH presente en las infecciones persistentes.	Cualitativa	Genotipo de VPH	Nominal
Carga viral en cérvix	Cuantificación de la infección por virus que se calcula por estimación de la cantidad de partículas virales en cérvix.	Número de partículas virales de VPH a nivel de cérvix.	Cuantitativa continua	Número de copias virales	Ordinal

8.4 Herramientas/ Instrumentos a utilizar

Para la construcción y evaluación de los modelos basados en aprendizaje de máquinas se utilizó el lenguaje de programación *Python 3.7*, el marco de trabajo *Anaconda 3.0* y *Jupyter notebook 6.3.0*. También se usaron bibliotecas como: *numpy, pandas, matplotlib, SHAP y scikit learn.*

8.5 Metodología

En la figura 12 se muestra el proceso para llevar a cabo la tarea de predicción de persistencia de infección por VPH a nivel del cérvix, el cual está compuesto por diversas tareas: proceso

de la recolección de datos, el preprocesamiento de datos, análisis de los datos, la división del conjunto de datos y la evaluación.

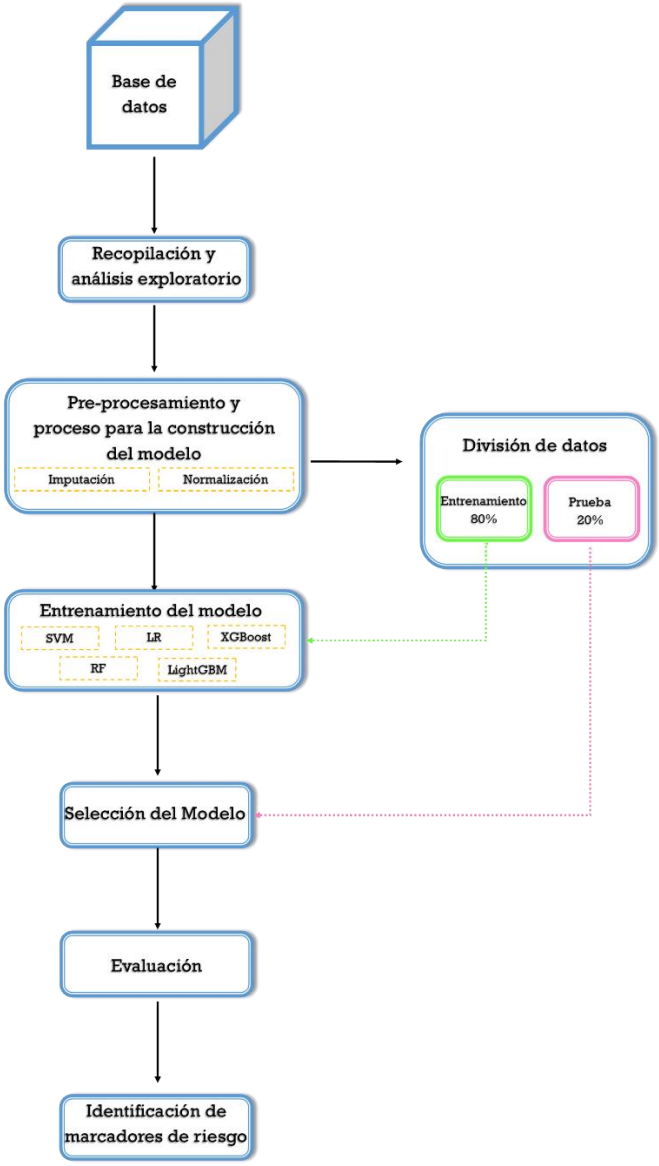


Figura 12. Flujo de trabajo de gestión de datos y desarrollo del modelo de predicción de persistencia de VPH

Modificado de Akter L. et al., 2021 (55).

9. PLAN DE ANÁLISIS

La metodología del modelo propuesto se dividió en las siguientes partes:

- Recopilación y análisis exploratorio de la base de datos disponible.

- Preprocesamiento y proceso para la construcción de modelos de predicción de persistencia de VPH
- Entrenamiento del modelo por medio del conjunto de modelos lineales y no lineales.
- Selección de modelos de predicción de persistencia de infección por el VPH.
- Evaluación e identificación del modelo que obtenga el mejor rendimiento.
- Finalmente, identificar los marcadores de riesgo para persistencia a partir de un enfoque de interpretabilidad.

10. RESULTADOS

10.1 Recopilación y análisis exploratorio de la base de datos de las mujeres del CAPASAM

Se obtuvo una base de datos que contenía la información de 267 mujeres, con 608 variables tanto sociodemográficas, de historia sexual y reproductiva, relacionadas al estilo de vida, intrínsecas del virus (genotipo) y carga viral. Se inició con el análisis exploratorio de los datos, dónde se realizó la limpieza de los mismos por medio de la imputación por mediana aritmética y la normalización de ellos en un rango numérico de 0 a 1.

En cuanto a la información recabada en el análisis exploratorio se describen a continuación los principales resultados encontrados:

De las 267 mujeres que ingresaron a la cohorte se encontró que solo 202 mujeres se encontraban positivas a VPH y 65 negativas en la toma basal del estudio, como se muestra en la tabla 6. Solo para el análisis descriptivo de la información, se consideraron aquellas mujeres que presentaban positividad del VPH y que se siguieron en el estudio, para poder comprender las características de este grupo de población.

Tabla 6. Resultado de PCR al VPH en la toma basal en las mujeres de CAPASAM 2015

Resultado de PCR a VPH en toma basal		
Resultado	**Número de pacientes**	**Porcentaje**
Negativo	65	24.34%
Positivo	202	75.65%

Las edades de las mujeres VPH positivas del CAPASAM en la toma basal se encontraban en el rango de los 15 años a los 68 años de edad, se presentó una media de 36.18 años de edad. Como se muestra en la figura 13, entre los 30 y 50 años se concentró 53.46% de la población a estudiar.

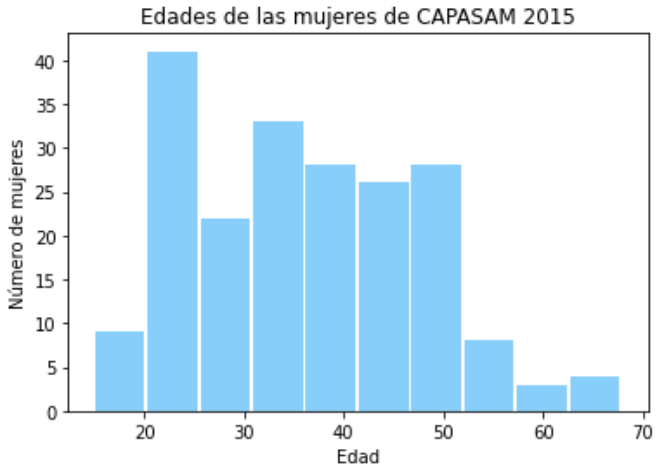


Figura 13. Edades de las mujeres VPH positivas en el estudio basal de la cohorte de CAPASAM 2015

En cuestión a las edades de debut sexual de las mujeres VPH positivas se encontraron en el rango de los 10 a los 31 años de edad, se presentó una media de 17.54 años de edad. 76.73% de las mujeres tuvieron su inicio de vida sexual entre los 15 y 20 años, como se muestra en la figura 14.

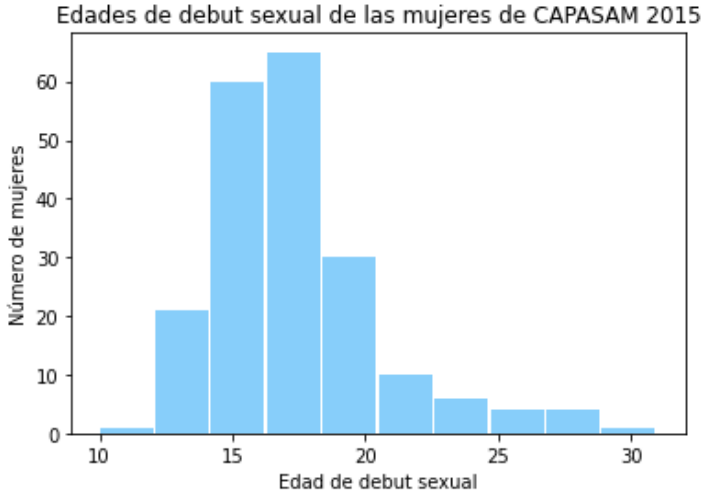


Figura 14. Edades de debut sexual de las mujeres VPH positivas en el estudio basal de la cohorte de CAPASAM 2015

Referente al número de parejas sexuales de las mujeres VPH positivas, en la figura 15, se observa que el 35.14% tuvo una sola pareja sexual, el 27.22% refirieron 2 parejas sexuales y el 37.62% restante de las mujeres tuvieron más de 3 parejas sexuales.

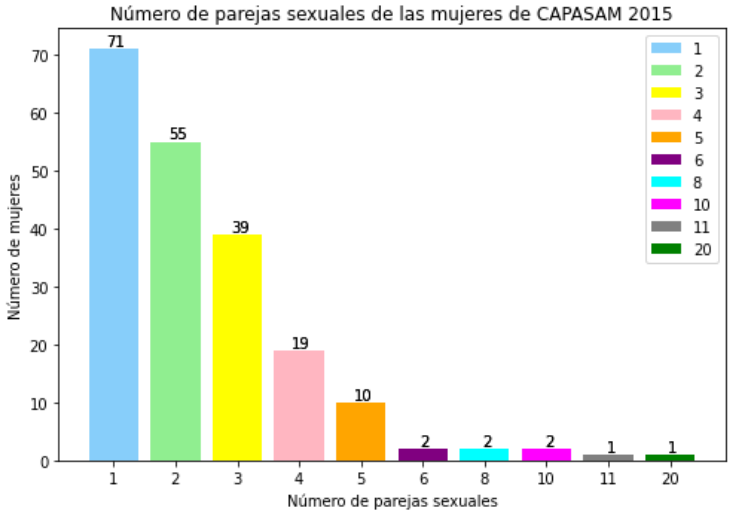


Figura 15. Número de parejas sexuales en las mujeres VPH positivas en el estudio basal de la cohorte de CAPASAM 2015

En cuanto al antecedente de tabaco referido por las pacientes VPH positivas en el cuestionario del estudio basal de la cohorte, se encontró que 147 mujeres (72.77%) reportaron no consumir tabaco, en comparación con el 27.22% de las mujeres que sí lo refirieron.

10.2 Resultados de coinfecciones por genotipos de VPH

A continuación, se muestran las coinfecciones por genotipos de VPH presentes tanto en la toma basal, como en el seguimiento anual y a los dos años; en la figura 16 se destacan las coinfecciones por los genotipos 53 y 58 (nueve casos) y por los genotipos 31 y 53 (ocho casos) en la toma basal. Mientras que en la figura 17, se muestra el seguimiento al año donde las coinfecciones más comunes fueron por los genotipos 31 y 58 con cuatro casos. Finalmente, en el seguimiento al segundo año (figura 18), se encontraron diversas coinfecciones, en los genotipos 16 y 39, en los genotipos de 16 y 58 y en los genotipos 39 y 58 con un caso cada uno.

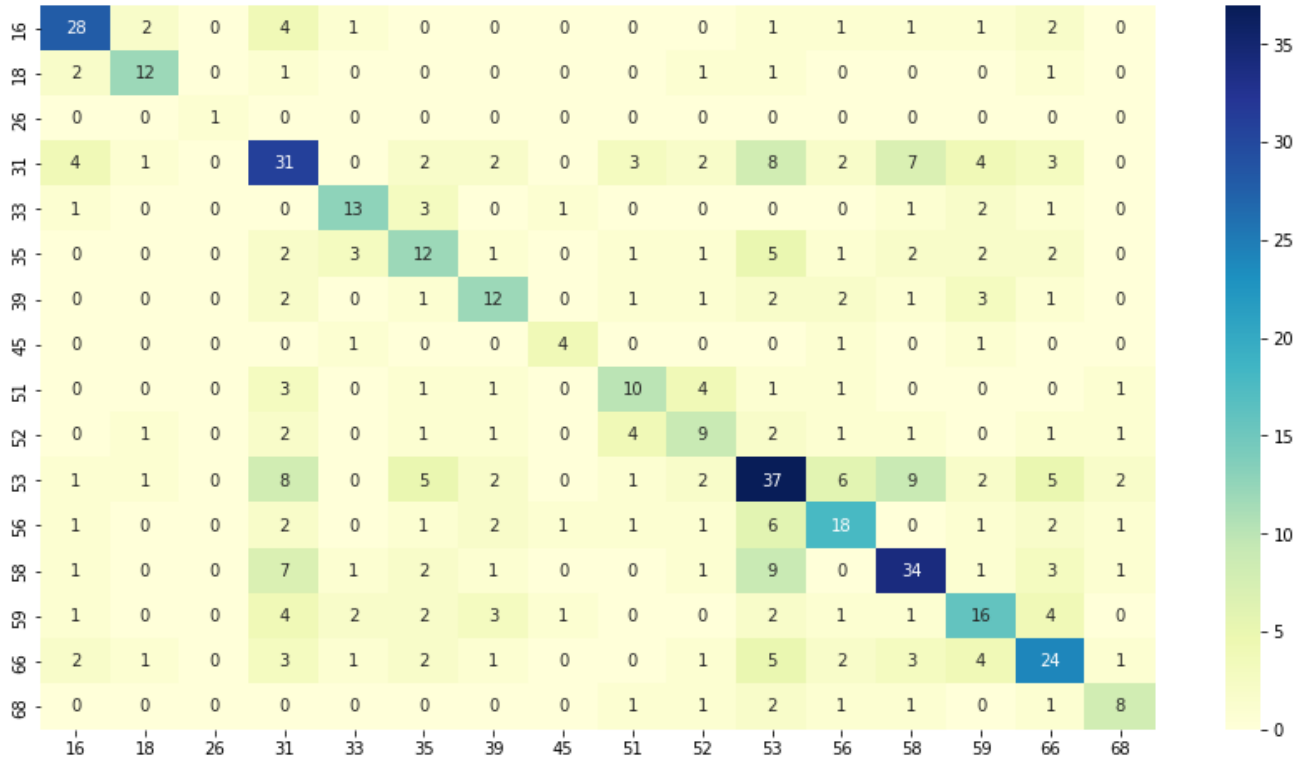


Figura 16. Genotipos y coinfecciones de VPH presentes en el estudio basal de la cohorte (CAPASAM 2015)

Respecto a los tres principales genotipos que se encontraron presentes en las mujeres positivas a VPH en la muestra basal, fueron el VPH 58 (34 casos), VPH 31 (31 casos) y VPH 16 (30 casos) (figura 16). En tanto que, al seguimiento anual, se puede apreciar en la figura 18, que destacaron el VPH 58 (17 casos) y VPH 31, 39 y 66 con 10 casos. Finalmente, en lo reportado al segundo año, el VPH 58 (en cuatro casos) fue el genotipo predominante, seguido de VPH 16,18 y 39 con dos casos (figura 18).

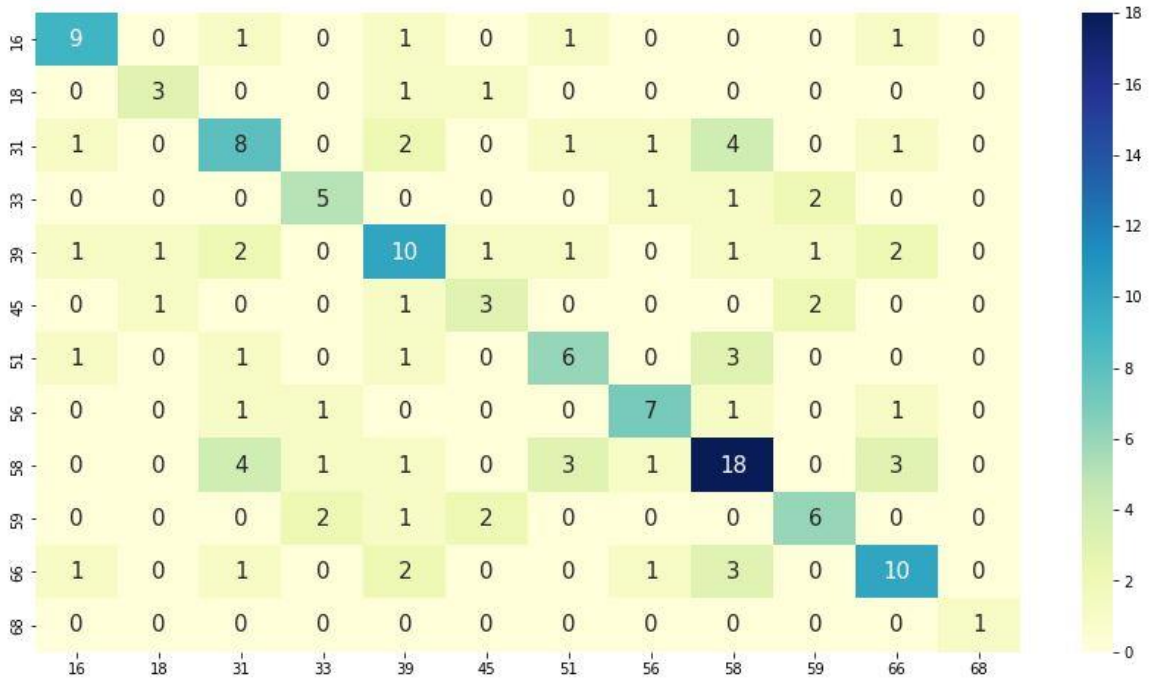


Figura 17. Genotipos y coinfecciones presentes en las mujeres VPH positivas en seguimiento al segundo año de la cohorte (CAPASAM 2016)

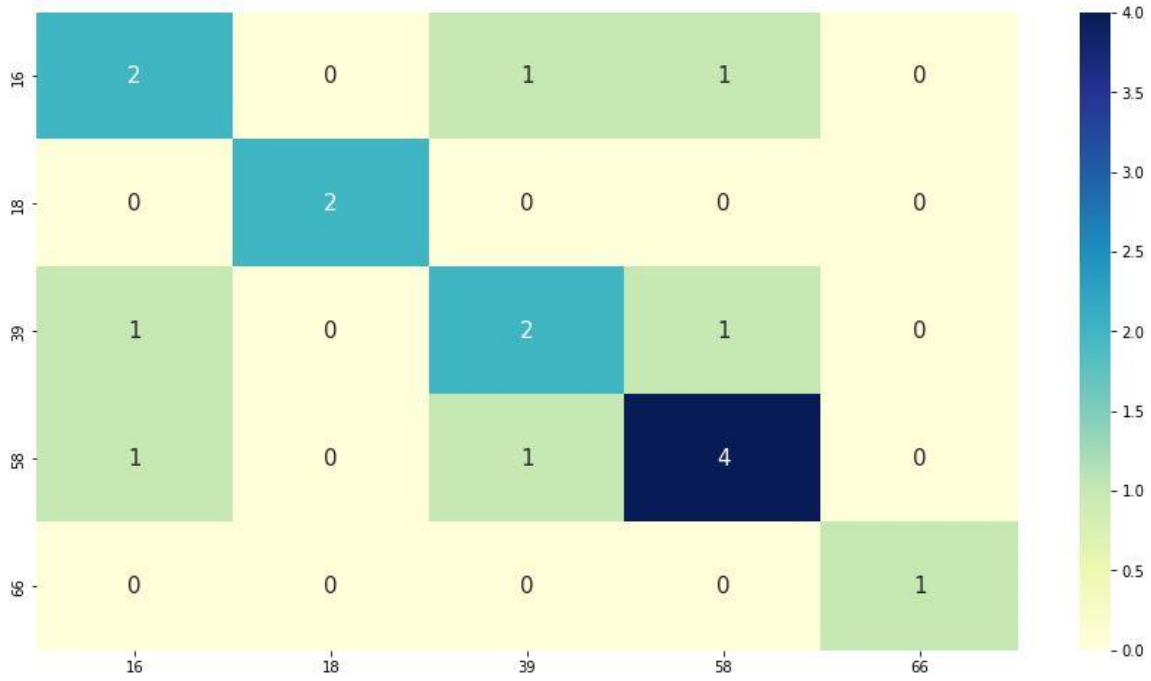


Figura 18. Genotipos y coinfecciones presentes en las mujeres VPH positivas en seguimiento al segundo año de la cohorte (CAPASAM 2017)

10.3 Análisis estadístico de persistencia viral

Para determinar la persistencia anual y al segundo año de VPH se crearon las variables considerando el mantenimiento del mismo genotipo viral en la muestra basal y al año (persistencia anual); y persistencia al segundo año con la presencia del mismo genotipo al año y al segundo año. Referente a la persistencia anual se encontraron 19 mujeres que desarrollaron persistencia, de éstas, dos mujeres presentaron dos casos de persistencia distinta, y se pudo detectar la persistencia de 21 genotipos, destacando en frecuencia al genotipo VPH 58. Mientras que al segundo año solo se detectaron cinco casos de mujeres que desarrollaron persistencia, siendo los genotipos más comunes el VPH 18 y 58, en la tabla 7 se puede apreciar el total de los casos y el genotipo presentado de persistencia al año y al segundo año.

Tabla 7. Número de casos de persistencia presentados en Mujeres de CAPASAM 2015-2017

Temporalidad	Persistencia viral	Número de pacientes
ANUAL	VPH 16	3
	VPH 18	1
	VPH 31	2
	VPH 33	1
	VPH 39	2
	VPH 45	1
	VPH 51	0
	VPH 56	2
	VPH 58	4
	VPH 59	2
	VPH 66	3
	VPH 68	0
SEGUNDO AÑO	VPH 16	0
	VPH 18	2
	VPH 31	0
	VPH 33	0
	VPH 39	1
	VPH 45	0
	VPH 51	0
	VPH 56	0
	VPH 58	2
	VPH 59	0
	VPH 66	0

	VPH 68	0

10.4 Preprocesamiento y proceso para la construcción de modelos de predicción de persistencia de VPH

Primero, se realizó la experimentación para la elección de las variables que se utilizarían para la construcción de los modelos referentes a la predicción de persistencia del primer año y el segundo año de VPH, se utilizaron 71 variables, en el anexo 2 se presenta la agrupación de dichas variables en categorías: sociodemográficas, de historia clínica y estilos de vida, historia sexual y reproductiva y relacionadas al virus y se encuentra el modelo de predicción usando estas variables.

Como resultados de los experimentos realizados se identificó que los cinco marcadores generales de persistencia anual fueron la coinfección por VPH general, el estado civil, el número de coinfecciones de alto riesgo, la carga viral de genotipo 58 y la presencia de resequedad vaginal. En comparación con los cinco marcadores generales de persistencia en el segundo año identificados que fueron la carga viral por el genotipo 18, la edad de debut sexual, el número de coinfecciones de VPH de alto riesgo, la presencia de genotipos de alto riesgo y la dispareunia.

Posteriormente, a partir de los resultados obtenidos en esta experimentación, se hizo una selección recursiva de las características, eligiendo a las variables con mayor plausibilidad biológica con el desenlace en estudio (persistencia de infección por VPH) que se muestran en la tabla 8, y que fueron utilizadas para los modelos de predicción finales y los marcadores de riesgo de persistencia viral al año y al segundo año de seguimiento de la cohorte. En la figura 19, se puede observar el diagrama de flujo que ilustra las fases del proceso para la construcción de modelos de predicción de persistencia de VPH, que más adelante se describirán a detalle.

Tabla 8. Selección de variables para el entrenamiento de los modelos de predicción de persistencia de VPH en el primer y segundo año

Variables utilizadas en la predicción de persistencia al primer año	Variables utilizadas en la predicción de persistencia al segundo año

Edad Antecedente de tabaquismo Edad debut sexual Número de parejas sexuales Antecedente de uso de anticonceptivos hormonales Coinfección de VPH general Número de embarazos Antecedente de ITS Presencia de carga viral de 1000 copias virales Presencia de carga viral de 2000 copias virales Presencia de carga viral de 3000 copias virales Persistencia anual	Se usan las mismas variables del primer año más: Coinfección de VPH general anual Carga viral al año de seguimiento de 1000 copias virales Carga viral al año de seguimiento de 2000 copias virales Carga viral al año de seguimiento de 3000 copias virales Persistencia al segundo año

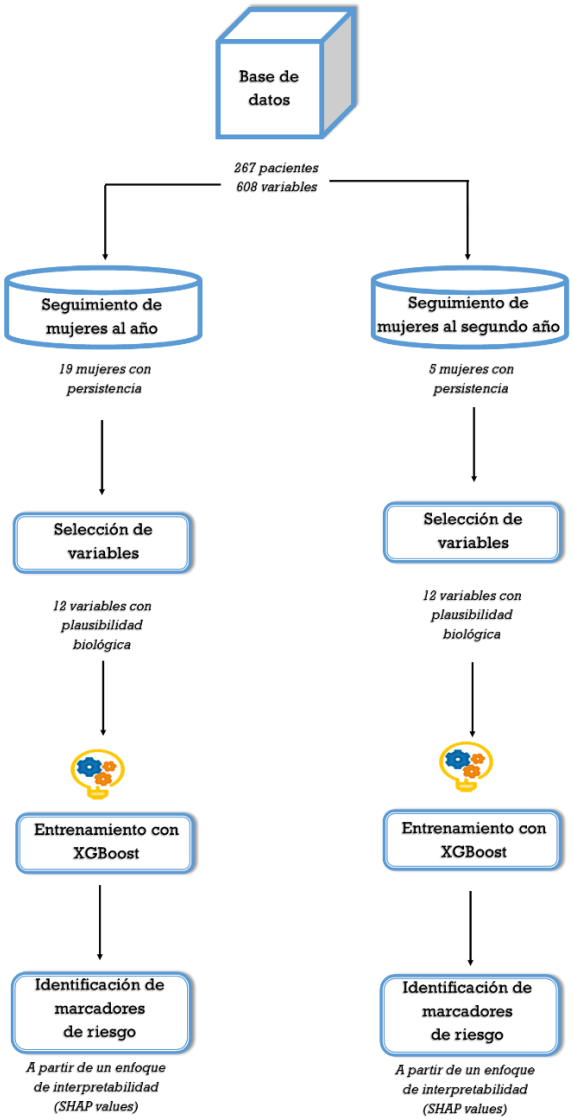


Figura 19. Diagrama de flujo de proceso para la construcción de modelos de predicción de persistencia de VPH

10.5 Entrenamiento y construcción del modelo de predicción de persistencia del primer año

En primera instancia se evaluaron a las técnicas de SMOTE y ADASYN para el tratamiento de desbalance de clases, es decir, se buscó identificar qué técnica contribuye a que el modelo obtenga un mejor rendimiento; en la tabla 9 y en la tabla 10, se muestran los rendimientos en el conjunto de entrenamiento de dichas técnicas, con SMOTE el modelo de XGBoost tuvo una sensibilidad de 0.93, especificidad de 1 y una AUC de 0.99; mientras que con ADASYN se obtuvieron los rendimientos para XGBoost de sensibilidad de 0.94, especificidad de 1 y una AUC de 0.99, se destaca que los rendimientos en ambas fueron muy similares, sin embargo, se eligió el método SMOTE, para continuar con las pruebas.

Tabla 9. Rendimientos de técnica SMOTE en la predicción de persistencia de VPH al primer año

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9923	0.9393	1	0.9687	0.9696
LR	0.8254	0.7828	0.7474	0.7692	0.7651
LightGBoost	0.9563	0.9242	0.8535	0.8926	0.8888
RF	0.9754	0.9393	0.8333	0.8920	0.8863
SVM	0.7474	0.7878	0.7070	0.7572	0.7474

Tabla 10. Rendimientos de técnica ADASYN en la predicción de persistencia de VPH al primer año

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9932	0.9444	1	0.9714	0.9772
LR	0.8196	0.8232	0.7474	0.7931	0.7853
LightGBoost	0.9685	0.9393	0.8484	0.8985	0.8939
RF	0.9837	0.9646	0.8434	0.9095	0.9040
SVM	0.7676	0.8030	0.7323	0.7756	0.7676

Una vez identificado a la mejor técnica de desbalance, se evaluaron los diferentes modelos de aprendizaje para identificar al modelo con el mejor rendimiento sobre el cual se probará el conjunto de prueba. El modelo XGBoost tuvo una sensibilidad de 0.93, especificidad de 1 y una AUC de 0.99, siendo el que presento los mejores resultados, en comparación con SVM que obtuvo los rendimientos más bajos con una sensibilidad de 0.78, especificidad de 0.70, y un AUC de 0.74. Estos rendimientos pueden también observarse en la figura 20, destacando

que el modelo basado en XGBoost logró los mejores rendimientos en comparación con el resto de modelos evaluados.

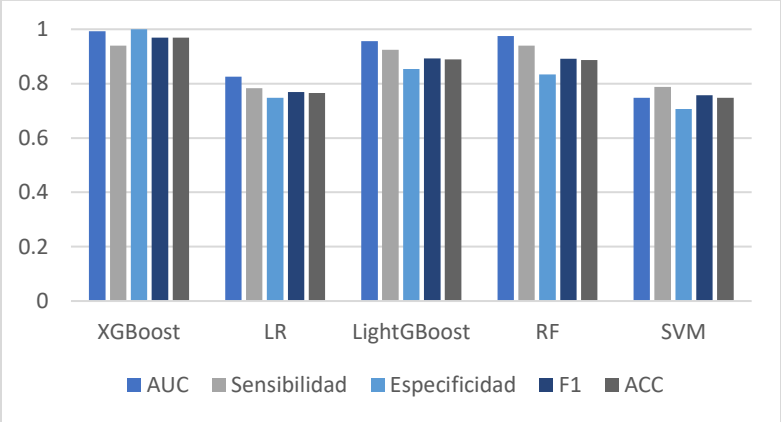


Figura 20. Rendimientos de los modelos de predicción de persistencia de VPH al primer año en las mujeres de CAPASAM 2015-2016

En la tabla 11 se presentan los rendimientos obtenidos en el conjunto de prueba con el modelo de predicción de persistencia viral anual en mujeres VPH positivas del CAPASAM en los años 2015-2016 usando XGBoost y SMOTE, presentando una especificidad de 1 y una sensibilidad del 0.9 y un AUC de 0.97, en la figura 21 se puede observar la curva ROC.

Tabla 11. Rendimientos del modelo de predicción de persistencia de VPH al primer año

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9772	0.9	1	0.9473	0.95

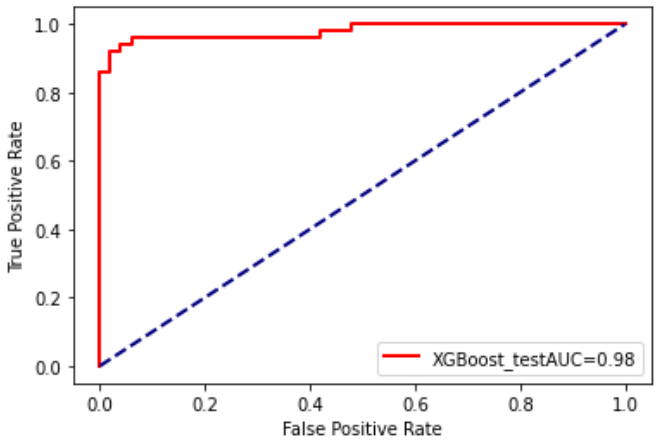


Figura 21. Curva ROC del modelo de XGBoost de predicción de persistencia de VPH al primer año

10.5.1 Marcadores de riesgo identificados para la presentación de persistencia de VPH a nivel del cérvix en el primer año

Posteriormente se utilizó el método de SHAP para identificar los 10 principales marcadores de riesgo para la presentación de persistencia de VPH a nivel del cérvix en el primer año. En la figura 22 se observan los marcadores, se destaca en primer lugar a la co-infección de VPH general como principal marcador, seguido del número de parejas sexuales y en tercer lugar se encontró al antecedente de tabaquismo.

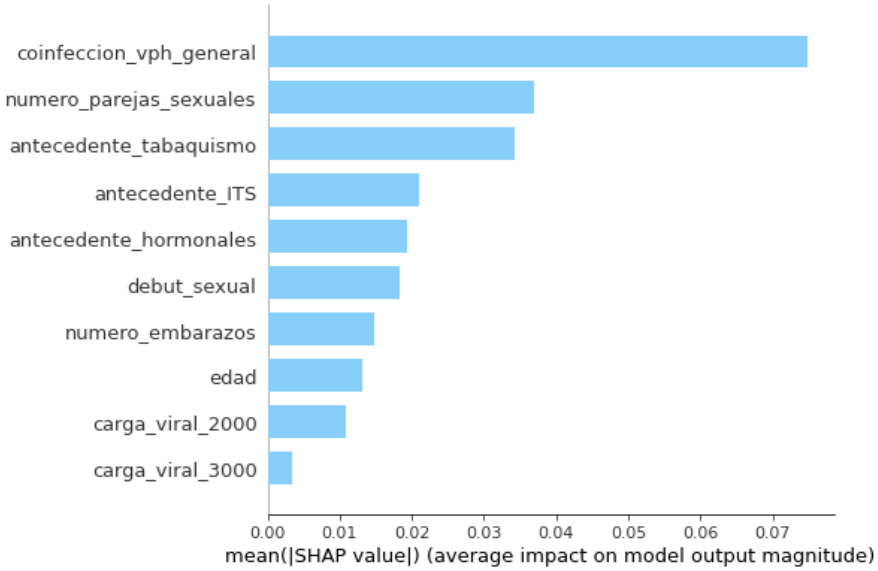


Figura 22. Marcadores de riesgo identificados para la persistencia viral por método SHAP las mujeres de CAPASAM 2015-2016

10.6 Entrenamiento y construcción del modelo de predicción de persistencia del segundo año

En la tabla 12, se muestran los resultados del modelo de predicción de persistencia de VPH en el segundo año en el conjunto de entrenamiento, se eligió también en este caso usar SMOTE y al modelo de XGBoost para continuar con las pruebas, de esta manera pueden ser comparables ambos modelos de predicción (anual y segundo año). En la figura 23, se muestran los resultados de forma gráfica de los modelos de predicción al segundo año, se destaca el modelo XGBoost con una sensibilidad 0.99, especificidad 0.98 y un AUC 0.99, en comparación con el resto de los modelos que tuvieron rendimientos más bajos.

Tabla 12. Rendimientos de los modelos de predicción de persistencia de VPH al segundo año usando SMOTE

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9968	0.9952	0.9856	0.9905	0.9904
LR	0.9712	0.9761	0.8851	0.9339	0.9307
LightGBoost	0.9954	0.9952	0.8995	0.9500	0.9474
RF	0.9990	0.9952	0.9090	0.9543	0.9522
SVM	0.9282	0.9857	0.8708	0.9324	0.9284

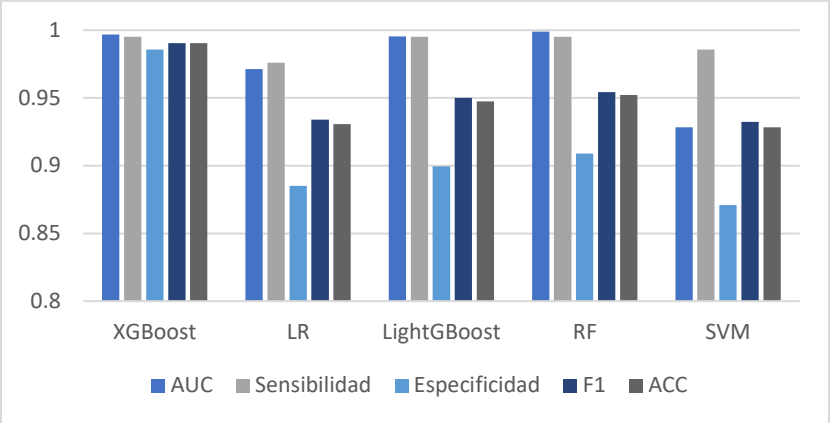


Figura 23. Rendimientos de los modelos de predicción de persistencia de VPH al segundo año en las mujeres de CAPASAM 2016-2017

En la tabla 13 se presentan los rendimientos obtenidos en el conjunto de prueba del modelo de predicción de persistencia viral al segundo año en mujeres del CAPASAM en los años 2016-2017 usando XGBoost y SMOTE, se obtuvo una especificidad de 1, una sensibilidad del 0.98 y un AUC de 0.99, en la figura 24 se puede observar la curva ROC.

Tabla 13. Rendimientos de los modelos de predicción de persistencia de VPH al segundo año

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9996	0.9807	1	0.9902	0.9904

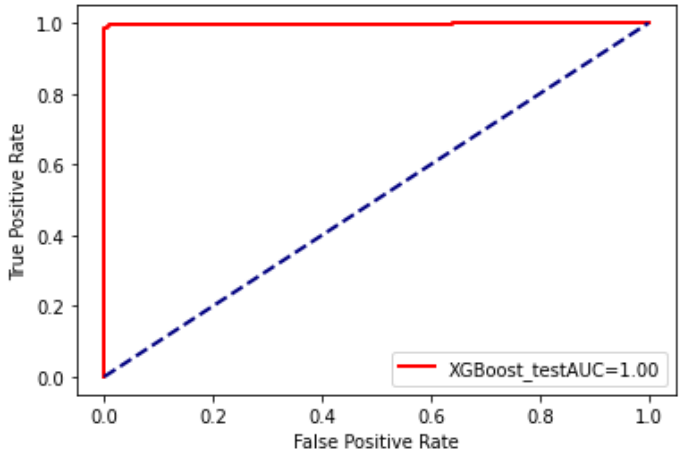


Figura 24. Curva ROC del modelo de XGBoost de predicción de persistencia de VPH al segundo año

10.6.1 Marcadores de riesgo identificados para la presentación de persistencia de VPH a nivel del cérvix en el segundo año

A continuación, se utilizó el método de SHAP para identificar los 10 principales marcadores de riesgo para el desarrollo de persistencia de VPH a nivel del cérvix al segundo año. En la figura 25 se pueden observar dichos marcadores, se destaca a la edad de debut sexual, como principal marcador de riesgo, seguido del número de parejas sexuales y en tercer lugar se encontró a la co-infección de VPH general.

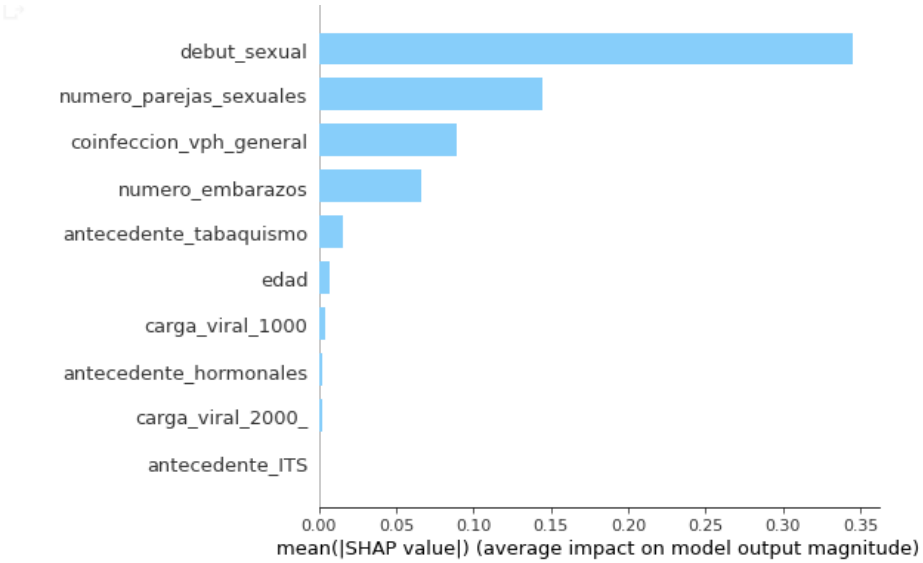


Figura 25. Marcadores de riesgo identificados para la persistencia viral por método SHAP las mujeres de CAPASAM 2016-2017

11. DISCUSIÓN

Los principales resultados de este estudio fueron la identificación de marcadores de riesgo de persistencia de infección por VPH en cérvix mediante modelos de aprendizaje de máquinas en las mujeres que recibieron atención en CAPASAM en el período de 2015-2016 y 2017-2018. Además, se encontraron que los genotipos más frecuentes en el desarrollo de persistencia anual y al segundo año fueron el VPH 58, 16 y 18.

Los 10 marcadores de persistencia anual identificados fueron la co-infección por genotipos de VPH general, el número de parejas sexuales, el antecedente de tabaquismo, el antecedente de ITS, el antecedente de uso de anticonceptivos hormonales, la edad de debut sexual, el número de embarazos, la edad de la paciente, la presencia de carga viral de 2000 copias virales y la presencia de carga viral de 3000 copias virales.

Mientras que los 10 marcadores de persistencia en el segundo año identificados fueron la edad de debut sexual, el número de parejas sexuales, la co-infección por genotipos de VPH general, el número de embarazos, el antecedente de tabaquismo, la edad de la paciente, la presencia de carga viral de 1000 copias virales, el antecedente de uso de anticonceptivos hormonales, la presencia de carga viral de 2000 copias virales y el antecedente de ITS.

Dentro de los estudios que nos hablan sobre la identificación de marcadores riesgo de desenlaces del VPH, se encuentran los que se enfocaron a la detección de marcadores de riesgo relacionados con la persistencia viral del VPH, pero que su metodología no es basada con modelos de aprendizaje de máquinas, y aquellos que utilizaron una metodología de aprendizaje de máquinas para obtener estos marcadores; sin embargo, se enfocaron en la predicción del desarrollo de CaCU, debido a que este estudio es de los primeros en su campo con una orientación en persistencia viral utilizando modelos de aprendizaje de máquinas para la identificación de factores de riesgo, la discusión se hará en torno a estos trabajos.

En las mujeres que participaron en la cohorte del CAPASAM 2015-2017, se encontraron que los genotipos más frecuentes en el desarrollo de persistencia anual y al segundo año fueron el VPH 58, 16 y 18. Mientras que en el trabajo elaborado por Swai, P et al. (67) que incluyó una cohorte de 462 mujeres africanas con VPH-AR, los genotipos 58, 35, 16, 31 y 52 fueron los más comunes a presentarse persistentes a los 18 meses, coincidiendo con los genotipos VPH 58 y 16 encontrados en el presente estudio.

Así mismo, Shi Nianmin et al. (68) realizaron un muestreo por conglomerados, para identificar mujeres asintomáticas con cepas del VPH-AR en China. Se incluyeron a 3000 mujeres en su estudio y en concordancia con el presente estudio los genotipos 58 y 16 tuvieron las tasas de infección más altas; sin embargo, hay diferencias en la definición de tiempo para la presentación de persistencia entre los estudios, el elaborado por Swai, P et al. (67) se determinó como persistencia a la presencia del mismo genotipo viral a los 18 meses, el propuesto por Shi Nianmin et al. (68) se estableció persistencia a los seis meses, mientras que en el presente estudio fue definida al transcurso de los 12 meses, además de que las características de las poblaciones de estudio varían sustancialmente pues las mujeres participantes en el trabajo de Shi Nianmin et al. (68) fueron sólo población asintomática y en rangos de edad de 18 a 30 años, a diferencia de la población estudiada en este estudio que también incluye mujeres mayores a 30 años.

Algunos de los marcadores de riesgo identificados en el presente estudio para persistencia de VPH en cérvix como el número de parejas sexuales, el antecedente de uso de anticonceptivos hormonales, la edad del debut sexual, el antecedente de tabaquismo y la edad han sido reportados también por la literatura para el desarrollo de CaCU, en donde se han utilizado técnicas de aprendizaje de máquinas. En el estudio realizado por Lilhore Kumar et al. (69) se desarrolló un modelo híbrido para la detección de CaCU utilizando técnicas de análisis causal y aprendizaje de máquinas, se analizaron los datos de 592 pacientes con 36 atributos, su mejor modelo fue con SVM con una precisión 91.2%, además en ese trabajo se hizo una revisión de la literatura sobre los factores de riesgo del CaCU y se realizaron múltiples hipótesis con la finalidad de comparar sus resultados con estos factores de riesgo, finalmente por medio del análisis de Boruta, el cual está diseñado para representar todas las características significativas dentro de un modelo de clasificación, se pudieron recabar como principales factores de riesgo para el desarrollo de CaCU al número de parejas sexuales, el estado de fumador, el uso de anticonceptivos hormonales, el número de ITS y la presencia de genotipos de VPH de alto riesgo.

Por otra parte, Mehmood Mevra et al. (70) propusieron un enfoque denominado CervDetect, que utiliza un enfoque híbrido mediante la combinación de RF y redes neuronales superficiales para evaluar los elementos de riesgo de la formación de CaCU, se analizaron los datos de 858 pacientes, que contenían 32 variables y se obtuvo una precisión de 93.6%; se detectaron como

factores de riesgo a la edad, el uso de anticonceptivos hormonales, la edad de debut sexual, el número de embarazos, el uso de DIU, el tiempo de exposición al tabaco, y el tiempo de diagnóstico de ITS, variables que coinciden con las encontradas en el presente estudio.

En la actualidad, la literatura acumulada en epidemiología de esta infección ha indicado que las conductas sexuales de alto riesgo (edad temprana del debut sexual, múltiples parejas sexuales y tener parejas con ITS) son los principales factores de riesgo para contraer una infección por VPH-AR (71). Así mismo, factores de riesgo como la edad avanzada ha sido asociada en estudios epidemiológicos recientes con persistencia y reducción en la eliminación de infección por VPH solo después de 400 días de infección (72), además de que factores relacionados con el virus como la co-infección también tuvo un efecto en la eliminación más lenta de infección recurrente por el mismo tipo viral, en concordancia con el presente estudio.

La variable edad, tiene una alta plausibilidad biológica para asociarse con el fenómeno de estudio de persistencia viral ya que se ha asociado que a mayor edad, mayor senescencia del sistema inmunitario, lo que hace más propensa la presentación de persistencia de infección por VPH a nivel del cérvix en mujeres con edad más avanzada (73); y puesto que la edad no es un factor que pueda ser modificable, se puede deducir la importancia que tiene la educación sexual en fomentar buenas conductas sexuales que disminuyan el efecto de los marcadores de riesgo previamente descritos.

En otro trabajo propuesto por Kaushik Manoj et al. (74) aplicaron un modelo basado en el aprendizaje automático utilizando variantes de genes de citocinas y características sociodemográficas como factores de riesgo para un mejor pronóstico y predicción del CaCU. Se incluyeron a 492 pacientes, de las cuales 246 presentaban CaCU y 246 se encontraban sanas; la técnica de LR logró la precisión promedio más alta de 82,25 % y la puntuación F1 promedio más alta de 82,58 %. Se detectaron como factores de riesgo nuevamente a la edad, el lugar de residencia, el nivel educativo, el estado socioeconómico, el número de embarazos, la edad del primer evento obstétrico, el ciclo menstrual, el uso de anticonceptivos, el hábito tabáquico y la presencia de genotipos de alto riesgo de VPH, principalmente (74).

Un estudio de cohorte de persistencia de VPH en cérvix realizado en mujeres danesas reportó que las mujeres que tenían dos o más tipos de VPH-AR en el momento de la toma basal tenían mayores probabilidades de persistencia que aquellas con un solo tipo de VPH (75), esto

concuerda con lo expuesto por Kaushik Manoj et al. (74) y con los resultados del presente trabajo dado que la coinfección por genotipos de VPH fue de los principales marcadores de riesgo identificados en la presentación de persistencia al año y al segundo año, esta situación se relaciona ya sea con una baja competencia en la mujer infectada para eliminar espontáneamente la infección por VPH a través de una adecuada respuesta inmune y/o por el fenómeno de posible reinfección por mayor exposición a nuevos genotipos de VPH durante la ventana de tiempo entre los muestreos de seguimiento (75).

Ijaz Muhammad et al. (76) realizaron un modelo de predicción de CaCU basado en datos con detección de valores atípicos y métodos de sobremuestreo, analizaron los datos de 858 pacientes, el mejor modelo fue RF con una precisión de 95-98%, de los factores de riesgo más relevantes que se pudieron extraer fueron la exposición al tabaco expresada en años, el uso de anticonceptivos hormonales, el número de ITS, la presencia de herpes genital, estar infectado por VIH, el número de diagnósticos de ITS, el diagnóstico previo de cáncer, antecedente de NIC, y el diagnóstico previo de VPH. Así mismo, Cicarrese et al (77) recientemente reportaron los factores de infección por VPH-AR y el tabaquismo como factores relacionados con persistencia viral.

Con los estudios previos se resalta al tabaquismo como un factor asociado con el desarrollo de CaCU, y aunque se desconoce el mecanismo exacto, se han propuestos algunos mecanismos moleculares de como el tabaco contribuye en la carcinogénesis, algunos de éstos explican que ocurre una la exposición directa del ADN en las células del epitelio del cérvix debido a la nicotina y cotinina, mientras que otras teorías abordan que la exposición de los metabolitos resultantes de la reacción de componentes como son hidrocarbonados policiclicos aromáticos, y las aminas aromáticas producen efectos potencialmente mutagénicos (78). Por lo que se puede hipotetizar que encontrar el marcador de riesgo de tabaquismo en las mujeres participantes de la presente cohorte es uno de los factores a modificar en las mujeres usuarias del CAPASAM para reducir el riesgo de persistencia viral y con ello el desarrollo de CaCU.

Por otra parte, hay que recalcar la importancia del emplear este tipo de herramientas (como modelos de aprendizaje de máquinas) que han funcionado en otras áreas de estudio y en otras poblaciones, como lo muestran en el trabajo propuesto por Desai Kanan T et al. (79) donde se desarrolló un método de evaluación visual automatizada (AVE) para la detección del CaCU por medio del aprendizaje profundo (Deep learning o DL por sus siglas en inglés), donde se

analizaron 2160 imágenes colposcópicas, por medio del método de Red Neuronal Convolucional VGG-16 y LR Multivariable, obteniendo una AUC 92.1% en relación a la detección de CaCU, este estudio destaca que la aplicación de nuevas tecnologías como es la implementación del método AVE, es una alternativa novedosa, de bajo costo para la detección y de mejora del triage en mujeres infectadas por VPH.

Finalmente, como se representa en la figura 26, con los resultados encontrados en el presente estudio se puede concluir que los marcadores de riesgo relacionados a través de un modelo de aprendizaje de máquinas con persistencia de VPH a nivel del cérvix son: factores relacionados con el virus (coinfección por genotipos de VPH general, carga viral), factores relacionados con antecedentes de salud sexual y reproductiva y el estilo de vida (antecedente de uso de anticonceptivos hormonales, antecedente de tabaquismo, número de embarazos, el antecedente de ITS, la edad de debut sexual y el número de parejas sexuales) y factores constitucionales de la mujer (edad).

Figura 26. Marcadores de riesgo relacionados a persistencia de VPH a nivel de cérvix a través de un modelo de aprendizaje de máquinas en mujeres de CAPASAM 2015-2017

En resumen, este trabajo se basó en el abordaje de modelos de aprendizaje de máquinas que permitieran realizar la predicción de persistencia viral en cérvix y la identificación de los factores de riesgo relacionados con ésta. Los resultados obtenidos en el presente trabajo de investigación son similares a los referidos en la literatura por diversos autores que utilizaron herramientas como la empleada en este proyecto para CaCU, así como también en aquellos que usaron otras metodologías para predecir persistencia de VPH.

12. LIMITACIONES DEL ESTUDIO

Una de las limitaciones en el presente estudio es la representatividad de la población a estudiar, ya que solo podrán extrapolarse los resultados a poblaciones con características similares a la población analizada, otra limitante es que no se tuvo una muestra tan grande y fue necesario equilibrar los datos para la compensación de la reducción de la población.

Además, de que al ser una base de datos ya previamente elaborada se tuvo que revisar a fondo para la correcta imputación de los datos perdidos o el ruido existente. También, se debe de considerar que algunos de los datos disponibles fueron recolectados por medio de auto reportes referidos por las pacientes participantes en la cohorte, lo que puede representar un sesgo de información.

Otra limitación importante en este estudio es el respaldo de la información, que es analizada referente al contexto de CaCU, mientras que son pocos los artículos académicos que tratan sobre la identificación de persistencia viral, por lo que la información recopilada durante el análisis y la discusión de este documento es sobre CaCU y no sobre la persistencia viral que es el tema central tratado en este trabajo.

Finalmente, otra de las limitantes presentes en este estudio es que las tomas realizadas en las pacientes para la detección de los genotipos de VPH fueron realizadas anualmente, por lo que no se puede determinar si hubo eliminación viral antes de los 12 meses o si ocurrió una reinfección por el mismo genotipo viral.

13. CONCLUSIONES

La infección por VPH es una ITS que representa una amenaza para la salud de las mujeres en todo el mundo. Aunque afecta a ambos sexos, es la mujer quien tiene el riesgo de desarrollar

el CaCU. Se tiene el conocimiento de que la persistencia del VPH es esencial, pero no suficiente para el desarrollo de neoplasias malignas de alto grado a nivel del cérvix y el desarrollo de CaCU.

En una minoría de casos se puede detectar la persistencia una vez transcurridos 12 meses; ya que cada infección por este virus es distinta (según el huésped infectado), la identificación de los factores que se han correlacionados con la presencia de persistencia es crucial, es por todo esto que este trabajo es una fuente de oportunidades que pueden optimizar tanto el diagnóstico como el tratamiento de cada paciente y mejorar sus resultados clínicos en el futuro, al haber logrado identificar los factores de riesgo que estarán relacionados con el desarrollo de la persistencia del VPH a nivel del cérvix.

El mundo que está en constante cambio y que avanza hacia el uso de mecanismos computacionales nos exige el crecimiento y la implementación de herramientas alternativas que sean confiables, eficientes y precisas, y que nos puedan apoyar al diagnóstico de persistencia viral, como son el uso de métodos de aprendizaje de máquinas. El abordaje de modelos de aprendizaje de máquinas pueden ser una herramienta útil que facilita el trabajo del personal de salud que atiende a población con infección por VPH.

La metodología propuesta en este trabajo para la predicción e identificación de marcadores de riesgo de persistencia de VPH a nivel de cérvix basado en aprendizaje de máquinas, mostró un buen rendimiento para realizar la predicción de persistencia en el primer año y segundo año. Además, los marcadores de riesgo, fueron muy similares a los reportados por la literatura: co-infección por genotipos de VPH, la edad del paciente, el antecedente de tabaquismo, el número de parejas sexuales, edad de inicio de vida sexual y el uso de anticonceptivos hormonales.

Este estudio puede ser considerado pionero en el abordaje de la predicción de persistencia viral por VPH en mujeres mexicanas y también en la búsqueda de nuevos marcadores de riesgo, no solo en infecciones por VPH. Sin embargo, es importante que en futuros estudios se aborden con una muestra mayor para poder obtener resultados que representen a toda la población mexicana.

14. CONSIDERACIONES ÉTICAS

Este estudio estuvo anidado al proyecto “Las citocinas inmunosupresoras Th2 y Th3 como predictores de persistencia viral y progresión a lesión pre-maligna en cérvix en mujeres infectadas con VPH-AR: estudio de cohorte”, el cual contó con la aprobación del Comité de Investigación CI- 342 -2016. No. de Proyecto 1287, Comité de Ética en Investigación con el número de aprobación CI: 1287 y Comité de de Bioseguridad CI: 1287 CB: 1278. Para tener acceso a la base de datos y utilizarla en el presente estudio, se solicitó autorización de la investigadora responsable, anexo 1.

15. RECURSOS MATERIALES Y FINANCIAMIENTO

El presente estudio estuvo anidado al proyecto de investigación: “Las citocinas inmunosupresoras Th2 y Th3 como predictores de persistencia viral y progresión a lesión pre-maligna en cérvix en mujeres infectadas con VPH-AR: estudio de cohorte”; el cual contó con financiamiento del Fondo Sectorial en Salud de CONACyT, Proyecto 233538 CONACYT-FONSEC SSA/IMSS/ISSSTE-2014.

16. BIBLIOGRAFÍA

1. Harden ME, Munger K. Human papillomavirus molecular biology [Internet]. Mutat Res Rev Mutat Res. 2017 [Consultado 12 Dec 2021];772:3-12. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5500221/
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries [Internet]. CA Cancer J Clin. 2021 [Consultado 12 Dec 2021];71(3):209–49. Disponible en: https://gco.iarc.fr/today/about
3. Rudolph SE, Lorincz A, Wheeler CM, Gravitt P, Lazcano-Ponce E, Torres-Ibarra L, et al. Population-based prevalence of cervical infection with human papillomavirus genotypes 16 and 18 and other high risk types in Tlaxcala, Mexico [Internet]. BMC Infect Dis. 2016 [Consultado 12 Dec 2021];16(1). Disponible en: https://pubmed.ncbi.nlm.nih.gov/27585544/
4. Torres-Poveda K, Ruiz-Fraga I, Madrid-Marina V, et al. High risk HPV infection prevalence and associated cofactors: a population-based study in female ISSSTE beneficiaries attending the HPV screening and early detection of cervical cancer

program [Internet]. BMC Cancer.2019 [Consultado 12 Dec 2021];19(1). Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6905062/

5. Arbyn M, Martin-Hirsch P, Buntinx F, et al. Triage of women with equivocal or low-grade cervical cytology results: a meta-analysis of the HPV test positivity rate [Internet]. J Cell Mol Med. 2009 [Consultado 12 Dec 2021];13(4):648–659.Disponible en: https://pubmed.ncbi.nlm.nih.gov/19166485/
6. Gravitt PE, Winer RL. Natural history of HPV infection across the lifespan: Role of viral latency. Viruses. 2017;9(10):1–10.
7. Usyk M, Zolnik CP, Castle PE, Porras C, Herrero R, Gradissimo A, et al. Cervicovaginal microbiome and natural history of HPV in a longitudinal study. PLoS Pathog [Internet]. 2020 [Consultado 12 Dec 2021];16(3):1–20. Disponible en: http://dx.doi.org/10.1371/journal.ppat.1008376
8. Monsalve Torra AE. Tesis Doctoral. Sistemas de Ayuda a la Decisión Clínica en Enfermedades de Diagnóstico Complejo [Internet]. España: Universidad de Alicante; 2017 [Consultado 12 Dec 2021]. Disponible en: https://rua.ua.es/dspace/bitstream/10045/65334/1/tesis_monsalve_torra.pdf
9. Burd EM. Human Papillomavirus and Cervical Cancer. Clin Microbiol Rev [Internet]. 2003 [Consultado 12 Dec 2021];16(1):1. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC145302/
10. Lizano M, et al. Infección por virus del Papiloma Humano: Epidemiología, Historia Natural y Carcinogénesis. Cancerol [Internet]. 2009 [Consultado 12 Dec 2021];4:2015–216. Disponible en: http://incan-mexico.org/revistainvestiga/elementos/documentosPortada/1272302572.pdf
11. Castro AA, Fournier M. Virus del papiloma humano. Ginecología Rev Med Cos Cen LXX [Internet]. 2013 [Consultado 12 Dec 2021]; 606: 211-217.Disponible en: https://www.medigraphic.com/pdfs/revmedcoscen/rmc-2013/rmc132d.pdf
12. Kaliterna V, Barisic Z. Genital human papillomavirus infections. Front Biosci - Landmark [Internet]. 2018 [Consultado 12 Dec 2021];23(9):1587–611. Disponible en: https://www.fbscience.com/Landmark/articles/10.2741/4662
13. Chow LT, Broker TR. Human Papillomavirus Infections: Warts or Cancer? Cold Spring Harb Perspect Biol [Internet]. 2013 [Consultado 12 Dec 2021];5(7):a012997.Disponible en: http://cshperspectives.cshlp.org/content/5/7/a012997.full

14. Maglennon GA, Doorbar J. Suppl 2: The Biology of Papillomavirus Latency. Open Virol J [Internet]. 2012 [Consultado 12 Dec 2021];6(1):190. Disponible en: /pmc/articles/PMC3547330/
15. OPS/OMS. Virus del Papiloma Humano (VPH) [Internet]. Preguntas Frecuentes: Virus del papiloma humano (VPH). 2018 [Consultado 12 Dec 2021]. Disponible en: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=14873:sti-human-papilloma-virus-hpv&Itemid=3670&lang=es
16. Serrano B, Brotons M, Bosch FX, Bruni L. Epidemiology and burden of HPV-related disease. Best Pract Res Clin Obstet Gynaecol. 2018;47:14–26.
17. Gravitt PE. The known unknowns of HPV natural history. J Clin Invest [Internet]. 2011 [Consultado 14 Jan 2022];121(12):4593. Disponible en: /pmc/articles/PMC3225991/
18. Peralta-Rodríguez R, Romero-Morelos P, Villegas-Ruíz V, Mendoza-Rodríguez M, Taniguchi-Ponciano K, González-Yebra B, et al. Prevalence of human papillomavirus in the cervical epithelium of Mexican women: meta-analysis. Infect Agent Cancer [Internet]. 2012 [Consultado 14 Jan 2022];7(1):34. Disponible en: /pmc/articles/PMC3586354/
19. Parada R, Morales R, Giuliano AR, Cruz A, Castellsagué X, Lazcano-Ponce E. Prevalence, concordance and determinants of human papillomavirus infection among heterosexual partners in a rural region in central Mexico. BMC Infect Dis [Internet]. 2010 [Consultado 14 Jan 2022];10(1):223. Disponible en: /pmc/articles/PMC2941497/
20. Torres-Poveda, Kirvis et al. “Molecular markers for the diagnosis of high-risk human papillomavirus infection and triage of human papillomavirus-positive women.” Revista de investigacion clinica; organo del Hospital de Enfermedades de la Nutricion vol. 72,4 (2020): 198-212. doi:10.24875/RIC.20000058
21. DOF. Modificación a la norma oficial mexicana NOM-014-SSA2-1994, para la prevención, detección, diagnóstico, tratamiento, control y vigilancia epidemiológica del cáncer cérvico uterino [Internet]. [Consultado 13 Dec 2021]. Disponible en: http://www.dof.gob.mx/normasOficiales/2383/SALUD/SALUD.htm
22. DOF. Norma oficial mexicana NOM-010-SSA2-2010 para la Prevención y el Control del VIH y el sida. [Internet]. [Consultado 13 Dec 2021]. Disponible en: http://dof.gob.mx/nota_detalle.php?codigo=5166864&fecha=10/11/2010
23. Secretaría de Salud. GPC: Prevención, detección, diagnóstico y tratamiento de lesiones

precursoras del cáncer del cuello cérvicouterino en primer y segundo nivel de atención. [Internet]. [Consultado 13 Dec 2021]. Disponible en: https://cenetec-difusion.com/gpc-sns/?p=525

24. Nilsson NJ. Introduction to machine learning an early draft of a proposed textbook. [Internet]. USA: University Stanford; 1998 [Consultado 14 Jan 2022]. Disponible en: https://ai.stanford.edu/~nilsson/MLBOOK.pdf
25. Viera AFG, Viera AFG. Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investig Bibl [Internet]. 2017 [Consultado 13 Dec 2021];31(71):103–26. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2017000100103&lng=es&nrm=iso&tlng=es
26. Molnar, C. Interpretable machine learning. A Guide for Making Black Box Models Explainable [Internet]. USA; 2019 [Consultado 22 Jan 2022]. Disponible en: https://christophm.github.io/interpretable-ml-book/.
27. Murphy KP. Machine Learning: A Probabilistic Perspective [Internet]. Inglaterra: Cambridge, Massachusetts; 2012 [Consultado 22 Jan 2022]. Disponible en: http://noiselab.ucsd.edu/ECE228/Murphy_Machine_Learning.pdf
28. León EC. Trabajo de fin de grado. Introducción a las máquinas de vector soporte (SVM) en aprendizaje supervisado [Internet].España: Universidad Zaragoza; 2012 [Consultado 22 Jan 2022].Disponible en: https://zaguan.unizar.es/record/59156/files/TAZ-TFG-2016-2057.pdf
29. Megan D. et al. Techniques for Interpretable Machine Learning. Communications of the ACM [Internet]. 2020 [Consultado 12 Jul 2022]. Disponible en: https://dl.acm.org/doi/fullHtml/10.1145/3359786
30. Betancourt GA. Las máquinas de soporte vectorial (SVMs). Scientia Et Technica [Internet]. 2005 [Consultado 22 Jan 2022];9(27):67-72. Disponible en: https://www.redalyc.org/pdf/849/84911698014.pdf
31. Scikit Learn. Support Vector Machines [Internet]. 2007-2021 [Consultado 14 Jan 2021]. Disponible en: https://scikit-learn.org/stable/modules/svm.html
32. Breiman L. Random Forests [Internet]. Berkeley: Universidad de California; 2001. [Consultado 22 Jan 2022]. Disponible en: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
33. Espinosa-Zúñiga JJ, Espinosa-Zúñiga JJ. Aplicación de algoritmos Random Forest y

XGBoost en una base de solicitudes de tarjetas de crédito. Ing Investig y Tecnol [Internet]. 2020 [Consultado 14 Jan 2022];21(3):1–16. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000300002&lng=es&nrm=iso&tlng=es

34. Lizares Castillo M. Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico [Internet]. Lima, Perú: Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos; 2017 [Consultado 14 Jan 2022]. Disponible en: http://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/7122/Lizares_cm.pdf?sequence=1&isAllowed=y
35. Jiménez JAA, Naranjo MAG. Tema 12: Arboles de decisión. Razonamiento Automático. Universidad de Sevilla: España; 2000. Curso 2000–2001.
36. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD [Internet]. 2016 [Consultado 14 Jan 2022]:785–794. Disponible en: http://dx.doi.org/10.1145/2939672.2939785
37. Ichi.Pro. ¿Cómo funciona XGBoost? [Internet]. 2016 [Consultado 14 Jan 2022]. Disponible en: https://ichi.pro/es/como-funciona-xgboost-128143693994154
38. Utrera RG, Pau C:, Balado A. Uso de algoritmos de aprendizaje automático aplicados a bases de datos genéticos [Internet]. Cataluña: España; 2017 [Consultado 14 Jan 2022]. Disponible en: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/65426/6/rgagoTFM0617memoria.pdf
39. Chávez Martínez R. Comparación entre regresión logística y redes neuronales para predecir cáncer de piel en perros [Internet]. Perú: Universidad de Lima; 2019 [Consultado 14 Jan 2022]. Disponible en: https://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/8401/Chavez_Martinez_Renato.pdf?sequence=3&isAllowed=y
40. Mahamat AA, Boukar MM, Ibrahim NM, Stanislas TT, Linda Bih N, Obianyo II, et al. Machine learning approaches for prediction of the compressive strength of alkali activated termite mound soil. Appl Sci [Internet]. 2021;11(11). [Consultado 15 Jan 2022]. Disponible en: https://www.mdpi.com/2076-3417/11/11/4754
41. Brownlee Jason. Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and

CatBoost [Internet]. 2021. [Consultado 02 Agosto 2022]. Disponible en: https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/

42. Microsoft. LightGBM [Internet]. 2016. [Consultado 02 Agosto 2022]. Disponible en: https://www.microsoft.com/en-us/research/project/lightgbm/
43. GeekforGeeks. LightGBM (Light Gradient Boosting Machine) [Internet sitio web]. 2021. [Consultado 02 Agosto 2022]. Disponible en: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/
44. Bengoechea Rodríguez, S. et al. Herramienta para la interpretabilidad y análisis de sesgos en modelos de ML [Internet]. 2021. [Consultado 02 Agosto 2022]. Disponible en: https://biblioteca.cunef.edu/files/documentos/TFM%20S.%20Bengoechea,%20B.%20Cardaba,%20M.%20Ferrin,%20L.%20Martinez.pdf
45. Scott Lundberg. Welcome to the SHAP documentation [Internet]. 2020. [Consultado 02 Agosto 2022]. Disponible en: https://shap.readthedocs.io/en/latest/index.html#
46. Christoph Molnar. Aprendizaje automático interpretable. SHAP (explicaciones aditivas SHapley) [Internet]. 2021. [Consultado 02 Agosto 2022]. Disponible en: https://fedefliguer.github.io/AAI/shap.html
47. Jason Brownlee. Applied Machine Learning Process [Internet]. Publicado 12 de Febrero 2014. [Consultado 26 Jul 2022]. Disponible en: https://machinelearningmastery.com/start-here/#process
48. Le T al et. Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. 10th International Conference on Knowledge and Systems Engineering (KSE) [Internet]. 2018. [Consultado 26 Jul 2022]. Disponible en: http://www.jaist.ac.jp/~razvan/publications/comparison_imputation_healthcare.pdf
49. Brita Inteligencia Articial. Las seis técnicas principales utilizadas en la Ingeniería de Machine Learning - Brita Inteligencia Artificial [Internet]. 2021 [Consultado 15 Jan 2022]. Disponible en: https://brita.mx/las-seis-tecnicas-principales-utilizadas-en-la-ingenieria-de-machine-learning
50. Ruiz Chávez Z, García Rodríguez J. Tesis de grado. Técnicas de Aprendizaje Automático Aplicadas al Procesamiento de Información Demográfica [Internet]. España: Universidad de Alicante; 2019 [Consultado 15 Jan 2022]. Disponible en: https://rua.ua.es/dspace/bitstream/10045/95608/1/tesis_zoila_ruiz.pdf

51. Power Data. Preprocesar y normalizar datos, 4 pasos para limpiar y mejorar datos [Internet]. Power Data; 2017 [Consultado 15 Jan 2022]. Disponible en: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/preprocesar-y-normalizar-datos-4-pasos-para-limpiar-y-mejorar-datos
52. Microsoft Docs. Normalización de datos: referencia de componente - Azure Machine Learning. Microsoft Docs [Internet]. [Consultado 15 Jan 2022]. Disponible en: https://docs.microsoft.com/es-es/azure/machine-learning/component-reference/normalize-data
53. Julián W, González G. Analysis of the input data processing for fault location in power distribution systems. Tecnura. 2014;64(41):64–75.
54. Chawla et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research [Internet]. 2002;16: 321-357. [Consultado 26 Jul 2022]. Disponible en: https://arxiv.org/abs/1106.1813
55. Jason Brownlee. SMOTE for Imbalanced Classification with Python [Internet]. Publicado 17 de Marzo 2021. [Consultado 26 Jul 2022]. Disponible en: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
56. H. He, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) [Internet]. 2008. [Consultado 26 Jul 2022]. Disponible en: https://doi.org/10.1109/IJCNN.2008.4633969
57. Ochi T., Murase K., Fujii T., et al. Predicción de supervivencia utilizando redes neuronales artificiales en pacientes con cáncer de cuello uterino tratadas solo con radioterapia . En t. J. Clin. oncol. 2002; 7: 0294–0300.
58. Yuan C, Yao Y, Cheng B, Cheng Y, Li Y, Li Y, et al. The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. Sci Reports 2020 101 [Internet]. 2020 [Consultado 13 Dec 2021];10(1):1–12. Disponible en: https://www.nature.com/articles/s41598-020-68252-3
59. Gupta S, Gupta MK. Computational Prediction of Cervical Cancer Diagnosis Using Ensemble-Based Classification Algorithm. Comput J [Internet]. 2021 [Consultado 13 Dec 2021]; Disponible en: https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxaa198/6153484

60. Asadi F, Salehnasab C, Ajori L. Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. J Biomed Phys Eng [Internet]. 2020 [Consultado 13 Dec 2021];10(4):513. Disponible en: /pmc/articles/PMC7416093/
61. Akter L, Ferdib-Al-Islam ·, Islam · Md Milon, Mabrook ·, Al-Rakhami S, Haque MR. Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques. SN Comput Sci 2021 23 [Internet]. 2021 [Consultado 13 Dec 2021];2(3):1–10. Disponible en: https://link.springer.com/article/10.1007/s42979-021-00551-6
62. Rayavarapu K, Krishna KKV. Prediction of Cervical Cancer using Voting and DNN Classifiers. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-5. Disponible en: https://ieeexplore.ieee.org/document/8551176
63. Ranganath R, Perotte A, Blei D. Deep Survival Analysis. Presented at 2016 Machine Learning and Healthcare Conference (MLHC 2016), Los Angeles, CA. Disponible en: https://arxiv.org/abs/1608.02158
64. Hemingway H, Feder GS, Fitzpatrick NK, Denaxas S, Shah AD, Timmis AD. Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAl disease research using LInked Bespoke studies and Electronic health Records (CALIBER) programme. Program Grants Appl Res [Internet]. 2017 [Consultado 13 Dec 2021];5(4):1–330. Disponible en: https://pubmed.ncbi.nlm.nih.gov/28151614/
65. Torres-Poveda K, Burguete-García AI, Bahena-Román M, Méndez-Martínez R, Zurita-Díaz MA, López-Estrada G, et al. Risk allelic load in Th2 and Th3 cytokines genes as biomarker of susceptibility to HPV-16 positive cervical cancer: a case control study. BMC Cancer [Internet]. 2016 May 24 [Consultado 16 Dec 2021];16(1). Disponible en: /pmc/articles/PMC4879749/
66. Torres-Poveda K, Bahena-Román M, Delgado-Romero K, Madrid-Marina V. A prospective cohort study to evaluate immunosuppressive cytokines as predictors of viral persistence and progression to pre-malignant lesion in the cervix in women infected with HR-HPV: Study Protocol. BMC Infect Dis 2018; 18(1):582-590.
67. Swai, P., Rasch, V., Linde, D.S. et al. Persistence and risk factors of high-risk human

papillomavirus infection among HIV positive and HIV negative tanzanian women: a cohort study. Infect Agents Cancer 17, 26 (2022). https://doi.org/10.1186/s13027-022-00442-2

68. Shi, Nianmin et al. “Analysis of risk factors for persistent infection of asymptomatic women with high-risk human papilloma virus.” Human vaccines & immunotherapeutics vol. 13,6 (2017): 1-7. doi:10.1080/21645515.2016.1239669

69. Lilhore, Umesh Kumar et al. “Hybrid Model for Detection of Cervical Cancer Using Causal Analysis and Machine Learning Techniques.” Computational and mathematical methods in medicine vol. 2022 4688327. 4 May. 2022, doi:10.1155/2022/4688327

70. Mehmood M, Rizwan M, Gregus ml Mand Abbas S. Machine Learning Assisted Cervical Cancer Detection. Front. Public Health (2021); 9:788376. doi: 10.3389/fpubh.2021.788376

71. Weida Z, Huihong L, Xiong W, Yuanhua C. The current situation of Hainan Li Nationality females human papilloma virus infection and related factors analysis.China Modern Med J 2016; 26(01):89-93

72. Li W, Meng Y, Wang Y, Cheng X, Wang C, Xiao S, Zhang X, Deng Z, Hu M, Shen P, Xu S, Fu C, Jiang W, Wu B, Li K, Chen G, Wei J, Xi L, Hu J, Ma D, Xue M, Xie X, Wu P. Association of age and viral factors with high-risk HPV persistence: A retrospective follow-up study. Gynecol Oncol. 2019 Aug;154(2):345-353. doi: 10.1016/j.ygyno.2019.05.026. Epub 2019 Jun 24. PMID: 31242966.

73. Oh HY, Kim MK, Seo S, Lee DO, Chung YK, Lim MC, et al. Alcohol consumption and persistent infection of high-risk human papillomavirus. Epidemiol Infect. 2015;143(7):1442–50.

74. Kaushik, Manoj et al. “Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: A machine learning approach.” Computers in biology and medicine vol. 134 (2021): 104559. doi:10.1016/j.compbiomed.2021.104559

75. Nielsen A, Kjaer SK, Munk C, Osler M, Iftner T. Persistence of high-risk human papillomavirus infection in a population-based cohort of Danish women. J Med Virol. 2010;82(4):616–23.

76. Ijaz, Muhammad Fazal et al. “Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods.” Sensors (Basel, Switzerland) vol. 20,10 2809. 15 May. 2020, doi:10.3390/s20102809

77. Ciccarese G, Herzum A, Pastorino A, Dezzana M, Casazza S, Mavilia MG, Copello F, Parodi A, Drago F. Prevalence of genital HPV infection in STI and healthy populations and risk factors for viral persistence. Eur J Clin Microbiol Infect Dis. 2021 Apr;40(4):885-888. doi: 10.1007/s10096-020-04073-6. Epub 2020 Oct 16. PMID: 33067736.

78. Núñez-Troconis José. Cigarrillo y cáncer de cuello uterino. Rev. chil. obstet. ginecol. [Internet]. 2017 Abr [citado 2022 Ago 19] ; 82(2): 232-240. Disponible en: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-75262017000200014&lng=es. http://dx.doi.org/10.4067/S0717-75262017000200014.

79. Desai, Kanan T et al. “The development of "automated visual evaluation" for cervical cancer screening: The promise and challenges in adapting deep-learning for clinical testing: Interdisciplinary principles of automated visual evaluation in cervical screening.” International journal of cancer vol. 150,5 (2022): 741-752. doi:10.1002/ijc.33879

17 ANEXOS

Anexo 1. Carta de Confidencialidad y autorización de uso de datos

Instituto Nacional de Salud Pública

Escuela de Salud Pública de México

Maestría en Salud Pública área de concentración en Enfermedades Infecciosas

Proyecto de Titulación

Carta de Confidencialidad y autorización de uso de datos

Título del proyecto académico: "PREDICCIÓN DE PERSISTENCIA DE INFECCIÓN POR VIRUS DEL PAPILOMA HUMANO EN CÉRVIX: BASADA EN MODELO DE APRENDIZAJE DE MÁQUINAS"

Alumna de la ESPM: Rodríguez Esquivel Jocelyn Isabel.

Fecha aprobación por el Comité de ética:

Estimado (a): Dra. Kirvis Torres Poveda

- Soy alumna de la Maestría en Salud Pública del Instituto Nacional de Salud Pública. Actualmente estoy realizando mi proyecto de titulación. El objetivo del estudio es analizar marcadores de riesgo de persistencia de infección por Virus del Papiloma Humano en cérvix mediante modelos de aprendizaje de máquinas en las mujeres que recibieron atención en CAPASAM en el período de 2015-2017. Los resultados esperados es el desarrollo de una metodología para predicción y la identificación de marcadores de riesgo de persistencia de VPH a nivel de cérvix basado en aprendizaje de máquinas.
- Por este medio le solicito su autorización para el uso de la base de datos recolectada en el estudio "Las citocinas inmunosupresoras Th2 y Th3 como predictores de persistencia viral y progresión a lesión pre-maligna en cérvix en mujeres infectadas con VPH-AR: estudio de cohorte", el cual cuenta con la aprobación del Comité de Investigación CI- 342 -2016. No. de Proyecto 1287, Comité de Ética en Investigación con el número de aprobación CI: 1287 y Comité de de Bioseguridad CI: 1287 CB: 1278

Confidencialidad: toda la información que usted proporcione para el estudio será de carácter estrictamente confidencial, será utilizada únicamente por el equipo de investigación del proyecto y no estará disponible para ningún otro propósito. Los resultados de este estudio serán publicados con fines científicos.

Aviso de Privacidad Simplificado: La información solicitada será utilizada exclusivamente para las

finalidades expuestas en este documento. Los datos proporcionados, serán protegidos conforme a lo dispuesto por el **Comité de Ética en Investigación del INSP.**

Declaración de la persona que da el consentimiento y autorización de uso de datos. ● He leído esta Carta de Confidencialidad. ● Autorizo el uso de datos proporcionados. ● Me han explicado el estudio el estudio de investigación incluyendo el objetivo y resultados esperados.
Nombre y Firma

Anexo 2. Experimentación con distintas variables en el modelaje de persistencia viral al año y al segundo año en mujeres de CAPASAM 2015-2017

Resultados de modelo predicción de persistencia del primer año

Se determinaron cuatro categorías para incluir a distintas variables: sociodemográficas, de historia clínica y estilos de vida, historia sexual y reproductiva y relacionadas al virus. Se realizo el modelaje al año y al segundo año con todas las características y por categoría.

Variables seleccionadas por categoría y agrupación de variables

Categoría	Variables utilizadas en la predicción de persistencia al primer año	Variables utilizadas en la predicción de persistencia al segundo año
Variables sociodemográficas	Edad Estado civil Ocupación Nivel educativo	Se usan las mismas variables del primer año
Variables de historia clínica y estilos de vida	Antecedente de cáncer familiar Tipo de cáncer familiar Nivel de consanguinidad de cáncer Dispareunia Disuria Prurito Resequedad vaginal Leucorrea Lavado vaginal postcoital Antecedente de tabaquismo Tabaquismo actual Número de cigarrillos al día Antecedente de alcoholismo Alcoholismo actual	Se usan las mismas variables del primer año
Variables de historia sexual y reproductiva	Antecedente de ITS Tipo de ITS Edad de menarca Edad de inicio de vida sexual Número de parejas sexuales Número de embarazos Tipo de método de planificación familiar Otros métodos de planificación familiar Diagnóstico colposcópico Resultado de citología Resultado de biopsia Antecedente de tratamiento ginecológico	Se usan las mismas variables del primer año
Variables relacionadas al virus	Presencia de coinfección general de VPH Presencia de genotipos de alto riesgo Número de coinfecciones de genotipos de alto riesgo	Presencia de coinfección general de VPH al año Presencia de genotipos de alto riesgo al año Número de infecciones de genotipos de alto riesgo al año

	Infección únicamente por genotipos 16 y 18 Infección por genotipo 16 Carga viral de genotipo 16 Infección por genotipo 18 Carga viral de genotipo 18 Infección por genotipo 26 Carga viral de genotipo 26 Infección por genotipo 31 Carga viral de genotipo 31 Infección por genotipo 33 Carga viral de genotipo 33 Infección por genotipo 35 Carga viral de genotipo 35 Infección por genotipo 39 Carga viral de genotipo 39 Infección por genotipo 45 Carga viral de genotipo 45 Infección por genotipo 51 Carga viral de genotipo 51 Infección por genotipo 52 Carga viral de genotipo 52 Infección por genotipo 53 Carga viral de genotipo 53 Infección por genotipo 56 Carga viral de genotipo 56 Infección por genotipo 58 Carga viral de genotipo 58 Infección por genotipo 59 Carga viral de genotipo 59 Infección por genotipo 66 Carga viral de genotipo 66 Infección por genotipo 68 Carga viral de genotipo 68 Persistencia anual	Infección únicamente por genotipos 16 y 18 al año Infección por genotipo 16 al año Carga viral de genotipo 16 al año Infección por genotipo 18 al año Carga viral de genotipo 18 al año Infección por genotipo 31 al año Carga viral de genotipo 31 al año Infección por genotipo 33 al año Carga viral de genotipo 33 al año Infección por genotipo 39 al año Carga viral de genotipo 39 al año Infección por genotipo 45 al año Carga viral de genotipo 45 al año Infección por genotipo 51 al año Carga viral de genotipo 51 al año Infección por genotipo 52 al año Carga viral de genotipo 52 al año Infección por genotipo 56 al año Carga viral de genotipo 56 al año Infección por genotipo 58 al año Carga viral de genotipo 58 al año Infección por genotipo 59 al año Carga viral de genotipo 59 al año Infección por genotipo 66 al año Carga viral de genotipo 66 al año Infección por genotipo 68 al año Carga viral de genotipo 68 al año Persistencia al segundo año al año

Después, debido al desbalance de clases presente en los datos, se utilizaron técnicas de sobremuestreo (oversampling) con SMOTE y ADASYN, con la finalidad de compensar los escasos casos de persistencia y elegir el mejor método para la imputación. En las tablas a continuación, se muestran los rendimientos en el conjunto de entrenamiento de dichas técnicas, se destaca que los rendimientos en ambas fueron muy similares, sin embargo, se eligió el método SMOTE para continuar con las pruebas.

Rendimientos de técnica SMOTE en la predicción de persistencia de VPH al primer año usando todas las características, en el conjunto de entrenamiento

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9935	0.9646	0.9949	0.9794	0.9797
LR	0.9956	0.9747	0.9747	0.9747	0.9747
LightGBoost	0.9752	0.8838	0.9646	0.921	0.9242
RF	0.9895	0.9595	0.9444	0.9523	0.952
SVM	0.9797	0.9949	0.9646	0.98	0.9797

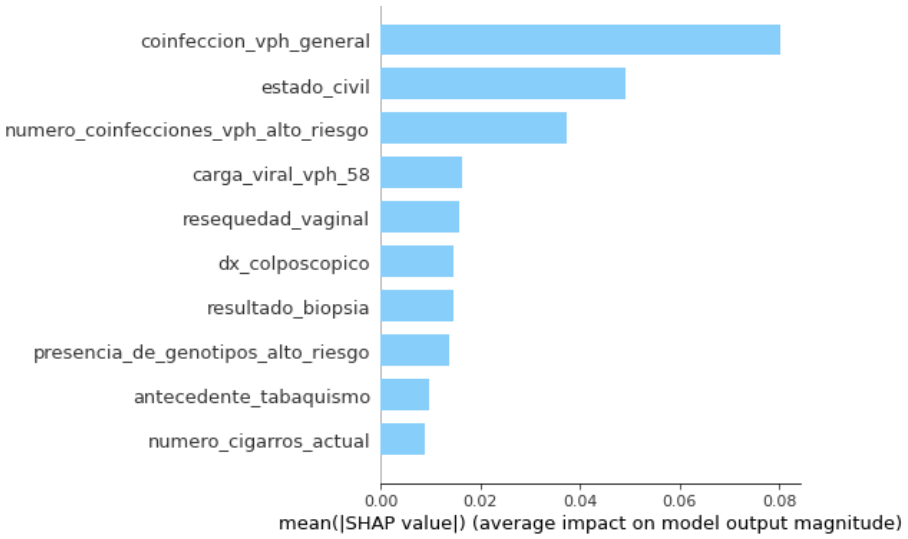
Rendimientos de técnica ADASYN en la predicción de persistencia de VPH al primer año usando todas las características, en el conjunto de entrenamiento

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9946	0.9648	0.9848	0.9746	0.9748
LR	0.994	0.9798	0.9797	0.9774	0.9773
LightGBoost	0.9818	0.9195	0.9646	0.9408	0.942
RF	0.9899	0.9597	0.9595	0.9597	0.9596
SVM	0.9773	0.9849	0.9696	0.9775	0.9773

Rendimientos de los modelos de predicción de persistencia de VPH al primer año usando todas las características, en el conjunto de prueba

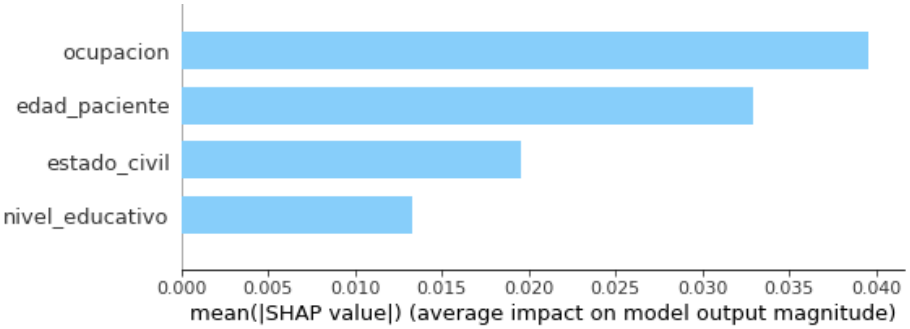
Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.984	0.94	0.92	0.9306	0.93

Marcadores de riesgo identificados para la persistencia viral por método SHAP las mujeres de CAPASAM 2015-2016

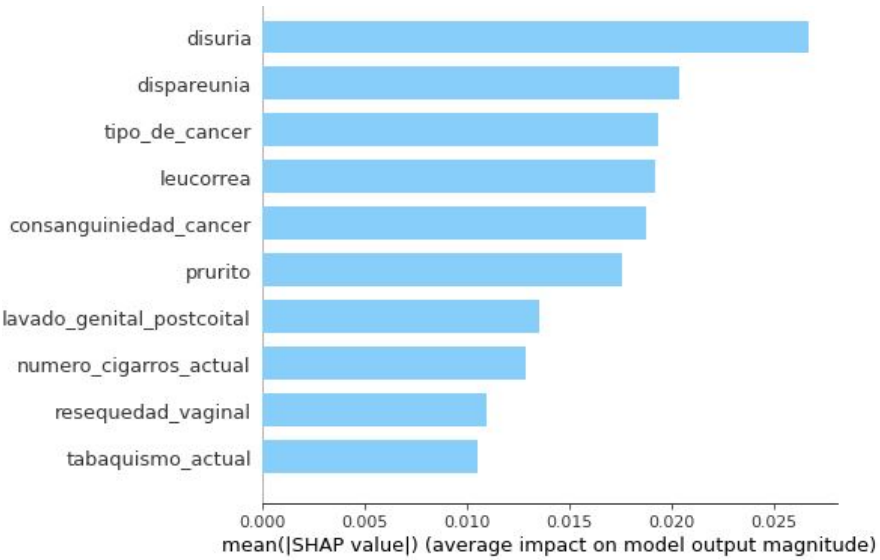


Posteriormente para poder identificar las variables más significativas para el desarrollo de la persistencia viral al año según las características presentes entre ellas, se construyó un modelo por cada categoría, sin el uso de SMOTE, con la intención que por medio del método de SHAP, se recolectaran las de mayor relevancia.

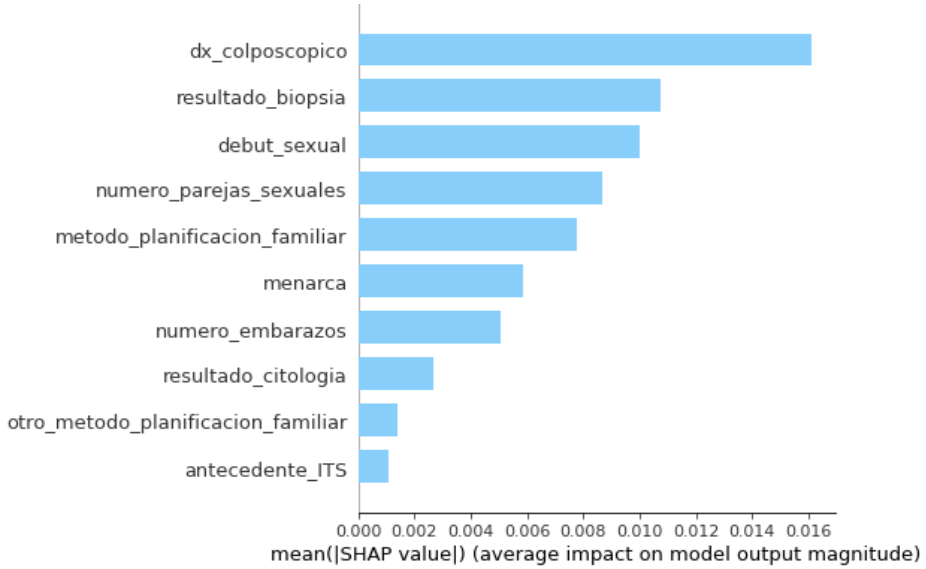
Marcadores de importancia identificados por categoría de variables sociodemográficas para el desarrollo de persistencia viral anual por método SHAP las mujeres de CAPASAM 2015-2016



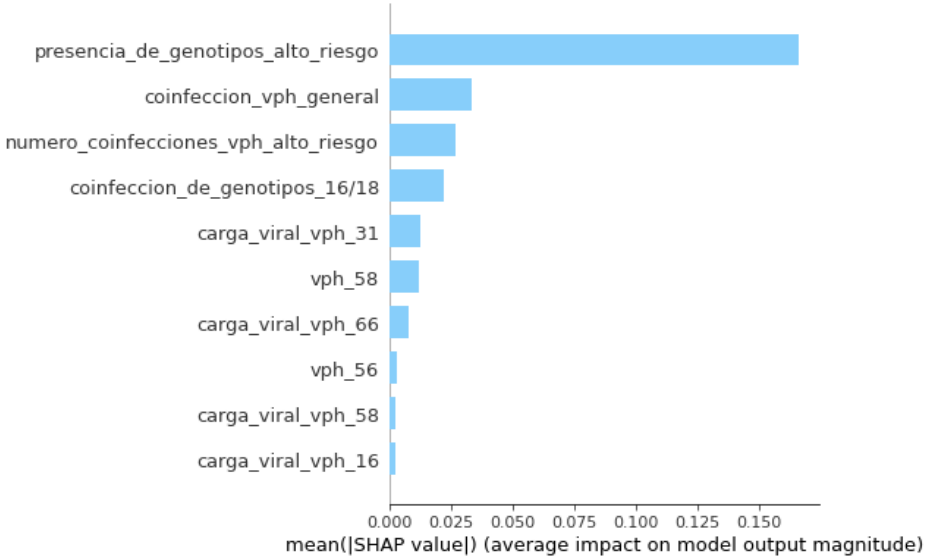
Marcadores de importancia identificados por categoría de variables de historia clínica y estilos de vida para el desarrollo de persistencia viral anual por método SHAP las mujeres de CAPASAM 2015-2016



Marcadores de importancia identificados por categoría de variables de historia sexual y reproductiva para el desarrollo de persistencia viral anual por método SHAP las mujeres de CAPASAM 2015-2016



Marcadores de importancia identificados por categoría de variables relacionadas al virus para el desarrollo de persistencia viral anual por método SHAP las mujeres de CAPASAM 2015-2016



Resultados de modelo predicción de persistencia del segundo año

En cuanto a la construcción del modelo del segundo año, se realizaron los mismos pasos antes descritos en la construcción del modelo del primer año.

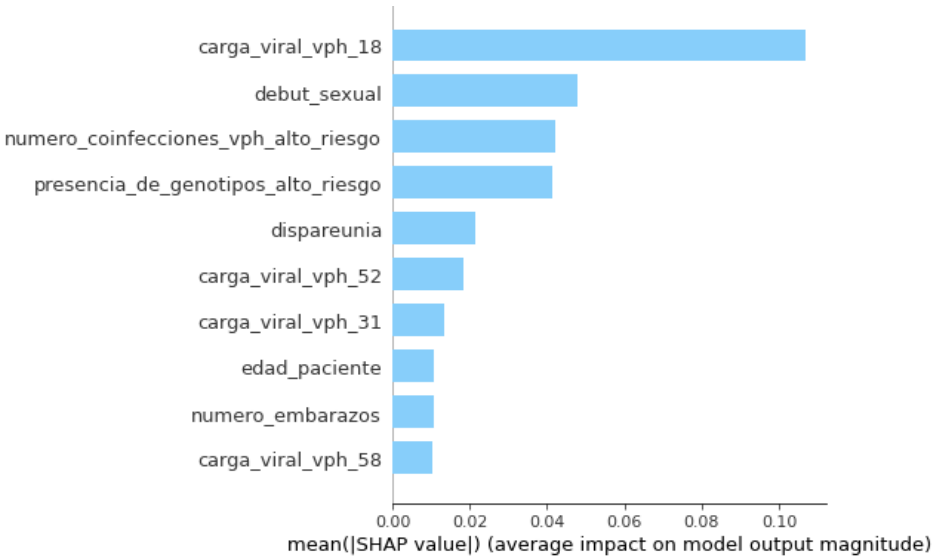
Rendimientos de los modelos de predicción de persistencia de VPH al segundo año con SMOTE usando todas las características, en el conjunto de entrenamiento

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9998	0.9904	0.9952	0.9928	0.9928
LR	1	1	0.9904	0.9952	0.9952
LightGBoost	0.9994	0.9952	0.9712	0.9835	0.9832
RF	0.9999	1	0.9952	0.9976	0.9976
SVM	0.9952	1	0.9904	0.9952	0.9952

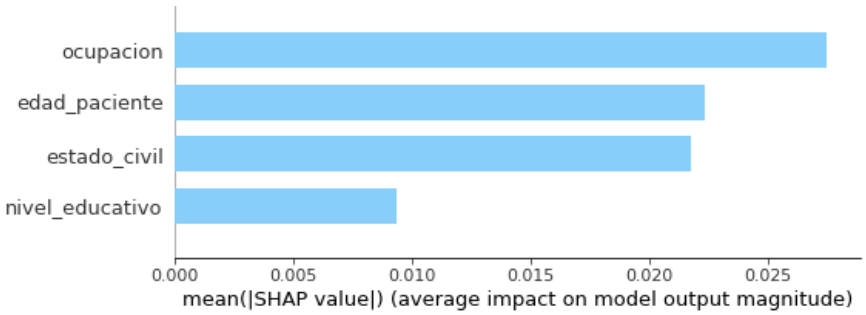
Rendimientos de los modelos de predicción de persistencia de VPH al segundo año usando todas las características, en el conjunto de prueba

Modelo	AUC	Sensibilidad	Especificidad	F1	ACC
XGBoost	0.9996	0.9807	1	0.9902	0.9904

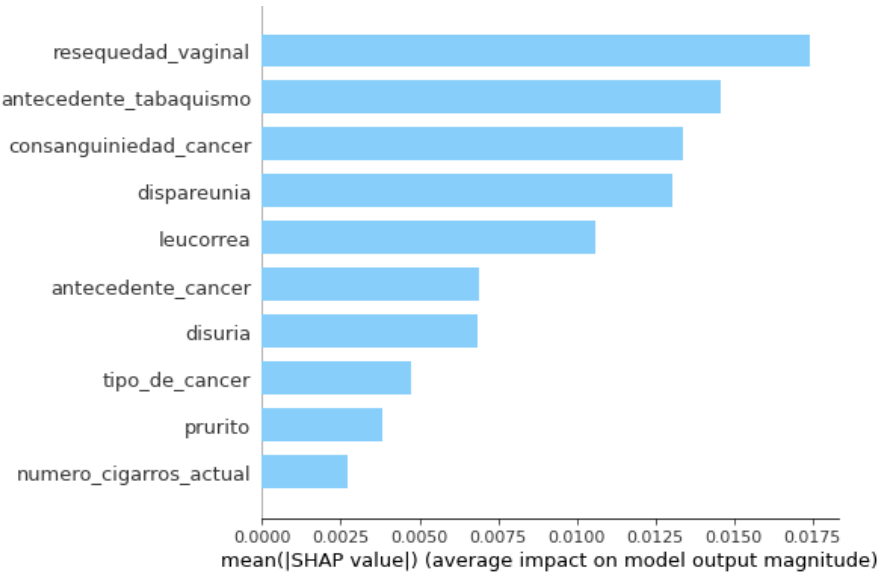
Marcadores de riesgo identificados para la persistencia viral al segundo año por método SHAP las mujeres de CAPASAM 2016-2017



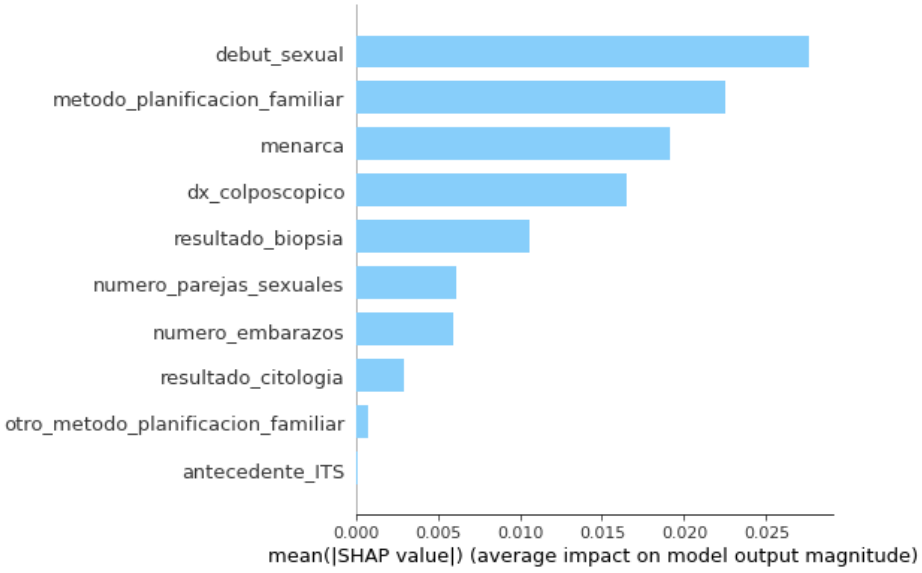
Marcadores de importancia identificados por categoría de variables sociodemográficas para el desarrollo de persistencia viral al segundo año por método SHAP las mujeres de CAPASAM 2016-2017



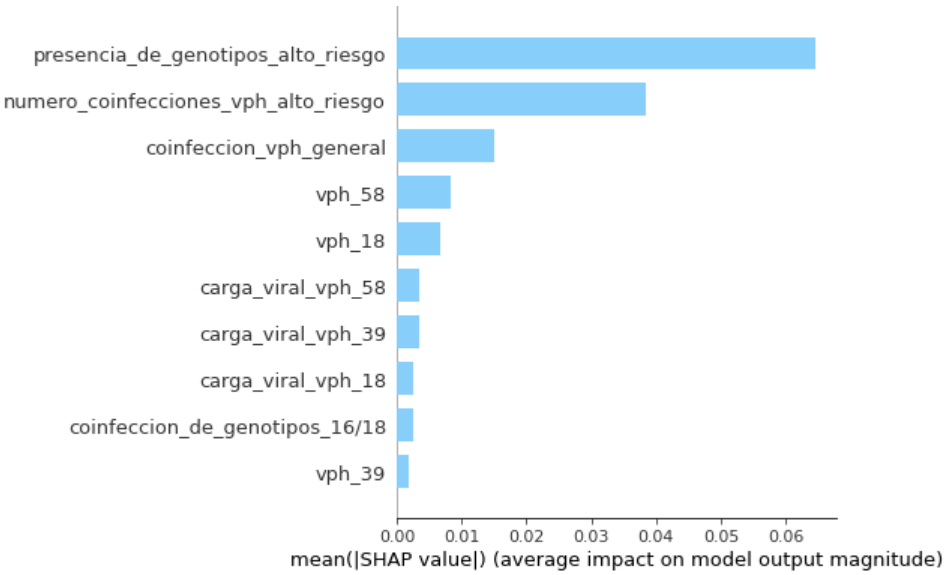
Marcadores de importancia identificados por categoría de variables de historia clínica y estilos de vida para el desarrollo de persistencia viral al segundo año por método SHAP las mujeres de CAPASAM 2016-2017



Marcadores de importancia identificados por categoría de variables de historia sexual y reproductiva para el desarrollo de persistencia viral al segundo año por método SHAP las mujeres de CAPASAM 2016-2017



Marcadores de importancia identificados por categoría de variables relacionadas al virus para el desarrollo de persistencia viral al segundo año por método SHAP las mujeres de CAPASAM 2016-2017


presencia_de_genotipos_alto_riesgo
numero_coinfecciones_vph_alto_riesgo
coinfeccion_vph_general
vph_58
vph_18
carga_viral_vph_58
carga_viral_vph_39
carga_viral_vph_18
coinfeccion_de_genotipos_16/18
vph_39
0.00
0.01
0.02
0.03
0.04
0.05
0.06
mean(|SHAP value|) (average impact on model output magnitude)