UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Abordaje de predicción de hipertensión en pacientes peruanos basado en algoritmos de Machine Learning

Tesis para obtener el Título Profesional de Ingeniero de Sistemas

Autor:

Gabriela Maria Auqui Aguilar

Jack Cosme Castillo Ramos

Luis Felipe Humberto Moran Nureña

Asesor:

PhD. Javier Linkolk Lopez Gonzales

Lima, setiembre de 2025

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Javier Linkolk Lopez Gonzales, docente de la Facultad de Ingeniería y Arquitectura, Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: "ABORDAJE DE PREDICCIÓN DE HIPERTENSIÓN EN PACIENTES PERUANOS BASADO EN ALGORITMOS DE MACHINE LEARNING" de los autores Gabriela Maria Auqui Aguilar, Jack Cosme Castillo Ramos, Luis Felipe Humberto Moran Nureña 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 04 días del mes de diciembre del año 2025.

PhD. Javier Linkolk Lopez Gonzales

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Ñaña, Villa Unión, a 18 día(s) del mes de Setiembre del año 2025 siendo las 09:00 horas, se reunieron los miembros del jurado en la Universidad Peruana Unión Campus Lima, bajo la dirección del (de la) presidente(a): Mg. Immer Elias Cuellar Rodriguez, el (la) secretario(a): Mg. Nemias Saboya Rios, y los demás miembros: Mg. Ferdinan Edgardo Pineda Anco, Dr. Juan Jesus Soria Quijaite y el (la) asesor(a) Ph. D. Javier LinKolk López Gonzales con el propósito de administrar el acto académico de sustentación de la tesis titulado: "Abordaje de predicción de hipertensión en pacientes peruanos basado en algoritmos de machine learning" del(los) bachiller/es: a) Luis Felipe Humberto Moran Nureña, b) Gabriela Maria Auqui Aguilar, c) Jack Cosme Castilla Ramos, conducente a la obtención del título profesional de: Ingeniero de Sistemas (Denominación del Título Profesional)

El Presidente inició el acto académico de sustentación invitando al (a la) / a (los) (las) candidato(a)/s hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del jurado a efectuar las preguntas, y aclaraciones pertinentes, las cuales fueron absueltas por al (a la) / a (los) (las) candidato(a)/s. Luego, se produjo un receso para las deliberaciones y la emisión del dictamen del jurado.

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

Bachiller (a): Luis Felipe Humberto Moran Nureña

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	14	C	Aceptable	Bueno

Bachiller (b): Gabriela Maria Auqui Aguilar

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	14	C	Aceptable	Bueno

Bachiller (c): Jack Cosme Castillo Ramos

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	14	C	Aceptable	Bueno

(*) Ver parte posterior

Finalmente, el Presidente del jurado invitó al (a la) / a (los) (las) candidato(a)/s a ponerse de pie, para recibir la evaluación final y concluir el acto académico de sustentación procediéndose a registrar las firmas respectivas.

Presidente/a

Secretario/a

Asesor/a

Miembro

Miembro

Bachiller (a)

Bachiller (b)

Bachiller (c)

Esta sustentación fue realizada de manera virtual u online sincrónica conforme al Reglamento General de Grados y Títulos.

INDICE

CONTENIDO

Abordaje de predicción de hipertensión en pacientes peruanos basado en algoritmos de Machine Learning

Hypertension prediction approach in Peruvian patients based on machine learning algorithms

RESUMEN

La hipertensión es un problema de salud pública con alta prevalencia en Perú, afectando significativamente la calidad de vida y la sostenibilidad del sistema de salud. Los métodos tradicionales de análisis de datos han demostrado ser insuficientes para procesar el gran volumen de información médica disponible. En este estudio, se exploró el uso de algoritmos de machine learning para predecir la hipertensión en pacientes peruanos, basándose en 1631 registros extraídos de historias clínicas electrónicas. Se implementaron tres modelos predictivos: Regresión Logística, Random Forest y Support Vector Machine (SVM). Los resultados evidencian la capacidad de los algoritmos para identificar patrones y factores de riesgo asociados a la hipertensión, lo que permite una detección temprana y una mejor planificación de estrategias preventivas. La integración de machine learning en el ámbito sanitario peruano representa una oportunidad para optimizar la toma de decisiones médicas y fomentar una cultura de prevención en la población. El modelo Random Forest mostró el mejor rendimiento, con una precisión (accuracy) de 92.05%, un ROC-AUC score de 0.9484, y un F1-Score de 0.85 para la clase de hipertensos. En comparación, el modelo SVM presentó una precisión de 79.20%, un ROC-AUC score de 0.8162, y un F1-Score de 0.57. Por otro lado, el modelo de Regresión Logística mostró un rendimiento inferior, con una precisión de 71.25% y un F1-Score de 0.04, destacándose por su menor capacidad para identificar correctamente los casos de hipertensión

ABSTRACT

Hypertension is a public health problem with high prevalence in Peru, significantly affecting quality of life and the sustainability of the healthcare system. Traditional data analysis methods have proven insufficient for processing the large volume of medical information available. In this study, we explored the use of machine learning algorithms to predict hypertension in Peruvian patients, based on 1,631 records extracted from electronic medical records. Three predictive models were implemented: Logistic Regression, Random Forest, and Support Vector Machine

(SVM). The results demonstrate the algorithms' ability to identify patterns and risk factors associated with hypertension, enabling early detection and better planning of preventive strategies. The integration of machine learning into the Peruvian healthcare system represents an opportunity to optimize medical decision-making and foster a culture of prevention among the population. The Random Forest model showed the best performance, with an accuracy of 92.05%, a ROC-AUC score of 0.9484, and an F1-Score of 0.85 for the hypertensive class. In comparison, the SVM model had an accuracy of 79.20%, a ROC-AUC score of 0.8162, and an F1-Score of 0.57. On the other hand, the Logistic Regression model showed inferior performance, with an accuracy of 71.25% and an F1-Score of 0.04, standing out for its lower ability to correctly identify cases of hypertension.

Palabras clave: Hipertensión, Machine Learning, Inteligencia Artificial, Historias clínicas, Salud Pública.

Keywords: Hypertension, Machine Learning, Artificial Intelligence, Medical Records, Public Health.

1. INTRODUCCIÓN

La sostenibilidad de los sistemas nacionales de salud y la economía de diversos países es afectada por la situación de enfermedades crónicas no transmisibles [1], [2], [3], [4] entre este grupo de afecciones se encuentra la hipertensión, considerada como desafío para la salud global [5], [6], [7] responsable de 8.5 millones de muertes a nivel mundial [8]. En Sudamérica [9], [10], [11] estudios reportan como primer causante de muerte a las enfermedades cardiovasculares (ECV), seguidas por el cáncer y las enfermedades respiratorias, asimismo, se evidencia que la hipertensión es el primer indicador para la aparición de ECV [12]. La organización mundial de salud en el año 2023 reporto que la prevalecía de hipertensión arterial en Perú excede el 20% [13], asimismo, en nuestro país, el sistema sanitario aplica tratamientos farmacológicos que consisten en dosis de medicamentos antihipertensivos, junto con tratamientos no farmacológicos como la restricción de alimentos, alcohol, tabaco, control regular de peso y la actividad física constante, etc. [14], [15], [16]. Por otro lado, el procesamiento de datos de esta enfermedad, solo se ha desarrollado estudios estadísticos, considerando que los métodos tradicionales de procesamiento son ineficientes para el gran volumen de datos que se producen a diario en Perú [17], [18].

La comunidad científica dedicada al estudio de la hipertensión, menciona que la aplicación de Machine Learning, brindara oportunidades de tratamiento y métodos de prevención cada vez más eficaces [19], [20], [21], manifestando ser una herramienta valiosa para la atención sanitaria y el manejo de diferentes

condiciones clínicas [18], [22], [23]. Para predecir o evaluar la hipertensión se han aplicado dos enfoques: clasificación basada en datos clínicos e imágenes [24] y características extraídas de señales de fotopletismografía (PPG) [25], además algunos estudios se basan en el tiempo de tránsito de pulso [26] o en la velocidad de la onda del pulso [27], también utilizan la relación de intensidad de PPG [28] y las características particulares en el dominio de tiempo [29]. En esta investigación, los estudios han sido agrupados de la siguiente forma: El primer grupo consiste en investigaciones desarrolladas sobre imágenes recopiladas. En este contexto, el estudio realizado en China predice la presión arterial desde videos faciales y diagnostico facial de la medicina tradicional del país, aplicando el modelo Lightweight (L-VGG16) con un coeficiente de determinación R^2\ del 79% [30]. Por su parte, el segundo grupo se caracteriza por ser de tipo texto, aquí destacan estudios, como el que se propuso en Estados Unidos [31], donde utilizaron 6 modelos predictivos para detectar enfermedades de las arterias coronarias, sobresaliendo la métrica Accuracy con un 93% para el modelo de red neuronal. Asimismo, el tercer grupo cumple la característica de tener señales de fotopletismografía y señales de presión arterial de forma invasivas y no invasivas. En este escenario, se destacan estudios, como el que fue desarrollado en Canadá [32] aplicando la fotopletismografía y presión arterial no invasivas en conjunto con Machine learning, para clasificar los estados de presión arterial tales como la Normotensión, Hipertensión e Hipotensión con un Accuracy del 70% sobresaliendo el modelo Support vector machine (SVM), a su vez estima la presión arterial media, sistólica, diastólica en pacientes enfermos crónicos.

En el contexto peruano el uso de la inteligencia artificial ha sido poco explorado en el ámbito sanitario. Por lo tanto, se ha determinado el uso de modelos predictivos para identificar los factores que aumentan la prevalencia de hipertensión mediante el análisis de historias clínicas electrónicas, con la finalidad de detectar y prevenir esta enfermedad a tiempo, así fomentar una cultura de prevención en la población peruana. En este estudio, se ha implementado tres algoritmos de machine learning tales como: Regresión logística, Random forest y Support vector machine (SVM) para modelar los datos extraídos de un hospital peruano y desarrollar la construcción, el diseño y el entrenamiento de los registros electrónicos.

2. METODOLOGÍA

En esta investigación se implementó la metodología Cross-Industry Standard Process for Data Mining (CRISP-DM) para un análisis detallado y la obtención de información importante para la predicción de hipertensión en pacientes peruanos. El objetivo principal de CRISP-DM es guiar a través de un proceso estructurado que incluye todas las etapas de un proyecto de análisis de datos [33], [34].

La metodología CRISP-DM tiene las siguientes fases (a) comprensión del negocio, (b) comprensión de datos, (c) preparación de datos, (d) modelado, (e) evaluación y (f) selección e interpretación. En las siguientes subsecciones desarrollamos cada fase.

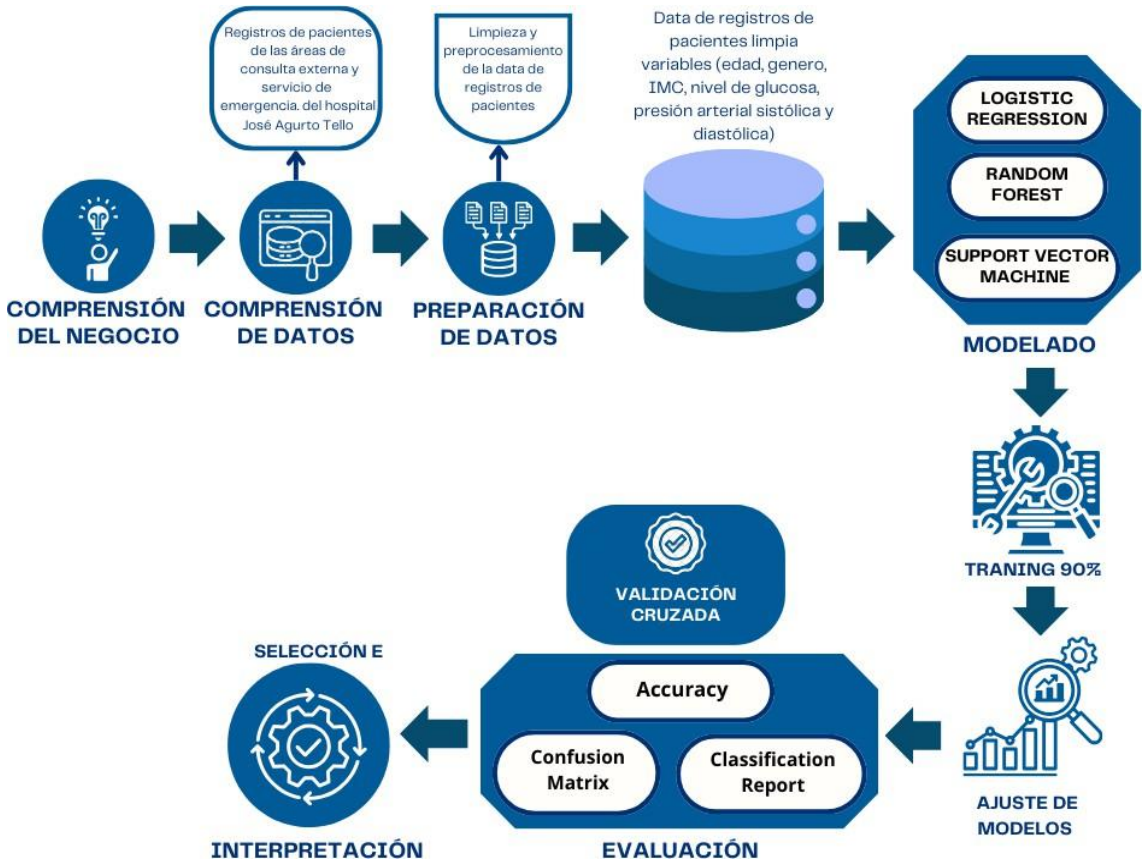


Figura 1 Metodología CRISP-DM aplicada a la investigación

2.1. COMPRENSIÓN DEL NEGOCIO

En esta primera etapa, describimos la situación actual de la hipertensión en Perú. El objetivo principal del estudio es desarrollar un modelo predictivo para evaluar el riesgo de hipertensión en pacientes peruanos utilizando modelos predictivos de machine learning.

La Organización mundial de la salud ha reportado que los países con ingresos altos invierten el 80% de los gastos e inversión en Salud a diferencia de los países subdesarrollados [35], dentro de estos países, se ha verificado que Perú en el periodo 2013-2022 invirtió el 12.7% de su presupuesto público demostrando que tiene uno de los más bajos índices de gasto en salud a nivel mundial [36], así mismo en un informe presupuestal reciente reporto que solo un 50,5% es para tratar enfermedades no transmisibles y el 47,3% es dirigido para el control y tratamiento de personas ya diagnosticadas con hipertensión [37]. Por otro lado, en el 2022 el

16,2% de personas de 15 años y/o mayores presento presión arterial alta, donde los mayores índices fueron registrados en la Provincia Constitucional del Callao (21,3%), Lima Metropolitana (21,1%), y el departamento de Lima (20,2%) [38].

2.2. COMPRENSIÓN DE DATOS

El conjunto de datos de pacientes peruanos incluye registro de atención, historial médico y resultados de pruebas de laboratorio, extraídos desde el 01 de enero hasta el 31 de diciembre del 2023, los datos ascendieron a 195 463 y fueron recopilados mediante consultas SQL. Además, se emplearon la Guía para el manejo de la hipertensión arterial [39] para identificar indicadores como el Índice de Masa Corporal (IMC), los niveles de glucosa y colesterol, así como la presión arterial diastólica, con el fin de realizar un análisis y procesamiento de los resultados estadísticos. (1631).

Tabla 1. Características de las Variables de Estudio

Variable	Definición	Medición
Edad	Edad de paciente	Numérica
Sexo	Genero de paciente	Nominal
Peso	Peso del paciente	Numérica
Índice de Masa Corporal (IMC)	Medida del IMC	Numérica
Talla	Talla del paciente	Numérica
Nivel de Glucosa	Nivel de Glucosa en la sangre	Numérica
Nivel de Colesterol	Nivel de Colesterol en la sangre	Numérica
Presión Arterial Sistólica	Medida de la presión	Numérica
Presión Arterial Diastólica	Medida de la presión	Numérica
Hipertensión	Hipertensión en pacientes	Nominal

2.3. PREPARACIÓN DE DATOS

Para esta etapa que precede a la fase de modelado, en cuanto a la recopilación de datos se realizó consultas SQL para obtener los registros de los pacientes del Sistema Historia clínica Electrónica de Consultorios externos. Los datos recolectados se exportaron a un archivo de Microsoft Excel, denominado “DataFinal”, que contiene tanto la variable dependiente (hipertensión) como las variables independientes.

Para el preprocesamiento de los datos, el entrenamiento con los algoritmos elegidos y el análisis de las métricas seleccionadas, se realizó en la plataforma Google Colaboratory. Para estos fines, ha sido necesario utilizar la librería Pandas para el manejo y análisis de datos, en primer lugar, se cambió los valores de sexo al

tipo entero, se eliminó también la columna paciente, adicionalmente se creó la columna IMC haciendo uso de las columnas Peso sobre Talla, por último, se verifico los valores faltantes por columna para finalmente obtener una correcta depuración de los registros y los datos.

2.4. MODELADO

Actualmente, los algoritmos de aprendizaje automático se utilizan para llevar a cabo tareas como clasificación, regresión, agrupamiento o reducción de dimensionalidad en grandes volúmenes de datos [40], [41]. Para analizar las restricciones de entrada y las variables de salida, se emplean distintas herramientas de aprendizaje automático que permiten identificar y analizar las variables de salida necesarias. Este proceso incluye el uso de varios modelos de aprendizaje automático [42], [43].

2.4.1. Regresión logística

La regresión logística es un modelo robusto distinguido por sus tareas de clasificación, aplica una función logística para modelar la probabilidad de que ocurra un evento en función de las variables de entrada, lo que la convierte en una herramienta valiosa en múltiples campos [44], un estudio logro evaluar el riesgo de cáncer de mama mediante la aplicación de regresión logística obteniendo en AUC (área bajo la curva receiver operating characteristic) de 0,669 [45].

Es una función exponencial:

$$f_\beta(x_i) \triangleq Pr(y_i = 1 | x_i ; \beta) = (1 + exp(-[1x_i^T]\beta))^{-1} \quad (1)$$

Para i=1, …, N donde $y_i \epsilon \{0,1\}$ representa la variable de respuesta binaria observada i-esima mutuamente independiente asociada al vector de covariables i-esimo x$i \in R^{P+1}$ indica los parametros de regresion que se van a estimar[46].

2.4.2. Random Forest

Random Forest es un enfoque de aprendizaje automático que se constituye por ser un conjunto de árboles de decisión [47], cada árbol de Random forest se entrena de manera independiente y en paralelo utilizando un subconjunto de datos. Este algoritmo forma parte de los métodos basados en reglas y es altamente eficaz tanto para tareas de clasificación como de regresión [48], [49], [50]. Random Forest

construye N árboles de decisión y combina sus predicciones. La predicción final se obtiene de la siguiente manera:

Para clasificación: Se usa el voto mayoritario de los árboles individuales:

$$\hat{y} = mode\{T_1(x), T_2(x), \dots, T_N(x)\} \quad (2)$$

Donde $T_i(x)$ es la predicción del i-ésimo árbol para la entrada X [51].

Para regresión: Se calcula el promedio de las predicciones de todos los árboles:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \quad (3)$$

Donde $T_i(x)$ representa la predicción del i-ésimo árbol [52].

2.4.3. Support vector machine

Es un método de aprendizaje automático que puede manejar grandes volúmenes de datos de entrenamiento sin limitaciones significativas. Asimismo, desafíos asociados con conjuntos de datos pequeños, problemas de no linealidad y tareas de reconocimiento [53].

Hiperplano de separación

El hiperplano en un espacio de dimensión d se define como:

$$w^T x + b = 0 \quad (4)$$

Donde:

$x \in \mathbb{R}^n$: vector de características (por ejemplo, presión, caudal, etc.).

$w \in \mathbb{R}^n$: vector normal al hiperplano, que determina su orientación.

$b \in \mathbb{R}$: término independiente o sesgo, que determina la posición del hiperplano respecto al origen.

$w^T x$: producto escalar entre los vectores w y x .

Margen de Separación

El objetivo principal de una Máquina de Vectores de Soporte (SVM) es encontrar el hiperplano óptimo que maximice el margen de separación entre las dos clases. Este margen se define como la distancia entre los hiperplanos que pasan por los vectores de soporte más cercanos de cada clase. Maximizar dicho margen mejora la capacidad del modelo para generalizar ante nuevos datos.

Maximizar el margen equivale a minimizar la norma del vector w , ya que el margen geométrico está inversamente relacionado con esta norma. Por conveniencia matemática, se suele minimizar $\frac{1}{1}||w||^2$.

$$\min_{w,b} \frac{1}{2}|w|^2 \text{ sujeto a: } y_i(w^T x_i + b) \geq 1, \forall i \quad (5)$$

Donde:

$x_i \in \mathbb{R}^n$: vector de características del ejemplo i .

$y_i \in \{-1, +1\}$: etiqueta de clase correspondiente.

w : es el vector normal al hiperplano.

b : es el término independiente o sesgo.

$||w||$: norma euclídea del vector w .

La restricción $y_i(w^T x_i + b) \geq 1$ asegura que todos los puntos estén correctamente clasificados y fuera del margen.

Función de Optimización (Hard Margin SVM)

En el caso ideal en el que los datos son linealmente separables, la SVM busca encontrar el hiperplano de separación óptimo que maximiza el margen sin permitir errores de clasificación. Este caso se conoce como SVM de margen duro (Hard Margin SVM).

El problema de optimización se formula como:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad (6)$$

Sujeto a las restricciones:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i = 1,2,\dots,N \quad (7)$$

Donde:

$x_i \in \mathbb{R}^n$: es el vector de características del i-ésimo ejemplo de entrenamiento.

$y_i \in \{-1, +1\}$: es la etiqueta de clase correspondiente.

w : vector de pesos o coeficientes del modelo.

b : sesgo o término independiente.

N : número total de observaciones en el conjunto de entrenamiento.

SVM con margen Suave

En escenarios del mundo real, los datos suelen no ser perfectamente separables de forma lineal. Para abordar esta situación, se introduce una versión más flexible del modelo conocido como SVM con margen suave (Soft Margin SVM).

Este enfoque permite que algunos puntos violen las restricciones del margen mediante la inclusión de variables de holgura $\xi_i \geq 0$, que mide el grado de error de clasificación de cada ejemplo.

Además, se incorpora un parámetro de penalización $C > 0$, que controla el equilibrio entre maximizar el margen y minimizar el error de clasificación.

$$\min_{w,b} \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \xi_i \qquad (8)$$

Sujeto a:

$$y_i(w^T x_1 + b) \geq 1 - \xi_i, \ \forall i \quad \xi_i \geq 0, \ \forall i \qquad (9)$$

Donde:

ξ_i : variable de holgura que permite errores en la clasificación del punto i .

C : parámetro de penalización que determina cuánto se penalizan los errores (valores grandes de C fuerzan una menor tolerancia al error).

El resto de los símbolos (w, x_i, y_i, b) conservan su significado anterior.

SVM con Núcleos (Kernel Trick)

Cuando los datos no son linealmente separables, se usa una función de transformación $\phi(x)$ y un Kernel K (x_i, x_j) para trabajar en un espacio de mayor dimensión:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (10)$$

Algunas funciones de kernel comunes son:

Kernel Lineal: K (x_i, x_j) = $x_i^T x_j$ (11)

Kernel Polinomico: K (x_i, x_j) = $(x_i^T x_j + c)^d$ (12)

Kernel RBF (Radial Basis Function): K (x_i, x_j) $= \exp(-\gamma \|x_i - x_j\|^2)$ (13)

2.5. EVALUACIÓN

Esta investigación utilizo validación cruzada y las siguientes métricas para evaluar el rendimiento y diferentes aspectos de la precisión en los modelos de Machine learning.

2.5.1. Precisión

La precisión (Accuracy, ACC) mide la proporción de predicciones correctas realizadas por el modelo en comparación con el total de observaciones. Su valor oscila entre 0 y 1, donde 1 representa una predicción perfecta y 0 una predicción completamente errónea. Se calcula mediante la ecuación [54]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Donde:

TP (True Positives): Casos correctamente clasificados como hipertensos.

TN (True Negatives): Casos correctamente clasificados como no hipertensos.

FP (False Positives): Casos que fueron incorrectamente clasificados como hipertensos.

FN (False Negatives): Casos de hipertensión que fueron clasificados erróneamente como sanos.

Sin embargo, en problemas médicos como la detección de hipertensión, donde la clase positiva suele estar desbalanceada, esta métrica puede ser engañosa, ya que un modelo que prediga la mayoría de los casos como negativos puede lograr una precisión alta, pero sin detectar correctamente los pacientes hipertensos [55].

2.5.2. Sensibilidad

La sensibilidad, también llamada tasa de verdaderos positivos (True Positive Rate - TPR), mide la capacidad del modelo para identificar correctamente los casos positivos (hipertensos). Su cálculo se expresa como [54]:

$$Recall = TPR = \frac{TP}{TP+FN} \quad (15)$$

Donde:

TP (True Positives): número de positivos correctamente clasificados.

FN (False Negatives): número de positivos incorrectamente clasificados como negativos.

Un valor de sensibilidad cercano a 1 indica que el modelo identifica la mayoría de los pacientes con hipertensión, reduciendo el número de falsos negativos (FN). En la detección de enfermedades como la hipertensión, la sensibilidad es una métrica crucial, ya que un FN alto puede llevar a la falta de tratamiento adecuado y aumentar el riesgo de complicaciones de salud [55].

2.5.3. Precisión

La precisión indica cuántas de las predicciones positivas realizadas por el modelo son realmente correctas. Se calcula como [54]:

$$precision = \frac{TP}{TP+FP} \quad (16)$$

Una precisión alta indica que el modelo realiza predicciones positivas con un bajo índice de falsos positivos. Esto es relevante en la predicción de hipertensión, pues una precisión baja significaría un alto número de falsos positivos, lo que podría

generar preocupaciones innecesarias en pacientes sanos y aumentar costos por tratamientos innecesarios [56].

2.5.4. Especificidad

La especificidad o tasa de verdaderos negativos (True Negative Rate - TNR) mide la capacidad del modelo para identificar correctamente a los pacientes sanos y se expresa como:

$$especificidad = \frac{TN}{TN+FP} \quad (17)$$

Un valor de especificidad alto indica que el modelo tiene una baja tasa de falsos positivos. Esto es importante porque un modelo con baja especificidad podría etiquetar erróneamente a personas sanas como hipertensas, generando preocupaciones innecesarias e incidiendo en costos adicionales por evaluaciones médicas innecesarias [55].

2.5.5. Medida

La F_1-score combina la precisión y la sensibilidad en una única métrica mediante su promedio armónico, de la siguiente manera [54]:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (18)$$

La F_1-score es útil en problemas con clases desbalanceadas, como la detección de hipertensión, donde el modelo podría tentarse a clasificar casi siempre como negativo para obtener una precisión artificialmente alta. Un F_1-score más alto indica un mejor balance entre precisión y sensibilidad, lo cual es deseable en este estudio [55], [56].

2.5.6. Área Bajo la Curva

El Área Bajo la Curva de la Característica Operativa del Receptor (AUC-ROC) mide la capacidad del modelo para distinguir entre clases, evaluando su desempeño en múltiples umbrales de clasificación. Un AUC-ROC de 1 indica una clasificación perfecta, mientras que un AUC de 0.5 sugiere una clasificación aleatoria [56].

Sin embargo, cuando se trata de la predicción de enfermedades poco frecuentes como la hipertensión, el Área Bajo la Curva de Precisión-Recall (AUC-PR) suele ser más informativa, ya que pone un mayor énfasis en la capacidad del modelo para detectar correctamente los casos positivos, en lugar de simplemente considerar todas las predicciones.

$$AUC_{PR} = \sum_{i=1}^{n-1}(Recall_{i+1} - Recall_i)\ x\ \frac{Precision_{i+1} + Precision_i}{2}$$

(19)

Un alto valor de AUC-PR indica que el modelo logra mantener una alta sensibilidad sin sacrificar demasiada precisión, lo cual es crucial en la detección temprana de hipertensión y la toma de decisiones clínicas oportunas [57].

2.6. SELECCIÓN E INTERPRETACIÓN

El análisis comparativo de los algoritmos de aprendizaje automático utilizados en la predicción de hipertensión mostró diferencias significativas en su desempeño:

Random Forest fue el modelo con mejor rendimiento, obteniendo una precisión del 92.05% y un ROC-AUC Score de 0.9484, lo que indica una alta capacidad de discriminación entre pacientes hipertensos y no hipertensos. Además, presentó una alta precisión (0.93) y un buen equilibrio entre sensibilidad (0.79) y especificidad, reduciendo la tasa de falsos negativos y falsos positivos. La matriz de confusión confirmó que Random Forest presenta un excelente desempeño en la identificación correcta de hipertensión.

SVM (Support Vector Machine) mostró un desempeño intermedio, con una precisión del 79.20% y un ROC-AUC Score de 0.8162. Aunque su capacidad predictiva fue aceptable, presentó un recall bajo (0.48), lo que indica que tuvo dificultades para identificar correctamente a los pacientes hipertensos, generando una mayor cantidad de falsos negativos.

Regresión Logística obtuvo el peor desempeño, con una precisión del 71.25% y un ROC-AUC Score de 0.5879, apenas superior al azar. Su bajo recall (0.02) indica una alta tasa de falsos negativos, lo que lo hace poco fiable para la detección de hipertensión en este estudio.

Random Forest demostró ser el modelo más efectivo para la predicción de hipertensión, con una combinación óptima de precisión, sensibilidad, capacidad de clasificación y un bajo número de falsos negativos y falsos positivos según su matriz de confusión.

3. RESULTADOS

Se evaluó la distribución de la edad según la presencia o ausencia de hipertensión. En el grupo de pacientes con hipertensión, la media de edad fue de 64.65 años (±12.39), mientras que en el grupo sin hipertensión fue de 61.54 años (±16.69). La Figura 1 presenta la distribución de esta variable.

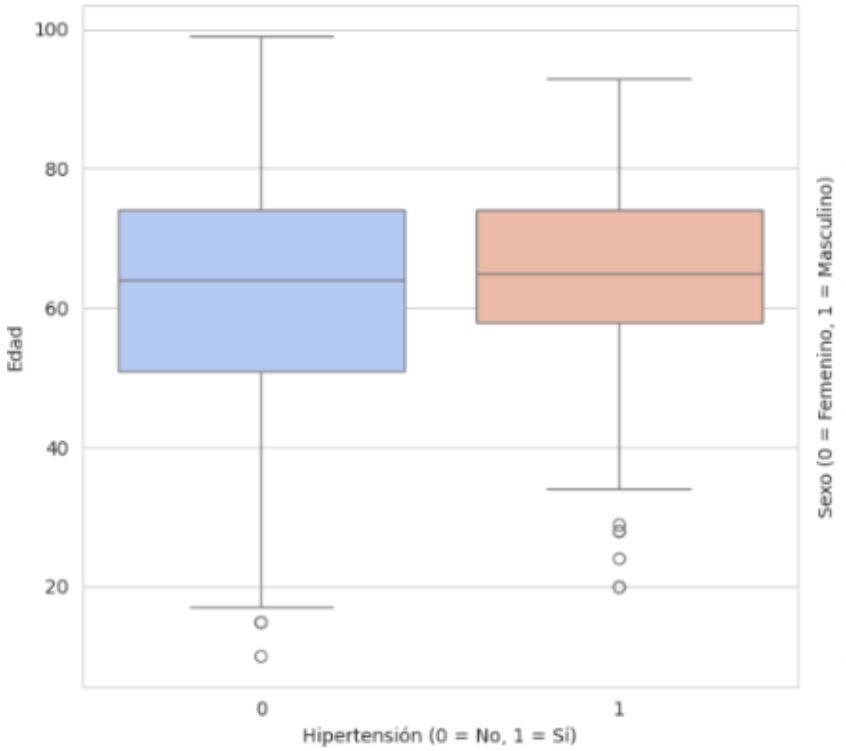


Figura 2. Distribución de edad según hipertensión

En cuanto a la variable "Sexo", se observó la proporción de pacientes masculinos y femeninos con y sin hipertensión. Como la variable está codificada (0 = Femenino, 1 = Masculino), la Figura 2 muestra la comparación de estas proporciones.

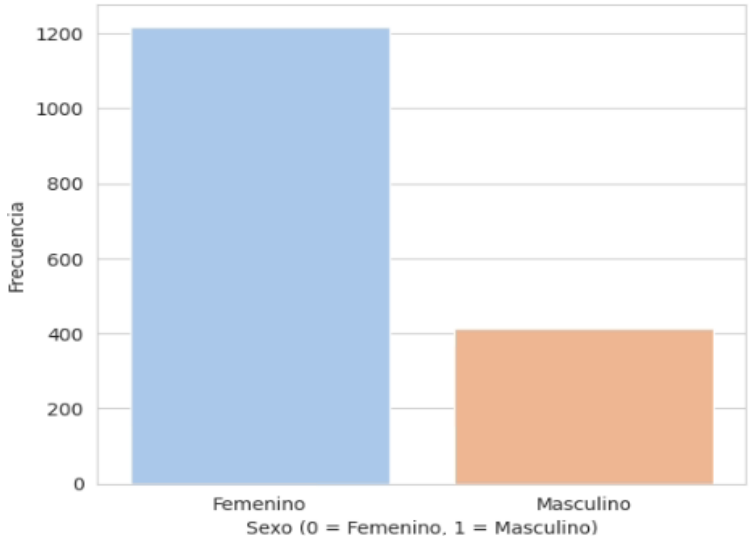


Figura 3. Distribución por sexo según hipertensión

La distribución entre el índice de masa corporal (IMC) y la presencia de hipertensión, con una segmentación por categorías de IMC (Peso insuficiente, Peso normal, Sobrepeso y Obesidad). Se observa la tendencia de los valores de IMC en ambos grupos de hipertensión, lo que permite analizar si un mayor IMC está asociado con una mayor prevalencia de hipertensión, tal como se muestra en la figura 4

Tabla 2. Categoría de IMC

Categoría de IMC	Descripción
Peso insuficiente	IMC < 18.5. Representa a individuos con un peso por debajo del recomendado. Se analiza si este grupo tiene una menor prevalencia de hipertensión en comparación con los demás.
Peso normal	IMC entre 18.5 y 24.9. Considerado el rango de peso saludable. Se observa la distribución de hipertensión dentro de este grupo y si presenta una baja proporción de casos.
Sobrepeso	IMC entre 25.0 y 29.9. Representa a individuos con un peso superior al normal. Se evalúa si existe una tendencia creciente en la presencia de hipertensión en este grupo.
Obesidad	IMC ≥ 30.0. Se analiza si este grupo presenta la mayor proporción de casos de hipertensión, dado que la obesidad es un factor de riesgo conocido para esta condición.

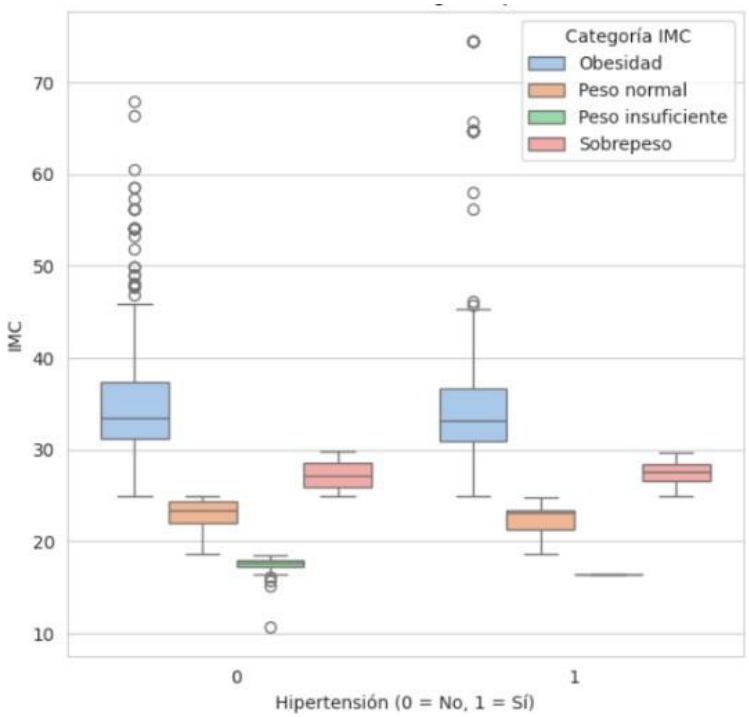


Figura 4. Distribución de IMC por hipertensión

Tabla 3. Tabla de resultados estadísticos

Característica	Media	Mediana	Moda	Desv. Estándar	Varianza	Mínimo	Máximo	Rango IQR
Edad (años)	62.5	64	64	15.56	242.05	10	99	21
Sexo (0=F, 1=M)	0.25	0	0	0.44	0.19	0	1	1
IMC (kg/m²)	29.5	28.24	23.4	7.58	57.43	10.68	74.51	7.15
Presión Sistólica (mmHg)	131.1	125	115	26.8	718.22	60	234	33
Presión Diastólica (mmHg)	73.22	71	60	14.29	204.23	20	133	22
Glucosa (mg/dL)	130.62	133	108.7	75.11	5641.65	8.5	613.41	135.65
Colesterol (mg/dL)	120.22	99.51	124.1	56.98	3247.23	34.19	521.4	38.72

Análisis de Predicciones

Las predicciones se realizaron utilizando diversos algoritmos y los datos asignados para la prueba de testeo, reflejando así la distribución de los datos. Los resultados obtenidos se presentan en la tabla 2.

Tabla 4. Análisis de predicciones

Métrica	Random Forest	SVM	Logistic Regression
Accuracy	92.05%	79.20%	71.25%
ROC-AUC Score	0.9484	0.8162	0.5879
Precision (Clase 1 - Hipertensos)	0.93	0.71	0.67
Recall (Clase 1 - Hipertensos)	0.79	0.48	0.02
F1-Score (Clase 1)	0.85	0.57	0.04

Análisis de evaluación por modelo Random Forest

La curva ROC del algoritmo Random Forest se muestra en la Figura 4, con un área bajo la curva (ROC-AUC Score) de 0.9484. Esto indica un excelente desempeño del modelo para diferenciar entre pacientes hipertensos y no hipertensos. Una curva más cercana a la esquina superior izquierda sugiere una mayor capacidad predictiva.

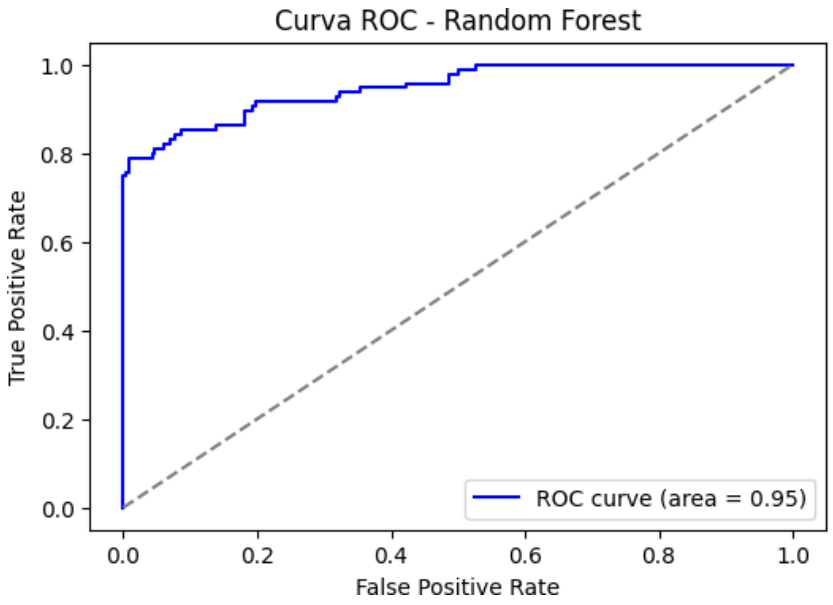


Figura 5. Curva ROC - Random Forest

La matriz de confusión del modelo Random Forest muestra una alta tasa de aciertos en la predicción de pacientes hipertensos y no hipertensos. La mayoría de los valores se concentran en la diagonal principal, indicando un bajo número de falsos positivos y falsos negativos.

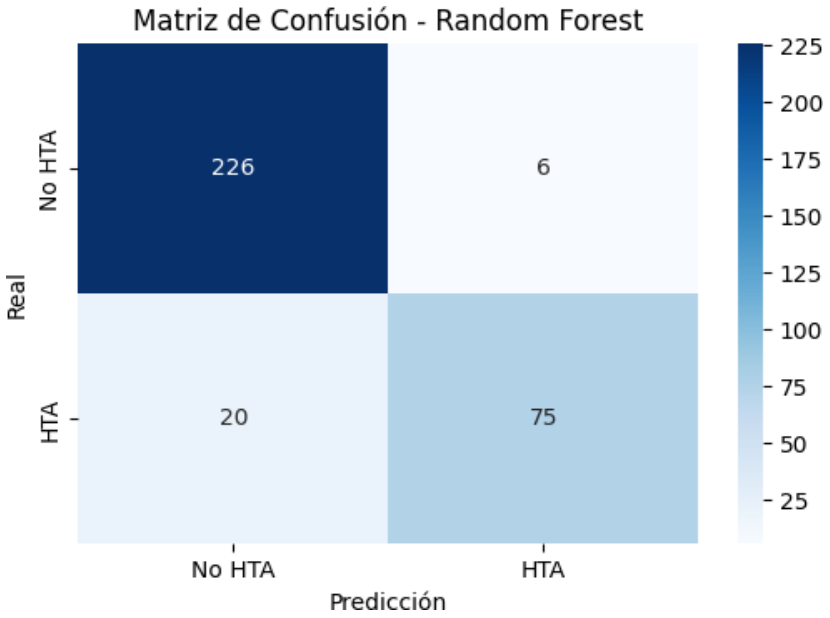


Figura 6. Matriz de confusión del modelo Random Forest

SVM (Support Vector Machine)

La curva ROC del modelo SVM se presenta en la Figura, con un ROC-AUC Score de 0.8162. Aunque muestra un rendimiento menor que el Random Forest, mantiene una buena capacidad de discriminación. Sin embargo, su desempeño es inferior en comparación con otros modelos evaluados.

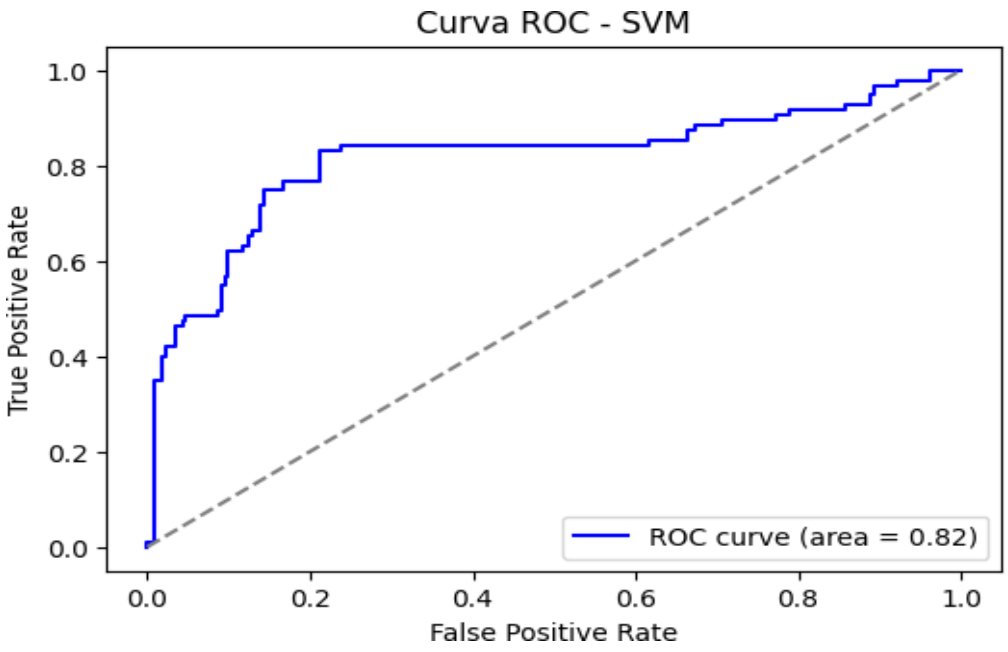


Figura 7. Curva ROC - Support Vector Machine

En la matriz de confusión del modelo SVM, se observa una mayor cantidad de falsos negativos, lo que indica que el modelo tiene dificultades para identificar correctamente a los pacientes hipertensos. Esto sugiere que el modelo puede estar más inclinado a predecir casos negativos.

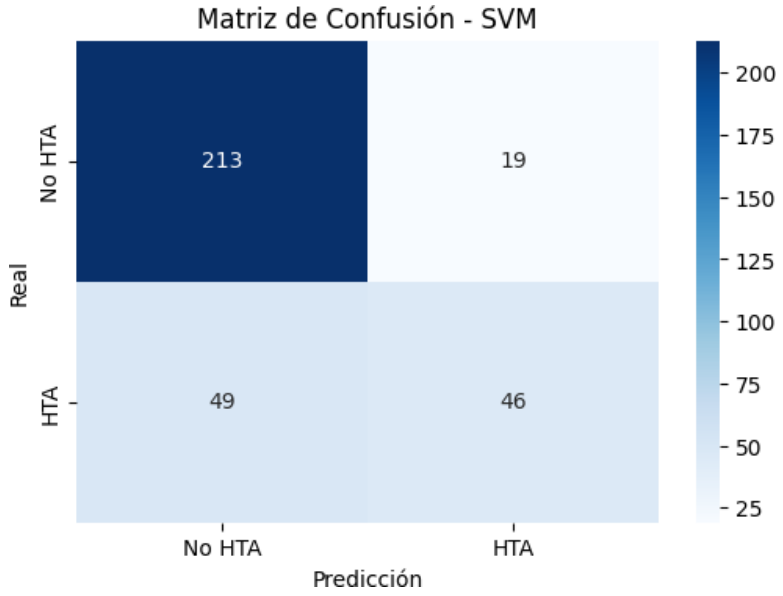


Figura 8. Matriz de confusión del modelo Support Vector Machine

Logistic Regression

La curva ROC del modelo Regresión Logística se muestra en la Figura, con un ROC-AUC Score de 0.5879. Este valor sugiere que el modelo tiene una capacidad predictiva baja y apenas supera el azar (0.5). La curva se encuentra más cercana a la diagonal, lo que indica que la clasificación no es óptima.

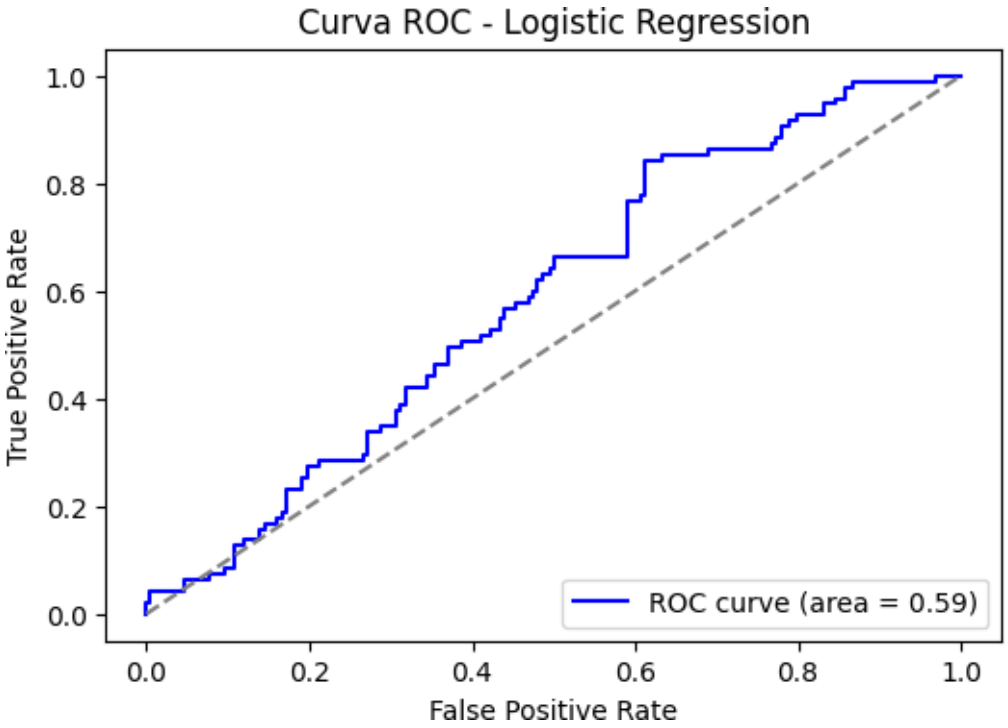


Figura G. Curva ROC - Logistic Regresion

La matriz de confusión del modelo Regresión Logística muestra una clara dificultad en la predicción de pacientes hipertensos, con un alto número de falsos negativos. Esto indica que el modelo no está logrando capturar correctamente la presencia de hipertensión en los datos.

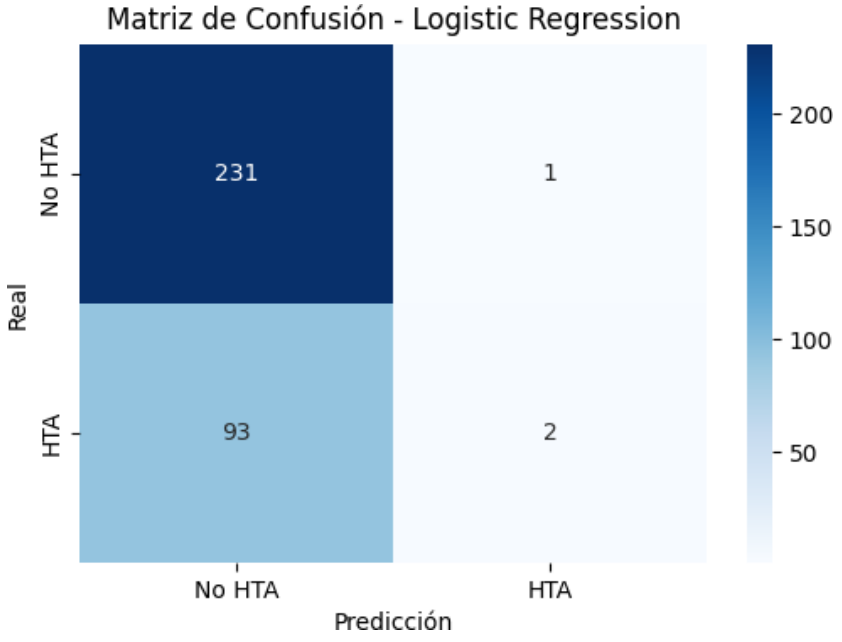


Figura 10. Matriz de confusión del modelo Logistic Regresion

4. CONCLUSIONES

Los resultados obtenidos sugieren que la edad sigue siendo un factor clave en la prevalencia de hipertensión, donde los pacientes hipertensos presentan una media de edad más alta en comparación con aquellos sin hipertensión. Esta tendencia se alinea con estudios previos que destacan el aumento del riesgo de hipertensión con la edad [58][59]. En relación con el sexo, observamos que existe una diferencia en la prevalencia de hipertensión entre hombres y mujeres, un hallazgo que podría estar influenciado por factores hormonales, estilo de vida y predisposiciones genéticas, como se ha observado en investigaciones previas [60] [61]. El análisis del índice de masa corporal (IMC) reveló que la obesidad tiene la mayor proporción de casos de hipertensión, lo cual coincide con la literatura que reconoce a la obesidad como un factor de riesgo importante para esta condición [59] [62] . En cuanto a la evaluación de modelos de predicción, se observó que el algoritmo Random Forest presentó el mejor desempeño con un ROC-AUC Score de 0.9484, indicando una excelente capacidad de clasificación. La matriz de confusión mostró que este modelo tiene un bajo número de falsos positivos y falsos negativos, lo que lo convierte en el más efectivo para la predicción de hipertensión [63] . Por otro lado, el modelo SVM, aunque tuvo un rendimiento aceptable con un ROC-AUC Score de 0.8162, mostró más falsos negativos en la matriz de confusión, lo que indica que tiene dificultades para identificar correctamente a los pacientes hipertensos. Esto podría estar relacionado con la dificultad de SVM para separar los datos en el espacio de características [64]. Finalmente, el modelo Regresión Logística tuvo el más bajo desempeño con un ROC-AUC Score de 0.5879, apenas superior al azar. La matriz de confusión mostró una alta tasa de falsos negativos, lo que indica que este modelo no es adecuado para la predicción de hipertensión en este conjunto de datos.

5. REFERENCIAS

[1] N. R. C. Campbell and M. L. Niebylski, 'Prevention and control of hypertension: Developing a global agenda', *Curr Opin Cardiol*, vol. 29, no. 4, pp. 324–330, 2014, doi: 10.1097/HCO.0000000000000067.

[2] A. Arredondo and R. Avilés, 'Hypertension and its effects on the economy of the health system for patients and society: Suggestions for developing countries', 2014, *Oxford University Press*. doi: 10.1093/ajh/hpu010.

[3] P. Sarkar, P. S. Aithal, and M. Rahman, 'Hypertension Hindrance: Unraveling the Impact on Bangladesh's Economic Development', *Economic Affairs (New Delhi)*, vol. 69, no. 2, pp. 799–807, Jun. 2024, doi: 10.46852/0424-2513.3.2024.3.

[4] P. Bhandari, 'Prevalence of cardiovascular risk factors among Asian migrant workers in South Korea', *PLoS One*, vol. 18, no. 7 JULY, Jul. 2023, doi: 10.1371/journal.pone.0288375.

[5] W. Gao *et al.*, 'A clinical prediction model of medication adherence in hypertensive patients in a Chinese community hospital in Beijing', *Am J Hypertens*, vol. 33, no. 11, pp. 1038–1046, 2020, doi: 10.1093/ajh/hpaa111.

[6] G. Mancia *et al.*, '2023 ESH Guidelines for the management of arterial hypertension', 2023. [Online]. Available: http://journals.lww.com/jhypertension

[7] K. P. Patel, R. Trivedi, and R. A. Maheshwari, 'An Overview of the Benefits of Indian Spices for High Blood Pressure', Oct. 01, 2023, *Informatics Publishing Limited and Society for Biocontrol Advancement*. doi: 10.18311/jnr/2023/33475.

[8] B. Zhou *et al.*, 'Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants', *The Lancet*, vol. 398, no. 10304, pp. 957–980, 2021, doi: 10.1016/S0140-6736(21)01330-1.

[9] P. López-Jaramillo and J. P. López-López, 'Factores de riesgo y muerte cardiovascular en América del Sur', *Clínica e Investigación en Arteriosclerosis*, vol. 35, no. 4, pp. 195–200, 2023, doi: 10.1016/j.arteri.2022.12.001.

[10] P. Lopez-Jaramillo, 'Editorial: Global excellence in cardiovascular medicine: Central and South America', 2024, *Frontiers Media SA*. doi: 10.3389/fcvm.2024.1429182.

[11] R. A. Ingaramo, 'Obesity, Diabetes, and Other Cardiovascular Risk Factors in Native Populations of South America', Jan. 01, 2016, *Current Medicine Group LLC 1*. doi: 10.1007/s11906-015-0613-6.

[12] J. P. Zila-Velasque *et al.*, 'Prevalence of hypertension in adults living at altitude in Latin America and the Caribbean: A systematic review and meta-analysis', *PLoS One*, vol. 18, no. 10 October, pp. 1–17, 2023, doi: 10.1371/journal.pone.0292111.

[13] World Health Organization, 'Global report on hypertension - The race against a silent killer', 2023.

[14] P. M. Calderon-Ramirez, E. Huamani-Merma, M. G. Mirano-Ortiz-de-Orue, D. Fernandez-Guzman, and C. J. Toro-Huamanchumo, 'Factors associated with poor adherence to medication in patients with diabetes and hypertension in Peru: findings from a pooled analysis of six years of population-based surveys', *Public Health*, vol. 231, pp. 108–115, Jun. 2024, doi: 10.1016/j.puhe.2024.03.012.

[15] ESSALUD, 'Guía de práctica clínica Para el manejo de la hipertensión arterial esencial', pp. 5–6, 2022.

[16] MINSA, 'Guía técnica: guía de práctica clínica para el diagnóstico, tratamiento y control de la enfermedad hipertensiva', *Minsa*, pp. 1–27, 2015, [Online]. Available: http://www.minsa.gob.pe/transparencia/dge_normas.asp.

[17] O. Ciobanu-Caraus, A. Aicher, J. M. Kernbach, L. Regli, C. Serra, and V. E. Staartjes, 'A critical moment in machine learning in medicine: on reproducible and interpretable learning', *Acta Neurochir (Wien)*, vol. 166, no. 1, 2024, doi: 10.1007/s00701-024-05892-8.

[18] A. N. Reiz *et al.*, 'Big data and machine learning in critical care: Opportunities for collaborative research', *Med Intensiva*, vol. 43, no. 1, pp. 52–57, 2019, doi: 10.1016/j.medin.2018.06.002.

[19] O. Ciobanu-Caraus, A. Aicher, J. M. Kernbach, L. Regli, C. Serra, and V. E. Staartjes, 'A critical moment in machine learning in medicine: on reproducible and interpretable learning', *Acta Neurochir (Wien)*, vol. 166, no. 1, 2024, doi: 10.1007/s00701-024-05892-8.

[20] A. T. Layton, 'AI, Machine Learning, and ChatGPT in Hypertension', Apr. 01, 2024, *Lippincott Williams and Wilkins*. doi: 10.1161/HYPERTENSIONAHA.124.19468.

[21] D. Amaratunga, J. Cabrera, D. Sargsyan, J. B. Kostis, S. Zinonos, and W. J. Kostis, 'Uses and opportunities for machine learning in hypertension research', *Int J Cardiol Hypertens*, vol. 5, no. February, p. 100027, 2020, doi: 10.1016/j.ijchy.2020.100027.

[22] S. Montagna *et al.*, 'Machine Learning in Hypertension Detection: A Study on World Hypertension Day Data', *J Med Syst*, vol. 47, no. 1, pp. 1–10, 2023, doi: 10.1007/s10916-022-01900-5.

[23] R. C. Deo, 'Machine learning in medicine', *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[24] E. Martinez-Ríos, L. Montesinos, M. Alfaro-Ponce, and L. Pecchia, 'A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data', *Biomed Signal Process Control*, vol. 68, no. June, 2021, doi: 10.1016/j.bspc.2021.102813.

[25] C. El-Hajj and P. A. Kyriacou, 'A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure', *Biomed Signal Process Control*, vol. 58, p. 101870, 2020, doi: 10.1016/j.bspc.2020.101870.

[26] L. Palmeri *et al.*, 'Photoplethysmographic waveform characteristics of newborns with coarctation of the aorta', *Journal of Perinatology*, vol. 37, no. 1, pp. 77–80, 2017, doi: 10.1038/jp.2016.162.

[27] T. Koivistoinen *et al.*, 'Pulse Wave Velocity Predicts the Progression of Blood Pressure and Development of Hypertension in Young Adults', *Hypertension*, vol. 71, no. 3, pp. 451–456, 2018, doi: 10.1161/HYPERTENSIONAHA.117.10368.

[28] M. Elgendi *et al.*, 'On time domain analysis of photoplethysmogram signals for monitoring heat stress', *Sensors (Switzerland)*, vol. 15, no. 10, pp. 24716–24734, 2015, doi: 10.3390/s151024716.

[29] X. Ding, B. P. Yan, Y. T. Zhang, J. Liu, N. Zhao, and H. K. Tsang, 'Pulse Transit Time Based Continuous Cuffless Blood Pressure Estimation: A New Extension and A Comprehensive Evaluation', *Sci Rep*, vol. 7, no. 1, pp. 1–11, 2017, doi: 10.1038/s41598-017-11507-3.

[30] W. Xing, Y. Shi, C. Wu, Y. Wang, and X. Wang, 'Predicting blood pressure from face videos using face diagnosis theory and deep neural networks technique', *Comput Biol Med*, vol. 164, no. April, p. 107112, 2023, doi: 10.1016/j.compbiomed.2023.107112.

[31] A. Akella and S. Akella, 'Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution', *Future Sci OA*, vol. 7, no. 6, 2021, doi: 10.2144/fsoa-2020-0206.

[32] E. Mejía-Mejía, J. M. May, P. A. Kyriacou, and M. Elgendi, 'Classification of blood pressure in critically ill patients using photoplethysmography and machine learning', *Comput Methods Programs Biomed*, vol. 208, 2021, doi: 10.1016/j.cmpb.2021.106222.

[33] R. Wirth and J. Hipp, 'CRISP-DM: Towards a Standard Process Model for Data Mining'.

[34] J. Han, M. Kamber, and J. Pei, 'Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)', 2011.

[35] World Health Organization (WHO), *Global spending on health: a world transition*. 2019. [Online]. Available: https://www.who.int/publications/i/item/who-his-hgf-hfworkingpaper-19.4

[36] Organización Panamericana de la Salud (OPS), 'Inversión Pública En Salud: ¿Mayor Presupuesto Implica Mayor Ejecución?', *ComexPerú*, pp. 1–5, 2023.

[37] MINSA, 'Evaluación de los programas presupuestales de salud año 2019', 2019.

[38] Instituto Nacional de Estadística e Informática, 'PERÚ: ENFERMEDADES NO TRANSMISIBLES Y TRANSMISIBLES, 2022', 2023.

[39] G. Mancia *et al.*, '2023 ESH Guidelines for the management of arterial hypertension the Task Force for the management of arterial hypertension of the European Society of Hypertension: Endorsed by the International Society of Hypertension (ISH) and the European Renal Association (ERA)', *J Hypertens*, vol. 41, no. 12, pp. 1874–2071, Dec. 2023, doi: 10.1097/HJH.0000000000003480.

[40] S. N. S. Z. Shah and M. M. Rosli, 'Clustering algorithms for analysing electronic medical record: A mapping study', *IAES International Journal of Artificial Intelligence*, vol. 12, no. 4, pp. 1784–1792, 2023, doi: 10.11591/ijai.v12.i4.pp1784-1792.

[41] G. Rebala, A. Ravi, and S. Churiwala, 'Machine Learning Definition and Basics', in *An Introduction to Machine Learning*, Springer International Publishing, 2019, pp. 1–17. doi: 10.1007/978-3-030-15729-6_1.

[42] V. S. Jatti *et al.*, 'Predicting specific wear rate of laser powder bed fusion AlSi10Mg parts at elevated temperatures using machine learning regression algorithm: Unveiling of microstructural morphology analysis', *Journal of Materials Research and Technology*, vol. 33, pp. 3684–3695, Nov. 2024, doi: 10.1016/j.jmrt.2024.09.244.

[43] A. V. Dorugade, 'Adjusted ridge estimator and comparison with Kibria's method in linear regression', *Journal of the Association of Arab Universities for Basic and Applied Sciences*, vol. 21, pp. 96–102, Oct. 2016, doi: 10.1016/J.JAUBAS.2015.04.002.

[44] M. K. Pawar and P. Patil, 'Logistic Regression for Enhancing Scalability of Blockchain System', *Procedia Comput Sci*, vol. 252, pp. 146–153, 2025, doi: 10.1016/j.procs.2024.12.016.

[45] M. Yamada *et al.*, 'Breast cancer risk assessment based on a predictive model: evaluation of risk factors among Japanese women', *BMC Cancer*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12885-025-13556-8.

[46] M. Cherifi, M. N. El Korso, S. Fortunati, A. Mesloub, and L. Ferro-Famil, ‘Robust inference with incompleteness for logistic regression model’, 2025.

[47] L. Breiman, ‘Random Forests’, 2001. Accessed: Mar. 04, 2025. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[48] M. Bhat and R. B. Keskar, ‘Self-supervised random forests for robust voice activity detection with limited labeled data’, *Applied Acoustics*, vol. 234, p. 110636, Apr. 2025, doi: 10.1016/j.apacoust.2025.110636.

[49] H. Qin *et al.*, ‘Interpretable machine learning approaches for children’s ADHD detection using clinical assessment data: an online web application deployment’, *BMC Psychiatry*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12888-025-06573-1.

[50] G. Alwakid, F. Ul Haq, N. Tariq, M. Humayun, M. Shaheen, and M. Alsadun, ‘Optimized machine learning framework for cardiovascular disease diagnosis: a novel ethical perspective’, *BMC Cardiovasc Disord*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12872-025-04550-w.

[51] T. Hastie, R. Tibshirani, and J. Friedman, ‘Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction’.

[52] Z. Guo, W. Du, C. Li, X. Guo, and Z. Liu, ‘Fault diagnosis of rotating machinery with high-dimensional imbalance samples based on wavelet random forest’, *Measurement (Lond)*, vol. 248, May 2025, doi: 10.1016/j.measurement.2025.116936.

[53] K. Feng *et al.*, ‘Gas kick and lost circulation risk identification method with multi-parameters based on support vector machine for drilling in deep or ultradeep waters’, *Engineering Science and Technology, an International Journal*, vol. 64, p. 102007, Apr. 2025, doi: 10.1016/j.jestch.2025.102007.

[54] M. Sokolova and G. Lapalme, ‘A systematic analysis of performance measures for classification tasks’, *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[55] D. Chicco and G. Jurman, ‘The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation’, *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[56] D. M. W. Powers and Ailab, ‘Evaluation: From precision, recall and f-measure to roc, informedness, markedness C correlation’.

[57] T. Saito and M. Rehmsmeier, ‘The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets’, *PLoS One*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[58] M. E. V. Bianchi, A. M. Cusumano, C. Torres, N. G. Rojas, and G. A. Velasco, 'Prevalence of obesity and arterial hypertension and their relationship with age and gender in the city of Resistencia, Argentina, in the years 2008-2014', *Hipertens Riesgo Vasc*, vol. 36, no. 1, pp. 14–20, Jan. 2019, doi: 10.1016/j.hipert.2018.04.003.

[59] Romero Giraldo Milagros, Avendaño-Olivares Jane, Vargas Fernández Rodrigo, and Runzer Colmenares Fernando, 'Diferencias según sexo en los factores asociados a hipertensión arterial en el Perú: Análisis de la Encuesta Demográfica y de Salud Familiar 2017', 2017.

[60] J. J. Song, Z. Ma, J. Wang, L. X. Chen, and J. C. Zhong, 'Gender Differences in Hypertension', Feb. 01, 2020, *Springer*. doi: 10.1007/s12265-019-09888-z.

[61] Y. Appelman, B. B. van Rijn, M. E. ten Haaf, E. Boersma, and S. A. E. Peters, 'Sex differences in cardiovascular risk factors and disease prevention', *Atherosclerosis*, vol. 241, no. 1, pp. 211–218, Nov. 2014, doi: 10.1016/j.atherosclerosis.2015.01.027.

[62] O. A. Shariq and T. J. Mckenzie, 'Obesity-related hypertension: A review of pathophysiology, management, and the role of metabolic surgery', *Gland Surg*, vol. 9, no. 1, pp. 80–93, Feb. 2020, doi: 10.21037/gs.2019.12.03.

[63] B. Liu, X. Chen, and Y. Chen, 'Research and design of multi-charge module based on average current method', in *Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 353–357. doi: 10.1109/IFEEA51475.2020.00080.

[64] E. Held, J. Cape, and N. Tintle, 'Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data', in *BMC Proceedings*, BioMed Central Ltd., 2016. doi: 10.1186/s12919-016-0020-2.

Jack Cosme Castillo Ramos, Egresado de la Universidad Peruana Unión (UPeU) de la carrera de Ingeniería de Sistemas ha desempeñado el cargo de encargado de la Estadística e Información en Secretaría General en la Universidad Peruana Unión (UPeU), también ha sido Desarrollador de Microservicios en el Ministerio de Salud del Perú (MINSA), también como Analista Programador en el Ministerio de Trabajo y Promoción del Empleo (MTPE) y Programador JAVA en la Superintendencia Nacional de los Registros Públicos (SUNARP).

Gabriela Maria Auqui Aguilar, Egresado de la Universidad Peruana Unión (UPeU) de la carrera de Ingeniería de Sistemas ha desempeñado el cargo de Asistente de Acreditación SINEACE y auxiliar de laboratorio- Universidad Peruana Unión (UPEU), también como Analista de sistemas y procesos en Quality C Innovation Services S.A.C, también como Analista de calidad y procesos en Cable Visión Perú, así también como Analista de procesos en CAAP Consulting.

Luis Felipe Humberto Moran Nureña, Egresado de la Universidad Peruana Unión (UPeU) de la carrera de Ingeniería de Sistemas ha desempeñado el cargo de Analista de Soporte Técnico y Asistente de Soporte Técnico en el Proyecto Especial de Inversión Publica Escuelas Bicentenario (PEIP EB), también como Asistente Administrativo en el Hospital José Agurto Tello, así también como Analista técnico en la Red de Salud Huarochirí.

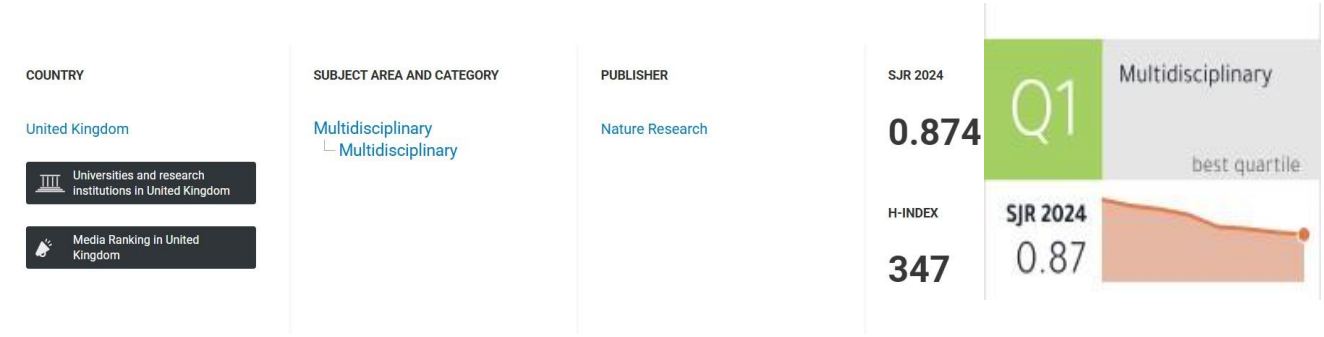
Javier Linkolk Lopez Gonzales, Miembro del ISI y del IEEE. Obtuvo la licenciatura en Ingeniería Estadística e Informática por la Universidad Peruana Unión (UPeU, Perú) y el máster en Metrología por la Pontificia Universidad Católica de Río de Janeiro (PUC-Rio, Brasil). Asimismo, obtuvo el doctorado en Estadística por la Universidad de Valparaíso (UV, Chile). Actualmente es profesor titular en la Universidad Peruana Unión. Además, está acreditado como investigador RENACYT. Sus principales intereses de investigación incluyen el reconocimiento de patrones en el aprendizaje automático, la contaminación atmosférica con técnicas de aprendizaje profundo y las series temporales con análisis espectral singular.

Soria Quijaite Juan Jesús, PhD en Ingeniería de Sistemas, magíster en Docencia Universitaria y Gestión Educativa, magíster en Matemática Aplicada y especialista en Estadística para la Investigación. Consultor estadístico del Ministerio de la Producción Cite Agroindustrial, docente investigador de la Universidad Peruana Unida, miembro de la red mundial de investigadores AUTHOR AID, miembro de la Red de Docentes de América Latina y el Caribe, miembro del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica RENACYT (CONCYTEC) en el Nivel VI, asesor de tesis con los enfoques

del método científico, buenas prácticas del PMBOK versión 7 y pensamiento sistémico en diferentes especialidades de pregrado y postgrado universitario. Puede ser contactado al correo electrónico: jesussoria@upeu.edu.pe.

EVIDENCIAS DE SUMISION

Scientific Reports



scientific reports

Manuscript ID	Title	Submission Date
scirep-7018419	Hypertension prediction approach in Peruvian patients based on machine learning algorithms	2025-07-04 11:18:08

Manuscript Status: Under review

"AÑO DEL BICENTENARIO, DE LA CONSOLIDACIÓN DE NUESTRA INDEPENDENCIA, Y DE LA CONMEMORACIÓN DE LAS HEROICAS BATALLAS DE JUNÍN Y AYACUCHO"

RESOLUCIÓN N° 0064-2024/UPeU-FIA-CF-T

Lima, Ñaña 20 de febrero de 2024

VISTO:

El expediente de **Jack Cosme Castillo Ramos**, identificado(a) con Código Universitario Nº 201011058, **Luis Felipe Humberto Moran Nureña**, identificado(a) con Código Universitario Nº 201310311 y **Gabriela Maria Auqui Aguilar**, identificado(a) con Código Universitario Nº 201320743, de la Escuela Profesional de Ingeniería de Sistemas de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión;

CONSIDERANDO

Que la Universidad Peruana Unión tiene autonomía académica, administrativa y normativa, dentro del ámbito establecido por la Ley Universitaria Nº 30220 y el Estatuto de la Universidad;

Que la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, mediante sus reglamentos académicos y administrativos, ha establecido las formas y procedimientos para la aprobación e inscripción del perfil de proyecto de tesis en formato artículo y la designación o nombramiento del asesor para la obtención del título profesional;

Que **Jack Cosme Castillo Ramos, Luis Felipe Humberto Moran Nureña y Gabriela Maria Auqui Aguilar**, han solicitado: la inscripción del perfil de proyecto de tesis titulado "Abordaje de predicción de hipertensión en pacientes peruanos basado en algoritmos de machine learning" y la designación del Asesor, encargado de orientar y asesorar la ejecución del perfil de proyecto de tesis en formato artículo;

Estando a lo acordado en la sesión del Consejo de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, celebrada el 20 de febrero de 2024, y en aplicación del Estatuto y el Reglamento General de Investigación de la Universidad;

SE RESUELVE:

Aprobar el perfil de proyecto de tesis en formato artículo titulado **"Abordaje de predicción de hipertensión en pacientes peruanos basado en algoritmos de machine learning"** y disponer su inscripción en el registro correspondiente, designar al (a la) **Ph.D. Javier Linkolk López Gonzales** como ASESOR para que oriente y asesore la ejecución del perfil de proyecto de tesis en formato artículo el cual fue dictaminado por: **Dr. Juan Jesus Soria Quijaite** y **Mg. Ferdinan Edgardo Pineda Anco**, otorgándoles un plazo máximo de doce (12) meses para la ejecución.

Regístrese, comuníquese y archívese.

Dra. Erika Inés Acuña Salinas
DECANA

Mg. Ketty Magaly Arellano Lino
SECRETARIA ACADÉMICA

cc:
-Interesado
-Asesor
-Dirección General de Investigación
-Archivo