UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Ingeniería y Arquitectura

Enfoque predictivo para la reactividad a la prueba del virus de la inmunodeficiencia humana (VIH) basado en algoritmos de Machine Learning

Tesis para obtener el Título de Segunda Especialidad Profesional de Ingeniería y Arquitectura: Estadística Aplicada para Investigación

Autor:

Mg. Urlish Kleyber Marroquin Marroquin

Asesor:

PhD. Javier Linkolk López Gonzales

Lima, noviembre de 2023

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo PhD. Javier Linkolk López Gonzales, docente de la Unidad de Posgrado de Ingeniería y Arquitectura, Escuela de Posgrado de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: “Enfoque predictivo para la reactividad a la prueba del virus de la inmunodeficiencia humana (VIH) basado en algoritmos de Machine Learning” del autor Urlish Kleyber Marroquin Marroquin tiene un índice de similitud de 14% verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima, a los 06 días del mes de Diciembre del año 2023.

Nombres y apellidos del asesor

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Ñaña, Villa unión a 30 días del mes de noviembre del año 2023, siendo las 8:40 horas, se reunieron de forma online sincrónica, bajo la dirección del presidente del jurado Dra. Damaris Susana Quinteros Zuñiga, el secretario Dr. Josué Edison Turpo Chaparro; los demás miembros: y la Mg. Lizeth Geanina Huanca López y el Mg. Esteban Tocto Cano y el asesor PhD. Javier Linkolk López Gonzales con el propósito de administrar el acto académico de sustentación de Tesis de la Segunda Especialidad titulada "Enfoque predictivo para la reactividad a la prueba del virus de la inmunodeficiencia humana (VIH) basado en algoritmos de Machine Learning", conducente a la obtención del Título de Segunda Especialidad Profesional de Ingeniería y Arquitectura: Estadística Aplicada para Investigación.

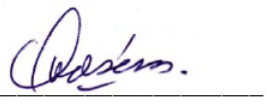
El presidente inició el acto académico de sustentación invitando al candidato hacer uso del tiempo determinado para su exposición. Concluido la exposición, el Presidente invitó a los demás miembros del Jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, los cuales fueron absueltos por el candidato. Luego se produjo un receso para las deliberaciones y la emisión del dictaminador del Jurado.

Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Candidato: Urlish Kleyber Marroquin Marroquin

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	19	A	Excelente	Excelencia

Finalmente, el Presidente del Jurado invitó al candidato a ponerse de pie, para recibir la evaluación final. Además, el Presidente del Jurado concluyó el acto académico de sustentación, procediéndose a registrar a registrar las firmas respectivas.



Presidente

Secretario

Asesor

Miembro

Miembro

Candidato

Agradecimientos

Agradezco a todas aquellas personas que, de alguna forma, me apoyaron en la culminación de este proyecto: mi familia, los docentes de la especialidad, mis compañeros de trabajo, mi asesor de tesis que creyó en mi tema de investigación y a las instituciones que me permitieron dar inicio y llevar adelante el presente proyecto, a todos ellos muchas gracias.

Dedicatoria

Dedico este proyecto a Dios por cuidarme y darme fortaleza, y a mi familia por su apoyo y comprensión, porque siempre me impulsaron a seguir adelante y terminar con este proyecto de investigación.

Índice

Resumen

En la actualidad existen muchos métodos de predicción que se utilizan en el campo de la salud; no obstante, los estudios actuales requiere mayor estudio e indagación en la reactividad de resultados de las pruebas del virus de la inmunodeficiencia humana (VIH) utilizando modelos predictivos basados en Machine Learning, sin embargo, si se lograra predecir la reactividad positiva de la prueba de VIH de una persona con anterioridad, se podría realizar las acciones oportunas con antelación para poder brindarle una atención oportuna a la persona, además así poder asignar la preparación del tratamiento antirretroviral (TARV) y esquema de tratamiento a utilizar adecuado. Por tal motivo, este estudio propone un enfoque predictivo para la reactividad a la prueba del virus de la inmunodeficiencia humana (VIH) basado en algoritmos de Machine Learning. Los datos utilizados proceden de los registros del personal de salud que pertenecen a una brigada móvil y que realizan funciones de tamizaje de VIH a través del Aplicativo Móvil de Tamizaje ITS (App VIH) del Ministerio de Salud del Perú (Minsa). La metodología utilizada consistió en evaluar cuatro modelos de machine learning con los algoritmos “Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier y Extreme Gradient Boosting”, con la intención de comparar sus resultados y elegir al mejor modelo que presente el mejor desempeño, para luego ser utilizado a través de una interfaz gráfica, que permita determinar si una persona posee Reactividad Positiva o Negativa al resultado de la Prueba de VIH. Los resultados demostraron que el mejor modelo de machine learning para el estudio fue Extra Trees Classifier con un Accuracy Score de 98.19% en comparación con Decision Tree Classifier con el menor Accuracy Score de 98.03%; calculado a través de la precisión promedio usando el puntaje de validación cruzada con 10 interacciones de los datos de entrenamiento; esto indica que el modelo predice con una precisión del 98% la prueba de VIH. Por tal motivo, este modelo podrá ser usado en la toma de decisiones relacionadas con las personas que puedan contraer el VIH y así brindarles un tratamiento antirretroviral (TARV) adecuado, además de la misma forma, apoyar a las labores realizadas por las brigadas móviles urbanas en temas relacionados con el VIH.

1. Introducción

A través de la historia de la sociedad humana, los seres humanos han sido afectados por múltiples enfermedades, estas impactan negativamente en la vida de las personas como el caso del Virus de la Inmunodeficiencia Humana (VIH) [1], el cual daña el sistema inmunitario al destruir un tipo de glóbulo blanco que ayuda al cuerpo a combatir las infecciones; por tal motivo, para combatir esta enfermedad el Organismo Mundial de la Salud (OMS) [2] creo estrategias mundiales del sector de la salud para 2022-2030 contra el VIH, con el objetivo de poner fin al Sida, estableciendo "Global health sector strategies on, respectively, HIV, viral hepatitis and sexually transmitted infections for the period 2022-2030" [3]; la cual describió varias estrategias: Prestar servicios de alta calidad, basados en la evidencia y centrados en las personas, Optimizar los sistemas, los sectores y las alianzas para lograr impacto, Generar y utilizar datos para orientar la toma de decisiones encaminadas a la acción, Implicar a la sociedad civil y las comunidades empoderadas y Fomentar la innovación para lograr impacto. De la misma forma, en el caso de Perú la organización que combate esta enfermedad es el Ministerio de Salud (MINSA) [4] a través de la generación de normativas para enfrentar al VIH "Norma Técnica de Salud de Atención Integral del Adulto con Infección por el Virus de la Inmunodeficiencia Humana (VIH)" [5], la cual es ayudada por otras organizaciones para brindar apoyo a las personas con VIH como el caso de ICAP de la Universidad de Columbia que trabaja con AID FOR AIDS (AFA) y los Centros para el Control y la Prevención de Enfermedades (CDC), que brindan atención en VIH a refugiados venezolanos en Colombia y Perú [6]; o el caso de Socios en Salud, el cual desarrollo el Proyecto País TB-VIH 2022-2025, que asegura el acceso a los servicios integrales de salud para poner fin a la tuberculosis y VIH al 2025 [7]. Esto demuestra que el VIH al ser un problema mundial, este requiere una acción oportuna para lograr prevenir la infección y su propagación, para lograr esto se puede reducir el riesgo de contagiarse o transmitir el VIH al realizarse alguna de las siguientes acciones: "Hacer la prueba del VIH, Practicar conductas sexuales menos riesgosas, Usar condones, Limitar el número de parejas sexuales, Recibir tratamiento para enfermedades de

transmisión sexual, Averiguar profilaxis preexposición (PrEP) y No inyectar drogas" [8]. En este contexto, entre las opciones mencionadas, la que destaca es la de realizarse una prueba de VIH [9], la cual consta en analizar una muestra de sangre para ver si ha sido infectado con VIH, ya que el saber si una persona tiene VIH le da información importante para que se mantenga sano: "Si la prueba da positivo, puede tomar medicamentos para tratar el virus, el tratamiento antirretroviral (TARV) [10] para el VIH reduce la cantidad de VIH en la sangre (carga viral); en caso contrario, Si la prueba le da negativo, puede tomar medidas para prevenir la infección por el VIH" [11]. Por tal motivo, en la actualidad existen muchos métodos de predicción que se utilizan en el campo de la salud [12]; no obstante, los estudios actuales no han profundizado en la reactividad de las personas que requieran realizar la prueba del virus de la inmunodeficiencia humana (VIH) utilizando modelos predictivos basados en Machine Learning [13] a través de los algoritmos "Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier y Extreme Gradient Boosting" [14]; ya que estos algoritmos son capaces de poder encontrar relaciones complejas no lineales en los datos utilizados para la generación de cada modelo [15]; logrando de esta forma reconocer con anticipación a una posible persona con altas probabilidades de tener una reactividad positiva a la prueba de VIH, permitiendo a los responsables poder tomas las acciones oportunas en relación con el seguimiento, toma de la prueba de VIH y campañas como el del Día Nacional de la Prueba de VIH [16]. Por tales motivos, el objetivo de la presente investigación es la de proponer un enfoque predictivo para la reactividad a la prueba del virus de la inmunodeficiencia humana (VIH) basado en algoritmos de Machine Learning, para apoyar la labor de las Brigadas Móviles Urbanas (BMU) que realicen actividades de consejería en VIH y tamizaje de VIH [17] utilizando el Aplicativo Móvil de Tamizaje ITS (App VIH) [18] para el registro de información, pudiendo así pronosticar la reactividad a la prueba del VIH y de esta forma beneficiar a las personas, porque podrán recibir una atención adecuada mientras esperan los resultados oficiales de la reactividad de la prueba de VIH [19] y gracias a esto poder asignar la preparación del tratamiento antirretroviral (TARV) y esquema de tratamiento a utilizar adecuado si es positivo [20].

2. Trabajos Relacionados

En la literatura, se encontró varios estudios relacionados con el Virus de la Inmunodeficiencia Humana (VIH) y el uso del Machine Learning (ML); estos fueron agrupados en tres grupos: "El primer grupo se centra en las investigaciones de modelos predictivos de Machine Learning realizadas en la ciudad de Lima; El segundo grupo se centra en las investigaciones de reactividad a la prueba del VIH realizadas en el Perú; El tercer grupo se centra en las investigaciones de modelos predictivos basados en Machine Learning relacionadas con el VIH y que fueran realizadas internacionalmente".

Iniciando con el primer grupo, este se centra en las investigaciones de modelos predictivos de Machine Learning realizadas en la ciudad de Lima. Aquí, se destacan tres trabajos investigativos, como el trabajo de modelamiento predictivo para deserción de la vacunación de VPH mediante algoritmos de Machine Learning [21]; Asimismo, otro trabajo es el modelamiento predictivo visual del éxito académico de estudiantes universitarios mediante algoritmos de Machine Learning y un enfoque de Análisis Visual-Predictivo de Datos [22]; Por último, el trabajo de modelamiento predictivo para la detección de mujeres maltratadas físicamente en el ámbito peruano mediante algoritmos de Machine Learning [23].

Continuando con el segundo grupo, este se centra en las investigaciones de VIH realizadas en el Perú. Aquí, se destacan cinco trabajos investigativos, como el trabajo de la conducta sexual y realización de la prueba del virus de la inmunodeficiencia humana en jóvenes universitarios [24]; Asimismo, otro trabajo es la experiencia peruana sobre el flujograma de diagnóstico del virus de inmunodeficiencia humana a través de los resultados de las pruebas de VIH [25]; Continuando, otro trabajo es la evaluación de tres marcas de pruebas rápidas frente a muestras de sangre para la detección de anticuerpos contra VIH [26]; Siguiendo, otro trabajo es la asociación entre el nivel educativo y conocimiento sobre la transmisión de VIH/Sida en mujeres adolescentes con asociación al índice de riqueza [27]; Por último, el trabajo de los desafíos en la continuidad de atención de personas viviendo con VIH en el Perú durante la pandemia de la COVID-19 y lograr prevenir de este modo su impacto en su salud mental [28].

Por último, con el tercer grupo, este se centra en las investigaciones de modelos predictivos basados en Machine Learning relacionadas con el VIH y que fueran realizadas internacionalmente. Aquí, se destacan siete trabajos investigativos, como el trabajo del papel de Machine Learning en la predicción del riesgo de VIH [29], que dio como resultado que el Machine Learning aplicado a la predicción del riesgo de VIH tiene el potencial de contribuir a poner fin al VIH; Asimismo, otro trabajo es la estimación del riesgo de VIH basada en el Machine Learning utilizando índices de tasa de incidencia [30], que dio como resultado un método de diagnóstico para identificar a los pacientes con riesgo de contraer VIH a partir de su historial clínico; Continuando con trabajo es la predicción del comportamiento relacionado con el VIH en adolescentes mediante el Machine Learning [31], que dio como resultado el poder identificar a los adolescentes que probablemente se involucren en comportamientos de riesgo de contraer el VIH; El siguiente trabajo es la predicción de la falta de respuesta de la intervención en adolescentes para la prevención del VIH mediante el Machine Learning [32], que dio como resultado la capacidad de identificar a los adolescentes que probablemente no responderán a una intervención; Prosiguiendo con el trabajo de predicción del estado serológico respecto al VIH en función de las características socio conductuales en África Oriental y Meridional [33], que dio como resultado el identificar la positividad del VIH en personas con alto riesgo de infección; Siguiendo con el trabado de uso de técnicas de Machine Learning para identificar predictores de VIH para detección en África subsahariana [34], que dio como resultado el identificar los predictores de VIH y predecir a las personas con alto riesgo de infección VIH; Por último, el trabajo del uso de una herramienta para evaluar el comportamiento del riesgo asociado con contraer el VIH en tres sitios en Sudáfrica [35], que dio como resultado el identificar el tratamiento adecuado utilizando datos de encuestas digitales para los recursos de salud.

3. Metodología

El proceso metodológico utilizado en la presente investigación fue inspirado en una metodología propuesta por mí, utilizada para la creación de modelos predictivos basados en Machine Learning aplicados al área de la salud [21], el cual fue adaptado en seis fases: "Fase 1: Obtención de Datos, Fase 2: Exploración de Datos, Fase 3: Preparación de Datos, Fase 4: Desarrollo de Modelados, Fase 5: Evaluación de Modelos y Fase 6: Despliegue del Modelo"; cada fase de la metodología posee diferentes actividades que son imprescindibles para la presente investigación. La metodología se muestra en la **Figura 1**.

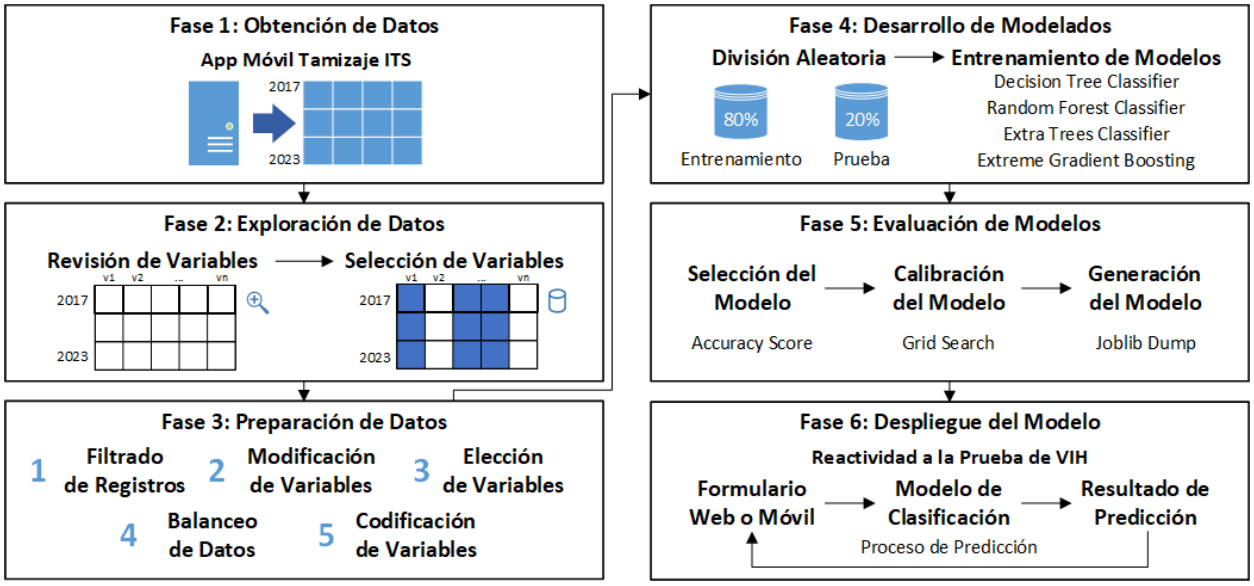


Figura 1. Proceso metodológico del modelo predictivo.

3.1. Fase 1: Obtención de Datos

Esta fase consistió en la obtención de los registros de las personas que se realizan una prueba de VIH y son registrados en el App Móvil de Tamizaje ITS (App VIH), pertenecientes al Ministerio de Salud del Perú (MINSA) y que sus variables estén validadas por personal experto en el área de la salud. Los registros fueron obtenidos en un archivo Excel, los cuales fueron generados por el Ministerio de Salud para ser utilizarlos en el presente estudio. Estos registros obtenidos contienen los datos de los registros originales de los tamizajes ITS realizadas a través del App Móvil de Tamizaje ITS entre los años del 2017 al 2023, con un total, 309273 registros encontrados en el archivo obtenido.

3.2. Fase 2: Exploración de Datos

3.2.1. Revisión de Variables

Esta subfase consistió en reconocer y analizar aquellas variables obtenidas de los registros originales de los tamizajes ITS (App VIH); dando como resultado un total de 36 variables, que son agrupadas en 14 grupos según sus características, para que de esta forma se pueda revisar todas las variables obtenidas, las cuales se detallan en la **Tabla 1**.

Tabla 1. Identificación de variables agrupadas por descripción.

Nro.	Descripción
1-5	Ubicación de la Brigada
6	Establecimiento de Salud
7-8	Datos de la Brigada
9-11	Fecha de Abordaje
12-13	Lugar de Abordaje
14-15	Datos del Brigadista
16-21	Datos del Paciente
22-23	Consentimiento y Consejería
24	Tipo de Tamizaje
25-26	Entrega de Preservativo
27-29	Reactividad a las Pruebas
30-31	Servicio de Pruebas
32-34	Tratamiento Antirretroviral
35-36	Migración al HISMINSA

3.2.2. Selección de Variables

Esta subfase consistió en la selección inicial de variables reconocidas e identificadas como pertinentes para la creación del modelo; en el que únicamente se eligieron 18 variables, para que de esta forma únicamente sean utilizadas las variables útiles para la preparación de datos, las cuales se detallan en la **Tabla 2**.

Tabla 2. Selección de las variables utilizadas en el estudio.

Variable	Tipo	Descripción
REGION	Cualitativa nominal	Región de la Brigada
PROVINCIA	Cualitativa nominal	Provincia de la Brigada
DISTRITO	Cualitativa nominal	Distrito de la Brigada

TIPO BRIGADA	Cualitativa nominal	Tipo de Brigada
FECHA ABORDADO	Cuantitativa continua	Fecha de Abordado
TIPO LUGAR ABORDAJE	Cualitativa nominal	Tipo Lugar Abordaje
PACIENTE TIPO DOCUMENTO	Cualitativa nominal	Paciente Tipo Doc.
PACIENTE EDAD	Cuantitativa discreta	Paciente Edad
PACIENTE SEXO	Cualitativa nominal	Paciente Sexo
PACIENTE TIPO POBLACION	Cualitativa nominal	Paciente Tipo Pob.
PACIENTE NACIONALIDAD	Cualitativa nominal	Paciente Nacionalidad
CONSENTIMIENTO INFORMADO	Cualitativa nominal	Consentimiento Inf.
TIPO TAMIZAJE	Cualitativa nominal	Tipo de Tamizaje
NUMERO CONDONES	Cuantitativa discreta	Número Condones
NUMERO LUBRICANTES	Cuantitativa discreta	Número Lubricantes
ES REACTIVO PRIMERA PRUEBA	Cualitativa nominal	Res. Primera Prueba
TIENE SEGUNDA PRUEBA	Cualitativa nominal	Tiene Segunda Prueba
ES REACTIVO SEGUNDA PRUEBA	Cualitativa nominal	Res. Segunda Prueba

3.3. Fase 3: Preparación de Datos

3.3.1. Filtrado de Registros

Esta subfase consistió en filtrar los registros por "Ser de Lima o Callao, ser mayor de edad y tamizaje de VIH"; además de eliminar los nulos en Tipo Población; por último se elimina la variable Tipo Tamizaje por solo ser utilizada para filtrar; reduciendo la cantidad de registros desde 309273, los cuales se detallan en la **Tabla 3**.

Tabla 3. Filtros realizados con los datos originales.

Variable	Filtro	Registros
TAMIZAJE ITS (ORIGINAL)	Todos	309273
REGION	Lima o Callao	113722
PACIENTE EDAD	Entre 18 a 60 años	110255
TIPO TAMIZAJE	VIH	79449
PACIENTE TIPO POBLACION	Valores No Nulos	75076

3.3.2. Modificación de Variables

Esta subfase consistió en generar nuevas variables en base de las existentes, realizar el escalado numérico de características (ENC) e ingeniería de características categóricas (ICC), para poder de esta forma mejorar el procesamiento del modelo; dando como resultado 6 variables nuevas, 1 variable ENC y 6 variables ICC, las cuales se detallan en la **Tabla 4**.

Tabla 4. Listado de la modificación de variables.

Original	Cambio	Cualitativa
REGION, PROVINCIA y DISTRITO	Nuevo - ESTRATO SOCIAL	Ordinal
REGION, PROVINCIA y DISTRITO	Nuevo - AREA ZONA	Nominal
TIPO BRIGADA	ICC - TIPO BRIGADA	Nominal
FECHA ABORDADO	Nuevo - MES ABORDADO	Ordinal
TIPO LUGAR ABORDAJE	ICC - TIPO LUGAR ABORDAJE	Nominal
PACIENTE TIPO DOCUMENTO	ICC - PACIENTE TIPO DOCUMENTO	Nominal
PACIENTE EDAD	ENC - PACIENTE GRUPO EDAD	Ordinal
PACIENTE SEXO	ICC - PACIENTE SEXO	Nominal
PACIENTE TIPO POBLACION	ICC - PACIENTE TIPO POBLACION	Nominal
PACIENTE NACIONALIDAD	Nuevo - TIPO NACIONALIDAD	Nominal
CONSENTIMIENTO INFORMADO	ICC - CONSENTIMIENTO INFORMADO	Nominal
NUMERO CONDONES y NUMERO LUBRICANTES	Nuevo - USA PRESERVATIVO	Nominal
ES REACTIVO PRIMERA PRUEBA, TIENE SEGUNDA PRUEBA y ES REACTIVO SEGUNDA PRUEBA	Nuevo - REACTIVIDAD	Nominal

3.3.3. Elección de Variables

Esta subfase consistió en la elección de las variables que obtuvieron un Valor P menor al alfa de 5% en la Prueba de Chi Cuadrado “chi2”; lo que indica que hay una asociación estadísticamente significativa entre las variables de estudio y la variable Reactividad, quedando 7 variables de todas las demás, las cuales se detallan en la **Tabla 5**.

Tabla 5. Resultado de la prueba de Chi Cuadrado en relación con la Reactividad.

Variable	Valor P	Selección
AREA ZONA	0.972891	No
PACIENTE SEXO	0.485906	No
PACIENTE GRUPO EDAD	0.379955	No
USA PRESERVATIVO	0.224902	No
CONSENTIMIENTO INFORMADO	0.069672	No
TIPO NACIONALIDAD	0.028902	Sí
TIPO LUGAR ABORDAJE	0.017659	Sí
PACIENTE TIPO DOCUMENTO	0.004377	Sí
PACIENTE TIPO POBLACION	0.000000	Sí
MES ABORDADO	0.000000	Sí
ESTRATO SOCIAL	0.000000	Sí
TIPO BRIGADA	0.000000	Sí

3.3.4. Balanceo de Datos

Esta subfase consistió en balancear los datos a través de la aplicación de la técnica de balanceo de datos “NearMiss”, la cual disminuye aleatoriamente el número de valores de la clase mayoritaria, lo que permitió reparar la brecha original de desigualdad de resultados de reactividad, reduciendo los registros de 75076 a 5466, los cuales se detallan en la **Tabla 6**.

Tabla 6. Resultado del proceso de la ejecución de NearMiss.

Variable	Original	NearMiss
REACTIVIDAD (Positiva)	2733	2733
REACTIVIDAD (Negativa)	72343	2733

3.3.5. Codificación de Variables

Esta subfase consistió en codificar las variables a través de una transformación en las opciones de respuesta de las variables categóricas “get_dummies”, dividiendo cada variable en otras nuevas según la cantidad de tipos de valores que tengan, en relación con el número de valores menos uno, las cuales se detallan en la **Tabla 7**.

Tabla 7. Resultado de la codificación de variables.

Variable	Valores	Nuevas Variables
ESTRATO SOCIAL	5	ES_01, ES_02, ES_03, ES_04
TIPO BRIGADA	2	TB_01
MES ABORDADO	12	MA_01, MA_02, MA_03, MA_04, MA_05, MA_06, MA_07, MA_08, MA_09, MA_10, MA_11
TIPO LUGAR ABORDAJE	12	LA_01, LA_02, LA_03, LA_04, LA_05, LA_06, LA_07, LA_08, LA_09, LA_10, LA_11
PACIENTE TIPO DOCUMENTO	5	TD_01, TD_02, TD_03, TD_04
PACIENTE TIPO POBLACION	8	TP_01, TP_02, TP_03, TP_04, TP_05, TP_06, TP_07
TIPO NACIONALIDAD	2	TN_01
REACTIVIDAD	2	RR_01

3.4. Fase 4: Desarrollo de Modelados

3.4.1. División Aleatorio

Esta subfase consistió en la distribución de los registros obtenidos del balanceo de los datos, dando como resultado, dos grupos aleatorios “train_test_split”, el primero que consta del 80% de los registros para entrenamiento y el segundo que consta del 20% de los registros para pruebas, las cuales se detallan en la **Tabla 8**.

Tabla 8. Resultado de la división aleatoria de los registros.

División	Porcentaje	Cantidad
Entrenamiento	80%	4372
Prueba	20%	1094

3.4.2. Entrenamiento de Modelos

Esta subfase consistió en el diseño de cuatro modelos de clasificación, los cuales son capaces de encontrar relaciones complejas no lineales, dividiendo el conjunto de tal de registros en dos partes; el 80% para entrenamiento y el 20% para prueba, los cuales se detallan en la **Tabla 9**.

Tabla 9. Detalle de los modelos de clasificación utilizados.

Modelo	Código	Tipo de Algoritmo
Decision Tree Classifier	DTS	Aprendizaje Supervisado para Clasificación
Random Forest Classifier	RFS	Aprendizaje Supervisado para Clasificación
Extra Trees Classifier	ETS	Aprendizaje Supervisado para Clasificación
Extreme Gradient Boosting	XGB	Aprendizaje Supervisado para Clasificación

4. Resultados de la Metodología

4.1. Fase 5: Evaluación de Modelos

4.1.1. Selección del Modelo

Esta subfase consistió en ver cuál de los modelos anteriormente entrenados tiene el mejor puntaje de precisión “accuracy”; usando el puntaje de validación cruzada “cross_val_score” con 10 interacciones de los datos de entrenamiento; se configuró con un n_splits = 10 y random state = 0, los cuales se detallan en la **Tabla 10**, el cual será posteriormente calibrado, los cuales se detallan en la **Tabla 11**. Además de mostrar la comparación de los algoritmos a través de gráficos de cajas, los cuales se ilustran en la **Figura 2**.

Tabla 10. Ejecución de cada interacción por cada modelo.

Ejecución	Decision Tree Classifier	Random Forest Classifier	Extra Trees Classifier	Extreme Gradient Boosting
1	0.981735	0.981735	0.981735	0.984018
2	0.984018	0.984018	0.984018	0.984018
3	0.977117	0.977117	0.979405	0.977117
4	0.981693	0.983982	0.983982	0.981693
5	0.974828	0.974828	0.977117	0.974828
6	0.972540	0.974828	0.974828	0.974828
7	0.983982	0.983982	0.983982	0.981693
8	0.983982	0.986270	0.986270	0.986270
9	0.983982	0.986270	0.988558	0.983982
10	0.979405	0.979405	0.979405	0.979405
Promedio	**0.980328**	**0.981244**	**0.981930**	**0.980785**

Tabla 11. Resultados de los modelos de clasificación.

Modelo	Accuracy Score	Selección
Decision Tree Classifier	0.980328	No
Random Forest Classifier	0.981244	No
Extra Trees Classifier	0.981930	Sí
Extreme Gradient Boosting	0.980785	No

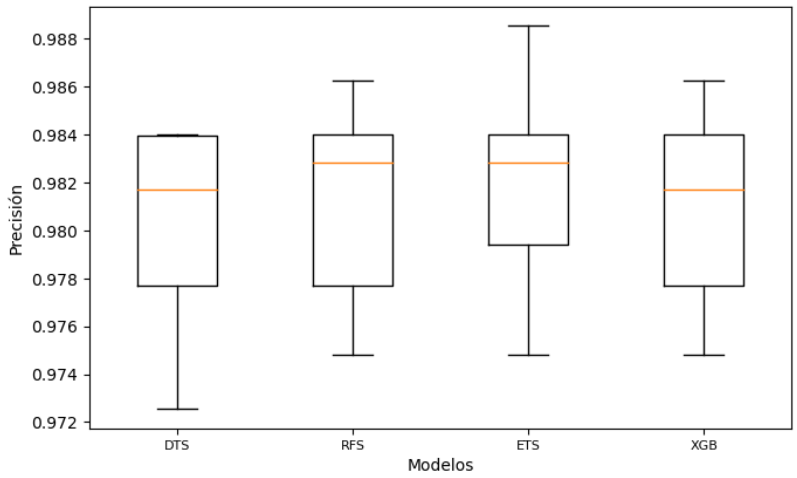


Figura 2. Comparación de modelos de Machine Learning.

4.1.2. Calibración del Modelo

Esta subfase consistió en la calibración del modelo seleccionado a través de Grid Search "GridSearchCV"; se configuró con un cv=10 y n_jobs=-1, que identifico cuáles son los mejores parámetros para optimizar el modelo, para luego rediseñarlo y dar como resultado la generación de un modelo óptimo, los cuales se detallan en la **Tabla 12**.

Tabla 12. Modelo de Machine Learning calibrado.

Variable	Valor
max_depth	20
max_features	sqrt
n_estimators	19
random_state	0

4.1.3. Generación del Modelo

Esta subfase consistió en verificar el Accuracy Score, Precision, Recall, F1-score y Area Under the Curve (AUC) del modelo para luego guardarlo con joblib, los cuales se detallan en la **Tabla 13**. Además de mostrar la Curva ROC en un gráfico, el cual se ilustra en la **Figura 3**.

Tabla 13. Resultados del modelo de Extra Trees Classifier.

Extra Trees Classifier	Valor	precision	recall	f1-score
Accuracy Score: 0.976234	No	0.99	0.96	0.98
AUC: 0.976143	Sí	0.96	0.99	0.98

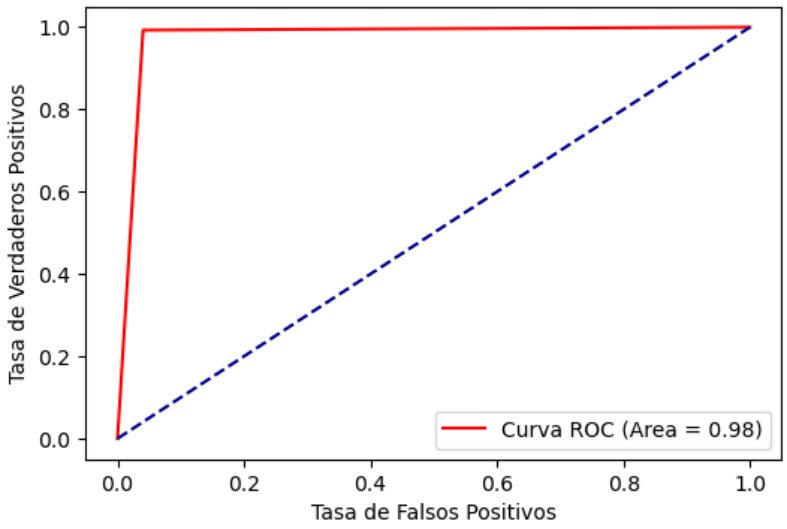


Figura 3. Curva ROC del modelo de Extra Trees Classifier.

El modelo Extra Trees Classifier obtuvo los siguientes resultados: Accuracy Score de 97.62%, el cual muestra un nivel muy alto de predicciones que el modelo realizó correctamente; Precision de 96%, el cual muestra un nivel muy alto en proporción a los positivos que son verdaderos positivos; Recall de 99%, el cual muestra un nivel muy alto en proporción de positivos reales correctamente clasificados; F1-score de 98%, el cual muestra un nivel muy alto en precisión y exhaustividad; Curve ROC muestra un nivel muy alto de verdaderos positivos contra los falsos positivos; y Area Under the Curve (AUC) de 97.61%, el cual muestra un nivel muy alto en qué tan bien se clasifican las predicciones.

4.2. Fase 6: Despliegue del Modelo

El despliegue de la aplicación de Reactividad a la Prueba de VIH fue elaborado en la plataforma web gratuita de Heroku. El proceso inicia con la visualización del formulario web, luego se debe ingresar los valores de las variables: Tipo Brigada, Lugar Abordaje, Mes Abordado, Tipo Nacionalidad, Tipo Documento, Tipo Población y Estrato Social, para posteriormente presionar el botón Predecir, el cual realiza el proceso de predicción de Reactividad a la Prueba de VIH y devuelve el resultado obtenido al formulario web, como se muestra en la **Figura 4**.

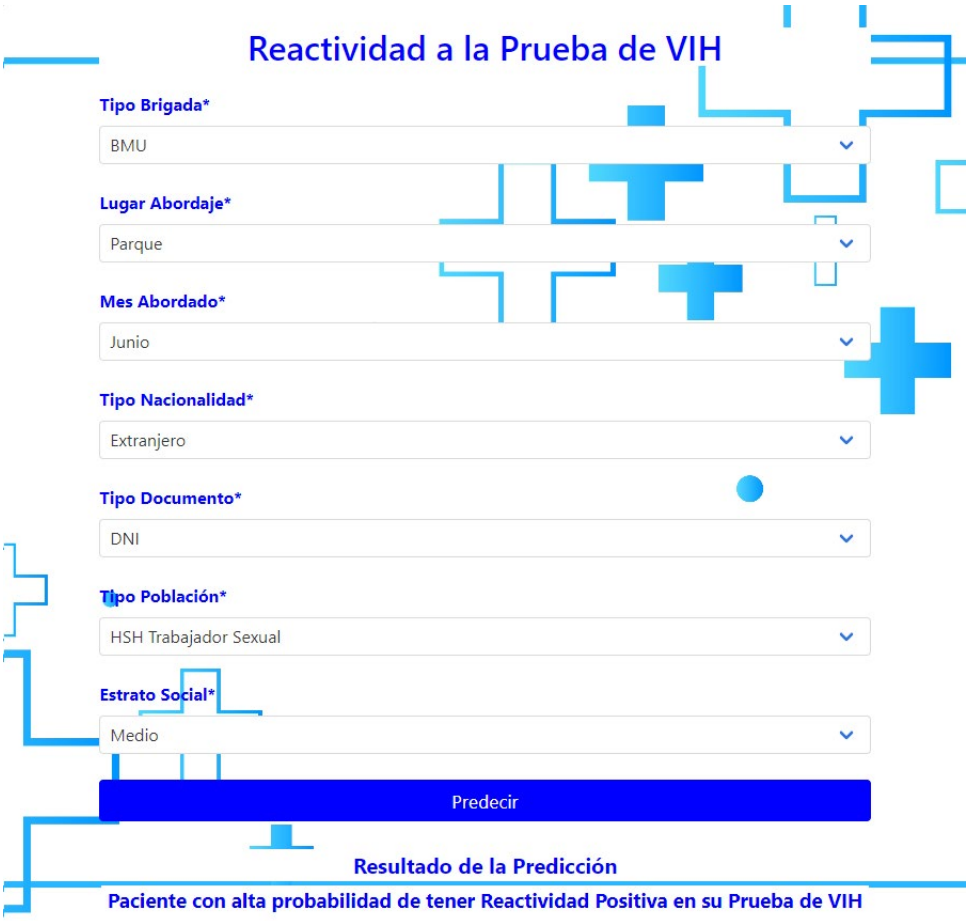


Figura 4. Curva ROC del modelo de Extra Trees Classifier.

4.3. Configuración

Para la realización del Modelo Predictivo de Machine Learning se utilizaron varias herramientas de desarrollo, de las cuales se destaca Anaconda Navigator, Python y otros programas especificados en la **Tabla 14**. Por otro lado, se utilizaron varias librerias para lograr la creación del modelo predictivo y la generación del formulario de consumo del modelo elegido; de las cuales destacan pandas, numpy, sklearn, imblearn, joblib y flask, que son detalladas en la **Tabla 15**.

Tabla 14. Programas utilizados en el modelado.

Herramienta	Versión
Anaconda Navigator	2.1.4
Anaconda Prompt	2.1.4
Spyder ID	5.1.5
Python	3.9.12

Tabla 15. Librerías utilizadas en el modelado.

Desde	Importa	Versión
pandas	pandas	1.4.2
numpy	numpy	1.21.5
matplotlib	pyplot	3.5.1
IPython.display	display	8.2.0
sklearn.preprocessing	LabelEncoder	1.0.2
sklearn.feature_selection	chi2	1.0.2
sklearn.model_selection	train_test_split	1.0.2
sklearn.tree	DecisionTreeClassifier	1.0.2
sklearn.ensemble	RandomForestClassifier	1.0.2
sklearn.ensemble	ExtraTreesClassifier	1.0.2
xgboost	XGBClassifier	1.7.3
sklearn.model_selection	GridSearchCV	1.0.2
sklearn.model_selection	StratifiedKFold	1.0.2
sklearn.model_selection	cross_val_score	1.0.2
imblearn.under_sampling	NearMiss	0.10.1
sklearn.metrics	accuracy_score	1.0.2
sklearn.metrics	confusion_matrix	1.0.2
sklearn.metrics	classification_report	1.0.2
sklearn.metrics	roc_curve	1.0.2
sklearn.metrics	roc_auc_score	1.0.2
joblib	dump	1.1.0
joblib	load	1.1.0
flask	Flask	1.1.2
flask	jsonify	1.1.2
flask	request	1.1.2
flask	render_template	1.1.2

5. Discusión

Existen algunos estudios que abordan temas relacionados con la prevención o tratamiento del VIH utilizando técnicas de Machine Learning [29] [30] [31] [32] [33] [34] [35], en los que se observa el gran impacto que ha tendido el Machine Learning en la predicción del riesgo de VIH y que estas pueden ayudar a eliminar el VIH [29] [30]. También se ha visto que se ha podido determinar que existen diversos factores que afecta a las personas con el riesgo de contraer VIH en su juventud [31] [32]. Por último, se ha comprobado en este estudio, que las variables socioculturales "Tipo Población o Estrato Social" son factores relevantes en la predicción a la reactividad del resultado de la prueba del VIH, como también se ha corroborado en numerosos estudios como [33] [34] [35].

En la **Tabla 13** se observa que el algoritmo Extra Trees Classifier proporciona una alta tasa en Accuracy Score 97.62%, AUC 97.61%, Precision 96%, Recall 99% y F1-score 98%. Además, es importante considerar que para la selección de variables del estudio se eligieron las que obtuvieran un Valor P menor al alfa de 5% en la Prueba de Chi Cuadrado "chi2"; Lo que indicaba que hay una asociación estadísticamente significativa entre las variables de estudio, que en su mayoría son variables sociales, y la variable Reactividad, visualizado en la Tabla 5. En este sentido, solo se conservaron para el estudio Tipo nacionalidad, Tipo Lugar Abordaje, Paciente Tipo Documento, Paciente Tipo Población, Mes Abordado, Estrato Social, Tipo Brigada y Reactividad.

En consecuencia, se identificó que las variables que tienen mayor importancia para el presente estudio son Paciente Tipo de Población, Mes Abordado, Estrato Social y Tipo Brigada, las cuales en su mayoría son sociales, al obtener un valor P de 0, al identificar que si existe concordancia perfecta entre las frecuencias observadas y las esperadas. Sin estas variables, el estudio no podría concretarse, ya que son importantes para el modelo. Este hallazgo permite utilizar el presente estudio para ser utilizado en diferentes ciudades del Perú, aparte de las utilizadas en este que son Lima y Callao, pudiendo de esta forma expandir su uso a nivel nacional.

Cabe mencionar que este estudio tiene algunas limitaciones. Es la primera vez que estos datos se exploran en un trabajo científico, por lo que no existen resultados previos en la literatura referente al uso de estos datos. Por otro lado, se hubiera deseado contar con más variables para el presente estudio como variables socioculturales, socio demográficas o socio conductual. El análisis y el enfoque debería ser más robusta con la inclusión de estas variables. En este sentido, el alcance sería más esclarecedor y los hallazgos podrían ayudar a mejorar la toma de decisiones en problemas relacionados con la reactividad de los resultados a la prueba de VIH.

6. Conclusiones

Los resultados obtenidos permiten afirmar que las técnicas de Machine Learning seleccionadas presentaron una capacidad predictiva eficiente en relación con la reactividad a la prueba de VIH, utilizando cuatro modelos diferentes para realizar las pruebas. Sin embargo, entre todos estos, se recomienda el modelo creado con el algoritmo Extra Trees Classifier, porque esta muestra mejores indicadores de rendimiento, tanto en Exactitud como en Precisión y Sensibilidad. Identificando los factores clave para predecir la reactividad del resultado de la prueba de VIH a través del Tipo de Brigada, Estrato Social, Mes de Abordado y Paciente Tipo Población, las cuales son determinantes, ya que dieron un valor de 0 en el estadístico chi-cuadrado, demostrando que, si existe concordancia perfecta entre las frecuencias observadas y la esperada, es decir, si se desea tener una buena ejecución en las actividades de las brigadas móviles urbanas se debe tener en consideración los mejores indicadores de rendimiento anteriormente mencionados y así ir a lugares donde posiblemente encuentren personas reactivamente positivas al VIH.

Existen estudios que han demostrado la existencia de factores externos a los utilizados en el presente estudio por los brigadistas que realizan pruebas de VIH, que influyen de manera importante en los predictores del VIH [33] [34] [35]. En este sentido, otra mirada a la contribución de este estudio corresponde al hecho de introducir factores adicionales, como tipo de brigada y tipo de población, al análisis. De esta forma, para trabajos futuros, también se podrá incorporar información no incluida como datos socio conductuales, datos demográficos o socioeconómicos. Además, la incorporación de estos factores mejorará el mecanismo de predicción de la reactividad al resultado de las pruebas de VIH, brindando información valiosa para desarrollar estrategias y actividades para la realización de las actividades de las brigadistas. Finalmente, el enfoque predictivo propuesto puede ser considerado como un apoyo para la toma de decisiones para identificar como realizar una campaña efectiva de realización de pruebas de VIH, para que de esta forma apoyar a las labores realizadas por las brigadas móviles urbanas en temas relacionados con el VIH.

Referencias

[1] MedlinePlus, “What is HIV?,” *MedlinePlus - National Library of Medicine*, 2020. https://medlineplus.gov/hiv.html (accessed Mar. 27, 2023).

[2] WHO, “HIV and AIDS,” *World Health Organization (WHO)*, 2023. https://www.who.int/news-room/fact-sheets/detail/hiv-aids (accessed Mar. 27, 2023).

[3] WHO, “Global health sector strategies on, respectively, HIV, viral hepatitis and sexually transmitted infections for the period 2022-2030,” *World Health Organization (WHO)*, 2022. https://www.who.int/publications/i/item/9789240053779 (accessed Mar. 27, 2023).

[4] MINSA, “¿Qué es el VIH?,” *Ministerio de Salud del Perú (MINSA) - Estado Peruano*, 2023. https://www.gob.pe/16439-que-es-el-vih (accessed Mar. 27, 2023).

[5] MINSA, “Norma Técnica de Salud de Atención Integral del Adulto con Infección por el Virus de la Inmunodeficiencia Humana (VIH),” *Ministerio de Salud del Perú (MINSA) - Estado Peruano*, 2020. https://www.gob.pe/institucion/minsa/normas-legales/1422592-1024-2020-minsa3 (accessed Mar. 27, 2023).

[6] ICAP, “ICAP in Peru,” *ICAP - Columbia Mailman School of Public Health*, 2021. https://icap.columbia.edu/where-we-work/peru/ (accessed May 15, 2023).

[7] SociosEnSalud, “Socios En Salud inicia nueva intervención para poner fin a la tuberculosis y VIH al 2025,” *Socios En Salud*, 2022. https://sociosensalud.org.pe/nueva-intervencion-para-poner-fin-a-la-tuberculosis-y-vih-al-2025/ (accessed May 15, 2023).

[8] HIVinfo, “The Basics of HIV Prevention,” *HIVinfo in collaboration with NIH’s Office of AIDS Research*, 2021. https://hivinfo.nih.gov/understanding-hiv/fact-sheets/basics-hiv-prevention (accessed Mar. 28, 2023).

[9] MedlinePlus, “HIV Screening Test,” *MedlinePlus - National Library of Medicine*, 2020. https://medlineplus.gov/lab-tests/hiv-screening-test/ (accessed Mar. 28, 2023).

[10] MedlinePlus, “HIV Medicines,” *MedlinePlus - National Library of Medicine*, 2020. https://medlineplus.gov/hivmedicines.html (accessed Mar. 28, 2023).

[11] CDC, “Getting Tested,” *Centers for Disease Control and Prevention (CDC)*, 2022. https://www.cdc.gov/hiv/basics/hiv-testing/getting-tested.html (accessed Mar. 28, 2023).

[12] J. M. Pineda, “Modelos predictivos en salud basados en aprendizaje de maquina (machine learning),” *Rev. Médica Clínica Las Condes*, vol. 33, no. 6, pp. 583–590, Nov. 2022, doi: 10.1016/j.rmclc.2022.11.002.

[13] IBM, “What is machine learning?,” *International Business Machines Corporation (IBM)*, 2023. https://www.ibm.com/topics/machine-learning (accessed Mar. 28, 2023).

[14] Microsoft, “Machine learning algorithms,” *Microsoft*, 2023. https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-machine-learning-algorithms (accessed May 16, 2023).

[15] Aurélien Géron, *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*, Second ed. United States of America: O'Reilly Media Inc, 2019.

[16] MINSA, "Día nacional de la prueba de vih: Minsa realiza campaña descentralizada de orientación, prevención y tamizaje del vih," *Ministerio de Salud del Perú (MINSA) - Estado Peruano*, 2022. https://www.gob.pe/institucion/minsa/noticias/620293-dia-nacional-de-la-prueba-de-vih-minsa-realiza-campana-descentralizada-de-orientacion-prevencion-y-tamizaje-del-vih (accessed May 17, 2023).

[17] MINSA, "Minsa: Brigadas Móviles Urbanas informaron a más de 25 000 personas sobre la viruela del mono y VIH," *Ministerio de Salud del Perú (MINSA) - Estado Peruano*, 2022. https://www.gob.pe/institucion/minsa/noticias/659202-minsa-brigadas-moviles-urbanas-informaron-a-mas-de-25-000-personas-sobre-la-viruela-del-mono-y-vih (accessed May 17, 2023).

[18] MINSA, "Manual de Usuario Historia Clínica Electrónica Atención Primaria," *Ministerio de Salud del Perú (MINSA) - Estado Peruano*, 2019. https://vih.minsa.gob.pe/ (accessed Mar. 28, 2023).

[19] CDC, "Types of HIV Tests," *Centers for Disease Control and Prevention (CDC)*, 2022. https://www.cdc.gov/hiv/basics/hiv-testing/test-types.html (accessed May 18, 2023).

[20] HIVinfo, "HIV Treatment: The Basics," *HIVinfo in collaboration with NIH's Office of AIDS Research*, 2021. https://hivinfo.nih.gov/understanding-hiv/fact-sheets/hiv-treatment-basics (accessed Mar. 28, 2023).

[21] U. Marroquin, N. Saboya, and A. A. Sullon, "Machine Learning-based predictive model for the prognosis of human papillomavirus (HPV) vaccination attrition," in *2021 4th International Conference on Robot Systems and Applications*, Apr. 2021, pp. 44–49, doi: 10.1145/3467691.3467695.

[22] D. Orrego Granados, J. Ugalde, R. Salas, R. Torres, and J. L. López-Gonzales, "Visual-Predictive Data Analysis Approach for the Academic Performance of Students from a Peruvian University," *Appl. Sci.*, vol. 12, no. 21, p. 11251, Nov. 2022, doi: 10.3390/app122111251.

[23] N. Saboya, A. A. Sullon, and O. L. Loaiza, "Predictive Model Based on Machine Learning for the Detection of Physically Mistreated Women in the Peruvian Scope," in *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, Oct. 2019, pp. 18–23, doi: 10.1145/3369114.3369143.

[24] M. P. Bermúdez, M. T. Ramiro, I. Teva, T. Ramiro-Sánchez, and G. Buela-Casal, "Sexual behaviour and human immunodeficiency virus testing in university students from Cuzco (Peru)," *Gac. Sanit.*, vol. 32, no. 3, pp. 223–229, May 2018, doi: 10.1016/j.gaceta.2017.07.002.

[25] E. F. Miranda Ulloa, S. E. Romero Ruiz, M. Acuña Barrios, R. Briceño Espinoza, G. Obregon Boltan, and D. V. Suárez Agüero, "Peruvian experience on the human immunodeficiency virus diagnostic flowchart," *Rev. la Fac. Med. Humana*, vol. 22, no. 2, pp. 428–430, Mar. 2022, doi: 10.25176/RFMH.v22i2.4401.

[26] E. F. Miranda Ulloa, R. Briceño Espinoza, S. Romero Ruiz, and F. Cárdenas Bustmante, "Evaluation Of Three Brands Of Rapid Tests Against Blood Samples For The Detection Of Antibodies Against HIV," *Rev. la Fac. Med. Humana*, vol. 21, no. 3, pp. 677–679, Jun. 2021, doi: 10.25176/RFMH.v21i3.3941.

[27] N. Amado Cornejo and C. Luna-Muñoz, “Association between educational level and knowledge on transmission of hiv/aids in adolescent women in Peru-ENDES 2019,” *Rev. la Fac. Med. Humana*, vol. 21, no. 4, pp. 804–810, Sep. 2021, doi: 10.25176/RFMH.v21i4.4266.

[28] J. L. Paredes, R. Navarro, D. M. Cabrera, M. M. Diaz, F. Mejia, and C. F. Caceres, “Challenges to the continuity of care of people living with HIV throughout the COVID-19 crisis in Peru,” *Rev. Peru. Med. Exp. Salud Publica*, vol. 38, no. 1, pp. 166–70, Mar. 2021, doi: 10.17843/rpmesp.2021.381.6471.

[29] J. Fieggen, E. Smith, L. Arora, and B. Segal, “The role of machine learning in HIV risk prediction,” *Front. Reprod. Heal.*, vol. 4, Dec. 2022, doi: 10.3389/frph.2022.1062387.

[30] O. Haas, A. Maier, and E. Rothgang, “Machine Learning-Based HIV Risk Estimation Using Incidence Rate Ratios,” *Front. Reprod. Heal.*, vol. 3, Dec. 2021, doi: 10.3389/frph.2021.756405.

[31] B. Wang *et al.*, “Adolescent HIV-related behavioural prediction using machine learning: a foundation for precision HIV prevention,” *AIDS*, vol. 35, no. Supplement 1, pp. S75–S84, May 2021, doi: 10.1097/QAD.0000000000002867.

[32] B. Wang *et al.*, “Predicting Adolescent Intervention Non-responsiveness for Precision HIV Prevention Using Machine Learning,” *AIDS Behav.*, Oct. 2022, doi: 10.1007/s10461-022-03874-4.

[33] E. Orel *et al.*, “Prediction of HIV status based on socio-behavioural characteristics in East and Southern Africa,” *PLoS One*, vol. 17, no. 3, p. e0264429, Mar. 2022, doi: 10.1371/journal.pone.0264429.

[34] C. K. Mutai, P. E. McSharry, I. Ngaruye, and E. Musabanganji, “Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa,” *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 159, Dec. 2021, doi: 10.1186/s12874-021-01346-2.

[35] M. Majam *et al.*, “Utility of a machine-guided tool for assessing risk behaviour associated with contracting HIV in three sites in South Africa,” *Informatics Med. Unlocked*, vol. 37, p. 101192, 2023, doi: 10.1016/j.imu.2023.101192.

Anexo

```
#Se importa el excel 'Tamizajes ITS' en la variable 'data'
data = pd.read_excel('Tamizajes ITS.xlsx',sheet_name='Hoja1', na_values=missing_values)

#Se realiza el filtro de la variable 'REGION': Lima o Callao
data_filtro_x_region = data_selecc
data_filtro_x_region = data_filtro_x_region[(data_filtro_x_region['REGION'] == 'LIMA') |
(data_filtro_x_region['REGION'] == 'CALLAO')]

#Asignación de Estratos Sociales a los Distritos de LIMA BARRANCA
data_estrato_social['ESTRATO SOCIAL'] = np.where(np.logical_and(data_estrato_social['PROVINCIA'] ==
'BARRANCA', data_estrato_social['DISTRITO'] == 'BARRANCA'), 3,
np.where(np.logical_and(data_estrato_social['PROVINCIA'] == 'BARRANCA',
data_estrato_social['DISTRITO'] == 'PARAMONGA'), 3,
np.where(np.logical_and(data_estrato_social['PROVINCIA'] == 'BARRANCA',
data_estrato_social['DISTRITO'] == 'PATIVILCA'), 4,
np.where(np.logical_and(data_estrato_social['PROVINCIA'] == 'BARRANCA',
data_estrato_social['DISTRITO'] == 'SUPE'), 4,
np.where(np.logical_and(data_estrato_social['PROVINCIA'] == 'BARRANCA',
data_estrato_social['DISTRITO'] == 'SUPE PUERTO'), 4, data_estrato_social['ESTRATO SOCIAL'])))))

#Se crea la variable 'PACIENTE GRUPO EDAD' con los valores de la variable 'PACIENTE EDAD'
data_genera_v_grupo_edad = data_genera_t_tipo_documento
data_genera_v_grupo_edad['PACIENTE GRUPO EDAD'] =
np.where(np.logical_and(data_genera_v_grupo_edad['PACIENTE EDAD'] >= 18,
data_genera_v_grupo_edad['PACIENTE EDAD'] <= 29), 1,
np.where(np.logical_and(data_genera_v_grupo_edad['PACIENTE EDAD'] >= 30,
data_genera_v_grupo_edad['PACIENTE EDAD'] <= 60), 2,
data_genera_v_grupo_edad['PACIENTE EDAD']))
data_genera_v_grupo_edad['PACIENTE GRUPO EDAD'].value_counts()

#Prueba de Chi-Cuadrado
data_chi_scores = chi2(data_eleccion_x,data_eleccion_y)
data_p_values = pd.Series(data_chi_scores[1],index = data_eleccion_x.columns)
data_p_values.sort_values(ascending = False , inplace = True)
data_p_values.plot.bar()

#Codificación de características categóricas: Reactividad
pd.value_counts(data_dummy['REACTIVIDAD'], sort = True)
data_dummy['REACTIVIDAD CAT'] = np.where(data_dummy['REACTIVIDAD'] == 1, 'RR_01',
np.where(data_dummy['REACTIVIDAD'] == 2, 'RR_02', 'RR_00'))
pd.value_counts(data_dummy['REACTIVIDAD CAT'], sort = True)
data_dummy_rr = pd.get_dummies(data_dummy['REACTIVIDAD CAT']).iloc[:,:-1]
```

```
#Proceso de validación cruzada con 10 interacciones de los datos de entrenamiento
for name, model in models:
        kfold = StratifiedKFold(n_splits=10, shuffle=True, random_state=0)
        cv_results = cross_val_score(model, data_x_train, data_y_train, cv=kfold, scoring='accuracy')
        results.append(cv_results)
        names.append(name)
        valores.append(msg)

#Se crea la variable Grid_Search que almacena el modelo de calibración
Grid_Search = GridSearchCV(ExtraTreesClassifier(), Params_Modelo, cv=10, n_jobs=-1)

#Se escoge el mejor modelo predictivo
ETS_Best_Params = ExtraTreesClassifier(n_estimators=19, max_depth=20, max_features='sqrt',
random_state=0)

#Se entrena el modelo calibrado con los datos de entrenamiento
ETS_Best_Params.fit(data_x_train, data_y_train)

#Se realiza las predicciones con los datos del array de prueba 'data_x_test'
ETS_Calibrado_Predic = ETS_Best_Params.predict(data_x_test)

#Prueba de uso del modelo predictivo
dataXnewValues = [['ES_01', 'ES_02', 'ES_03', 'ES_04', 'TB_01', 'MA_01', 'MA_02',
                   'MA_03', 'MA_04', 'MA_05', 'MA_06', 'MA_07', 'MA_08', 'MA_09',
                   'MA_10', 'MA_11', 'LA_01', 'LA_02', 'LA_03', 'LA_04', 'LA_05',
                   'LA_06', 'LA_07', 'LA_08', 'LA_09', 'LA_10', 'LA_11', 'TD_01',
                   'TD_02', 'TD_03', 'TD_04', 'TP_01', 'TP_02', 'TP_03', 'TP_04',
                   'TP_05', 'TP_06', 'TP_07', 'TN_01'],
                   [0, 0, 1, 0, 1, 0, 0,
                    0, 0, 0, 1, 0, 0, 0,
                    0, 0, 0, 0, 0, 0, 0,
                    0, 0, 1, 0, 0, 0, 0,
                    0, 1, 0, 0, 1, 0, 0,
                    0, 0, 0, 1]]

dataXnewColumns = dataXnewValues.pop(0)
dataXnewDf = pd.DataFrame(dataXnewValues, columns=dataXnewColumns)

#Se crea la variable 'Ynew' que recibe el resultado de la predicción del modelo
Ynew = ETS_Best_Params.predict(dataXnewDf)

#Se lista el resultado de la Predicción del Módelo
print("X=%s, Predicción=%s" % (Xnew[0], Ynew[0]))

#Generación de modelo predictive para usarlo luego
dump(ETS_Best_Params, 'ReactividadVIH.joblib')
```