UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

CENTRO DE CIENCIAS BÁSICAS

TESIS

Análisis del desempeño de distintos métodos de clasificación aplicados a la predicción de la anatomía vascular cervical en pacientes con hipertensión, diabetes o dislipidemia.

PRESENTA

Mtro. Juan Manuel Marquez Romero

PARA OBTENER EL GRADO DE DOCTOR EN CIENCIAS APLICADAS Y TECNOLOGÍA CON ORIENTACIÓN EN CIENCIAS DE LA COMPUTACIÓN

COTUTORES

Dr. Rogelio Salinas Gutiérrez

Dra. Svetlana Vladislavovna Doubova

INTEGRANTE DEL COMITÉ TUTORAL

Dr. Ángel Eduardo Muñoz Zavala

AGUASCALIENTES, AGS, 27 DE JULIO DE 2025

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

DICTAMEN DE LIBERACION ACADEMICA PARA INICIAR LOS TRAMITES DEL EXAMEN DE GRADO

Fecha de dictaminación dd/mm/aaaa: 06/08/2025

NOMBRE: Juan Manuel Marquez Romero **ID** 20016

PROGRAMA: DOCTORADO EN CIENCIAS APLICADAS Y TECNOLOGÍA **LGAC (del posgrado):** CIENCIAS DE LA COMPUTACIÓN

TIPO DE TRABAJO: (X) Tesis ( ) Trabajo Práctico

TITULO: Análisis del desempeño de distintos métodos de clasificación aplicados a la predicción de la anatomía vascular cervical en pacientes con hipertensión, diabetes o dislipidemia

IMPACTO SOCIAL (señalar el impacto logrado): Mediante algoritmos de aprendizaje automático, este trabajo puede optimizar la toma de decisiones clínicas, especialmente en contextos con recursos limitados, pudiendo impactar positivamente en la calidad de vida de los pacientes y en la equidad en el acceso a la salud.

INDICAR SI NO N.A. (NO APLICA) SEGÚN CORRESPONDA:	
Elementos para la revisión académica del trabajo de tesis o trabajo práctico:	
SI	El trabajo es congruente con las LGAC del programa de posgrado
SI	La problemática fue abordada desde un enfoque multidisciplinario
SI	Existe coherencia, continuidad y orden lógico del tema central con cada apartado
SI	Los resultados del trabajo dan respuesta a las preguntas de investigación o a la problemática que aborda
SI	Los resultados presentados en el trabajo son de gran relevancia científica, tecnologíca o profesional según el área
SI	El trabajo demuestra más de una aportación original al conocimiento de su área
SI	Las aportaciones responden a los problemas prioritarios del país
SI	Generó transferecia del conocimiento o tecnológica
SI	Cumple con la ética para la investigación (reporte de la herramienta antiplagio)
El egresado cumple con lo siguiente:	
SI	Cumple con lo señalado por el Reglamento General de Docencia
SI	Cumple con los requisitos señalados en el plan de estudios (créditos curriculares, optativos, actividades complementarias,estancia, predoctoral, etc)
SI	Cuenta con los votos aprobatorios del comité tutoral, en caso de los posgrados profesionales si tiene solo tutorpodrá liberar solo el tutor
N.A.	Cuenta con la carta de satisfacción del Usuario
SI	Coincide con el título y objetivo registrado
SI	Tiene congruencia con cuerpos académicos
SI	Tiene el CVU del Conacyt actualizado
SI	Tiene el artículo aceptado o publicado y cumple con los requisitos institucionales (en caso que proceda)
En caso de Tesis por artículos científicos publicados	
N.A.	Aceptación o Publicación de los articulos según el nivel del programa
N.A.	El estudiante es el primer autor
N.A.	El autor de correspondencia es el Tutor del Núcleo Académico Básico
N.A.	En los artículos se ven reflejados los objetivos de la tesis, ya que son producto de este trabajo de investigación.
N.A.	Los articulos integran los capítulos de la tesis y se presentan en el idioma en que fueron publicados
N.A.	La aceptación o publicación de los artículos en revistas indexadas de alto impacto

Con base a estos criterios, se autoriza se continúen con los trámites de titulación y programación del examen de grado:

Si x

No

FIRMAS

Elaboró:

* NOMBRE Y FIRMA DEL CONSEJERO SEGÚN LA LGAC DE ADSCRIPCION: DR. FRANCISCO JAVIER ÁLVAREZ RODRÍGUEZ

NOMBRE Y FIRMA DEL SECRETARIO TÉCNICO: DR. FRANCISCO JAVIER ÁLVAREZ RODRÍGUEZ

*** En caso de conflicto de intereses, firmará un revisor miembro del NAB de la LGAC correspondiente distinto al tutor o miembro del comité tutoral, asignado por el Decano**

Revisó:

NOMBRE Y FIRMA DEL SECRETARIO DE INVESTIGACIÓN Y POSGRADO: DR. ALEJANDRO PADILLA DÍAZ

Autorizó:

NOMBRE Y FIRMA DEL DECANO: M. EN C. JORGE MARTÍN ALFÉREZ CHÁVEZ

Nota: procede el trámite para el Depto. de Apoyo al Posgrado

En cumplimiento con el Art. 105C del Reglamento General de Docencia que a la letra señala entre las funciones del Consejo Académico: . . . Cuidar la eficiencia terminal del programa de posgrado y el Art. 105F las funciones del Secretario Técnico, llevar el seguimiento de los alumnos.

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

Mtro. en C. Jorge Martín Alférez Chávez
DECANO (A) DEL CENTRO DE CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como ***CO-TUTOR*** designado del estudiante ***JUAN MANUEL MARQUEZ ROMERO*** con ID *20016* quien realizó *la tesis* titulado: ***ANÁLISIS DEL DESEMPEÑO DE DISTINTOS MÉTODOS DE CLASIFICACIÓN APLICADOS A LA PREDICCIÓN DE LA ANATOMÍA VASCULAR CERVICAL EN PACIENTES CON HIPERTENSIÓN, DIABETES O DISLIPIDEMIA***, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a imprimir*la* así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 25 de julio de 2025.

Dr. Rogelio Salinas Gutiérrez
Co-tutor de *tesis*

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

Mtro. en C. Jorge Martín Alférez Chávez
DECANO (A) DEL CENTRO DE CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como ***TUTOR*** designado del estudiante ***JUAN MANUEL MARQUEZ ROMERO*** con ID *20016* quien realizó *la tesis* titulado: ***ANÁLISIS DEL DESEMPEÑO DE DISTINTOS MÉTODOS DE CLASIFICACIÓN APLICADOS A LA PREDICCIÓN DE LA ANATOMÍA VASCULAR CERVICAL EN PACIENTES CON HIPERTENSIÓN, DIABETES O DISLIPIDEMIA***, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a imprimir/*a* así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de julio de 2025.

Dra. Svetlana Vladislavovna Doubova
Tutor de *tesis*

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

Mtro. en C. Jorge Martín Alférez Chávez
DECANO (A) DEL CENTRO DE CIENCIAS BÁSICAS

P R E S E N T E

Por medio del presente como ***ASESOR*** designado del estudiante ***JUAN MANUEL MARQUEZ ROMERO*** con ID *20016* quien realizó *la tesis* titulado: ***ANÁLISIS DEL DESEMPEÑO DE DISTINTOS MÉTODOS DE CLASIFICACIÓN APLICADOS A LA PREDICCIÓN DE LA ANATOMÍA VASCULAR CERVICAL EN PACIENTES CON HIPERTENSIÓN, DIABETES O DISLIPIDEMIA***, un trabajo propio, innovador, relevante e inédito y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que *él* pueda proceder a imprimir*la* así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 28 de julio de 2025.

Dr. Ángel Eduardo Muñoz Zavala
Asesor de tesis

c.c.p.- Interesado
c.c.p.- Secretaría Técnica del Programa de Posgrado

Turk Neurosurg 34(4):695-700, 2024

Original Investigation

DOI: 10.5137/1019-5149.JTN.45018-23.2

Received: 20.07.2023
Accepted: 01.12.2023

Published Online: 13.06.2024

Effect of Size and Location of Unruptured Intracranial Aneurysms on Self-Reported Headache

Juan M MARQUEZ-ROMERO[1,4], Dulce A ESPINOZA-LÓPEZ[2], Juan M CALLEJA-CASTILLO[3], Fernando ZERMEÑO-PÖHLS[2], Rogelio SALINAS-GUTIÉRREZ[4]

[1]Hospital General de Zona #2, IMSS, OOAD Aguascalientes
[2]Departamento de Neurología, Instituto Nacional de Neurología "Manuel Velasco Suárez", Mexico
[3]Centro Neurológico ABC, Mexico
[4]Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico

Corresponding author: Juan M CALLEJA-CASTILLO ✉ juancalleja@me.com

ABSTRACT

AIM: To describe the relationship between aneurysm size and location with the prevalence of headache at diagnosis and three- and six-month follow-up in a sample of patients with UIA.

MATERIAL and METHODS: In this cohort study, patients were diagnosed with UIAs by digital subtraction angiography (DSA). Follow-up visits occurred three and six months after the diagnosis. Headache presence was registered, and headache was further classified by phenotypes. After DSA, the recorded variables were aneurysm number, morphology, location, and size (diameter [W], neck [N], and dome-neck distance [H]). The aspect ratio (H/N) and the dome/neck ratio (W/N) were calculated. The outcome of this study was the self-reported headache status at follow-up.

RESULTS: Data from 42 patients and 46 aneurysms were available; 81.0% of patients were women, with a mean age of 57.4±14.3 years. Headache was reported by 61.9% of the patients. The pain phenotype was tension-type in 38.1%, migraine in 11.9%, neuralgia in 2.4%, and unclassifiable in 9.5%. The median (min–max) measurements were W=5.05 (0.89–22.9); N=3.02 (0.52–17.9); H=5.08 (0.92–23.0); aspect ratio 1.59 (0.68–17.69) and W/N ratio 1.65 (0.62–16.92). Thirty-three patients (37 aneurysms) received treatment, 47.8% by surgical clipping and 32.6% by endovascular occlusion. In the treated patients, headaches had persisted in 14.3% until the first visit and in 9.5% until the second visit. There were no differences in any registered variables between patients with and without headaches at follow-up.

CONCLUSION: In this study, data was found that support that headaches in patients with UIAs improve after treatment and that such improvement is probably unrelated to the size and shape of the UIAs.

KEYWORDS: Intracranial aneurysm, Prevalence, Headache, Outcome

INTRODUCTION

The overall prevalence of incidental unruptured intracranial aneurysms (UIAs) ranges from 3.8% (95% CI 3.0% to 4.8%) to 8.3% (95% CI 7.1% to 9.7%), depending on which definitions are used, and is according to size and location (8). Up to one third of patients with UIAs will present with headaches (2,9). Although there are defined criteria for secondary headaches attributed to UIAs, (6) headaches have generally been considered unrelated to the presence of UIAs (2).

While the true nature of the relationship between UIAs and headaches is still being studied, a meta-analysis suggests that headache intensity significantly decreases after treatment (3). To further increase the knowledge regarding the course of

Juan M MARQUEZ-ROMERO : 0000-0002-5059-1174
Dulce A ESPINOZA-LÓPEZ : 0000-0001-8920-3901
Juan M CALLEJA-CASTILLO : 0000-0002-0423-9719
Fernando ZERMEÑO-PÖHLS : 0000-0002-1694-8395
Rogelio SALINAS-GUTIÉRREZ : 0000-0002-1669-4460

Original Article

Access this article online

Quick Response Code:

Website: http://www.braincirculation.org

DOI: 10.4103/bc.bc_28_25

Predictors of unfavorable aortic arch anatomy on computed tomography angiography in patients with stroke risk factors

Juan Manuel Marquez-Romero[1], Svetlana V. Doubova[2], Dulce M. Bonifacio-Delgadillo[3], Ángel E. Muñoz-Zavala[4], Rogelio Salinas-Gutiérrez[4]

Abstract:

CONTEXT: Identifying clinical variables associated with unfavorable aortic arch anatomy (AAA) is a seldom explored area with a high potential to increase endovascular stroke treatment safety, success, and efficiency.

AIMS: This study aims to explore the association between clinical variables and the occurrence of unfavorable AAA in patients with stroke risk factors (SRFs).

SETTINGS AND DESIGN: This is a retrospective cross-sectional study of computed tomography (CT) angiographies and electronic health records data.

SUBJECTS AND METHODS: The study involved classifying AAA into favorable (type I) or unfavorable (type II/III) categories using three-dimensional reconstructions. Unfavorable anatomic anomalies were bovine configuration, tortuosity, coiling, kinking, and atherosclerotic plaque.

STATISTICAL ANALYSIS USED: Clinical and laboratory variables were analyzed to identify predictors of unfavorable AAA through logistic regression.

RESULTS: We report data from 108 CT angiographies; 81 (75%) showed unfavorable AAA (49 were type II and 32 were type III). Images belonged to 108 patients (mean age 68.1 ± 13.8 years), and 57 (52.8%) were females. The most common SRF was hypertension (70.4%, with a mean duration of 7.64 years). Participants with unfavorable AAA had higher serum creatinine (0.91 ± 0.18 vs. 0.81 ± 0.13, $P = 0.002$). Unfavorable AAA was associated with three predictors: serum creatinine (odds ratio [OR] 5.0, 95% confidence interval [CI] 1.9–8.4), duration of hypertension (OR 0.11, 95% CI 0.03–0.20), and duration of hypertriglyceridemia (OR – 0.31, 95% CI – 0.62––0.07).

CONCLUSIONS: This study found that unfavorable AAA and anatomic anomalies are highly prevalent in patients with SRF. The results suggest three variables independently associated with unfavorable AAA.

Keywords:

Anatomy, aorta, computed tomography angiography, ischemic stroke, risk factors

[1]Hospital General de Zona #2, Instituto Mexicano Del Seguro Social, [4]Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Aguascalientes, [2]Unidad de Investigación Epidemiológica y Servicios de Salud Del CMN Siglo XXI, Instituto Mexicano del Seguro Social, [3]Departamento de Terapia Endovascular, Centro Médico Nacional 20 de Noviembre, ISSSTE, Ciudad de México, México

Address for correspondence:
Dr. Rogelio Salinas Gutiérrez, Av. University No. 940. University City, Zip Code 20100. Aguascalientes, Mexico.
E-mail: rogelio.salinas@edu.uaa.mx

Submission: 17-02-2025
Revised: 28-04-2025
Accepted: 14-05-2025
Published: ***

Introduction

Endovascular therapy (EVT) is the standard care for acute ischemic stroke (AIS) treatment. It involves thrombectomy with stent retrievers, aspiration catheters, or a combination.[1] The indications for EVT and procedures performed have increased significantly in recent years.[2] As EVT indications become wider, new challenges have emerged, particularly in terms of the technical success of EVT in achieving reperfusion. The reperfusion time is strongly linked to the clinical outcome of the patients treated with EVT. Consequently, factors

How to cite this article: Marquez-Romero JM, Doubova SV, Bonifacio-Delgadillo DM, Muñoz-Zavala ÁE, Salinas-Gutiérrez R. Predictors of unfavorable aortic arch anatomy on computed tomography angiography in patients with stroke risk factors. Brain Circ 2025;XX:XX-XX.

Código :	GMM/0200/25
Título :	Uso de algoritmos de aprendizaje automático para detectar trastornos neurocognitivos previos en adultos hospitalizados por COVID-19: Un estudio de clasificación
Título breve:	Aprendizaje Automático y Cognición
Estado:	Pendiente de dictamen de revisor
Tipo:	Artículo Original
Categoría:	Tecnología aplicada a la salud
Subcategoría:	Inteligencia artificial (IA) y aprendizaje automático
Resumen:	Antecedentes La pandemia por COVID-19 ha suscitado preocupación sobre la aparición de trastornos neurocognitivos (TNC). Identificar déficits cognitivos preexistentes en pacientes hospitalizados puede brindar información valiosa sobre los efectos a largo plazo del virus en la salud cognitiva. Objetivos Desarrollar y validar un enfoque de aprendizaje automático para clasificar a pacientes hospitalizados con COVID-19 como con o sin TNC, utilizando información de expedientes electrónicos (EE). Material y Métodos Se analizaron EE de clínicas de medicina familiar, enfocándose en registros de pacientes mayores de 18 años. Se utilizó Python para extraer palabras clave relacionadas con TNC, validadas mediante un panel Delphi modificado. Estas palabras se usaron para entrenar clasificadores basados en regresión logística (LR), k-vecinos más cercanos (k-NN) y análisis discriminante regularizado (RDA). Se aplicó validación cruzada de diez iteraciones para evaluar el desempeño. Resultados La muestra incluyó 205 pacientes, con edad media de 80.7 ± 6.7 años; 127 con TNC y 78 sin. Los modelos lograron precisiones balanceadas entre 79% y 81% y valores AUC de 0.78 a 0.87. RDA mostró el mayor AUC (0.87), mientras que k-NN presentó la mejor precisión balanceada. Conclusiones El enfoque propuesto permitió clasificar de manera efectiva a pacientes con TNC utilizando EE. Aunque RDA alcanzó el mejor AUC, k-NN destacó por su precisión. Este método representa una herramienta útil para la evaluación de TNC en entornos clínicos. Futuras investigaciones deberían incluir datos post-pandemia para mejorar su rendimiento.
Palabras clave:	Trastornos Neurocognitivos, COVID-19, Aprendizaje automático, Expedientes Electrónicos
Editor jefe:	Ana Carolina Sepúlveda Vildósola
Comentarios:	
Financiación:	No
Conflicto de intereses:	No
Carta de solicitud:	Carta_solicitud.pdf
DOI:	--------
Cronologia:	24-05-2025 En proceso de creación 24-05-2025 Artículo nuevo para validar 30-05-2025 Pendiente de que el autor complete el artículo 03-06-2025 Artículo completado por autor 16-06-2025 Pendiente de que el autor complete el artículo 16-06-2025 Artículo completado por autor 24-06-2025 Enviado a editor, pendiente de asignación de revisor 17-07-2025 Asignado a revisor, pendiente de que acepte la invitación 25-07-2025 Pendiente de dictamen de revisor

Agradecimientos:

Al Instituto Mexicano del Seguro Social por haber otorgado beca completa por 24 meses durante la realización del programa de Doctorado.

A los miembros del Comité Tutoral por su guía y apoyo académico durante la duración del programa de Doctorado.

INDICE GENERAL

Índice de Tablas 5
Índice de Figuras 7
Índice de Ecuaciones 9
Resumen 11
Abstract 13
1. Presentación 15
1.1 Problema 15
1.2 Justificación 17
1.2.1 Relevancia Médica 17
1.2.2 Relevancia Académica 18
1.2.3 Relevancia para el Instituto Mexicano del Seguro Social 19
1.3 Hipótesis 19
1.3.1 De trabajo 19
1.3.2 Nula 19
1.4 Objetivo 20
1.4.1 General 20
1.4.2 Específicos 20
1.5 Organización del Documento 21
2. Marco Teórico 23
2.1 Conceptos Generales 23
2.1.1 Inteligencia 23
2.1.2 Inteligencia artificial 23
2.1.3 Historia de la inteligencia artificial 24
2.1.4 Algoritmo 25
2.1.5 Aprendizaje Automático 25
2.1.6 Algoritmos de aprendizaje automático 26
2.2 Métodos para Establecer un Clasificador 26
2.2.1 Modelar directamente una regla de clasificación 27
2.2.2 Modelar la probabilidad con base en los datos de entrada 28
2.2.3 Hacer un modelo probabilístico de datos dentro de cada clase 28
2.3 Regla de Clasificación de la Estimación del Máximo a Posteriori (MAP) 29
2.4 Elección de Algoritmos de Aprendizaje Automático 30
2.5 Clasificador ingenuo de Bayes 30
2.5.1 Teoría básica 31
2.5.2 Atributos discretos y continuos en el clasificador ingenuo de Bayes 32
2.5.3 Ventajas 33
2.5.4 Desventajas 34
2.6 k-vecinos más cercanos 35
2.6.1 Teoría básica 35
2.6.2 Ejemplos del uso de K vecinos más cercanos 36
2.6.3 De un solo vecino a k vecinos 37
2.6.4 Ventajas 38
2.6.5 Desventajas 39
2.7 Arboles de Decisión 39
2.7.1 Teoría básica 40

2.7.2 Ejemplo de uso de árboles de decisión. 42
2.7.3 Ventajas 44
2.7.4 Desventajas 44
2.8 Regresión logística. 45
2.8.1 Antecedentes. Regresión Lineal 45
2.8.1.1 Ejemplo del uso de regresión lineal simple 47
2.8.1.2 Ventajas 49
2.8.1.3 Desventajas 49
2.8.2 Método de máxima verosimilitud aplicado a la regresión lineal 49
2.8.3 Teoría básica. Regresión Logística 51
2.8.3.1 Ventajas 51
2.8.3.2 Desventajas 51
2.9 Máquinas de vectores de soporte 52
2.9.1 Teoría básica 52
2.9.1.1 Cálculo de la Distancia de un Punto a la Frontera 54
2.9.1.2 Problema de Optimización Primal 54
2.9.1.3 Casos No Linealmente Separables y el Uso de Kernels 55
2.9.2 Ventajas 55
2.9.3 Desventajas 55
2.10 Redes neuronales 56
2.10.1 Antecedentes. El perceptrón 56
2.10.1.1 Modelo matemático 56
2.10.1.2 Algoritmo de aprendizaje del perceptrón 57
2.10.1.3 Limitaciones del perceptrón 57
2.10.2 Función de Activación Sigmoidea 58
2.10.3 Pérdida por Error Cuadrático Medio 58
2.10.4 Pase hacia Adelante 59
2.10.5 Retropropagación del Error 59
2.10.6 Teoría básica. Red Neuronal Artificial 60
2.10.6.1 Proceso de Entrenamiento de una Red Neuronal 61
2.10.7 Ventajas 62
2.10.8 Desventajas 62
2.11 Análisis discriminante 63
2.11.1 Teoría básica 63
2.11.1.1 Construcción de la función discriminante 65
2.11.2 Regla de decisión 65
2.11.2.1 Caso bidimensional 65
2.11.3 Ventajas 66
2.11.4 Desventajas 66
2.12 Cópulas 67
2.12.1 Teoría básica 67
2.12.2 Desarrollo del paquete MLCOPULA 69
2.12.2.1 Definiciones 69
2.12.2.2 Desarrollo del Paquete 69
2.12.3 Envío a CRAN 70
2.12.4 Aprobación y Publicación 70
2.12.5 Descripción del Paquete 70

2.12.5.1 Valor Devuelto ... 71
2.12.5.2 Funcionalidad ... 71
2.12.6 Ventajas ... 71
2.12.7 Desventajas ... 72
3. Conceptos Médicos ... 73
3.1 Enfermedad Vascular Cerebral ... 73
3.1.1 Factores de Riesgo para Enfermedad Vascular Cerebral ... 73
3.1.1.1 Factores de Riesgo No Modificables ... 73
3.1.1.2 Factores de Riesgo Modificables ... 75
3.1.1.2.1 Hipertensión Arterial ... 75
3.1.1.2.2 Dislipidemia ... 76
3.1.1.3 Alcohol ... 76
3.1.1.4 Diabetes ... 77
3.1.1.5 Tabaquismo ... 77
3.1.2 Valores séricos y hematológicos como factores de riesgo cardiovascular ... 77
3.2 Algoritmos de aprendizaje automático en la predicción de la anatomía del arco aórtico. 80
3.2.1 Trombectomía mecánica ... 81
3.2.1.1 Asociación del tiempo a la reperfusión con la eficacia ... 81
3.3 Anatomía vascular cervical ... 83
3.3.1 Alteraciones Morfológicas ... 84
4. Propuesta Metodológica ... 87
4.1 Diseño ... 87
4.2 Universo de Trabajo ... 87
4.3 Fuentes de información del estudio ... 87
4.4 Lugar donde se desarrolló el estudio ... 87
4.5 Aspectos Éticos ... 88
4.6 Descripción general del estudio ... 88
4.6.1 Primera fase ... 88
4.6.1.1 Conformación del conjunto de datos ... 88
4.6.1.2 Selección y tamaño de la muestra ... 89
4.6.1.3 Parámetros para el cálculo ... 89
4.6.1.4 Criterios de inclusión y exclusión ... 91
4.6.1.4.1 Criterios de inclusión ... 91
4.6.1.4.2 Criterios de no inclusión ... 92
4.6.1.4.3 Criterios de exclusión ... 92
4.6.1.5 Variables independientes ... 92
4.6.1.5.1 Demográficas: ... 92
4.6.1.5.2 Antropométricas ... 92
4.6.1.5.3 Clínicas ... 93
4.6.1.5.4 De laboratorio ... 95
4.6.1.6 Variable dependiente ... 98
4.6.2 Preprocesamiento de los datos ... 100
4.6.3 Segunda fase ... 100
4.6.3.1 Análisis del conjunto de datos ... 100
4.6.3.1.1 Lenguaje R ... 100
4.6.3.1.2 Preprocesamiento de los datos ... 101

4.6.3.1.3 Clasificación 102
4.6.4 Tercera fase 104
4.6.4.1 Comparación del desempeño 104
4.6.4.1.1 Métricas de desempeño 104
4.6.4.1.2 Aclaración acerca de la codificación y cálculo de métricas 106
4.6.4.1.3 Análisis Estadístico 106
5. Resultados 107
5.1 Características del conjunto de datos 107
5.2 Balance de clases en el conjunto de datos 107
5.3 Atributos 107
5.4 Identificación de predictores de la anatomía vascular favorable 108
5.5 Desempeño de los Clasificadores 109
5.6 Desempeño individual 113
5.7 Desempeño general 114
5.8 Abordaje del desbalance de clases 115
5.9 Desempeño posterior al abordaje del desbalance de clases 116
5.10 Desempeño individual 120
5.10.1 Regresión Logística 120
5.10.2 Clasificador Ingenuo de Bayes 121
5.10.3 K-vecinos más cercanos 121
5.10.4 Árboles de decisión 122
5.10.5 Análisis Discriminante 123
5.10.6 Red Neuronal 124
5.10.7 Máquinas de Vectores de Soporte 125
5.10.8 Cópula Gaussiana 126
6. Discusión 129
6.1 Limitaciones de la tesis 135
7. Conclusiones 137
7.1 Trabajo Futuro 138
Bibliografía 140
Anexo A. Galería de imágenes de Angiotomografía de vasos supraórticos de los pacientes estudiados 159
Anexo B. Código en lenguaje R utilizado para los análisis 198

Índice de Tablas

Tabla 1 Subcampos de la Inteligencia Artificial 25
Tabla 2 Tabla de datos. Ejemplo 1 de aplicación K vecinos más cercanos. 37
Tabla 3 Tabla de datos. Ejemplo de aplicación Arboles de decisión 42
Tabla 4 Ejemplo 1 de aplicación de regresión lineal 47
Tabla 5 Clasificación TICI 82
Tabla 6 Características relacionadas con acceso difícil a los vasos supra aórticos 84
Tabla 7 Características relacionadas con acceso difícil a los vasos cervicales 85
Tabla 8 Matriz de confusión para el cálculo de los resultados finales 104
Tabla 9 Atributos y valores incluidos en el conjunto de datos. 107
Tabla 10 Modelo final, predictores asociados a anatomía vascular cervical favorable 109
Tabla 11 Métricas de desempeño obtenidas después de la validación cruzada de 5 iteraciones. 109
Tabla 12 Comparación del tiempo de computación por clasificador. 113
Tabla 13 Métricas de desempeño obtenidas por los clasificadores después de validación cruzada de 5 iteraciones y corrección del desbalance. 116

Índice de Figuras

Figura 1 Representación gráfica del proceso de clasificación por k-vecinos más próximos 36
Figura 2 Ejemplo 2 de aplicación de k-NN 38
Figura 3 Representación gráfica del árbol de decisión del ejemplo 1. 44
Figura 4 Representación gráfica del error 46
Figura 5 Grafica de la ecuación lineal del ejemplo 1 48
Figura 6 Representación gráfica de la función logística 51
Figura 7 Representación gráfica de la clasificación por máquinas de vectores de soporte 52
Figura 8 Ejemplo 1 Máquinas de vectores de soporte 53
Figura 9 Representación gráfica de una red neuronal artificial de 3 capas 61
Figura 10 Representación gráfica del análisis discriminante 64
Figura 11 : Función de densidad de la cópula gaussiana con parámetro ρ = 0,5 68
Figura 12 Morfología normal y anormal de la arteria carótida interna. 58 84
Figura 13 Imagen de angiotomografía de vasos supra aórticos.[164] 89
Figura 14 Clasificación del tipo de arco aórtico.[164] 99
Figura 15 Ejemplos de imágenes de angiotomografía. 99
Figura 16 Esquema general de la Fase 2 del estudio. 103
Figura 17 Exactitud obtenida por los clasificadores estudiados. 111
Figura 18 Exactitud balanceada obtenida por los clasificadores estudiados. 111
Figura 19 Área bajo la curva ROC obtenida por los clasificadores. 112
Figura 20 Índice F1 obtenido por los clasificadores. 112
Figura 21 Orden del desempeño de los clasificadores. 113
Figura 22 Área bajo la curva ROC posterior al ajuste del desbalance 118
Figura 23 Orden del desempeño posterior al ajuste del desbalance. 119
Figura 24 Desempeño Final de la Regresión Logística 120
Figura 25 Desempeño Final del Clasificador Ingenuo de Bayes 121
Figura 26 Desempeño Final de k vecinos más próximos 122
Figura 27 Desempeño Final de los Árboles de Decisión 123
Figura 28 Desempeño Final del Análisis Discriminante 124
Figura 29 Desempeño Final de la Red Neuronal 125
Figura 30 Desempeño Final de las Máquinas de Vectores de Soporte 126
Figura 31 Desempeño Final de la Cópula Gaussiana 127

Índice de Ecuaciones

Ecuación 1 Regla de Clasificación de la Estimación del Máximo a Posteriori 29
Ecuación 2 Entropía 41
Ecuación 3 Impureza de Gini 41
Ecuación 4 Error de Clasificación 41
Ecuación 5 Ganancia 42
Ecuación 6 Forma general de las ecuaciones lineales 45
Ecuación 7 Sistema de ecuaciones lineales 46
Ecuación 8 Cálculo de pendiente e intercepto. 47
Ecuación 9 Exactitud 104
Ecuación 10 Sensibilidad/*Recall* 104
Ecuación 11 Especificidad 105
Ecuación 12 Valor predictivo positivo/Precisión 105
Ecuación 13 Valor predictivo negativo 105
Ecuación 14 F1 105
Ecuación 15 Exactitud Balanceada 105
Ecuación 16 Tasa de Falsos Positivos 106

Resumen

El uso de aprendizaje automático en datos clínicos ha ganado interés en los últimos años, especialmente en tareas de clasificación supervisada aplicadas a la medicina. Sin embargo, su implementación enfrenta retos importantes cuando se busca predecir características estructurales internas a partir de variables clínicas periféricas. Esta tesis tuvo como objetivo analizar el desempeño de ocho algoritmos de aprendizaje automático para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical, una variable habitualmente determinada por imagenología.

Se trabajó con un conjunto de datos clínicos estructurados, afectado por un notable desbalance entre clases (relación 1:3), lo cual se abordó mediante la técnica SMOTE. Los algoritmos evaluados incluyeron: Regresión Logística, Bayes Ingenuo, k-vecinos más cercanos, Máquinas de Vectores de Soporte, Árboles de Decisión, Análisis Discriminante Regularizado, Redes Neuronales y Cópula Gaussiana. Se evaluaron ocho métricas de desempeño mediante validación cruzada estratificada con cinco iteraciones.

Los resultados revelaron diferencias estadísticamente significativas entre clasificadores en la mayoría de las métricas. Algunos modelos como Análisis Discriminante Regularizado y Cópula Gaussiana destacaron por su exactitud, mientras que k-vecinos más cercanos y Redes Neuronales mostraron un mejor equilibrio entre sensibilidad, F1 y exactitud balanceada. La Regresión Logística y las Máquinas de Vectores de Soporte obtuvieron los desempeños más bajos en sensibilidad. En todos los algoritmos, la exactitud balanceada fue menor que la exactitud bruta, reflejando el impacto del desbalance de clases.

Estos hallazgos subrayan la importancia de utilizar múltiples métricas para evaluar clasificadores en contextos clínicos. Asimismo, sugieren que no existe un modelo único óptimo, y que la selección del algoritmo debe guiarse por las prioridades clínicas específicas del problema, como la necesidad de minimizar falsos negativos.

Abstract

The use of machine learning in clinical data analysis has gained increasing attention, particularly for supervised classification tasks in medical applications. However, its implementation faces significant challenges when attempting to predict internal anatomical features using peripheral clinical variables. This thesis aimed to evaluate the performance of eight machine learning algorithms in classifying patients with hypertension, diabetes, or dyslipidemia according to their cervical vascular anatomy, a variable typically assessed through imaging studies.

The study utilized a structured clinical dataset characterized by notable class imbalance (3:1 ratio), which was addressed using the SMOTE technique. The evaluated algorithms included Logistic Regression, Naive Bayes, k-Nearest Neighbors, Support Vector Machines, Decision Trees, Regularized Discriminant Analysis, Neural Networks, and Gaussian Copula. Eight performance metrics were assessed using stratified cross-validation.

The results showed statistically significant differences among classifiers across most metrics. Some models, such as Regularized Discriminant Analysis and Gaussian Copula, stood out in terms of accuracy, while k-Nearest Neighbors and Neural Networks demonstrated better balance across sensitivity, F1 score, and balanced accuracy. Logistic Regression and Support Vector Machines yielded the lowest sensitivity scores. All classifiers showed a drop in balanced accuracy compared to raw accuracy, highlighting the effect of class imbalance.

These findings emphasize the importance of evaluating multiple performance metrics in clinical classification problems. They also suggest that no single model is universally optimal, and that algorithm selection should be guided by specific clinical priorities, such as minimizing false negatives.

1. Presentación

Este documento de tesis explora el análisis del desempeño de distintos métodos de clasificación aplicados a la predicción de la anatomía vascular cervical en pacientes con hipertensión, diabetes, o dislipidemia. En el contexto de la medicina, el uso de aprendizaje automático (AA) ha demostrado ser útil para el análisis y clasificación de datos complejos, aunque enfrenta retos como la baja prevalencia de ciertas enfermedades y el desequilibrio entre clases en los conjuntos de datos médicos.

El objetivo principal del estudio es comparar la efectividad de ocho algoritmos de AA en la clasificación de la anatomía vascular cervical, lo que podría mejorar la comprensión y el manejo clínico de los pacientes con factores de riesgo cardiovascular. Para lograr este objetivo, se aplicaron y analizaron algoritmos como Bayes Ingenuo, Regresión Logística, k-vecinos más cercanos, Máquinas de Vectores de Soporte, Árboles de Decisión, Análisis Discriminante Regularizado, Redes Neuronales y Cópula Gaussiana.

1.1 Problema

La resolución de problemas de clasificación constituye un tópico relevante en medicina. En el contexto de un acto médico los problemas de clasificación se presentan en múltiples escenarios clínicos durante el curso de un proceso de salud y enfermedad, por ejemplo:[1]

- Anticipación de la aparición de una enfermedad (prevención)
- Presencia de una enfermedad en base a síntomas (diagnóstico)
- Respuesta a un tratamiento (eficacia)
- Anticipación de un desenlace de salud (pronóstico)

Históricamente, estas tareas han sido resueltas en medicina a través de la regresión logística como herramienta para establecer la relación entre la información conocida de un paciente (input) y su desenlace desconocido (output). Así, la regresión logística ha permitido el desarrollo de múltiples modelos predictivos que proveen a los médicos con reglas de decisión precisas para clasificar a los pacientes bajo su cuidado [2].

Sin embargo, en los últimos años, el avance de la tecnología en imagen médica y de dispositivos de captura de datos de salud ha producido grandes cantidades de información, sin que hasta ahora sea posible extraer conclusiones prácticas para la medicina clínica.

Desde su introducción en 1961 por Confield, Gordon y Smith, la regresión logística ha sido la herramienta estadística más utilizada para el análisis de datos en clínica y epidemiología.[3] A pesar de los recientes avances en el uso de técnicas de aprendizaje automático [4] se puede hipotetizar alternativamente, que el desempeño de la regresión logística como clasificador no será inferior al de las técnicas de aprendizaje automático. Lo anterior no significa necesariamente que estas técnicas no representen un avance de suprema importancia para la medicina sino que como la misma literatura lo demuestra las aplicaciones más importantes en medicina son en el reconocimiento de patrones y en la interpretación de imágenes médicas.[5]

Tanto en medicina como en otras áreas no es raro que las herramientas estadísticas no realicen las predicciones con la exactitud óptima. Lo anterior puede deberse a la utilización del modelo incorrecto, falta de predictores adecuados o tamaños de muestra reducidos. Si bien la falta de predictores y el tamaño de las muestras son problemas que pueden resolverse al incrementar el volumen de datos recogidos, elegir la herramienta estadística óptima para realizar la clasificación no es tan sencillo. Debido a lo anterior, para obtener conclusiones clínicamente relevantes, se ha confiado tradicionalmente en métodos de estadística multivariable y de programas estadísticos comerciales. Sin embargo, dentro del ámbito de las matemáticas y las ciencias de la computación, los problemas de clasificación se pueden abordar desde perspectivas adicionales a la estadística, como por ejemplo a través de la inteligencia artificial. [6]

En épocas recientes, con el propósito de resolver problemas de la vida real, se han desarrollado algunos algoritmos de aprendizaje automático.[7] Estos algoritmos de aprendizaje automático tienen la capacidad de autocorregirse por lo que tienen el potencial de mejorar con el tiempo al explorar cantidades de datos cada vez mayores con mínima supervisión humana.

El problema que se aborda en esta tesis se plantea tomando en cuenta la necesidad de un método que permita la correcta clasificación de la anatomía vascular cervical de pacientes con factores de riesgo cardiovascular del tipo hipertensión, diabetes o dislipidemia. Ya que los avances en computación han permitido el desarrollo de nuevos métodos de clasificación basados en aprendizaje automático, se puede teorizar que estos nuevos métodos aplicados a los problemas de clasificación en medicina podrían tener un mejor desempeño que el abordaje tradicional a través de la regresión logística. El problema esbozado anteriormente se puede enunciar en la forma de la siguiente pregunta de investigación:

- En pacientes con hipertensión, diabetes o dislipidemia, ¿Cuál es el desempeño de distintos métodos de clasificación aplicados a la predicción de la anatomía vascular cervical?

1.2 Justificación

1.2.1 Relevancia Médica

La publicación de los estudios MR CLEAN (Multicenter Randomized Clinical trial of Endovascular treatment for Acute ischemic stroke in the Netherlands), ESCAPE (Endovascular treatment for Small Core and Anterior circulation Proximal occlusion with Emphasis on minimizing CT to recanalization times), SWIFT PRIME (Solitaire™ FR With the Intention For Thrombectomy as Primary Endovascular Treatment for Acute Ischemic Stroke), EXTEND-IA (Extending the Time for Thrombolysis in Emergency Neurological Deficits — Intra-Arterial) y REVASCAT (Randomized Trial of Revascularization with Solitaire FR Device versus Best Medical Therapy in the Treatment of Acute Stroke Due to Anterior Circulation Large Vessel Occlusion Presenting within Eight Hours of Symptom Onset) en 2015 cambio por completo el panorama terapéutico de los pacientes con infarto cerebral agudo [8-12].

A partir de entonces, la terapia endovascular se ha convertido en la intervención más efectiva para lograr disminuir la discapacidad derivada del infarto cerebral.

Sin embargo, la adopción de las técnicas endovasculares como el nuevo estándar de tratamiento para el infarto cerebral agudo ha creado nuevos retos, especialmente concernientes a la nueva infraestructura y a los costos que se requieren para su aplicación

de forma generalizada [13]; sin embargo, las barreras que constituyen el costo y la creación de infraestructura para proveer de tratamiento endovascular a los pacientes que así lo requieren han sido lentamente abordados y resueltos particularmente en países industrializados.

No obstante, conforme aumenta el número de procedimientos endovasculares que se llevan a cabo en el mundo, han surgido a su vez nuevos problemas que interfieren con el éxito técnico del procedimiento en aquellos pacientes que tienen acceso al mismo [14].

1.2.2 Relevancia Académica.

En el ámbito académico y de generación de conocimiento en el área de estadística computacional, a través del Centro de Ciencias Básicas de la Universidad Autónoma de Aguascalientes, en este proyecto se pretende estudiar el desempeño de múltiples algoritmos de clasificación tanto probabilísticos (regresión logística, clasificador ingenuo de Bayes, análisis discriminante y copulas) como no probabilísticos (k-vecinos más próximos, redes neuronales, máquinas de vectores de soporte, arboles de decisión). [15]

El propósito final de lo anterior es generar nuevos conocimientos que permitan seleccionar los mejores métodos para clasificar pacientes basándose en datos clínicos. Lo anterior permitiría la selección del abordaje endovascular más adecuado a la anatomía de cada paciente, antes de la realización de un procedimiento de terapia endovascular, conduciendo a disminución en los tiempos de atención y a la reducción en el costo del procedimiento al disminuir el número de dispositivos utilizados por procedimiento.

Adicionalmente, y enfocándose particularmente en la línea de investigación de estadística computacional, en el presente proyecto se incluirá un enfoque novedoso mediante la utilización de cópulas como clasificadores.[16]

Este abordaje es novedoso no solo en su aplicación para el área médica, [17] sino también dentro del área de la estadística computacional, [18] por lo que se espera que los productos científicos que se deriven de este proyecto tengan un impacto significativo sobre la comunidad de investigadores teóricos en el área.

1.2.3 Relevancia para el Instituto Mexicano del Seguro Social

A nivel local y nacional, el Instituto Mexicano del Seguro Social (IMSS), busca coordinar acciones en los tres niveles de atención para reducir la discapacidad y mortalidad derivada del infarto cerebral. Mediante el Protocolo de Atención Integral[19] se trabaja en la prevención, control de factores de riesgo, diagnóstico y tratamiento oportuno con la administración de terapias de reperfusión aguda y a partir de 2022 el IMSS lanzó el programa "Código Cerebro" para coordinar acciones multidisciplinarias en los tres niveles de atención, prevenir, diagnosticar y tratar oportunamente, de esta manera, mejorar el tiempo de respuesta del personal médico en casos de infarto cerebral.

El programa "Código Cerebro" pretende reducir la mortalidad por infarto cerebral, así como la discapacidad y reducir los estados de discapacidad posteriores, mediante una atención oportuna y la aplicación adecuada de procesos médicos y de terapia endovascular, para mejorar la calidad de vida de los pacientes. En base a lo anterior, el presente proyecto presenta una oportunidad para optimizar el funcionamiento del Protocolo de Atención Integral del IMSS a través del programa "Código Cerebro" lo cual puede redundar en una mejoría en los tiempos de atención y costos para el IMSS.[20]

1.3 Hipótesis

1.3.1 De trabajo

El desempeño de los algoritmos de aprendizaje automático aplicados para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical será diferente al compararlos entre sí.

1.3.2 Nula

El desempeño de los algoritmos de aprendizaje automático aplicados para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical no será diferente al compararlos entre sí.

1.4 Objetivo

1.4.1 General

Analizar el desempeño de ocho algoritmos de aprendizaje automático aplicados para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.

1.4.2 Específicos

- Caracterizar epidemiológicamente a los pacientes atendidos en el Hospital General de Zona #2 del IMSS, OOAD Aguascalientes con hipertensión, diabetes o dislipidemia.
- Determinar los factores predictores que permitan clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño de la regresión logística para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño de k-vecinos más próximos para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño de máquinas de vectores de soporte para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño de redes neuronales para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño de análisis discriminante para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño de cópulas para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Determinar el desempeño del clasificador ingenuo de Bayes para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.

- Determinar el desempeño de árboles de decisión para clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical.
- Comparar el desempeño de los distintos algoritmos de aprendizaje automático entre sí.

1.5 Organización del Documento

El documento se organiza de la siguiente manera: se inicia con un marco teórico que contextualiza la importancia del AA en medicina y explica los algoritmos seleccionados. Posteriormente, se describen los materiales y métodos empleados en el estudio, incluyendo la conformación del conjunto de datos y los procedimientos de análisis. Los resultados presentan una comparación detallada del desempeño de los algoritmos, y finalmente, la discusión y conclusiones abordan las implicaciones de los hallazgos y proponen futuras líneas de investigación.

2. Marco Teórico

2.1 Conceptos Generales

2.1.1 Inteligencia

La inteligencia se define como una mezcla de habilidades que se utilizan para desenvolverse en el medio ambiente[21] e incluyen:

1. Capacidad de entender o comprender.
2. Capacidad de resolver problemas.
3. Conocimiento, comprensión, acto de entender.
4. Habilidad, destreza y experiencia. [22]

Las habilidades anteriores permiten sacar conclusiones acerca de lo aprendido previamente y resolver problemas nuevos.

2.1.2 Inteligencia artificial

El termino inteligencia artificial se puede definir como: [23]

1. Capacidad de los robots, computadoras y otras máquinas de exhibir competencias parecidas a las humanas para razonar y resolver problemas. Se contrasta con la inteligencia que exhiben los humanos y otros animales, denominada inteligencia humana e inteligencia animal, respectivamente.
2. Teoría y desarrollo de sistemas computacionales que son capaces de realizar tareas que normalmente requieren inteligencia humana, tales como percepción visual, reconocimiento del lenguaje, toma de decisiones y traducción entre lenguajes.
3. Combinación de ciencias de la computación, fisiología y filosofía con el fin de facilitar el entendimiento y la realización de tareas inteligentes por parte de la tecnología. [24]

La inteligencia artificial permite a las máquinas aprender de la experiencia, ajustarse a nuevas circunstancias y llevar a cabo tareas humanas. Actualmente los usos prácticos de la inteligencia artificial dependen de manera muy importante de técnicas de aprendizaje profundo y procesamiento natural del lenguaje.[25] Estas técnicas han permitido que la

inteligencia artificial tenga en nuestros días, múltiples aplicaciones que van desde los algoritmos de búsqueda en internet hasta aplicaciones militares como las armas autónomas. [26]

2.1.3 Historia de la inteligencia artificial

La inteligencia artificial tiene una larga historia que data de la antigüedad, ya que existen menciones de robots inteligentes y seres artificiales en la mitología.[27] Sin embargo, el primer ejemplo claro de una "máquina pensante" lo constituye la visión del matemático británico Charles Babbage, quien en 1830 describió su "máquina inteligente". [28] Babbage, quien es conocido típicamente como el padre de la computación, intentó diseñar una máquina analítica para realizar cálculos que permitieran la creación de tablas de navegación y la lectura de símbolos distintos a los números. Sin embargo, dicha máquina nunca se construyó. [29]

El concepto de utilizar computadoras para simular el comportamiento inteligente y el pensamiento crítico fue descrito por Alan Turing en 1950. [30] En su libro Computadoras e Inteligencia, describió una prueba simple, la cual posteriormente sería conocida como la prueba de Turing, para determinar si las computadoras eran capaces de tener inteligencia humana. [31] Seis años después, en 1956 durante la conferencia de Darmouth expertos como Marvin Minsky, John McCarthy, Allen Newell, Herbert Simon y otros; afirmaron que: "cada aspecto del aprendizaje o de cualquier otra característica de la inteligencia puede ser descrito de forma tan precisa que se puede construir una máquina que lo simule";[32] y acuñaron el termino de inteligencia artificial definida como la ciencia e ingeniería dedicada a crear maquinas inteligentes. [33] Posterior a esta conferencia donde la inteligencia artificial obtuvo su nombre y su misión, el desarrollo de la inteligencia artificial ha dado lugar a múltiples aplicaciones prácticas en áreas tan diversas como la economía, la mercadotecnia y la medicina.

Si bien la inteligencia artificial comenzó como una serie simple de instrucciones condicionales del tipo: si, entonces; con el paso del tiempo se ha desarrollado para incluir

algoritmos complejos que llevan a cabo tareas de forma parecida al cerebro humano y al igual que en medicina, existen especialidades dentro de la inteligencia artificial (
Tabla 1).

Tabla 1 Subcampos de la Inteligencia Artificial

Aprendizaje Automático, identificación y análisis de patrones. Mejora la experiencia adquirida en conjuntos de datos.
Aprendizaje profundo, compuesto de redes neuronales multicapa. Permite a las maquinas tomar decisiones por sí mismas.
Procesamiento Natural del Lenguaje, permite a las computadoras extraer datos del lenguaje humano y tomar decisiones basándose en esa información.
Visión por computadora permite a las computadoras obtener información y entender el ambiente a través de imágenes o videos.

2.1.4 Algoritmo

Robin K. Hill define algoritmo como: “una estructura de control compuesta finita, abstracta, efectiva e imperativamente creada para lograr un propósito bajo ciertas provisiones”.[34] La definición anterior se justifica teóricamente en elementos de las ciencias de la computación y de la filosofía pero resulta impráctica para las ciencias aplicadas. Debido a lo anterior se ha propuesto una definición planteada en elementos de aplicación directa:
Un algoritmo (para que un ejecutor E logre una meta G) es: Un procedimiento o método (P), por ejemplo, un conjunto (o secuencias) finitos de afirmaciones (o reglas o instrucciones) de tal forma que cada afirmación (A) está compuesta por un numero finito de símbolos de un alfabeto finito y sin ambigüedades para el ejecutor. Ejemplo: E sabe cómo realizar A, A puede realizarse en un tiempo finito y después de lograr A, E sabe cuál es el siguiente paso. P toma un tiempo finito y se detiene una vez que se alcanza G. [35]

2.1.5 Aprendizaje Automático

El aprendizaje automático, es una disciplina derivada de la inteligencia artificial que fusiona métodos estadísticos con informática para elaborar algoritmos capaces de clasificar muestras, predecir resultados y realizar inferencias en base a la información que se les proporciona previamente como entrenamiento.[36] Los modelos analíticos creados a partir de algoritmos permiten a la computadora "aprender" de los datos que se les proporcionan. El aprendizaje automático ocurre cuando el programa se cambia a si mismo de forma que pueda obtener mejores resultados en el futuro.[37]

2.1.6 Algoritmos de aprendizaje automático

El aprendizaje automático es un subconjunto de la inteligencia artificial y es una técnica que sirve para entrenar computadoras/sistemas para que realicen tareas de forma independiente sin necesidad de ser programadas explícitamente. Durante el proceso de entrenamiento y aprendizaje, se utilizan varios algoritmos los cuales ayudan al sistema a mejorar su auto entrenamiento a lo largo del tiempo, a estos algoritmos se les llama por lo tanto algoritmos de aprendizaje automático.[38]

Los algoritmos de aprendizaje automático trabajan dentro de un marco de referencia de tres modelos de aprendizaje generalizado los cuales constituyen esencialmente los tipos de aprendizaje automático:[36]

- Aprendizaje supervisado. Se despliega en los casos en los que los datos pueden ser etiquetados como pertenecientes a una base de datos específica.
- Aprendizaje no supervisado. Se implementa en caso de que exista dificultad para etiquetar la pertenencia de los datos o sus conexiones implícitas a una base de datos específica.
- Aprendizaje reforzado. Selecciona una acción basada en cada punto de datos y, después de eso, aprende qué tan buena fue la acción.

2.2 Métodos para Establecer un Clasificador

En el ámbito del aprendizaje automático y la estadística, los clasificadores se pueden establecer utilizando diversos métodos. Estos métodos se pueden categorizar en tres enfoques principales:

2.2.1 Modelar directamente una regla de clasificación

Este método implica la creación de reglas que asignan directamente los datos de entrada a una clase específica. Los algoritmos que siguen este enfoque suelen buscar una regla de decisión que divida el espacio de características en regiones, cada una asociada con una clase particular. [39] Los modelos discriminativos se centran en aprender la probabilidad condicional $P(C|X)$, donde C es la clase y X representa las características de entrada. Este enfoque se orienta a encontrar la frontera de decisión entre las clases, en lugar de modelar la distribución completa de los datos.[40] Algunos ejemplos notables de este enfoque incluyen:

- k-vecinos más cercanos: Pertenece a este tipo de clasificación ya que asigna una clase a un nuevo punto de datos basándose en las clases de los vecinos más cercanos, sin asumir ninguna distribución subyacente.
- Árboles de Decisión: Clasifica los datos al dividirlos recursivamente basándose en reglas de decisión, enfocándose directamente en aprender las fronteras de decisión.
- Perceptrón: Clasifica mediante una combinación lineal de características, aprendiendo una regla de decisión que separa las clases en el espacio de características.
- Máquinas de Vectores de Soporte: Encuentra un hiperplano óptimo para separar las clases, enfocándose en maximizar el margen entre las clases, sin modelar explícitamente las probabilidades.
- Análisis Discriminante: Pertenece a este tipo porque busca directamente la combinación lineal de características que mejor separa las clases, sin modelar las probabilidades de las clases.

- Regresión Logística: Clasifica al modelar la relación entre las características y la clase como una función logística, pero sin modelar explícitamente la distribución de los datos.

Este enfoque es conocido como clasificación discriminatoria porque el modelo se enfoca directamente en aprender las fronteras de decisión que separan las diferentes clases en el conjunto de datos.[41]

2.2.2 Modelar la probabilidad con base en los datos de entrada

Este enfoque implica la creación de modelos que calculan la probabilidad de que un punto de datos pertenezca a una clase específica, dados los datos de entrada.[42] El modelo genera probabilidades para cada clase y asigna el punto de datos a la clase con la mayor probabilidad. Un ejemplo de la probabilidad condicional se expresa como $P(C|X)$, que es directamente modelada por este tipo de algoritmos. Algunos ejemplos de este enfoque incluyen:

- Redes Neuronales (Perceptrón Multicapa con el Costo de Entropía Cruzada): Pertenece a este tipo de clasificación porque ajusta los pesos para calcular probabilidades precisas de pertenencia a una clase, modelando directamente la probabilidad condicional.
- Regresión Logística: Aunque puede clasificarse como un modelo discriminatorio, la regresión logística también modela la probabilidad de pertenencia a una clase dada una entrada específica, lo que le permite asignar probabilidades a las diferentes clases.

Este enfoque también se considera clasificación discriminatoria, ya que el modelo se enfoca en aprender a distinguir entre clases al modelar las probabilidades condicionales directamente.

2.2.3 Hacer un modelo probabilístico de datos dentro de cada clase

Este enfoque implica modelar la distribución de datos dentro de cada clase por separado. Luego, se utiliza el Teorema de Bayes para calcular la probabilidad de que un nuevo punto de datos pertenezca a una clase específica.[42] Los modelos generativos modelan la

probabilidad $\mathrm{P}(\mathrm{C}|\mathrm{X})$, que es la probabilidad de los datos de entrada dado que pertenecen a una clase específica, y utilizan la probabilidad de la clase $\mathrm{P}(\mathrm{C})$ para calcular la probabilidad posterior mediante la regla de Bayes. Ejemplos de este enfoque incluyen:

- Bayes Ingenuo: Es un ejemplo de este tipo de clasificación porque modela la probabilidad de los datos dentro de cada clase, asumiendo independencia entre las características, y luego calcula la probabilidad posterior utilizando el Teorema de Bayes.
- Clasificadores Basados en Modelos: Este tipo de clasificadores modela explícitamente la distribución de los datos para cada clase, utilizando diferentes distribuciones, lo que les permite ser considerados como ejemplos de clasificación generativa.
- Clasificadores Basados en Funciones Cópula: Pertenece a este tipo porque utilizan cópulas para modelar la dependencia entre variables dentro de cada clase, permitiendo estimar la distribución conjunta de las variables para cada clase, y así calcular las probabilidades posteriores para la clasificación.

Este enfoque se denomina clasificación generativa porque el modelo se enfoca en aprender la distribución generativa de los datos para cada clase y luego utiliza esta información para realizar la clasificación.

2.3 Regla de Clasificación de la Estimación del Máximo a Posteriori (MAP)

La regla MAP se utiliza tanto en modelos discriminativos como en generativos. Esta regla asigna una nueva observación x a la clase c^* si la probabilidad posterior $P(C = c^*|X = x)$ es mayor que para cualquier otra clase c.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \propto P(X|C)P(C)$$

Ecuación 1 Regla de Clasificación de la Estimación del Máximo a Posteriori

Esto significa que se necesita conocer la distribución de los datos $P(X|C)$ y la probabilidad de la clase $P(C)$para aplicar la regla MAP.

En resumen, los métodos: 2.2.1 reglas de clasificación directa y 2.2.2 modelado de la probabilidad condicional, son ejemplos de clasificación discriminatoria, donde el objetivo es aprender las fronteras o probabilidades que distinguen entre clases. [40] El método 2.2.3 modelado de datos dentro de cada clase, es un ejemplo de clasificación generativa, donde se modela cómo se generan los datos dentro de cada clase.

Tanto el enfoque 2.2.2 como el 2.2.3 son ejemplos de clasificación probabilística, ya que ambos implican calcular probabilidades, ya sea de pertenencia a una clase o de la distribución de datos dentro de las clases.[43]

2.4 Elección de Algoritmos de Aprendizaje Automático

Tanto la regresión logística como los demás algoritmos de aprendizaje automático propuestos son excelentes instrumentos de clasificación y resolución de problemas de regresión. Sin embargo, con el objetivo de disminuir el tiempo de computación es importante saber cuándo utilizar cada uno. Generalmente, se sugiere que primero se intente utilizar la regresión logística para ver que tan bien se desempeña el modelo y si falla se pueden utilizar los otros algoritmos basándose en las características del set de datos que se quiera clasificar. Estas características incluyen el grado de aleatoriedad de las relaciones entre las variables, sus coeficientes de correlación, el tamaño de la muestra, entre otros. Tomando como estándar en medicina a la regresión logística, experiencias previas han mostrado en múltiples escenarios que los algoritmos de aprendizaje automático tienden a manifestar una eficiencia similar, pero bajo circunstancias particulares alguno puede ser más potente que otro dependiendo de las características del conjunto de datos.[44]

2.5 Clasificador ingenuo de Bayes

El clasificador ingenuo de Bayes es un algoritmo de clasificación supervisada ampliamente utilizado en el campo del aprendizaje automático. Se basa en el teorema de Bayes y asume que todas las características del conjunto de datos son independientes entre sí, una suposición que rara vez se cumple en la práctica, pero que aun así permite obtener buenos resultados en muchas aplicaciones. [45] Es especialmente útil cuando se trabaja con

grandes volúmenes de datos y categorías bien definidas. [46] Debido a su simplicidad, eficiencia y facilidad de implementación, el clasificador ingenuo de Bayes se ha convertido en una herramienta fundamental en el área de aprendizaje automático. [47]

2.5.1 Teoría básica

El clasificador ingenuo de Bayes se fundamenta en el teorema de Bayes, una expresión de probabilidad condicional que permite actualizar la probabilidad de una hipótesis a la luz de nueva evidencia. Dado un conjunto de características $X = (x_1, x_2, \dots, x_n)$ y una clase C_k perteneciente al espacio de clases $C = \{C_1, C_2, \dots, C_n\}$, el objetivo del clasificador es estimar la probabilidad posterior $P(C_k|\, X)$, es decir, la probabilidad de que la observación pertenezca a la clase C_k dado el vector de características X.

El teorema de Bayes establece que:

$$P(C_k|\, X) = \frac{P(X|C_k) \cdot P(C_k)}{P(X)}$$

Donde:

- $P(C_K)$ es la probabilidad a priori de la clase C_k
- $P(X|C_k)$es la verosimilitud, es decir, la probabilidad de observar X dado que la clase es C_k
- $P(X)$ es la evidencia, común a todas las clases y por lo tanto irrelevante para la clasificación
- $P(C_k|X)$ es la probabilidad posterior, la cantidad que se debe maximizar

El clasificador asignará a la instancia X la clase con mayor probabilidad posterior.

$$\hat{C} = arg \max_{C_k \in C} P(C_k|X)$$

Aplicando el teorema de Bayes y omitiendo el denominador $P(X)$, constante para todas las clases, se obtiene:

$$\hat{C} = arg \max_{C_k \in C} P(C_k) \cdot P(X|C_k)$$

Asumiendo independencia condicional, es decir se asume que las características $x_1, x_2, \dots, x\_n)$ son condicionalmente independientes dado C_k. Bajo esta suposición, la verosimilitud se factoriza como:

$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

Por tanto, la regla de clasificación se reduce a:

$$\hat{C} = arg \max_{C_k \in C} P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k)$$

Esta fórmula permite estimar las probabilidades a partir de datos etiquetados, generalmente mediante frecuencias relativas en el conjunto de entrenamiento.

Aunque el supuesto de independencia rara vez se cumple estrictamente en problemas reales, el clasificador ingenuo de Bayes suele ofrecer un desempeño competitivo, sobre todo en contextos donde las relaciones entre variables son débiles o el ruido es significativo.

2.5.2 Atributos discretos y continuos en el clasificador ingenuo de Bayes

En contextos donde las variables predictoras son discretas, la estimación de probabilidades condicionales mediante frecuencias relativas es una estrategia eficaz, siempre y cuando se cuente con un conjunto de entrenamiento suficientemente grande y balanceado. Sin embargo, esta aproximación pierde validez cuando el número de posibles combinaciones de características es muy alto o cuando algunas combinaciones no aparecen en los datos, lo que lleva a problemas de sobreajuste o a la estimación de probabilidades nulas. Este último escenario es particularmente problemático, ya que, al calcular el producto de probabilidades condicionales, una sola probabilidad cero anula toda la expresión.

Sin embargo, cuando los atributos son continuos, la noción de frecuencia relativa deja de ser aplicable, ya que la probabilidad de observar exactamente un valor real particular es prácticamente cero. En este caso, se recurre a la función de densidad de probabilidad (fdp), denotada como $f(x)$, que describe la distribución de probabilidad de una variable continua. Aunque $f(x)$, no representa una probabilidad en sí misma, su integral en un intervalo da la

probabilidad de que la variable caiga dentro de dicho rango. Formalmente, para una variable continua X, a probabilidad de que su valor se encuentre entre a y b se define como:

$$Pa \leq X \leq b) = \int_{b}^{a} f(x)\, dx$$

En el contexto del clasificador ingenuo de Bayes, el uso de una fdp permite modelar la verosimilitud condicional $P(x_i|C_k)$ como una densidad, típicamente bajo el supuesto de distribución normal:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{k,i}^2}}\, exp\left(-\frac{\left(x_i - \mu_{k,i}\right)^2}{2\sigma_{k,i}^2}\right)$$

Donde $\mu_{k,i}$ y $\sigma_{k,i}^2$ on la media y varianza estimadas del atributo $x_i$en la clase C_k. Esto puede entenderse como un proceso de discretización por partición del espacio de valores reales. Al dividir el dominio de una variable continua en intervalos finitos, y calcular frecuencias en cada uno, se aproxima la densidad subyacente. Al refinar indefinidamente estos intervalos, el límite del cociente entre la frecuencia relativa en cada subintervalo y su ancho tiende a la función de densidad $f(x)$ Así, el uso de una fdp en dominios continuos se interpreta como una generalización del enfoque por frecuencias aplicado a variables discretas, permitiendo extender el clasificador a dominios con atributos discretos y continuos.

2.5.3 Ventajas

Una de sus principales fortalezas es su eficiencia computacional, ya que tanto el entrenamiento como la clasificación de nuevas instancias pueden realizarse de forma muy rápida, con complejidad lineal respecto al número de características y al tamaño del conjunto de entrenamiento. [47]

Otra ventaja destacada es su robustez frente al sobreajuste, particularmente en dominios con ruido o cuando el número de características supera ampliamente el número de observaciones. Esto se debe a que el modelo se basa en una estructura probabilística simple y no intenta ajustar de manera exacta la complejidad de los datos. Además, el clasificador ingenuo de Bayes requiere pocos datos para estimar sus parámetros, especialmente cuando

se utilizan atributos discretos o se modelan atributos continuos bajo supuestos paramétricos como la normalidad. [45] También, cuando existe alta dimensionalidad, como la clasificación de texto o el análisis de datos genómicos, este algoritmo suele mantener un desempeño competitivo, incluso frente a modelos más complejos, debido a que maneja eficientemente la independencia entre atributos y no requiere selección o reducción de variables como paso previo. [48]

Finalmente, su interpretabilidad es otra ventaja importante particularmente en áreas médicas, dado que los cálculos se basan en probabilidades explícitas y el modelo es transparente en su funcionamiento, resulta sencillo identificar qué atributos influyen más en la clasificación y cómo se combinan. [49]

2.5.4 Desventajas

La principal desventaja radica en su supuesto de independencia condicional entre las características dado el valor de la clase. En la práctica, esta suposición rara vez se cumple, especialmente en dominios donde existen correlaciones estructurales entre atributos, como en imágenes médicas, señales fisiológicas o variables clínicas interdependientes. [50]

Otra limitación es que, en presencia de atributos continuos que no siguen una distribución normal, la aproximación mediante una función de densidad gaussiana puede resultar inadecuada y conducir a estimaciones sesgadas de las probabilidades condicionales.

Si bien es posible emplear otras familias de distribuciones o incluso técnicas no paramétricas como el estimador de Parzen, esto compromete la simplicidad y eficiencia del algoritmo. [51]

El clasificador ingenuo de Bayes también puede verse afectado por el problema de probabilidades nulas, que ocurre cuando una combinación específica de clase y valor de atributo no está presente en el conjunto de entrenamiento. En tales casos, la probabilidad posterior se anula completamente, lo que puede llevar a errores de clasificación graves. Si bien este problema puede mitigarse mediante técnicas de suavizamiento (por ejemplo, Laplace). [52]

Finalmente, su incapacidad para modelar interacciones complejas entre variables lo limita en tareas donde el desempeño depende críticamente de la combinación no lineal de múltiples atributos, como ocurre en algunos modelos de percepción visual, procesamiento de lenguaje natural avanzado o diagnósticos médicos con múltiples escalas. [50]

2.6 k-vecinos más cercanos

Se trata de un abordaje no paramétrico que se usa para la clasificación y regresión. Es una de las técnicas de aprendizaje automático más sencillas. También se le conoce como modelo de aprendizaje perezoso, por su uso de aproximaciones locales. [53]

Dada una muestra de prueba, encontrar las k muestras de entrenamiento más cercanas en función de cierta métrica de distancia y luego usar estos k vecinos para hacer predicciones. Por lo general, para problemas de clasificación, se puede usar la votación para predecir la etiqueta de clase más frecuente en los k vecinos; para problemas de regresión, se puede usar el promedio para predecir la muestra de prueba como el promedio de los valores reales de k-desenlaces.

Además, las muestras pueden ser ponderadas por las distancias, de forma que a una muestra más cercana se le asigna un mayor peso.

2.6.1 Teoría básica

La lógica fundamental detrás de k-vecinos más cercanos (k-NN) es la de explorar la vecindad de cada punto de datos asumiendo que son comparables para extraer un resultado.

En k-NN (Figura 1) se busca un pronóstico en base a los datos que se encuentran en la vecindad. En el caso de la clasificación, se utiliza un voto de pluralidad sobre los k puntos de datos más cercanos, mientras que la media de los k puntos de datos más cercanos se calcula como la salida en la regresión.

Como regla general, se seleccionan números impares como k. El método k-NN es un modelo de aprendizaje considerado lento, ya que el tiempo de ejecución de la computación del resultado es amplio. Lo anterior se debe a la naturaleza no paramétrica del método, ya que,

a diferencia de los métodos paramétricos que se basan en un modelo, en k-NN no existe un modelo. [54]

2.6.2 Ejemplos del uso de K vecinos más cercanos.

La similitud entre dos pasteles se puede establecer contando los atributos en que se diferencian: cuantas menos diferencias, mayor la semejanza. La primera fila en la Tabla 2 muestra los valores de los atributos del objeto x.

Las filas que siguen muestran los valores de los atributos para cada uno de los doce ejemplos de entrenamiento, la columna más a la derecha especifica el número de diferencias en los valores de atributo del ejemplo dado y x.

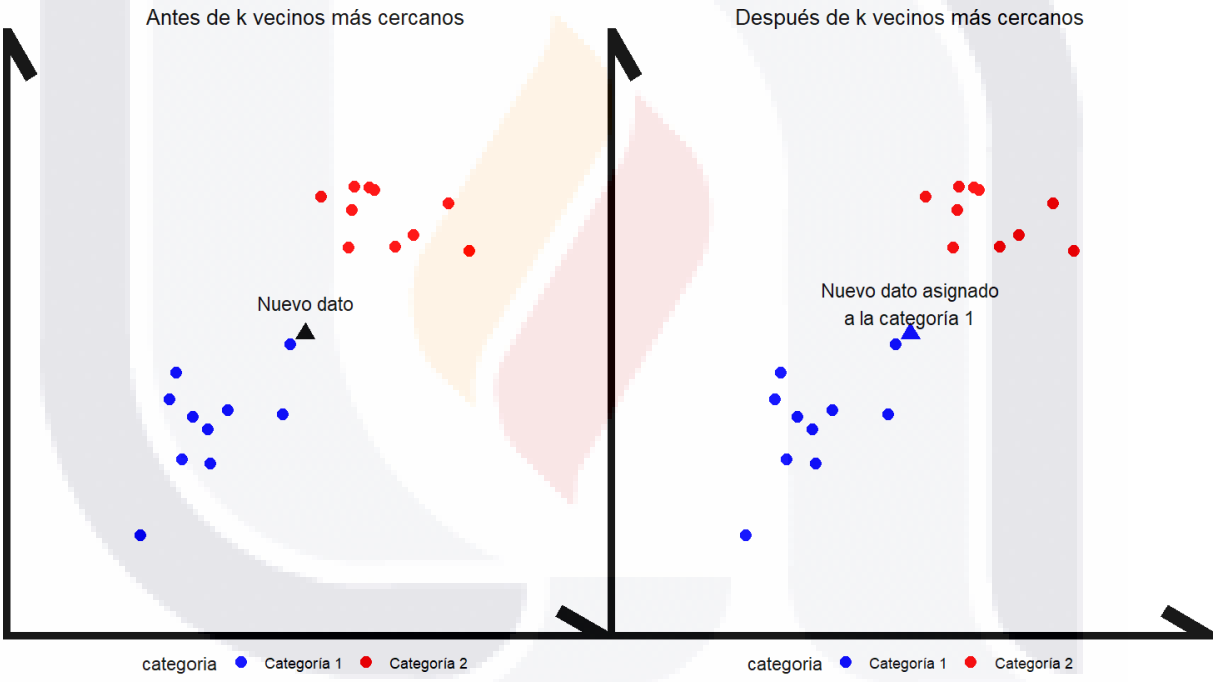


Figura 1 Representación gráfica del proceso de clasificación por k-vecinos más próximos

El valor más pequeño que se encuentra en el caso de ex5, concluimos que este es el ejemplo de entrenamiento más similar a x, y, por lo tanto, la clase de x debe etiquetarse como clase positiva (pos). En la Tabla 2, todos los atributos son discretos, pero si se tratara de atributos continuos, el proceso es el mismo. Dado que cada ejemplo se puede representar por un punto en un espacio n-dimensional, podemos usar la distancia euclidiana o alguna otra

fórmula geométrica; y, de nuevo, cuanto más pequeña sea la distancia, mayor semejanza. Esto, por cierto, es cómo el clasificador k-NN obtuvo su nombre: el ejemplo de entrenamiento con la distancia más pequeña de x en el espacio de instancia es, geométricamente hablando, el vecino más cercano de x.

Tabla 2 Tabla de datos. Ejemplo 1 de aplicación K vecinos más cercanos.

		Corteza		Relleno			
Ejemplo	Forma	Tamaño	Color	Tamaño	Color	Clase	# de diferencias
x	Cuadrado	Grueso	Gris	Delgado	Blanco	?	–
ex1	Círculo	Grueso	Gris	Grueso	Oscuro	pos	3
ex2	Círculo	Grueso	Blanco	Grueso	Oscuro	pos	4
ex3	Triangulo	Grueso	Oscuro	Grueso	Gris	pos	4
ex4	Círculo	Delgado	Blanco	Delgado	Oscuro	pos	4
ex5	Cuadrado	Grueso	Oscuro	Delgado	Blanco	pos	1
ex6	Círculo	Grueso	Blanco	Delgado	Oscuro	pos	3
ex7	Círculo	Grueso	Gris	Grueso	Blanco	neg	2
ex8	Cuadrado	Grueso	Blanco	Grueso	Gris	neg	3
ex9	Triangulo	Delgado	Gris	Delgado	Oscuro	neg	3
ex10	Círculo	Grueso	Oscuro	Grueso	Blanco	neg	3
ex11	Cuadrado	Grueso	Blanco	Grueso	Oscuro	neg	3
ex12	Triangulo	Grueso	Blanco	Grueso	Gris	neg	4

Al contar el número de diferencias entre doce ejemplos de entrenamiento, ex5 es el más similar a x.

2.6.3 De un solo vecino a k vecinos

En dominios ruidosos, no se puede confiar en un único vecino más cercano. ¿Qué pasa si su etiqueta de clase es incorrecta debido al ruido? Un enfoque más robusto identificará no uno, sino varios vecinos más cercanos, y permitirá que voten. Esta es la esencia del clasificador k-NN, donde k es el número de vecinos votantes (generalmente un parámetro especificado por el usuario). La Figura 2 resume el enfoque. Nótese que, cuando se aplica

un clasificador 4-NN a un dominio de 2 clases, puede resultar en una situación donde dos vecinos son positivos y dos negativos. En este evento, no está claro cómo clasificarlo.

Los empates de este tipo se evitan utilizando para k un número impar. En dominios con más de dos clases, sin embargo, un número impar de vecinos más cercanos no ayuda. Por ejemplo, un clasificador 7-NN puede darse cuenta de que tres de los vecinos pertenecen a la clase C1, tres vecinos a la C2 y un vecino a la C3. En la situación anterior, se debe escoger de antemano el mecanismo para escoger entre clases empatadas.

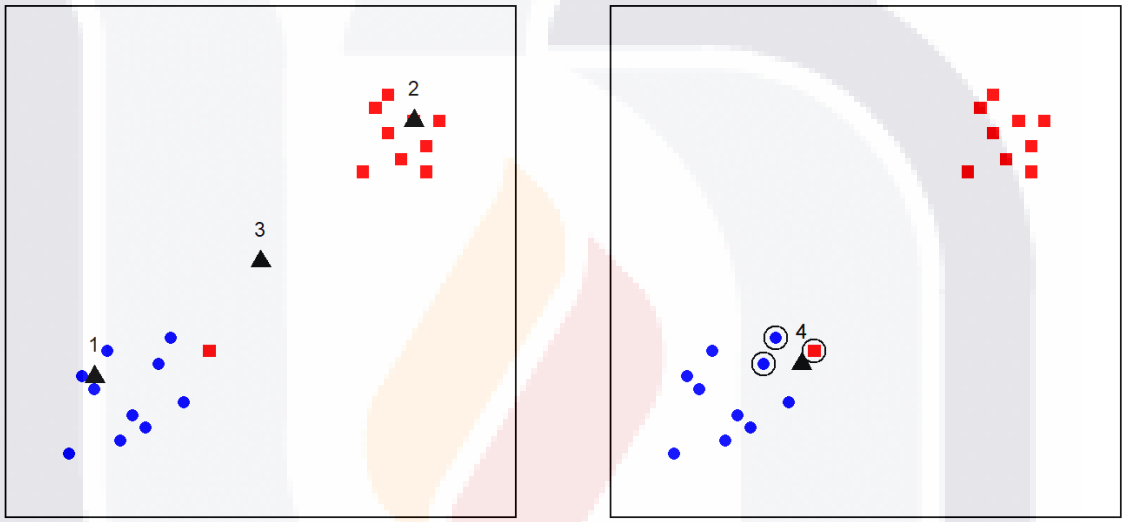


Figura 2 Ejemplo 2 de aplicación de k-NN

En la Figura 2, en la gráfica de la izquierda, los objetos 1 y 2 se encuentran en lo profundo de las áreas de "círculos" y "cuadrados", respectivamente, y por lo tanto son fáciles de clasificar. El objeto 3 se encuentra en la región límite entre las dos clases, y su clase es, por lo tanto, incierta. En el dominio de ruido de clase de la derecha, el clasificador 1-NN clasificará incorrectamente el objeto 4, pero el error se corrige si se usa el clasificador 3-NN.

2.6.4 Ventajas

k-NN es un modelo rápido y directo de aprendizaje automático. Cuenta con múltiples parámetros individualizables.[55] En comparación con otros métodos de aprendizaje automático, k-NN tiene algo único: no tiene un proceso de entrenamiento explícito. Es un

representante de aprendizaje perezoso, en el que simplemente se almacenan los datos de entrenamiento y no se hace nada con ellos hasta que se reciben las muestras de prueba.

2.6.5 Desventajas

La principal desventaja de k-vecinos más próximos es la dificultad para escoger el valor de k. Otras desventajas incluyen el requerimiento de una escala adecuada para un tratamiento equivalente de cada punto de datos y que los tiempos de computación son grandes si la muestra es muy numerosa.[55]

La elección del parámetro k juega un papel importante, ya que diferentes valores de k pueden conducir a resultados de clasificación muy diferentes. Además, diferentes cálculos de distancia también pueden conducir a un "vecindario" significativamente diferente, y en consecuencia, diferentes resultados de clasificación.

Los clasificadores de vecinos más cercanos están algo desactualizados y, como tal, ya rara vez se usan. k-NN es propenso a sufrir por escasez de datos, por atributos irrelevantes y por escalado de atributos inapropiados.

La elección concreta depende de los requisitos específicos de la aplicación que se requiera.

2.7 Arboles de Decisión

Un árbol de decisión es un clasificador simple en forma de estructura de árbol jerárquico, que realiza una clasificación supervisada utilizando una estructura ramificada dirigida con una serie de preguntas. Las preguntas se colocan en nodos de decisión; cada prueba evalúa el valor de un atributo particular (característica) del patrón (objeto) y proporciona una división binaria o multidireccional.

El nodo inicial se conoce como nodo raíz, que es considerado el padre de todos los demás nodos. Las ramas corresponden a las posibles respuestas. Se visitan nodos de decisión sucesivos hasta que se accede a un nodo terminal o hoja, donde se lee (asigna) la clase (categoría).

La clasificación se realiza siguiendo una ruta desde el nodo raíz hasta llegar a un nodo hoja. La estructura del árbol no es fija *a priori,* sino que el árbol crece y se ramifica durante el aprendizaje dependiendo de la complejidad del problema.

2.7.1 Teoría básica

Se puede pensar en la información como la reducción de la incertidumbre, y los atributos informativos serán los que resulten en la mayor reducción de dicha incertidumbre.[56] El contenido de información de un único estado de mensaje en unidades de información está dado por:

$$I(E) = log * \frac{1}{P(E)} = -logP(E)$$

Donde, $P(E)$ es la probabilidad previa de aparición del mensaje. Intuitivamente, la cantidad de información que transporta un mensaje está inversamente relacionada con la probabilidad de que ocurra. Los mensajes con alta probabilidad de ocurrir contienen poca información; por el contrario, los mensajes menos esperados contienen la mayor cantidad de información.[57]

Si sólo son posibles dos eventos $(0\ y\ 1)$, la base del logaritmo es 2 y la unidad de información resultante es el bit. Si los dos eventos son igualmente probables, se transmite 1 bit de información cuando ocurre uno de los dos posibles eventos igualmente probables. Sin embargo, si los dos eventos posibles no son igualmente probables, entonces la información transmitida por el evento menos común es mayor que la transmitida por el evento más común.

La entropía es una medida del desorden o imprevisibilidad de un sistema. (Se utiliza para variables discretas, mientras que la varianza es la métrica de variable continua). Dada una clasificación binaria C y un set de ejemplos S, la distribución de clase en cualquier nodo puede ser escrita como (p_0, p_1), donde $p_1 = 1 - p_0$ y la entropía H, de S, es la suma de la información.

$$H(S) = -p_0 \log_2 p_0 - p_1 \log_2 p_1$$

Si el atributo da como resultado una clasificación que separa los ejemplos en $(0.5, 0.5)$, la entropía (incertidumbre) de esa característica es máxima (igual a 1). Este no es un atributo útil. Si otro atributo divide los ejemplos en $(0.6, 0.4)$, la entropía relativa a esta nueva clasificación es $-(0.6)\ log_2(0.6)\ -(0.4) \log_2(0.4) = 0.97$. Si todos los ejemplos de prueba de un tercer atributo son de la misma clase [es decir, la división es $(0, 1)$ o $(1, 0)$,

entonces la entropía (incertidumbre) de esa característica es cero y proporciona una buena clasificación. Se puede considerar que la entropía describe la cantidad de impureza en un conjunto de características en un nodo. Cuanto menor sea el grado de impureza, más sesgada será la distribución de clases (y más útil será el nodo). Por ejemplo, un nodo con distribución de clases $(0, 1)$ tiene impureza cero (y entropía cero) y es un buen clasificador; mientras que un nodo con distribución de clases uniforme $(0.5, 0.5)$ tiene la impureza más alta (y entropía = 1) y es un clasificador inútil. En el caso general, el atributo objetivo puede tomar C valores diferentes (es decir, a división multidireccional) y la entropía de S relativa a esta clasificación basada en C está dada por:

$$H(p) = \sum_{i=1}^{c} p_1 \log_2 p_i$$

Ecuación 2 Entropía

Donde $p_i$es la proporción de S que pertenece a la clase i. La base del logaritmo es 2 ya que se mide la entropía en bits; y la entropía máxima posible relativa a este atributo es $\log_2 c$. Otras medidas de impureza que se pueden utilizar para determinar la mejor manera de dividir una serie de registros incluyen la impureza de Gini y el error de clasificación:

$$Gini(p) = 1 - \sum_{i} p_i^2$$

Ecuación 3 Impureza de Gini

$$error\ de\ clasificación(p) = 1 - max(p_i)$$

Ecuación 4 Error de Clasificación

La impureza de Gini es en realidad la tasa de error esperada si la etiqueta de clase se selecciona aleatoriamente de la distribución de clases presente. Las tres medidas alcanzan un valor máximo para una distribución uniforme $(p = 0.5)$ y un mínimo cuando todos los ejemplos pertenecen a la misma clase $(p = 0\ o\ 1)$.[56]

Para determinar el mejor atributo a elegir para cada nodo de decisión del árbol, la medida utilizada es la ganancia, que es la reducción esperada de impureza causada al dividir los ejemplos según este atributo. Más precisamente, la ganancia, $Ganancia(S, A)$, de un atributo A, relativa a una colección de muestras S, se define como:

$$Ganancia(S, A) = Impureza(S) - \sum_{i=1}^{k} \frac{|S_{vi}|}{|S|} Impureza(S_{vi})$$

Ecuación 5 Ganancia

donde el atributo A tiene un conjunto de valores $v_1, v_2, v_3, \ldots, v_k$, y el número de ejemplos dentro de S con el valor vi es $|Svi|$. El primer término es solo la impureza de la colección original S y el segundo término es el valor esperado de la impureza después de dividir S usando el atributo A. El segundo término es simplemente la suma de las impurezas de cada subconjunto S_{vi}, ponderadas por la fracción de ejemplos que pertenecen a S_{vi}. Si se utiliza la entropía como medida de impureza, entonces la ganancia se conoce como ganancia de información.[57]

2.7.2 Ejemplo de uso de árboles de decisión.

Hay cuatro cosas que a una persona le gusta hacer por la noche: ir a un pub, ver televisión, ir a una fiesta o estudiar. A veces, la elección ya está hecha: si tiene una tarea que entregar al día siguiente, necesita estudiar; si se siente perezoso, entonces el pub no es una opción, y si no hay fiesta, no puede ir a una. Esta persona está buscando un árbol de decisiones que le ayude a decidir qué hacer cada noche. A continuación, se muestra una lista de todo lo que ha hecho en los últimos 10 días. Ejemplo tomado de Dougherty, G. (2012).[58]

Tabla 3 Tabla de datos. Ejemplo de aplicación Arboles de decisión

¿Fecha Límite?	¿Fiesta?	¿Perezoso?	Actividad
Urgente	Sí	Sí	Fiesta
Urgente	No	Sí	Estudiar
Cercana	Sí	Sí	Fiesta
No	Sí	No	Fiesta
No	No	Sí	Pub
No	Sí	No	Fiesta
Cercana	No	No	Estudiar
Cercana	No	Sí	TV
Cercana	Sí	Sí	Fiesta
Urgente	No	No	Estudiar

Pseudocódigo para resolución del ejemplo de árbol de decisión

1. Calcular la entropía total:
   a. Inicializar un vector **`e`** de longitud igual al número de clases en los datos.
   b. Contar la frecuencia de cada clase en el conjunto de datos.
   c. Para cada clase, calcular su contribución a la entropía:
      - e[i] = (frecuencia de clase / total de instancias) * log2(frecuencia de clase / total de instancias) * -1
   d. Sumar todas las contribuciones para obtener la entropía total.
2. Definir una función para calcular la entropía de una variable:
   a. Inicializar vectores auxiliares para almacenar los cálculos intermedios.
   b. Para cada valor de la variable y cada clase:
      i. Calcular la proporción de cada clase para ese valor de la variable.
      ii. Calcular la contribución de cada clase a la entropía para ese valor.
      iii. Sumar las contribuciones de ese valor y guardarlas.
   c. Sumar las entropías calculadas para cada valor de la variable y devolver la entropía total.
3. Tabular cada variable con respecto a la clase objetivo:
   a. Crear tablas de contingencia entre cada variable y la clase objetivo.
4. Identificar el nodo raíz:
   a. Evaluar las tablas de las variables para encontrar si alguna tiene un único valor asociado a una clase específica.
   b. Si alguna variable tiene un valor con un solo desenlace posible, elegirla como nodo raíz.
5. Calcular la entropía para un valor específico de una variable:
   a. Filtrar los datos para el valor de interés de la variable.
   b. Calcular la entropía de ese valor usando la función definida anteriormente.
6. Calcular las ganancias de información:
   a. Para el valor de la variable en cuestión, calcular la ganancia de información restando la entropía de las otras variables.
      i. ganancia = entropía del valor de la variable - entropía de las otras variables.
7. Seleccionar la siguiente variable:
   a. Comparar las ganancias de información.
   b. Seleccionar la variable con la mayor ganancia de información como el siguiente nodo del árbol.
8. Dividir los datos según el valor de la variable seleccionada:
   a. Tabular los datos de la variable seleccionada con respecto a la clase objetivo.
   b. Si algún valor tiene una sola clase asociada, marcar ese nodo como hoja y detener la división para ese valor.
   c. Continuar con los valores que no tengan un desenlace único, aplicando las mismas reglas.
9. Evaluar las variables restantes:
   a. Para cada subconjunto de datos resultante, repetir el proceso de calcular entropías, ganancias de información y seleccionar variables hasta que se lleguen a nodos hoja.
10. Finalizar el árbol de decisión:
    a. Una vez que todas las ramas llevan a una clase específica, el árbol está completo.

La representación gráfica de este árbol se muestra en la Figura 3.

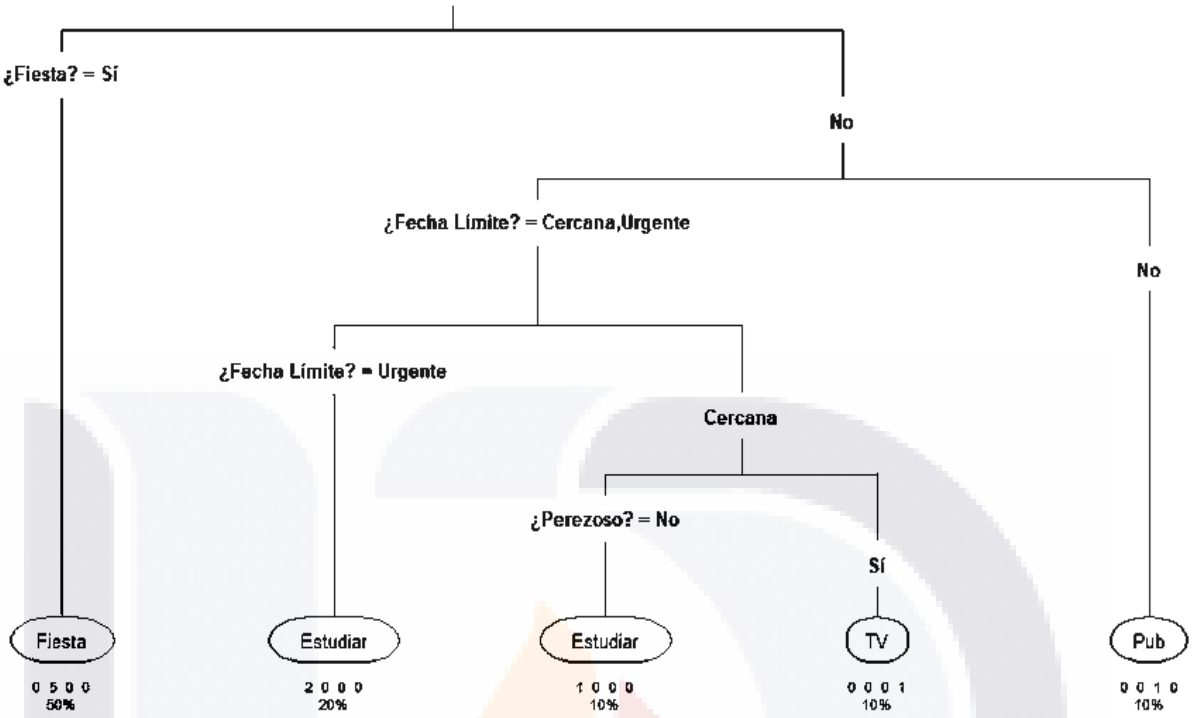


Figura 3 Representación gráfica del árbol de decisión del ejemplo 1.

2.7.3 Ventajas

Se puede utilizar datos categóricos (no continuos), incluidos datos nominales sin orden natural (aunque también puede adaptarse para utilizar datos cuantitativos).

Clara interpretabilidad, proporcionando una forma natural de incorporar conocimientos previos (al convertir las pruebas en expresiones lógicas).

Una vez construidos, requieren muy poco poder de computación.

2.7.4 Desventajas

Existe un número exponencial de árboles de decisión que pueden ser construidos a partir de un conjunto dado de características, pero encontrar el árbol óptimo no es computacionalmente factible.

El clasificador depende de los algoritmos utilizados para crear o "hacer crecer" el árbol de decisión.

Los árboles de decisión son razonablemente precisos, pero siempre subóptimos

Los algoritmos suelen emplear una estrategia codiciosa que crece el árbol utilizando el atributo (característica) más informativo en cada paso, pero no permiten retroceder en el árbol. El atributo más informativo será el que divida el conjunto que llega al nodo en los subconjuntos más homogéneos.

2.8 Regresión logística.

La regresión logística se considera el primer y principal algoritmo de clasificación aplicado a la medicina.[59] Cuando se habla de regresión, se habla en realidad de un modelo de clasificación el cual enmarca un modelo de salida dicotómico.[60] Este modelo utiliza una función logística.

El resultado de la regresión logística será una probabilidad $(0 \leq P(x) \leq 1)$, la cual se puede adoptar para predecir un resultado binario como cero o uno (si, $P(x) < 0.5$, resultado = 0, de lo contrario resultado = 1.[2]

2.8.1 Antecedentes. Regresión Lineal

La regresión lineal se refiere a la técnica matemática de ajustar datos dados a una función de cierto tipo, principalmente líneas rectas.

La Ecuación 6 muestra la forma general de las ecuaciones lineales:

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b$$

Ecuación 6 Forma general de las ecuaciones lineales

donde: $a_1, a_2, \ldots, a_n, x_1, x_2, \ldots, x_n, \beta \in R$. Los números a_i y β son conocidos, a_i se llama coeficiente que multiplica a x_i y a β se le llama lado derecho. Los números x_i son desconocidos y se les llama variables. La meta consiste en encontrar un conjunto de números $x_1, \ldots, x_n$ que satisfagan la ecuación. Un sistema de ecuaciones lineales es un conjunto de ecuaciones lineales.

Si se asume un set de datos $(x1, y1), (x2, y2), \ldots, (xn, yn)$, donde: $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$, son todos números conocidos. A cada par (xi, yi) para $(1 \leq i \geq n)$ se le llama punto.

La meta de la regresión lineal simple es encontrar los números a_i y β tales que la línea $y = ax + \beta$ aproxime mejor los datos. Cualquier línea recta está dada por una ecuación de la forma $y = ax + \beta$, donde a y β son números llamados respectivamente: la pendiente a y el intercepto-y β. La pendiente a y el intercepto-y β de la línea que mejor se aproxima a los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n y_n)$ son las soluciones al siguiente sistema de dos ecuaciones lineales (Ecuación 7):

$$(x_1^2 + x_2^2 + \cdots + x_n^2)a + (x_1 + x_2 + \cdots + x_n)\beta = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$
$$(x_1 + x_2 + \cdots + x_n)a + n\beta = y_1 + y_2 + \cdots + y_n$$

Ecuación 7 Sistema de ecuaciones lineales

La forma en la que se sabe que la línea creada por estas soluciones es la que mejor se aproxima a los datos es a través del método de los mínimos cuadrados. Como se muestra en la Figura 4, la diferencia vertical entre la línea y cada punto de datos representa una desviación de la media estimada (en la línea) y los valores reales observados de y para cada valor específico de x. A esta desviación se le llama error.

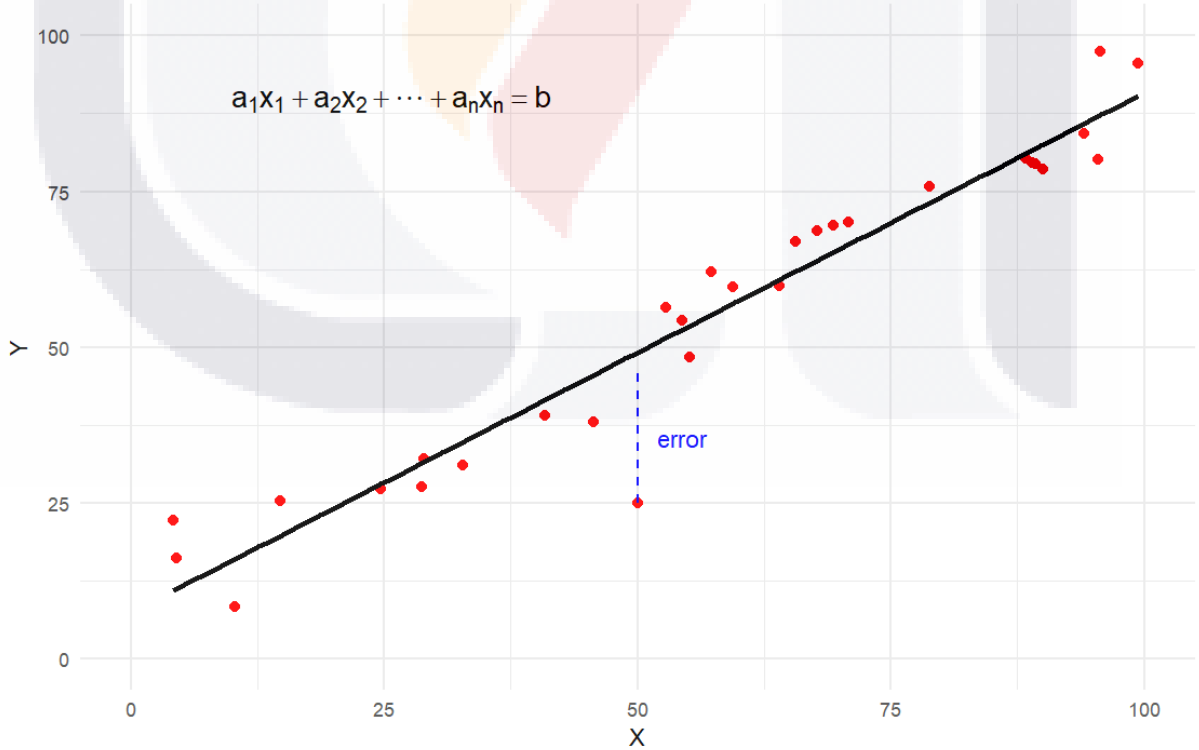


Figura 4 Representación gráfica del error.

La distribución teórica de los errores se asume como normal. La media de errores en un subgrupo es cero y la varianza del grupo se establece a un valor, por ejemplo, s^2.

Los errores se elevan al cuadrado y se suman para obtener la suma de los errores cuadrados. El método tiene como objetivo minimizar la suma al cuadrado de los errores y obtener los valores de la pendiente y del intercepto que cumplan el propósito. El mejor ajuste se refiere a que la línea estimada debe mostrar que se encuentra más cerca de los puntos de datos observados. Específicamente, se usan los primeros derivados de la suma de errores al cuadrado y se establece el valor en cero para obtener la pendiente y el intercepto que producen los valores mínimos de suma de errores al cuadrado (Ecuación 8).

$$\beta_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Ecuación 8 Cálculo de pendiente e intercepto.

2.8.1.1 Ejemplo del uso de regresión lineal simple.

Se asume que una agencia solo vende autos del mismo año y modelo, por lo que en teoría el precio de venta solo depende del kilometraje de los autos. Se adquiere un nuevo automóvil con 55,000 km y se nos hace la pregunta de cuál debe ser el precio de venta. Responder esta pregunta seria sencillo si se tuviera una función $y = f(x)$ que calcule el precio de venta y en función del kilometraje x. En este caso solo se reemplazaría x por 55,000 en la fórmula $y = f(x)$ para obtener el precio de venta. Sin embargo, no se cuenta con la función $f(x)$ y de hecho dicha función puede no existir. Con lo que sí se cuenta son las experiencias pasadas. Se asume que previamente se han vendido 5 automóviles cuyos kilometrajes y precios de venta se resumen en la Tabla 4:

Tabla 4 Ejemplo 1 de aplicación de regresión lineal

Kilómetros	Precio
20,000	13,000
30,000	11,000
40,000	11,500
60,000	9,500
70,000	10,000

Entonces la pregunta se vuelve: Dada la experiencia previa resumida en la Tabla 4, ¿A qué precio se debería vender un automóvil con 55,000 km? Aplicando los principios de la regresión lineal:

$$(x_1^2 + x_2^2 + \cdots + x_n^2)a + (x_1 + x_2 + \cdots + x_n)\beta = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Queda:

$$1.14 * 10^{10}a + 220000\beta = 2.32 * 10^{09},$$

y para

$$(x_1 + x_2 + \cdots + x_n)a + n\beta = y_1 + y_2 + \cdots + y_n$$

$$220000\ a\ +\ 5\beta\ =\ 55000$$

La solución a este sistema de ecuaciones es:

$$a = -0.05735 \quad \beta = 13{,}580$$

al reemplazar x por 55,000, se obtiene el precio de venta y:

$$y = -0.05735(55{,}000) + 13{,}580 = 10{,}426$$

Lo cual puede observarse gráficamente en la Figura 5.

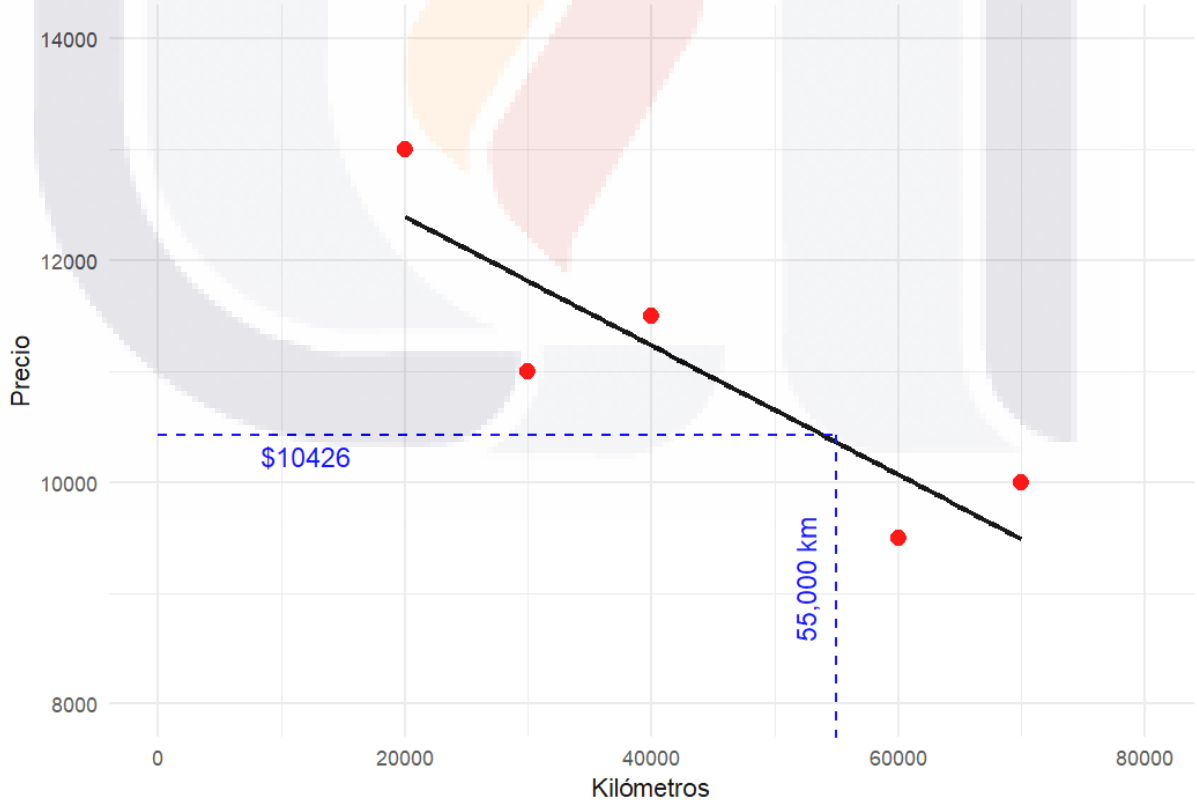


Figura 5 Grafica de la ecuación lineal del ejemplo 1

2.8.1.2 Ventajas

Es una técnica estadística simple que permite predecir la relación entre dos variables.

Su implementación práctica es sencilla y ampliamente disponible.

Es el fundamento de la regresión lineal múltiple.

2.8.1.3 Desventajas

Solo se puede usar para variables dependientes que sean continuas.

Su alcance es limitado cuando se aplica de forma única, por lo que existe la regresión lineal múltiple.

2.8.2 Método de máxima verosimilitud aplicado a la regresión lineal

El método de los mínimos cuadrados se puede aplicar para estimar los parámetros en un modelo de regresión lineal, independientemente de la forma de la distribución de los errores $\bar{\varepsilon}$. Con los mínimos cuadrados se obtienen los mejores estimadores lineales insesgados de β_0 y β_1.

Otros procedimientos estadísticos, como prueba de hipótesis y construcción de intervalo de confianza, suponen que los errores se distribuyen normalmente. Si se conoce la forma de distribución de los errores, un método alternativo para estimar parámetros es el método de máxima verosimilitud o máxima posibilidad.

Se tienen datos $(y_i, x_i), i = 1,2, \dots, n.$ Si se supone que los errores en el modelo de regresión son variables aleatorias normales e independientemente distribuidas (NID)(o, σ^2), las observaciones y_1 en esa muestra son NID con promedio $\beta_0 + \beta_1, x_i,$ y varianza σ^2.

La función de verosimilitud o de posibilidad se determina con la distribución conjunta de las observaciones.

Si se considera esta distribución conjunta con las observaciones dadas y los parámetros $\beta_0, \beta_1 \, y \, \sigma^2$ son constantes desconocidas, se tiene la función de verosimilitud. Para el modelo de regresión lineal simple con errores normales, la función de verosimilitud es:

$$L(y_i, x_i, \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} (2\,\pi\,\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 - \beta_1 x_i)^2\right]$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 - \beta_1 x_i)^2\right]$$

Los estimadores de máxima posibilidad son los valores de los parámetros, por ejemplo $\tilde{b}_0$, $\tilde{b}_1\ y\ \sigma^2$, que maximizan a L, o lo que es lo mismo, a lnL. Así:

$$lnL\ (y_i, x_i, \beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right) ln\ 2\pi - \left(\frac{n}{2}\right) \ln \sigma^2 - \left(\frac{1}{2\sigma^2}\right) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

y los estimadores de posibilidad máxima $\tilde{b}_0,\ \tilde{b}_1\ y\ \sigma^2$, deben satisfacer:

$$\frac{\partial lnL}{\partial \beta_0}\Big|_{\overline{\beta}_0, \overline{\beta}_1, \overline{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0$$

$$\frac{\partial lnL}{\partial \beta_1}\Big|_{\overline{\beta}_0, \overline{\beta}_1, \overline{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)\ x_i = 0$$

$$\frac{\partial lnL}{\partial \sigma^2}\Big|_{\overline{\beta}_0, \overline{\beta}_1, \overline{\sigma}^2} = -\frac{n}{2\tilde{\sigma}^2} + \frac{2}{\tilde{\sigma}^4}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 = 0$$

La solución de las ecuaciones determina los estimadores de máxima verosimilitud:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 \tilde{x}_i\right)^2}{n}$$

Obsérvese que los estimadores de máxima posibilidad de la ordenada al origen y la pendiente $\beta_0\ y\ \beta_1$, son idénticos a los obtenidos con los mínimos cuadrados.

También $\tilde{\sigma}^2$ es un estimador sesgado de $\sigma 2$. El estimador sesgado se relaciona con el estimador insesgado $\hat{\sigma}^2$ mediante

$$\tilde{\sigma}^2 = \left[\frac{n-1}{n}\right] \hat{\sigma}^2.$$

El sesgo es pequeño cuando n es moderadamente grande, por lo general se usa el estimador insesgado $\tilde{\sigma}^2$.

2.8.3 Teoría básica. Regresión Logística.

La regresión logística se comporta de manera similar a la regresión lineal. Sin embargo, en esta, el resultado lineal se acompaña de una función oculta sobrepuesta al resultado de la regresión. La función logística que se utiliza es la función sigmoidea. La función sigmoidea (Figura 6) permite describir muchos procesos naturales que manifiestan una progresión temporal desde unos niveles bajos al inicio, hasta acercarse a un punto máximo transcurrido un cierto tiempo; la transición entre el punto bajo y el máximo se produce en una región caracterizada por una aceleración intermedia muy marcada.[61]

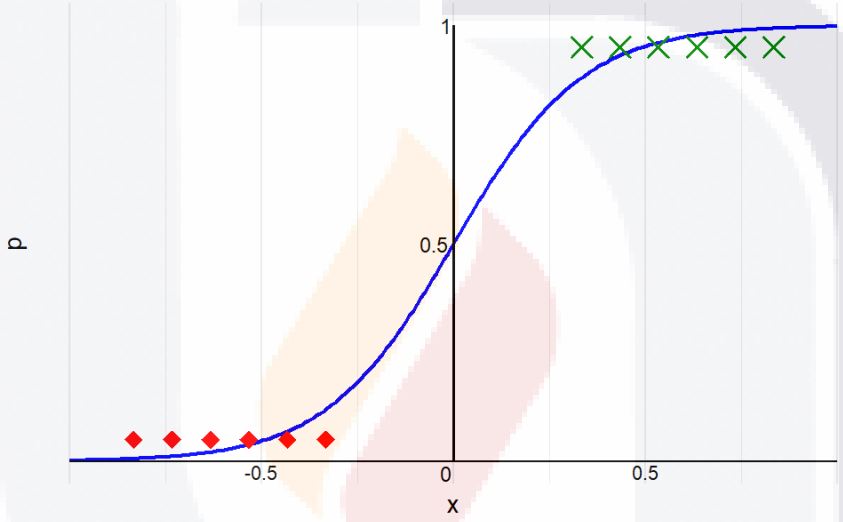


Figura 6 Representación gráfica de la función logística

2.8.3.1 Ventajas

La regresión logística es un método de clasificación rápido y directo. Sus parámetros explican la dirección y la intensidad del significado estadístico del efecto de las variables independientes sobre la variable dependiente. También se puede utilizar para clasificaciones multiclase.[62]

2.8.3.2 Desventajas

No se puede aplicar a problemas de clasificación no lineal. Se requiere una adecuada selección de las variables. Se requiere una alta tasa de señal/ruido en los datos. La precisión se ve comprometida por la colinealidad y los valores atípicos.[63]

2.9 Máquinas de vectores de soporte

Las máquinas de vectores de soporte son una técnica de aprendizaje automático que puede ser utilizada tanto para la clasificación como para la regresión. Para mejorar las adecuaciones lineales y no lineales, tiene dos variantes principales, las máquinas de vectores de soporte lineales las cuales no tienen núcleo y buscan una solución lineal al problema con un margen mínimo y las máquinas de vectores de soporte con núcleo para los casos en que la solución no es linealmente separable.[64]

2.9.1 Teoría básica

Las máquinas de vectores de soporte son herramientas de aprendizaje supervisado que se utilizan principalmente para la clasificación de imágenes y otros datos bio informáticos.[65] En el caso de las máquinas de vectores de soporte lineales el espacio del problema debe segregarse linealmente. El modelo produce un hiperplano que maximiza el margen de clasificación. Donde hay N características presentes, el hiperplano tendrá un subespacio dimensional N-1.[66] Los vectores de soporte se denominan nodos de límite en el espacio de esas características. El margen máximo se extrae en función de su posición relativa y se dibuja un hiperplano óptimo (Figura 7) en el punto medio.[64]

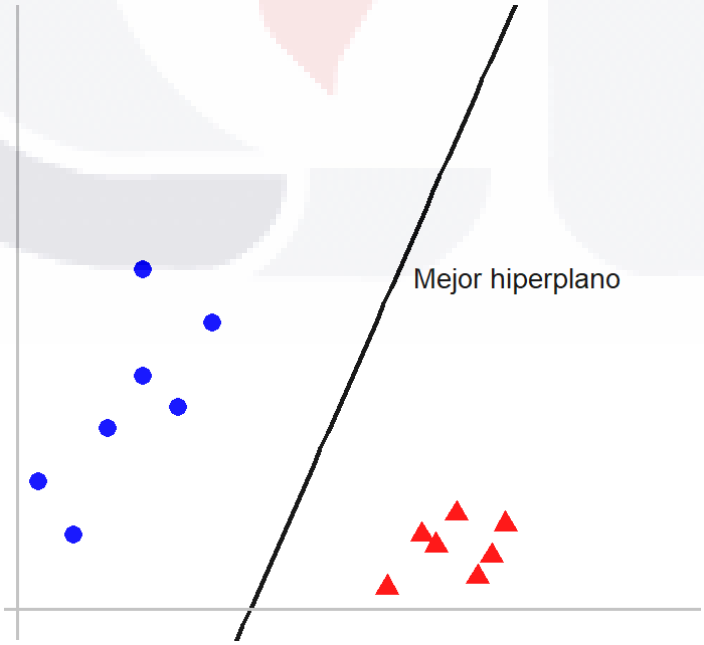


Figura 7 Representación gráfica de la clasificación por máquinas de vectores de soporte

En términos más simples, la idea es encontrar o determinar una recta que separe (linealmente) los datos de cada clase de forma que el margen generado por la recta y los puntos más cercanos sea el máximo. A esta recta se le llama frontera de decisión.

En la Figura 8 se muestran 4 puntos pertenecientes a dos grupos distintos (colores azul y rojo) y las distintas rectas (en verde) que los pueden separar por grupos.

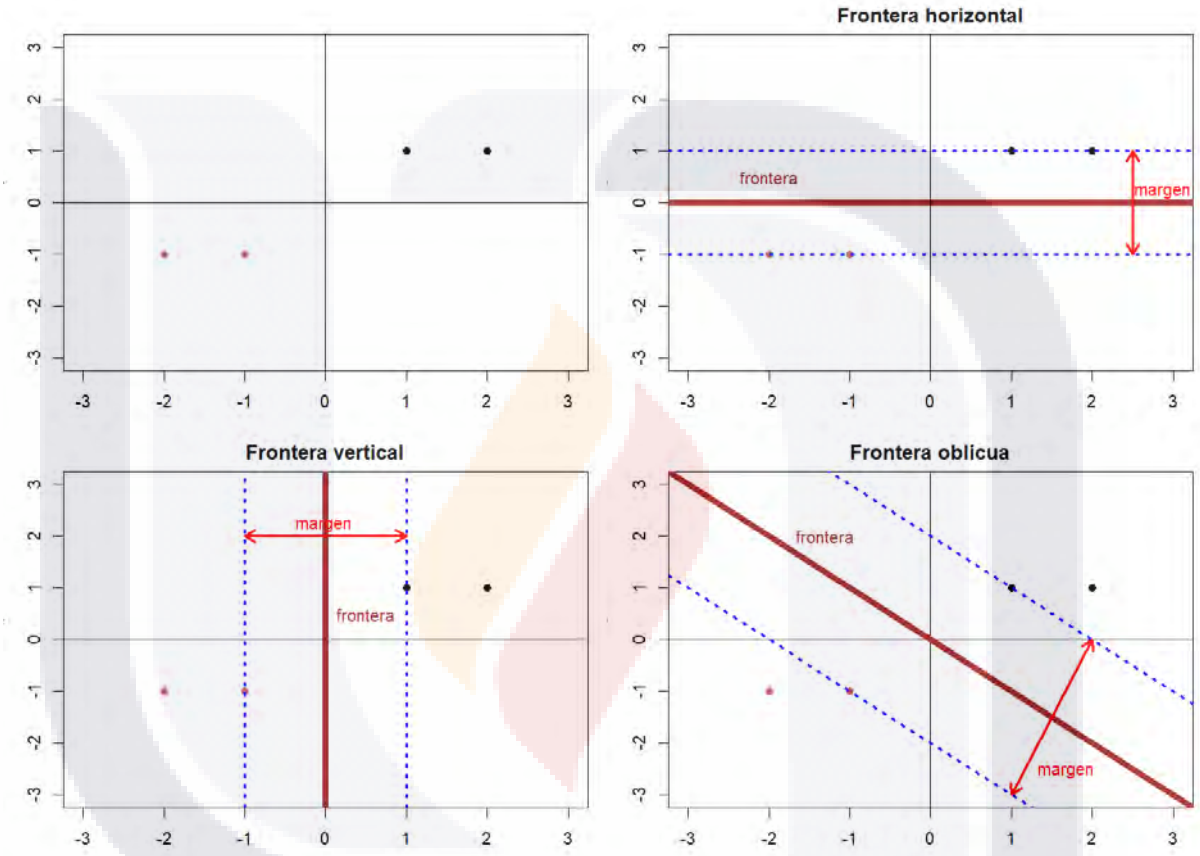


Figura 8 Ejemplo 1 Máquinas de vectores de soporte

Una frontera de decisión lineal puede expresarse mediante la siguiente ecuación:

$$w_1 x_1 + w_2 x_2 + b = 0$$

En notación matricial:

$$w^T x + b = 0$$

donde:

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

El objetivo de una máquina de vectores de soporte es encontrar la frontera de decisión que maximice el margen entre las dos clases, es decir, la distancia mínima entre la recta y los puntos más cercanos de cada clase (llamados vectores de soporte).

Este problema puede formularse como una optimización:

$$\arg\max_{w,b} \frac{1}{||w||} \min_i [y_i(w^T \cdot x_i + b]$$

Para facilitar su resolución, se adopta la convención de que los puntos más cercanos a la frontera (los vectores de soporte) cumplen:

$$y_i(w^T \cdot x_i + b) = 1$$

2.9.1.1 Cálculo de la Distancia de un Punto a la Frontera

La distancia entre un punto arbitrario (s, t) y la recta $w_1x_1 + w_2x_2 + b = 0$ se calcula como:

$$\frac{|w_1(s) + w_2(t) + b|}{\sqrt{w_1^2 + w_2^2}} \quad \rightarrow \quad \frac{|w_1(s) + w_2(t) + b|}{||w||} \; en\ notación\ matricial$$

Esta fórmula también aplica a un punto x_i cualquiera:

$$distancia_{x_1 frontera} = \frac{|w_1(x_{i1}) + w_2(x_{i2}) + b|}{||w||} = \frac{|w^T \cdot x_i + b|}{||w||} = \frac{y_i(w^T \cdot x_i + b)}{||w||}$$

Donde $y_i = 1$ ó $y_1 = 0$

La última igualdad se justifica porque, bajo la convención anterior, el término $y_i \in \{-1, +1\}$ hace que el numerador sea siempre positivo, facilitando la optimización.

2.9.1.2 Problema de Optimización Primal

El problema de encontrar la frontera óptima se reescribe como:

$$\min \frac{1}{2}||w||^2$$

Sujeto a:

$$y_i(w^T \cdot x_i + b) > 1 \; para\ todo\ i$$

Esta formulación cuadrática convexa es eficiente de resolver mediante métodos de optimización convexa.

2.9.1.3 Casos No Linealmente Separables y el Uso de Kernels

Cuando los datos no son separables linealmente (es decir, hay traslape entre clases), se puede aplicar una transformación no lineal $\phi(x)$ que proyecte los datos a un espacio de mayor dimensión donde sí puedan separarse linealmente.

Sin necesidad de calcular explícitamente $\phi(x)$, esto se logra mediante funciones kernel, que permiten evaluar el producto escalar $\phi(x_i)^T\phi(x_j)$ de forma eficiente. Algunos ejemplos comunes de funciones kernel son:

- Kernel lineal: $K(x, x') = x^T x'$
- Kernel polynomial: $K(x, x') = (x^T x' + c)\text{^}d$
- Kernel gaussiano: $K(x, x') = \exp\left(-\gamma\left||x, x'|\right|^2\right)$

El uso de kernels permite que las máquinas de vectores de soporte sean modelos altamente flexibles, adecuados tanto para problemas lineales como no lineales.

2.9.2 Ventajas

Las máquinas de vectores de soporte se pueden utilizar para resolver soluciones complejas con la ventaja del uso de núcleos para soluciones no lineales. Utilizan una función de optimización convexa, con la que es posible lograr mínimos globales. El manejo de las pérdidas con método de bisagra mejora la precisión. Existen técnicas como la de margen suave con constante C que ayudan a disminuir el efecto de los valores atípicos.[67]

2.9.3 Desventajas

Para una precisión adecuada, los hiperparámetros y los núcleos deben ajustarse cuidadosamente. La pérdida de las bisagras contribuye a la escasez en cuanto a las predicciones. En el caso de conjuntos de datos grandes, se requiere un período de entrenamiento más prolongado.

2.10 Redes neuronales

Las redes neuronales se basan en el proceso de aprendizaje que ocurre en el cerebro humano. Consisten en una red artificial de funciones llamadas parámetros que permiten a la computadora aprender y sintonizarse automáticamente al analizar datos nuevos. [68] A cada parámetro se le da el nombre de neurona y está constituido por una función que produce un resultado después de recibir uno o varios datos de entrada. Los resultados se pasan entonces a la siguiente capa de neuronas, las cuales los utilizan como datos de entrada para producir nuevos datos de salida.

Estos nuevos resultados se pasan a una nueva capa de neuronas y el proceso continúa hasta que la última capa de neuronas ha dado su resultado. A estas neuronas se les conoce como neuronas terminales y emiten el resultado final del modelo. [69]

2.10.1 Antecedentes. El perceptrón

El perceptrón es uno de los modelos más antiguos y fundamentales de clasificación supervisada, introducido por Frank Rosenblatt en 1958. Aunque conceptualmente sencillo, sentó las bases para el desarrollo posterior de las redes neuronales artificiales.

El perceptrón es esencialmente un clasificador lineal binario que toma como entrada un vector de características y genera una salida binaria (por ejemplo, 0 o 1) en función de una combinación lineal de las entradas.

2.10.1.1 Modelo matemático

Sea $x = [x_1, x_2, \dots, x \quad n]^T$ el vector de entrada (características del ejemplo a clasificar), y $w = [w_1, w_2, \dots, w_n]^T$ el vector de pesos asociados a cada característica. Además, se incluye un término de sesgo o umbral b (también denotado a veces como $-\theta$). La salida del perceptrón se define como:

$$y = \begin{cases} 1, & si\ w^T x + b > 0 \\ 0, & en\ otro\ caso \end{cases}$$

2.10.1.2 Algoritmo de aprendizaje del perceptrón

El aprendizaje en el perceptrón consiste en ajustar los pesos con base en ejemplos de entrenamiento etiquetados, utilizando una regla de actualización iterativa:

$$w^{t+1} = w^t + \eta \cdot (y_i - \hat{y}^i) \cdot x_i$$

$$b^{t+1} = b^t + \eta \cdot (y_i - \hat{y}^i)$$

donde:

- y_i es la clase verdadera del ejemplo,
- $\hat{y}^i$ es la salida predicha por el modelo,
- η es la tasa de aprendizaje (un parámetro positivo),
- x_i es el vector de entrada correspondiente.

La actualización ocurre solo cuando hay un error de clasificación. El algoritmo converge si los datos son linealmente separables; en caso contrario, el modelo puede no estabilizarse. [70]

2.10.1.3 Limitaciones del perceptrón

Aunque fue un avance crucial, el perceptrón tiene limitaciones importantes, por ejemplo, sólo puede resolver problemas linealmente separables, no puede modelar funciones lógicas simples como XOR y su arquitectura es de una sola capa y salida binaria.

Estas restricciones llevaron a una disminución del interés por las redes neuronales durante un tiempo. Sin embargo, la introducción de modelos multicapa (perceptrones multicapa o MLP) y el desarrollo del algoritmo de retropropagación del error en los años 80 permitieron superar estas limitaciones. [71]

Una de las claves para lograr este avance fue reemplazar la función escalón binaria del perceptrón por funciones de activación continuas y diferenciables, que permitieran calcular derivadas y propagar gradientes a través de múltiples capas.

Entre estas funciones, destaca la sigmoidea logística, que marcó un punto de inflexión en el entrenamiento efectivo de redes neuronales profundas.

2.10.2 Función de Activación Sigmoidea

El perceptrón clásico utiliza una función escalón (o Heaviside) como función de activación, lo que implica que su salida es estrictamente binaria. Sin embargo, esta función es no diferenciable y no permite una representación gradual ni una optimización basada en gradientes. [72]

Para resolver este problema, las redes neuronales introducen funciones de activación suaves y derivables, siendo la más común en los primeros modelos la función sigmoidea logística, definida como:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Donde $z = w^T x + b$ es la entrada de la neurona.

Esta función transforma el valor de entrada en un rango continuo entre 0 y 1, lo que permite interpretar la salida como una probabilidad, usar métodos de optimización basados en derivadas (como el descenso del gradiente), e introducir no linealidades que permiten aprender relaciones más complejas. La derivada de la función sigmoidea, esencial para el entrenamiento de redes neuronales, es:

$$\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$$

2.10.3 Pérdida por Error Cuadrático Medio

Para que una red neuronal pueda aprender a clasificar o predecir correctamente, necesita un mecanismo para cuantificar qué tan lejos está su salida del valor deseado. Este mecanismo es la función de pérdida.

Una función comúnmente utilizada en los primeros modelos es el Error Cuadrático Medio ([73] MSE, por sus siglas en inglés) :

$$E = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde:

- y_i es la salida esperada (etiqueta del dato),
- $\hat{y}_i$ es la salida producida por la red,
- La suma recorre todos los ejemplos del conjunto de entrenamiento.

El factor $\frac{1}{2}$ se incluye para simplificar la derivada durante la retropropagación. Aunque actualmente otras funciones como la entropía cruzada son más comunes en clasificación, el MSE tiene valor pedagógico y se mantiene útil en tareas de regresión.

2.10.4 Pase hacia Adelante

El proceso de pase hacia adelante consiste en calcular la salida de una neurona o de toda una red, dado un vector de entrada. En su forma más simple (una sola neurona con activación sigmoidea), el procedimiento implica calcular la entrada ponderada $z = w^T x + b$ y aplicar la función de activación $\hat{y} = \sigma(z)$**.** Este procedimiento se repite capa por capa en redes multicapa, utilizando los pesos y sesgos correspondientes a cada conexión.

El pase hacia adelante permite calcular la salida final de la red, la cual será comparada con la etiqueta real para calcular el error. [74]

2.10.5 Retropropagación del Error

Una vez calculado el error entre la salida deseada y la salida real, el siguiente paso es ajustar los pesos de la red para minimizar ese error. Esto se logra mediante el algoritmo de retropropagación del error, una aplicación del descenso del gradiente que propaga las correcciones desde la capa de salida hacia las capas ocultas. [75]

La idea central es aplicar la regla de la cadena para calcular el gradiente del error respecto a cada peso. Para una red con activación sigmoidea y una sola salida, el cálculo del error en la capa de salida es:

$$\delta = (y - \hat{y}) \cdot \sigma'(z)$$

donde δ representa el "error local" que será utilizado para actualizar los pesos. La actualización de pesos se hace con:

$$w_j^{t+1} = w_j^t + \eta \cdot \delta \cdot x_j$$

$$b^{t+1} = b^t + \eta \cdot \delta$$

En redes multicapa, se calcula un δ para cada neurona, y estos valores se propagan hacia atrás desde la salida hasta la primera capa oculta. Así, cada peso se ajusta en proporción a su contribución al error final.

La función de activación diferenciable, la función de pérdida, el pase hacia adelante y la retropropagación, permiten que una red neuronal artificial aprenda de los datos, superando las limitaciones del perceptrón original. A continuación, se explicará el modelo de red neuronal artificial, su estructura y su entrenamiento completo.

2.10.6 Teoría básica. Red Neuronal Artificial

Una red neuronal artificial puede entenderse como una función compuesta de múltiples subfunciones organizadas en capas, que recibe un conjunto de datos de entrada y produce un resultado final. Las funciones intermedias se agrupan en lo que se conoce como capas ocultas (Figura 9). Cada unidad (o neurona) dentro de la red realiza una operación sobre sus entradas: calcula una combinación lineal de los valores de entrada multiplicados por pesos específicos, y le suma un término de sesgo propio. Este resultado pasa por una función de activación, generando una salida que se transmite a la siguiente capa.

El flujo de información desde la entrada hasta la salida se conoce como pase hacia adelante, y constituye la primera fase del funcionamiento de la red.

Para que la red pueda aprender a resolver una tarea, se define una función de costo (o función de pérdida), que mide qué tan lejos está la salida producida por la red respecto al valor deseado. Existen diversas funciones de costo, cada una con ventajas y desventajas; la elección depende del tipo de problema y del comportamiento esperado de la red. Una vez definida la función de costo, el entrenamiento consiste en ajustar los pesos y sesgos para minimizar dicha función.

Las redes neuronales pueden entrenarse bajo cualquiera de los tres paradigmas del aprendizaje automático: supervisado, no supervisado y por refuerzo. Sin embargo, el más común y sencillo de implementar es el aprendizaje supervisado, en el cual la red neuronal recibe datos de entrada etiquetados (es decir, con su respectiva clase o resultado esperado). A partir de estos datos, la red aprende patrones generales que luego puede aplicar para clasificar o predecir nuevos datos no vistos.[6]

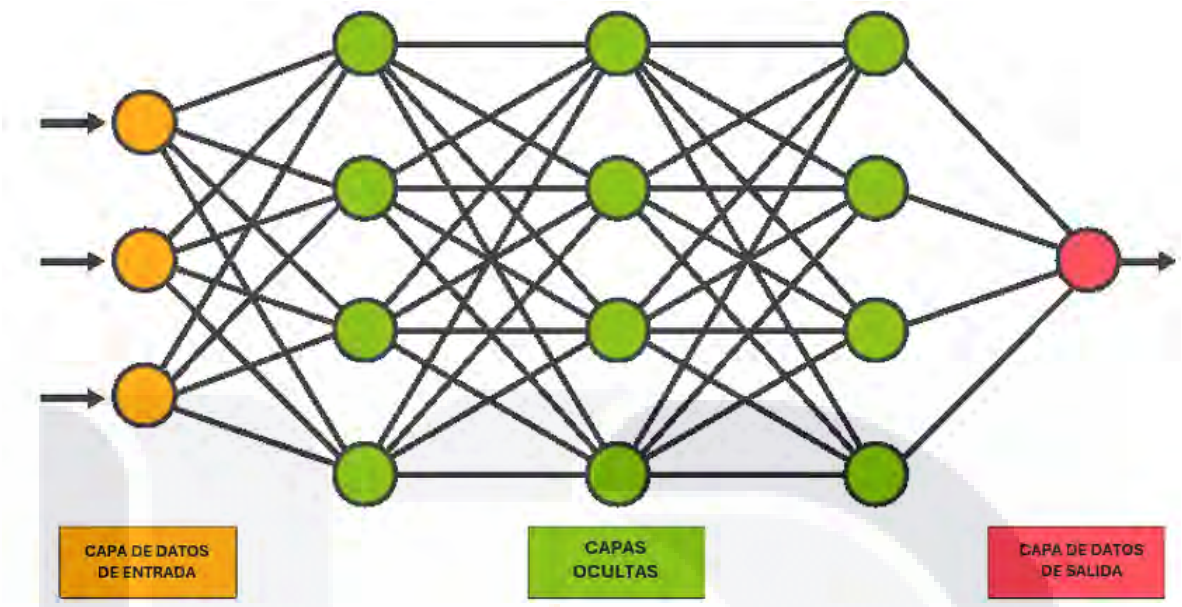


Figura 9 Representación gráfica de una red neuronal artificial de 3 capas

2.10.6.1 Proceso de Entrenamiento de una Red Neuronal

El entrenamiento de una red neuronal es un proceso iterativo que requiere equilibrio entre la complejidad del modelo y la cantidad y calidad de los datos disponibles. A grandes rasgos, el proceso sigue las siguientes etapas:

1) Preparación de los datos: Se recopilan y preprocesan los datos de entrenamiento, que consisten en pares de entrada y salida esperada.
2) Inicialización del modelo: Se asignan valores iniciales aleatorios a los pesos y sesgos de la red.
3) Pase hacia adelante: Se introducen los datos de entrada en la red y se calcula una salida estimada, propagando la información desde la capa de entrada hasta la salida.
4) Cálculo del error: Se mide la diferencia entre la salida estimada y la salida real utilizando la función de costo.
5) Ajuste de parámetros (retropropagación): Se utiliza un algoritmo de optimización, típicamente el descenso del gradiente, para calcular cómo deben

modificarse los pesos y sesgos con el fin de reducir el error. Esta información se propaga hacia atrás desde la salida hasta la entrada.

6) Iteración del proceso: Los pasos 3 a 5 se repiten muchas veces (épocas) hasta que el error sea aceptablemente bajo.

Evaluación: Finalmente, la red se prueba con un conjunto de datos no utilizados durante el entrenamiento (datos de validación o prueba) para verificar su capacidad de generalización.

2.10.7 Ventajas

La principal ventaja de las redes neuronales es su capacidad de optimizarse de forma automática, al correr múltiples veces la misma red. En la corrida inicial las predicciones serán necesariamente aleatorias, pero después de cada iteración la función de costo se analiza para determinar cómo se desempeñó el modelo y cómo puede mejorarse.

La información obtenida de la función de costo se traslada a una función de optimización, la cual calcula nuevos valores de sesgo y de peso específico. Una vez que los nuevos valores son integrados al modelo, este se vuelve a correr.

Este proceso se continua hasta que no existan nuevas mejorías en el desempeño a pesar de cambiar los valores de la función de costo.[76]

2.10.8 Desventajas

La principal dificultad al establecer una red neuronal es determinar los valores óptimos de sesgo y de peso específico para cada paso entre las capas de la red.

Otra potencial desventaja es el tiempo de computación, dada la necesidad de correr la red múltiples veces. Finalmente, la utilización de métodos de aprendizaje no supervisados, en los que la red neuronal tiene que inferir las reglas y sus funciones a partir de datos no etiquetados, limita severamente el tipo de predicciones que puede llevar a cabo la red neuronal.

En general, se considera a estos modelos incapaces de clasificar y se limitan a agrupamientos. [77]

2.11 Análisis discriminante

El Análisis Discriminante Lineal (LDA) es un método estadístico clásico utilizado para resolver problemas de clasificación supervisada. Su objetivo principal es encontrar una proyección lineal del espacio de características que maximice la separación entre clases y, al mismo tiempo, minimice la dispersión dentro de cada clase. [78]

Además de su uso principal en tareas de clasificación supervisada, el LDA tiene otras aplicaciones relevantes en el análisis de datos multivariados. Una de las más destacadas es la reducción de dimensionalidad, donde LDA se utiliza para proyectar datos de alta dimensión en un subespacio de menor dimensión que preserva al máximo la separabilidad entre clases. A diferencia de métodos no supervisados como PCA, LDA incorpora información de las etiquetas de clase, lo que lo hace especialmente útil en contextos donde el objetivo es mejorar la interpretabilidad o el rendimiento de modelos posteriores. [79]

Otra aplicación importante es como técnica exploratoria, para visualizar la distribución de los datos y evaluar qué tan bien separadas están las clases en un espacio reducido. Esto es particularmente útil en conjuntos de datos con muchas variables, como en genética, visión por computadora o neuroimagen. [80] También puede emplearse en sistemas de reconocimiento facial, detección de emociones y selección de variables relevantes en estudios clínicos o de mercado. [81]

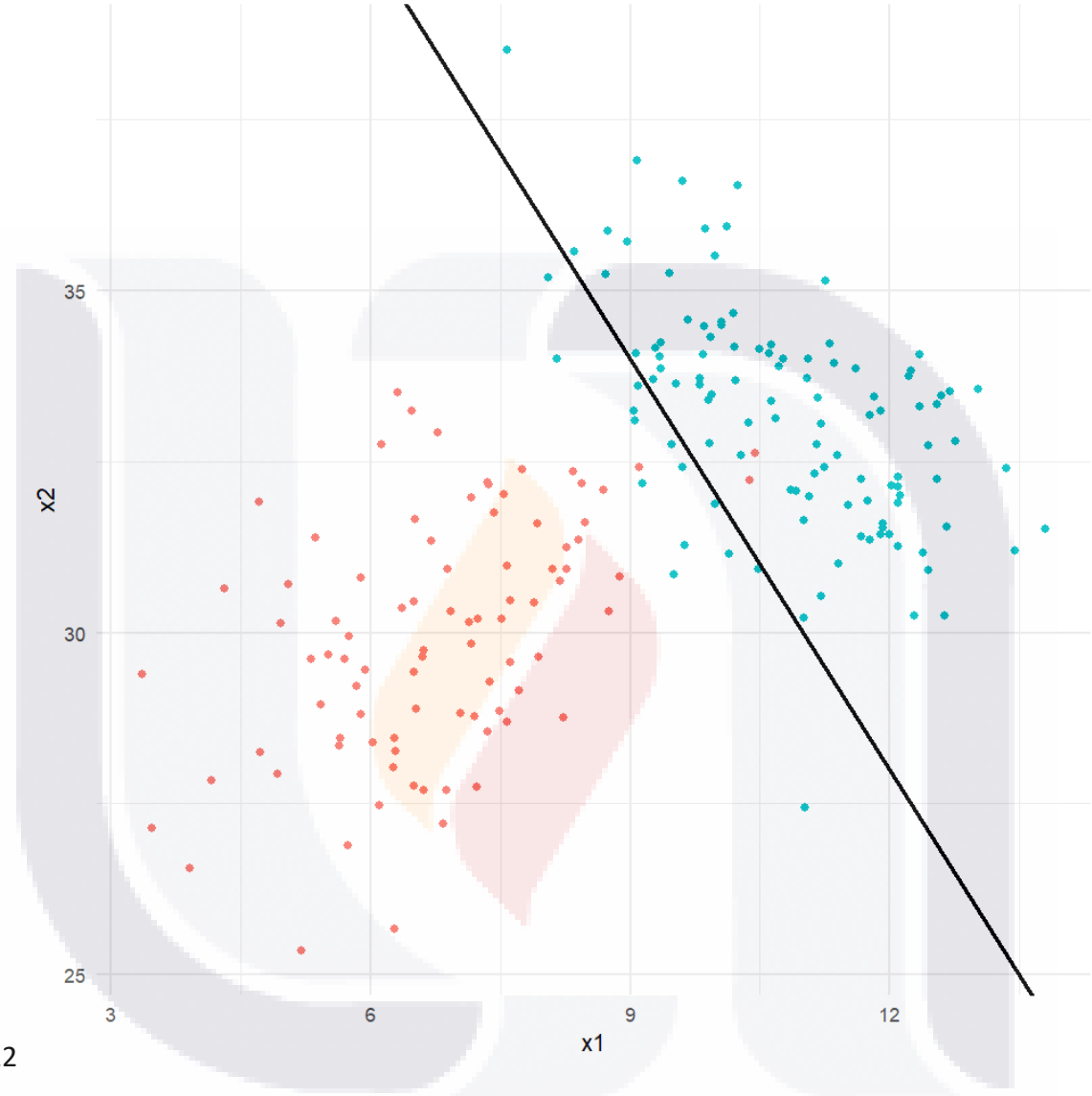
2.11.1 Teoría básica

El LDA se basa en dos supuestos fundamentales sobre la distribución de los datos dentro de cada grupo:

1) Las variables explicativas X siguen una distribución normal multivariada dentro de cada clase.
2) Todos los grupos comparten una misma matriz de covarianza Σ, es decir, se asume homogeneidad de varianzas y covarianzas.

Bajo estas condiciones, el análisis discriminante lineal construye una función discriminante lineal que permite clasificar nuevas observaciones.

Esta función tiene como objetivo separar los grupos mediante una frontera lineal que pasa, en promedio, por los centroides (vectores de medias) de cada grupo. (Figura 10) [82]



12

Figura 10 Representación gráfica del análisis discriminante

El modelo estándar de LDA asume que la distribución condicional de las variables explicativas dado el grupo al que pertenecen sigue una distribución normal multivariada:

$$X \mid y = k \sim N(\mu k, \Sigma)$$

donde μk es el vector de medias del grupo k y Σ es la matriz de covarianza común a ambos grupos. Esta estructura permite derivar una regla de clasificación lineal basada en la comparación de funciones discriminantes evaluadas para cada clase.

2.11.1.1 Construcción de la función discriminante

Supongamos dos poblaciones con vectores de entrada $x = (x_1, x_2, \cdots, x_d)^T$. El análisis discriminante busca una proyección lineal:

$$z = a_1x_1 + a_2x_2 + \cdots + a_dx_d = w^Tx$$

El vector *w* que maximiza la separación entre clases se obtiene como:

$$w = S^{-1}(\bar{x}_2 - \bar{x}_1)$$

Donde $\bar{x}_2\ y\ \bar{x}_1$ son las medias de los grupos 1 y 2, respectivamente, y S es la matriz de covarianza combinada:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 * n_2 - 2}$$

con S_1 y S_2 como matrices de varianzas y covarianzas estimadas para cada grupo.

2.11.2 Regla de decisión

La función discriminante permite establecer una frontera de decisión lineal que separa el espacio en dos regiones. Esta frontera puede expresarse como:

$$w^Tx = \frac{1}{2}w^T(\mu_1 + \mu_2) - \log\left(\frac{c(2|1)\pi_1}{c(1|2)\pi_2}\right)$$

Donde $\pi 1, \pi 2$ son las probabilidades a priori de pertenecer a cada grupo y $c(1 \mid 2), c(2 \mid 1)$ son los costos de clasificación errónea.

2.11.2.1 Caso bidimensional

En el caso particular de dos dimensiones $(d = 2)$, el vector de proyección es $w = (a_1, a_2)$, y la frontera de decisión toma la forma:

$$a_1x_1 + a_2x_2 = u$$

Donde

$$u = \frac{1}{2}[a_1(\mu_{12} - \mu_{11}) + a_2(\mu_{22} - \mu_{21})] - \log\left(\frac{c(2|1)\pi_1}{c(1|2)\pi_2}\right)$$

De esta forma, la ecuación puede reescribirse como:

$$x_2 = \frac{u - a_1 x_1}{a_2} = \frac{\mathrm{u}}{\mathrm{a}_2} - \frac{a_1}{a_2} x_1$$

lo que representa la ecuación de una recta en forma pendiente-ordenada al origen $x_2 = b + m x_1$.

2.11.3 Ventajas

Preservación de la información discriminatoria: El análisis discriminante busca maximizar la separación entre clases, lo que significa que preserva la información más relevante para la clasificación. Al proyectar los datos en un espacio de menor dimensión, el análisis discriminante puede capturar eficazmente las características discriminantes, lo que conduce a una mayor precisión en la clasificación.

Capacidad de Clasificación Multiclase: El análisis discriminante puede manejar conjuntos de datos con múltiples clases o grupos, lo que lo hace adecuado para problemas de clasificación multiclase. Construye funciones discriminantes que consideran explícitamente las relaciones entre las diferentes clases, lo que resulta en una mejor separación y rendimiento de clasificación en comparación con las técnicas de clasificación binaria.

Sensibilidad Reducida al Sobreajuste: El análisis discriminante opera bajo la suposición de separabilidad lineal, lo que le permite producir resultados confiables con un número limitado de muestras de entrenamiento. A diferencia de otras técnicas de reducción de dimensionalidad que pueden ser propensas al sobreajuste en espacios de baja dimensión, el análisis discriminante ofrece estabilidad y robustez en escenarios con tamaños de muestra pequeños. [83]

2.11.4 Desventajas

Suposición de Separabilidad Lineal: El análisis discriminante asume que los datos pueden ser separados por fronteras de decisión lineales. Si las clases exhiben relaciones no lineales complejas, el análisis discriminante puede no ser capaz de capturar la estructura subyacente de manera efectiva, lo que resulta en una reducción del rendimiento de

clasificación. En tales casos, pueden ser más adecuadas las técnicas de reducción de dimensionalidad no lineales como los métodos de kernel.

Sensibilidad a Valores Atípicos: El análisis discriminante es sensible a los valores atípicos, ya que busca maximizar la separación entre las clases. Los valores atípicos pueden tener un impacto sustancial en la estimación de las medias de clase y las matrices de covarianza, lo que lleva a una separación distorsionada entre las clases. Se pueden utilizar variantes robustas del análisis discriminante u otros métodos robustos frente a valores atípicos para mitigar este problema.

Limitaciones en escenarios con muestra de tamaño pequeño: El análisis discriminante puede sufrir de inestabilidad y resultados poco confiables cuando el número de muestras es menor que la dimensionalidad de los datos. Este problema, conocido como "problema de muestra de tamaño pequeño", puede provocar sobreajuste y un bajo rendimiento de generalización. Pueden ser preferibles técnicas de regularización u otros métodos de reducción de dimensionalidad específicamente diseñados para escenarios con muestra de tamaño pequeño.

2.12 Cópulas

Las cópulas son funciones de distribución que contienen la norma reguladora de la dependencia entre varias variables. La cópula es la expresión que describe la forma en que dependen unas variables de otras.[16]

2.12.1 Teoría básica

En cualquier función (distribución bivariante o multivariante) hay dos aspectos que definen el comportamiento estadístico de sus variables: 1) Su distribución marginal (cómo se comporta estadísticamente cuando se le observa de forma aislada), y 2) La estructura de dependencia (comportamiento conjunto de ambas variables, Figura 11). La cópula es entonces, la función que regula el comportamiento conjunto sin tomar en cuenta el comportamiento individual de cada variable.

La función se obtiene sustrayendo las distribuciones marginales de la función de distribución.

Para el caso de una función de distribución de dos variables $F(x,y)$ con funciones de distribución marginal $r = s(x)$, $p = q(y)$ las funciones s y q determinan la forma de la función $F(x,y)$.

Para despejar la cópula, debe eliminarse el efecto que s y q sobre $F(x,y)$.

Despejar *x* y *y* de las funciones inversas s^{-1} y p^{-1}, quedando:

$$x = s^{-1}(r)\ ,\ y = q^{-1}(p)$$

, tanto r como p oscilan entre 0 y 1. Se insertan en $F(x,y)$ las expresiones anteriores. Resultando la función a la que se llama cópula:

$$F(x,y) = F[s^{-1}(r), q^{-1}(p)] = C(r,p)$$

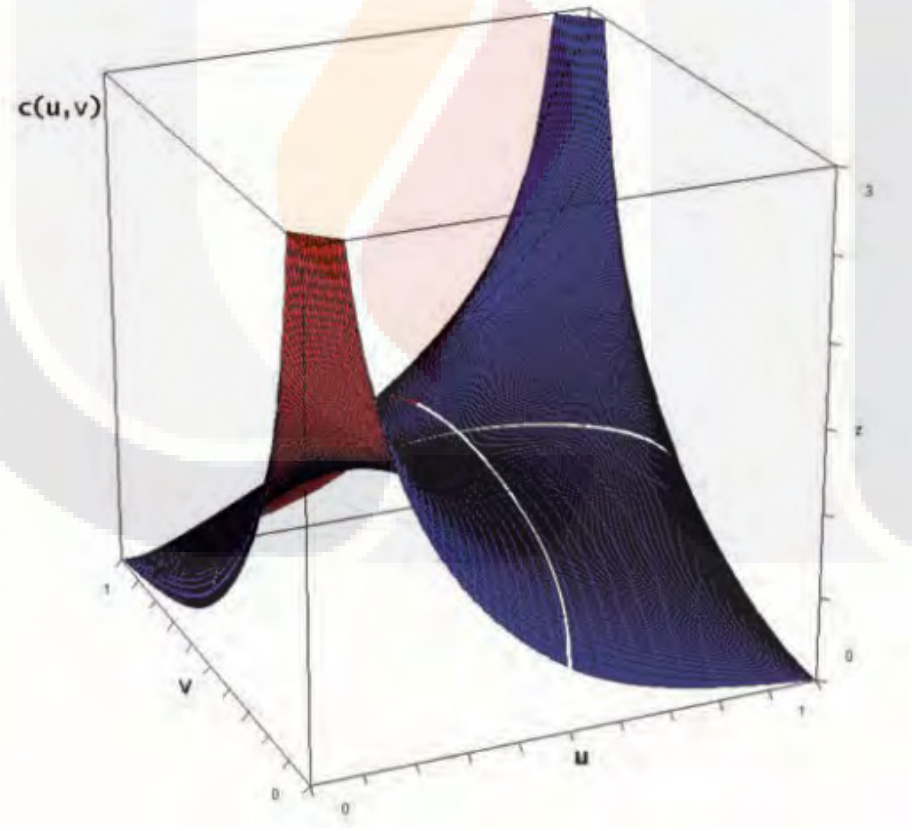


Figura 11 : Función de densidad de la cópula gaussiana con parámetro ρ = 0,5

2.12.2 Desarrollo del paquete MLCOPULA

Dado que para la implementación del clasificador con cópulas no existía un paquete que se pudiera utilizar, el doctorante de esta tesis participó como parte del equipo que desarrolló un paquete para el lenguaje de programación R que cumpliera esta función. A continuación, se detalla el proceso y funcionalidad de dicho paquete.

- Proceso General para Desarrollar, Enviar y Aprobar un Paquete en CRAN

2.12.2.1 Definiciones

Paquete: Un paquete en R es una colección de funciones, datos y documentación que permite a los usuarios realizar tareas específicas. Los paquetes son fundamentales para ampliar las capacidades de R, proporcionando herramientas adicionales para análisis de datos, modelado estadístico, visualización, y más.

CRAN: El Comprehensive R Archive Network (CRAN) es un repositorio en línea donde los usuarios pueden encontrar y descargar paquetes R. CRAN también sirve como plataforma para que los desarrolladores envíen sus paquetes y los pongan a disposición de la comunidad R.

2.12.2.2 Desarrollo del Paquete

1. Crear el Paquete: El primer paso es desarrollar el paquete. Esto incluye escribir las funciones, crear conjuntos de datos (si es necesario), y escribir la documentación utilizando roxygen2 o directamente en archivos Rd.

2. Pruebas y Validación: Antes de enviar el paquete a CRAN, es crucial realizar pruebas exhaustivas para asegurarse de que todas las funciones funcionan correctamente y que el paquete cumple con los estándares de CRAN. Usualmente, esto se hace con herramientas como testthat.

3. Escribir la Documentación: La documentación debe ser clara y completa, incluyendo ejemplos de uso de las funciones. Además, se debe crear un archivo DESCRIPTION y un archivo NAMESPACE.

2.12.3 Envío a CRAN

4. Preparar para el Envío: Una vez que el paquete está listo, se debe construir y comprobar utilizando herramientas como devtools. Es importante asegurarse de que no hay errores, advertencias ni notas en el chequeo del paquete.

5. Enviar a CRAN: El siguiente paso es enviar el paquete a CRAN. Esto se hace generalmente a través del correo electrónico a una de las direcciones de mantenimiento de CRAN, junto con un formulario de envío. Es esencial seguir las políticas y directrices de envío de CRAN.

2.12.4 Aprobación y Publicación

6. Revisión por CRAN: Después del envío, el paquete es revisado por los mantenedores de CRAN. Pueden requerir modificaciones o aclaraciones. Es crucial responder rápidamente a cualquier solicitud de cambios.

7. Aprobación y Publicación: Una vez que el paquete cumple con todos los requisitos de CRAN, se aprueba y se publica en el repositorio. El paquete estará entonces disponible para que los usuarios lo descarguen e instalen.

- Capacidades y Funcionalidad del Paquete MLCOPULA

2.12.5 Descripción del Paquete

El paquete MLCOPULA entrena un modelo de clasificación basado en cópulas. La densidad conjunta de las cópulas se construye con un modelo gráfico de árbol o cadena, como se muestra en la publicación a cargo del tutor principal de esta tesis. [84]

Uso de la Función Principal

```
copulaClassifier(X, y, distribution = "kernel", copula = "frank", weights = "likelihood", graph_model = "tree", k = 7, m = 7, method_grid = "ml")
```

Argumentos

- X: Data frame con n muestras y d > 1 variables predictoras.
- y: Vector de tamaño n, con las clases a predecir.

- distribution: Distribución marginal a usar: "normal" o "kernel", por defecto "kernel".
- copula: Nombre de la cópula a usar: "frank", "gaussian", "clayton", "joe", "gumbel", "amh", "grid", por defecto "frank". Para cópulas paramétricas, se pueden seleccionar una o más. Para la cópula no paramétrica, solo se puede seleccionar "grid".
- weights: Método de construcción de pesos: "likelihood" o "mutual_information", por defecto "likelihood".
- graph_model: Estructura del modelo gráfico: "tree" o "chain", por defecto "tree".
- k: Solo para la cópula "grid". Entero positivo indicando el número de subintervalos para la variable U2.
- m: Solo para la cópula "grid". Entero positivo indicando el número de subintervalos para la variable U1.
- method_grid: Método de ajuste, mínimos cuadrados "ls" o máxima verosimilitud "ml", por defecto "ml".

2.12.5.1 Valor Devuelto

La función regresa una lista con el modelo entrenado.

2.12.5.2 Funcionalidad

El paquete MLCOPULA es especialmente útil para la clasificación en contextos donde la dependencia entre las variables predictoras puede ser modelada eficazmente mediante cópulas. Esto permite capturar estructuras de dependencia complejas que otros algoritmos de clasificación podrían no manejar adecuadamente.

2.12.6 Ventajas

La importancia de la utilización de cópulas se deriva de las limitaciones inherentes a los coeficientes de correlación (τ de Kendall y ρ de Spearman) las cuales dan idea de la co-monotonicidad [85] y a que otras modalidades para considerar la dependencia entre variables como el coeficiente de dependencia en la cola de la distribución (*tail dependence*) tiene también limitaciones,[86] las cópulas permiten encontrar las expresiones que nos

describan la dependencia entre variables, y por lo tanto, determinar la estructura de la relación entre las variables.[87] Una ventaja adicional es que puede obtenerse una cópula desde una distribución multinomial conocida (Gauss, t de Student).[88]

2.12.7 Desventajas

La desventaja principal del uso de cópulas para describir la relación entre variables es que si la cópula cambia (algo que es dependiente de quien establece la cópula), entonces la dependencia de las variables entre sí también cambia. Lo anterior hace preferente su utilización como modelos de la dependencia entre variables aleatorias.

3. Conceptos Médicos

3.1 Enfermedad Vascular Cerebral

De acuerdo con experiencias previas,[89] se sabe que la prevalencia y la incidencia de la enfermedad vascular cerebral (EVC) son mayores en las poblaciones méxico-americanas, sobre todo en el subtipo hemorrágico.[90] Entre los factores que explican esto se han implicado los genéticos y el pobre control de los factores de riesgo cardiovascular en los países en desarrollo y sus poblaciones migrantes, entre otros. Se considera que para el año 2030, siete de cada diez muertes serán causadas por enfermedades no transmisibles, que la cardiopatía coronaria sea la causa principal y que la EVC ocupe el segundo lugar.[91]

En Latinoamérica y el Caribe hay dos veces más muertes por enfermedades no transmisibles que por enfermedades transmisibles. Son bien conocidos los factores de riesgo de la EVC y es importante recordar que estos factores se potencian entre sí. Si coexisten, por ejemplo, el tabaquismo con hipertensión y dislipidemia, el riesgo se puede potenciar 16 veces.

Las tasas de mortalidad por enfermedades del aparato circulatorio, según la Organización Panamericana de la Salud, son muy altas en Latinoamérica y no han sufrido modificaciones muy significativas a lo largo del tiempo. Algunos ejemplos (tasas por 100 mil habitantes comparando el año de 1979 y el 2015): Argentina: 46.6 - 46.4, Chile 29.4 – 29.0, Honduras 15.0 - 13.9, Puerto Rico 40.5 – 34.0 y Uruguay 44.1 - 41.4. En ese mismo período en México, la tasa de mortalidad de 38% disminuyó solo 0.7 (intervalo de confianza del 95 %, de 0.1 a 1.4). 92, Latinoamérica ha enfrentado y sigue enfrentando altas tasas de migración que han influido en sus características genéticas y hábitos de vida y nutricionales. Una publicación de la Fundación Latinoamericana del Corazón [92] proporciona en forma detallada estadísticas de enfermedades cardíacas y cerebrovasculares en las Américas. Estos datos demuestran la alta tasa de mortalidad por EVC en los países de América latina.

3.1.1 Factores de Riesgo para Enfermedad Vascular Cerebral

3.1.1.1 Factores de Riesgo No Modificables

A medida que el número de personas mayores de edad o ancianos incrementa en la población mundial, demuestra que la edad es el factor de riesgo más importante para

presentar EVC; llegando a duplicarse el riesgo cada 10 años después de los 35 años. [93, 94] También con el aumento en la longevidad, se observa un incremento lineal de pacientes con hipertensión arterial sistémica (HAS), entre los cuáles muchos de ellos no tienen un buen control de su presión arterial, o usa medicaciones antitrombóticas y/o anticoagulantes que les aumenta el riesgo de sangrado intracraneal y del número de ingresos hospitalarios; el cual se ha incrementado en un 18% en los últimos 10 años [95, 96].

Existen además ciertas poblaciones en las cuales se ha observado un incremento en el riesgo de EVC, estas incluyen:

- Hombres
- Mexicano–americanos y Afroamericanos viviendo en Estados Unidos de América
- Población latinoamericana, japonesa y china [97-102]

En el estudio NOMAS (*The Northern Manhattan Study*), se investigó la incidencia de EVC en la comunidad multiétnica del norte de Manhattan y se comparó la presencia de hemorragias lobares y/o profundas en pacientes negros, hispanos y blancos americanos. Se observó una incidencia elevada de 30.9/100000, y el riesgo fue mayor en los hombres en comparación con las mujeres (riesgo relativo 1.5, intervalo de confianza del 95%, 1.2 - 1.8). También se documentó que las hemorragias profundas fueron más frecuentes que las lobares en pacientes negros (riesgo relativo 4.8 vs 2.8) e hispanos (riesgo relativo 3.7 vs 1.4) comparados con los blancos estadounidenses (riesgo relativo 1.8 vs 1.0), sin embargo, estos hallazgos no tuvieron una asociación estadísticamente significativa. [98]

En Estados Unidos (E.U.), el riesgo de un primer EVC es mayor en pacientes negros, llegando a ser de 2 a 5 veces más alto este riesgo en menores de 65 años comparados con los blancos americanos de la misma edad. [93] En la población negra y la hispana, el mayor riesgo de EVC se encuentra en los jóvenes y adultos. [89, 103]

Esta información se corroboró en un estudio con 1038 pacientes con EVC realizado en área metropolitana de Cincinnati y el norte de Kentucky, donde la mayor incidencia fue en negros (48.9/100000) vs blancos (26.6/100000) y el riesgo más alto fue en los pacientes jóvenes (35 y 54 años) con hemorragias de tallo cerebral (riesgo relativo, 9.8; 95% CI, 4.2 - 23.0) y

profundas (riesgo relativo, 4.5; intervalo de confianza del 95%, 3.0 - 6.8); siendo la hipertensión arterial el factor de riesgo más importante para estas localizaciones [103].

En el Proyecto BASIC *(Brain Attack Surveillance in Corpus Christi)*, se observó mayor incidencia de EVC en mexicanoamericanos con riesgo relativo ajustado para la edad de 1.63 (intervalo de confianza del 95%, 1.24 - 2.16), y además los adultos entre 45 y 59 años triplicaron este riesgo comparado con los blancos no hispanos de la misma edad. [90]

Con respecto al género, se ha documentado mayor riesgo de EVC en los hombres frente a las mujeres; sin embargo, después de los 64 años, esta diferencia tiende a desaparecer y a igualarse. [93, 104] En la población masculina negra americana, se observa un riesgo de 2.3/1000 vs. 1.9/1000 en las mujeres, a diferencia de la población blanca, donde se reporta un riesgo de 0.9/1000 en hombres vs. 0.6/1000 en las mujeres. [93]

3.1.1.2 Factores de Riesgo Modificables

3.1.1.2.1 Hipertensión Arterial

El más importante y prevalente factor de riesgo modificable es la hipertensión arterial sistémica. En el estudio poblacional de Cincinnati, se observó una tasa elevada de HAS con una distribución homogénea entre blancos (69%) y afroamericanos (67%), al igual que entre hombres (72%) y mujeres (73%) [105]. Sin embargo, el mayor riesgo se encuentra en los pacientes hipertensos sin tratamiento regular o con descontinuación de la medicación frente a los que reciben tratamiento regular. En el estudio de Woo y colaboradores; con 322 pacientes de la población de Cincinnati, se documentó que la hipertensión arterial no tratada fue un factor de riesgo significativo para EVC (razón de momios 3.5, intervalo de confianza del 95%, 2.3-5.2; valor $p < 0.0001$), y en menor proporción con los que recibían tratamiento (razón de momios 1.4; intervalo de confianza del 95%, 1-1.9, valor $p < 0.003$).

Con estos resultados, se estimó que un cuarto de las EVC (17 – 28%), se podrían prevenir si todos los pacientes hipertensos recibieran tratamiento médico. [106] En el estudio de Thrift et al; se observó que la presencia de hipertensión duplica el riesgo de EVC (razón de momios 2.45, intervalo de confianza del 95%, 1.61-3.73), y también la razón de momios aumentaba en los pacientes que suspendían la medicación antihipertensiva (razón de momios 4.98,

intervalo de confianza del 95%, 2.25 – 11.02) comparados con los que no la suspendían (1.95, intervalo de confianza del 95%,1.20-3.16). De igual forma, se incrementó el riesgo si eran < 55 años (razón de momios 7.68, intervalo de confianza del 95%, 2.65 - 22.5) y en los fumadores activos (razón de momios 6.12 intervalo de confianza del 95%, 2.29 - 16.35). [107]

3.1.1.2.2 Dislipidemia

Mientras la hipercolesterolemia claramente es un factor de riesgo para enfermedad isquémica coronaria y cerebral, sucede todo lo contrario con la EVC donde aparentemente la hipocolesterolemia incrementa el riesgo, manteniéndose una relación inversa entre los niveles de colesterol y el riesgo de hemorragia intracerebral según lo reportado en estudios de casos y controles. [106, 107]

Esto también se documentó en un estudio contra placebo en prevención secundaria de pacientes con ataque isquémico transitorio o permanente que recibieron altas dosis de atorvastatina y durante su seguimiento presentaron una tendencia a eventos hemorrágicos recurrentes [108], aunque no en todos los subtipos de EVC. [109]

El estudio GRFHSS *(Genetic and Environmental Risk Factors of Hemorrhagic Stroke Study)*, observó que la hipercolesterolemia fue asociada a bajo riesgo de EVC, sin documentar incremento del riesgo por uso de estatinas [110]. Otros estudios experimentales con estatinas en prevención primaria y secundaria cardiovascular tampoco reportan mayor riesgo de sangrado intracraneal [111, 112]. En apoyo a esta teoría recientemente se publicó el estudio NASIS *(Prospective Data from the National Acute Stroke Israeli Surveys),* donde sugiere un efecto protector de las estatinas en pacientes con EVC que previamente las recibían [113].

3.1.1.3 Alcohol

El alcohol es otro factor de riesgo relacionado con EVC; probablemente, su mecanismo es dosis dependiente y se observa en alcoholismo pesado. También se asocia a expansión

temprana del hematoma [114], lo cual posiblemente se explique por alteración en la función plaquetaria y hepática [115] .

3.1.1.4 Diabetes

La diabetes mellitus está asociada a un gran riesgo de EVC en algunos estudios de casos y controles. Ariesen et al. en una revisión sistemática observó una leve significancia estadística con una razón de momios de 1.3 (intervalo de confianza del 95%, 1.02 - 1.67). [108] También en el análisis de datos realizado por Feldmann et al, del HSP *(Hemorrhagic Stroke Project)*, los pacientes entre 18 a 49 años y diabetes mellitus tenían un OR, 2.40 (intervalo de confianza del 95%, 1.15 - 5.01). [116]

3.1.1.5 Tabaquismo

El antecedente de tabaquismo (previo y/o actual al EVC), ha mostrado ser un débil factor de riesgo independiente tanto en pacientes jóvenes como en mayores de edad [108]. Esto se evidencio en un análisis de 10 estudios de casos y controles que reportaron una razón de momios de 1.25 (intervalo de confianza del 95%, 0.94 – 1.66) para tabaquismo actual y un riesgo relativo de 1.06 (intervalo de confianza del 95%,0.89 – 1.26) cuando se juntaron los estudios de cohorte con los casos y controles. Otros estudios muestran en cambio, una relación lineal de mayor riesgo en hombres que fuman >20 cigarrillos/día con una razón de momios de 2.06 (intervalo de confianza del 95%, 1.08 – 3.96), y en mujeres que fuman >15 cigarrillos/día con una razón de momios de 2.67 (intervalo de confianza del 95%, 1.04 – 6.90). [117] En el estudio de Ruiz-Sandoval y colaboradores, se observó que el 20% de los mexicanos menores de 40 años tenían antecedente de tabaquismo, sin embargo, no tuvo significancia estadística en el análisis multivariado. [118]

3.1.2 Valores séricos y hematológicos como factores de riesgo cardiovascular

Más allá del claro papel del tabaquismo, el consumo de alcohol, la hipertensión, la diabetes y la dislipidemia en el riesgo de padecer enfermedad cardiovascular en general y EVC en particular, [119] reportes recientes informan que algunos biomarcadores séricos como la

glucosa, [120, 121] el colesterol de alta y el de baja densidad,[122-124] los triglicéridos, [120, 122, 124] el ácido úrico,[125] la creatinina, [121, 126] la hormona estimulante de la tiroides [119] y la homocisteína, [123] se asocian con el riesgo de enfermedad cardiovascular. Sin embargo, cuando se intenta comparar los resultados entre estudios estos han sido inconsistentes ya que la mayoría se enfoca en un solo biomarcador, dejando de lado la posibilidad de un efecto multi biomarcador.

En un estudio realizado en China, [127] se exploró la relación entre múltiples marcadores séricos y la incidencia de enfermedad cardiovascular, los autores construyeron dos modelos, el primero considerando los valores séricos individuales y un segundo multi biomarcador y compararon la sensibilidad de los dos modelos. Sus resultados mostraron que los niveles séricos de hormona T4 libre y totales, los niveles séricos de glucosa, creatinina, triglicéridos y colesterol de baja densidad se asociaron con enfermedad cardiovascular en el modelo de valores individuales. Particularmente los triglicéridos y el colesterol de baja densidad séricos tuvieron una relación lineal con el riesgo. En el modelo multi biomarcador, la hormona T4 libre, los triglicéridos, la creatinina y el colesterol de baja densidad se asociaron positivamente con el riesgo cardiovascular.

Los autores de este trabajo notan el caso particular de la creatinina la cual no tuvo una relación lineal con el riesgo sino solo con el tercil superior (razón de momios ajustada 2.48; intervalo de confianza del 95 %, 1.17–5.27), al respecto citan otros autores que han encontrado la misma relación de la creatinina sérica con el riesgo cardiovascular [126] y aclaran que el mecanismo fisiopatológico no es conocido. En la comparación de sensibilidad, las áreas bajo la curva para la hormona T4 libre, glucosa, triglicéridos y colesterol de baja densidad en el modelo de valor único fueron: 0.569 (0.506, 0.632), 0.636 (0.575, 0.697), 0.628 (0.568, 0.689) y 0.580 (0.518, 0.642); en comparación, para el modelo multi biomarcador fue de 0.660 (0.601, 0.720) para una diferencia de 0.038.

Trabajos como el anterior permiten teorizar que la conjunción de varios biomarcadores séricos tiene potencial predictor del riesgo cardiovascular, particularmente del riesgo de enfermedad de vasos supra aórticos, particularmente de ambas carótidas. [128] Con respecto a variables hematológicas en una revisión de la literatura, Loeffen y colaboradores

[129] reportaron que gracias a la investigación reciente en ratones transgénicos propensos a la enfermedad de grandes vasos cruzados con ratones propensos a la ateroesclerosis, se pudo demostrar una relación positiva entre la hipercoagulabilidad y la ateroesclerosis.

Por lo anterior, concluyen que existe un efecto terapéutico de los fármacos inhibidores de enzimas de la coagulación sobre los vasos sanguíneos, probablemente evitando o enlenteciendo la ateroesclerosis. Sin embargo, como en casos anteriores, otros autores citan inconsistencias en los resultados y cuestionan seriamente el papel terapéutico de los anticoagulantes en la génesis de la enfermedad vascular.[130] De forma similar, otros autores han relacionado la cuenta total de leucocitos con la incidencia y mortalidad derivada de enfermedad cardiovascular, en este caso separada también en enfermedad coronaria y EVC. En el estudio de Lee y colaboradores [131] se encontró que después de ajustar para edad, sexo, lugar de reclutamiento y factores de riesgo vascular tradicionales, hubo una asociación directa entre la cuenta total de leucocitos y la incidencia de enfermedad coronaria (1.9; intervalo de confianza del 95 %, 1.19–3.09), EVC (1.9; intervalo de confianza del 95 %, 1.03–3.34) y la mortalidad cardiovascular (2.3; intervalo de confianza del 95 %, 1.38–3.72). Finalmente, estudios recientes en pacientes con COVID-19 han encontrado asociaciones significativas entre el riesgo de EVC y las cuentas totales de eritrocitos, plaquetas y neutrófilos, así como con los niveles séricos de dímero-D y otros productos de degradación de la fibrina. [132]

Con base en los datos anteriores y tomando en cuenta que los estudios sanguíneos de rutina que se realizan a pacientes que requieren la realización de estudios radiológicos contrastados incluyen la determinación de los niveles séricos de glucosa, urea, creatinina, nitrógeno ureico, ácido úrico y biometría hemática; y a la vez no incluyen estudios considerados especiales como perfil de hormonas tiroideas y marcadores de inflamación sistémica como la homocisteína, en el presente estudio se registraran solo los valores ya disponibles en el expediente del paciente para evitar solicitar exámenes sanguíneos por las implicaciones de costo y bioéticas que conllevaría.

Por otro lado, es importante agregar que si bien utilizando el lenguaje R es completamente posible aplicar las técnicas de aprendizaje automático con variables nominales del tipo

ordinal o no ordinal (por ejemplo: diabetes presente/ausente, hombre/mujer, etc.), se considera que estas aplicaciones son muy costosas en cuanto a tiempo de computación sobre todo cuando se trata de conjuntos de datos grandes.[133] El costo computacional se incrementa también debido a la necesidad de realizar pasos adicionales en el pretratamiento de los datos para posibilitar el entrenamiento de los modelos con datos que sean entendibles para los algoritmos de aprendizaje automático (variables continuas).[134] Estos pasos adicionales incluyen, por ejemplo: la creación de variables ficticias (dummy), cálculo de los predictores de varianza cero y casi cero, la identificación de predictores correlacionados y de las dependencias lineales, centrado, escalado e imputación de valores y finalmente la transformación de los predictores para poder llevar a cabo cálculos de distancia de clase. [135]

Lo anterior implica la utilización de nuevos algoritmos de aprendizaje que ayuden a los clasificadores a entender y organizar la información contenida en el conjunto de datos. Esto colocaría el presente proyecto dentro del campo transdisciplinario de la minería de datos lo cual excede el ámbito del estudio. [136] Quizá la experiencia más directa que evidencia las dificultades de incluir variables predictoras dicotómicas, ordinales y categóricas en algoritmos de aprendizaje automático viene de los análisis de riesgo que llevan a cabo las aseguradoras, de donde se sabe que la incorporación de variables continuas en su forma continua ayuda a la clasificación y desempeño de los modelos. [137]

3.2 Algoritmos de aprendizaje automático en la predicción de la anatomía del arco aórtico.

Se han aplicado una serie de algoritmos de aprendizaje automático para predecir la anatomía vascular de la aorta. Hahn [138] destaca el uso de arquitecturas U-Net para la segmentación aórtica, mientras que Ullah [139] analizó el potencial de los enfoques de aprendizaje supervisado para predecir el crecimiento de los aneurismas de aorta abdominal. Huo [140] presento un modelo de red bayesiana para clasificar a los pacientes con disección aórtica en la fase de diagnóstico precoz, y Liang [141] demostró la viabilidad del uso de redes neuronales profundas para predecir la hemodinámica de la aorta torácica

humana. Estos estudios subrayan colectivamente las diversas aplicaciones del aprendizaje automático en este campo, pero como se puede observar, ninguno de los ejemplos proporcionados se aplica directamente al tratamiento endovascular del infarto cerebral.

3.2.1 Trombectomía mecánica

3.2.1.1 Asociación del tiempo a la reperfusión con la eficacia

Las importantes mejoras técnicas en la trombectomía mecánica (TM) han resultado en tasas significativamente más altas de revascularización y mejores resultados clínicos en el infarto cerebral agudo (arteria carótida interna) debido a la oclusión intracraneal de grandes vasos (OIGV).

Nueve importantes ensayos controlados aleatorios han establecido la eficacia clínica de la reperfusión mecánica en comparación con el tratamiento médico solo en ventanas de tiempo tempranas y tardías. Sin embargo, la morbilidad y la mortalidad siguen siendo considerables en los pacientes con OIGV a pesar de la TM, como lo demuestran las altas tasas de dependencia funcional a largo plazo (puntuación de 3 a 6 en la escala de Rankin modificada [mRS] de 90 días: 28 % a 67.4 %) y mortalidad (9 % a 24%) en estos estudios.

El beneficio clínico de la TM en oclusiones de la arteria carótida interna está estrechamente vinculado al éxito técnico del procedimiento, específicamente a la capacidad de lograr una reperfusión rápida y completa del territorio cerebral afectado.

Este éxito se evalúa habitualmente mediante la escala Thrombolysis in Cerebral Infarction (TICI), la cual clasifica el grado de recanalización angiográfica posterior al tratamiento.

A pesar de los avances tecnológicos y la creciente experiencia en el tratamiento endovascular, los ensayos clínicos recientes reportan tasas de reperfusión favorable, definida como TICI ≥2b, que varían entre el 68 % y el 88 %.

La **¡Error! No se encuentra el origen de la referencia.** resume la clasificación TICI modificada habitualmente utilizada en estudios clínicos:

Tabla 5 Clasificación TICI

Grado TICI	Definición
0	Sin perfusión
1	Perfusión mínima, sin llenado distal
2a	Perfusión parcial con llenado <50 % del territorio distal del vaso afectado
2b	Perfusión parcial con llenado ≥50 % del territorio distal
2c	Perfusión casi completa con flujo lento o defectos de llenado distales
3	Perfusión completa, sin defectos de llenado

Lograr una reperfusión favorable a menudo requiere múltiples intentos de trombectomía y el uso de dispositivos y fármacos de rescate. La reperfusión completa rara vez se logra durante el primer paso del dispositivo. [142] Además de prolongar el tiempo del procedimiento,[143] los pases múltiples del dispositivo pueden promover la lesión del endotelio arterial,[144] lo que podría reducir la eficacia clínica [145] y aumentar la aparición de eventos adversos.[146]

Los estudios de reperfusión altamente eficaz mediante dispositivos endovasculares múltiples (HERMES) agruparon el análisis de los primeros cinco ensayos contemporáneos de TM informaron que la probabilidad de independencia funcional (mRS 0-2) a los 3 meses disminuyó del 64.1 % con un tiempo de aparición de síntomas a reperfusión de 180 minutos al 46.1% con un tiempo de inicio de los síntomas a la reperfusión de 480 minutos.[147]

El tiempo de procedimiento (TP) podría representar una proporción significativa del tiempo de inicio a la reperfusión en escenarios técnicamente desafiantes.[148] Muchos estudios han informado que cuanto más largo es el TP, mayor es el riesgo de transformación hemorrágica y menores las probabilidades de un buen resultado clínico.[145, 149-153]

Si bien falta una definición precisa de trombectomía refractaria (TR), puede definirse como un procedimiento que dura demasiado, requiere más de tres pases o no logra obtener un grado aceptable de reperfusión (TICI ≥2b). La TR es un problema multifacético que comprende factores relacionados con el paciente (anatomía vascular cervical y la naturaleza subyacente de la lesión oclusiva, incluida la composición del trombo y la presencia de placa aterosclerótica) y el procedimiento (selección adecuada del dispositivo y la técnica).[154-156]

3.3 Anatomía vascular cervical

La anatomía vascular cervical desfavorable es un desafío común para los procedimientos de neurointervención. Esta anatomía se puede encontrar de forma aislada o en combinación en muchos niveles diferentes, incluido el arco aórtico, la arteria carótida común, la arteria carótida interna cervical, el sifón carotídeo y la circulación intracraneal. El acceso difícil del catéter transfemoral a la arteria objetivo se relaciona con un TP más largo, una tasa más baja de recanalización y una tasa más baja de resultados favorables.

Kaesmacher y colaboradores [155] informaron que una de las razones del fracaso de la reperfusión puede ocurrir en hasta uno de cada tres pacientes (20 de 63 pacientes; 31.7%; intervalo de confianza del 95%, 20.3% a 43.2%). La imposibilidad de alcanzar el vaso ocluido generalmente se relacionaba con una anatomía arterial compleja a nivel del cayado aórtico o la tortuosidad del vaso cervical.

La anatomía compleja del arco aórtico y de la arteria carótida se asocia con carga aterosclerótica y se presenta con más frecuencia en pacientes de edad avanzada.[157] Ribo y colaboradores [148] comunicaron una mediana de tiempo desde la punción inguinal hasta el cateterismo carotídeo de 20 minutos. Propusieron una puntuación de riesgo para identificar a los pacientes de alto riesgo para el acceso supra aórtico desafiante (punción en la ingle para cateterismo carotídeo >30 minutos) que incluía hipertensión, edad >75 años, dislipidemia y accidente cerebrovascular en la circulación anterior izquierda. Una puntuación >2 predijo un acceso difícil, con una sensibilidad del 84 % y una especificidad del 74 %.[146]

Snelling y colaboradores [151] propusieron una puntuación basada en criterios anatómicos que incluían el tipo de arco aórtico, así como la presencia de arco bovino y dólico arteriopatía de la arteria carótida interna.

Las puntuaciones más altas predijeron de forma independiente un mayor tiempo desde la punción inguinal hasta el primer paso, puntuaciones más bajas de reperfusión (TICI) y resultados clínicos desfavorables después de la trombectomía.

El análisis preoperatorio de la angiografía por tomografía computarizada puede identificar rápida y consistentemente las características vasculares que pueden estar relacionadas con el acceso difícil a los vasos supra aórticos (Tabla 6)

Tabla 6 Características relacionadas con acceso difícil a los vasos supra aórticos

Síndrome de Leriche
Arco Aórtico tipo I
Arco Aórtico tipo II
Arco Aórtico tipo III
Variante Bovina del Arco Aórtico

3.3.1 Alteraciones Morfológicas

Las anomalías morfológicas de la arteria carótida interna cervical están asociadas con enfermedades cerebrovasculares y pueden representar un desafío importante para el tratamiento endovascular de enfermedades cervicales e intracraneales. Si bien la prevalencia de estas deformidades varía según los criterios morfológicos y las técnicas diagnósticas utilizadas, se identifican hasta en el 85.8% de los pacientes; Nagata y colaboradores [158] consideraron que una arteria carótida interna era tortuosa si el ángulo entre la arteria carótida común y las líneas centrales de la arteria carótida interna era >15° o si el trayecto de la arteria carótida interna tenía forma de S o de C. La arteria carótida interna se consideró recta si el ángulo era <15°. La morfología de la arteria carótida interna se clasificó además como enrollada (una curva en forma de S exagerada o una configuración circular; 3% de los casos) o torcedura (ángulo entre los segmentos de los vasos <90° y asociado con estenosis; 1.0 % de los casos) (Figura 12).

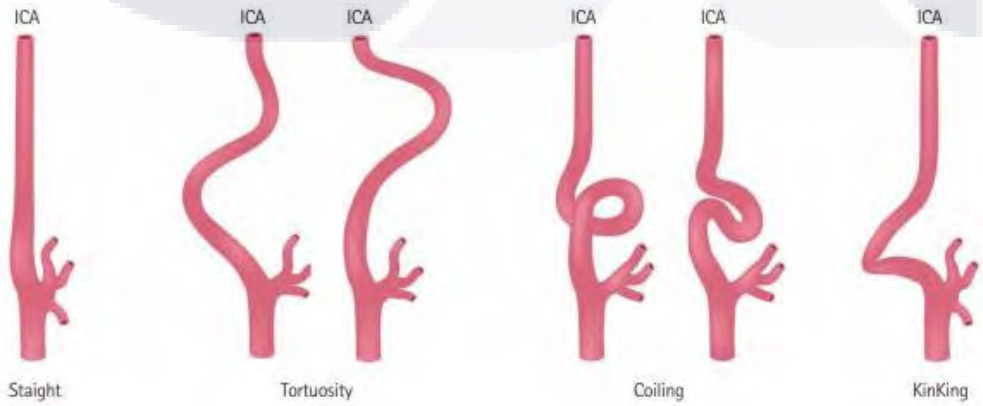


Figura 12 Morfología normal y anormal de la arteria carótida interna. 58

Estas deformidades pueden estar relacionadas con enfermedades congénitas (p. ej., displasia fibromuscular) o adquiridas. En general, las anomalías vasculares son adquiridas y se asocian con factores de riesgo de aterosclerosis como hipertensión, hiperlipidemia y tabaquismo. La prevalencia de la elongación de los vasos aumenta con la edad y se cree que está relacionada con la pérdida de elasticidad con el envejecimiento.[159]

La elongación de la arteria carótida interna en el cuello generalmente no es un obstáculo importante durante la TM con la colocación adecuada del catéter guía. Sin embargo, pueden presentarse algunas dificultades anatómicas más allá de la elongación y la tortuosidad extrema, como torceduras y enrollamientos, que pueden influir en las posibilidades de una recanalización exitosa.

La anatomía cervical tortuosa puede impedir el posicionamiento óptimo del acceso distal o los catéteres de aspiración intracraneal, que luego pueden herniarse de nuevo en el arco aórtico durante el intento subsiguiente de lograr el acceso intracraneal. De hecho, una posición de arco aórtico por debajo del margen inferior del cuerpo vertebral C1 y la presencia de tortuosidad carotídea se han asociado con tasas de recanalización más bajas con TM.[157] Esto podría estar relacionado con el riesgo de pérdida del acceso debido a una posición inestable del catéter o una aspiración menos eficiente. El segmento petroso de la arteria carótida interna está firmemente adherido al hueso petroso a diferencia de la arteria carótida interna extracraneal, que está rodeada de tejido blando. Por lo tanto, la aspiración a través del catéter, cuando se localiza proximalmente, aumenta el riesgo de colapso arterial, lo que atenúa la fuerza de eliminación del coágulo y favorece la embolia distal.[157]

La clasificación de los pacientes con acceso cervical técnicamente difícil durante la TM es fundamental, ya que permite una mejor selección de dispositivos y estrategias alternativas en una etapa temprana del procedimiento (Tabla 7).

Tabla 7 Características relacionadas con acceso difícil a los vasos cervicales

Dólico-arteriopatía	Curvatura el vaso a tratar
Tortuosidad del Sifón Carotídeo	Coágulo en bifurcación

La reperfusión endovascular proporciona beneficios abrumadores tanto en ventanas de tiempo tempranas como tardías. No obstante, el fracaso para lograr una reperfusión efectiva ocurrió en hasta el 25% de los casos en ensayos recientes. Las razones de la falta de respuesta al tratamiento son multifactoriales y se manifiestan de manera diferente, incluido un mayor número de pases, TP más largos y reperfusión fallida o incompleta. Si bien la tecnología de tratamiento endovascular está evolucionando rápidamente, sigue siendo fundamental comprender mejor los mecanismos específicos que subyacen a los desafíos individuales que se enfrentan en la práctica diaria, ya que esto debería guiar nuestras elecciones de dispositivos y técnicas. Los factores anatómicos desfavorables son un obstáculo importante durante el tratamiento endovascular de la arteria carótida interna. La identificación rápida de las características asociadas podría permitir estrategias adecuadas para superar los desafíos de la anatomía vascular cervical, ahorrando tiempo y asegurando mejores resultados angiográficos y clínicos.

4. Propuesta Metodológica

4.1 Diseño

Se llevó a cabo un estudio transversal analítico de prueba diagnóstica.

4.2 Universo de Trabajo

Todos los pacientes con hipertensión, diabetes o dislipidemia a quienes se les realice durante el período de reclutamiento establecido en el cronograma del presente protocolo, una angiotomografía de vasos supra aórticos atendidos en el Hospital General de Zona #2 del IMSS, OOAD Aguascalientes.

4.3 Fuentes de información del estudio

Los datos del estudio provinieron de dos fuentes de información.

1. Expediente clínico electrónico (plataforma ECE, disponible en https://www.ece.imss.gob.mx/ de la intranet del IMSS.
2. Expediente radiológico electrónico (plataforma PACS del Hospital General de Zona #2, OOAD Aguascalientes, disponible en https://172.47.1.7/PACS de la intranet del mismo hospital.

Tomando en cuenta que todos los estudios radiológicos contrastados tienen como prerrequisito la realización de exámenes de sangre y que las indicaciones de angiotomografía de vasos supra aórticos son todas asociadas a patología vascular (periférica, cardiaca o cerebral); en el presente estudio no se llevaron a cabo entrevistas ni se tomaron nuevas muestras de laboratorio a ningún participante.

4.4 Lugar donde se desarrolló el estudio

- Servicio de Medicina Interna del Hospital General De Zona # 2 del IMSS, OOAD Aguascalientes.
- Campus Central de la Universidad Autónoma de Aguascalientes.

4.5 Aspectos Éticos

En la realización del presente proyecto de investigación, se consideraron y se tomaron en cuenta los diferentes principios éticos para la investigación en humanos, los cuales están disponibles en la Declaración de Helsinki (En su versión revisada por la 64ª Asamblea Médica Mundial de Fortaleza, Brasil, de 2013, y por la Asamblea General de la OMS). [160] De acuerdo con lo estipulado en el Reglamento de la Ley General de Salud en Materia de Investigación para la Salud, en su título segundo, capítulo I, artículo 17, el presente estudio queda clasificado como: I.- Investigación sin riesgo. Ya que los métodos y técnicas fueron retrospectivos y no se realizó ninguna intervención o modificación intencionada en las variables fisiológicas, psicológicas y sociales de los individuos que participaron en el estudio (revisión de expedientes clínicos). [161]

De acuerdo con el Centro de Investigación Clínica General de los Institutos Nacionales de Salud de EE. UU. se le consideró una investigación con riesgo grado I (Riesgo mínimo) en una escala de I a IV. [162]

El protocolo de investigación fue revisado y aprobado por los Comités Locales de Investigación y Ética en Investigación en Salud del IMSS previo a su instrumentación. Esta investigación no involucrará poblaciones vulnerables como niños, mujeres embarazadas o grupos subordinados. No se registraron datos personales que permitan identificar a los participantes. Los investigadores asociados al proyecto no tuvieron conflictos de interés que declarar.

4.6 Descripción general del estudio

El presente estudio constó de tres fases:

4.6.1 Primera fase

4.6.1.1 Conformación del conjunto de datos.

Los datos se obtuvieron de pacientes atendidos en el Hospital General De Zona # 2 del IMSS, OOAD Aguascalientes. Se recabaron datos demográficos, antropométricos, clínicos y radiológicos (imágenes médicas) de todos los pacientes con hipertensión, diabetes o dislipidemia a quienes se les realice un estudio de angiotomografía de vasos supra aórticos.

Como se muestra en la Figura 13, la angiotomografía de vasos supra aórticos es una prueba radiológica que consiste en obtener imágenes de alta definición anatómica de las arterias que emergen del arco aórtico mediante el empleo de un equipo de tomografía computarizada y la inyección de contraste intravenoso. Posteriormente, las imágenes son reconstruidas en tres dimensiones.[163]

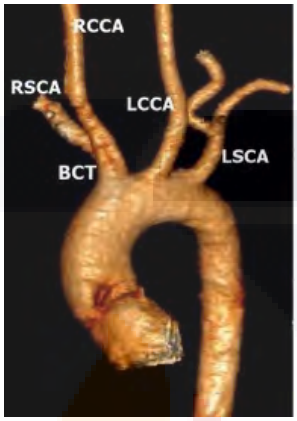


Figura 13 Imagen de angiotomografía de vasos supra aórticos.[164]

La indicación del estudio radiológico fue responsabilidad de los médicos tratantes, ninguno de los investigadores asociados al proyecto tuvo injerencia en la decisión de solicitar las angiotomografías.

4.6.1.2 Selección y tamaño de la muestra

- Tipo de muestreo

Consecutivo por conveniencia

- Tamaño de la muestra

4.6.1.3 Parámetros para el cálculo

Para el cálculo del tamaño de la muestra, se utilizaron los supuestos para tamaño de la muestra y poder estadístico propuestos por Guo y colaboradores [165] para estudios de algoritmos de clasificación.

El tamaño de la muestra (n) para el caso de conjuntos de datos de alta dimensionalidad varía entre 50 y 200 sujetos por clase para casos cuyas mediciones incluyen hasta 1000 características por sujeto.

Se asume que los sujetos pertenecen a una de dos clases en este caso, con anatomía vascular cervical favorable (AVF) o con anatomía vascular cervical no favorable (AVNF).

Mientras no se especifique lo contrario cada característica se toma como una variable independiente aleatoria. Se asume también que un porcentaje k, de entre las características de cada sujeto, se expresa de forma diferencial en AVF comparado con AVNF; a estas características se les denomina marcadores, y el resto de las características (100 - k) son características no discriminantes. El porcentaje de marcadores también puede variar entre 0.5 y 5%. Se asumen una distribución Gaussiana o una distribución mixta de las características en cada clase, de tal forma que los datos se generan de la siguiente manera:

Sea:

$$X|"AVF" \sim N(0, \quad \sigma^2 = 0.20),$$
$$X|"AVNF" \sim N(\delta, \quad \sigma^2 = 0.20)$$

donde δ varía entre 0.05 y 0.5 para los marcadores. δ se asume 0 para las características del ruido. En las simulaciones donde se asumió que las características seguían una distribución no Gaussiana, los datos se generaron de acuerdo con lo siguiente:

$$X|"AVF" \sim I[U \leq 0.9]Z_1 + I[U > 0.9]Z_2,$$
$$X|"AVNF" \sim I[U \leq 0.9]Z_{1*} + I[U > 0.9]Z_{2*}$$

donde se asume que U, Z_1, Z_2, Z_{1*} y Z_{2*} son variables independientes aleatorias. Las distribuciones de las variables aleatorias U, Z_1, Z_2, Z_{1*} y Z_{2*} fueron:

$$U \sim Uniforme(0,1)$$
$$Z_1 \sim N(\mu, \sigma^2)$$
$$Z_2 \sim Uniforme(\mu + c_{1\sigma}, \mu + c_{2\sigma})$$
$$Z_{1*} \sim N(\mu^*, \sigma^2)$$
$$Z_{2*} \sim Uniforme(\mu^* + c_{1\sigma}, \mu^* + c_{2\sigma})$$

Aquí $N(\mu, \sigma^2)$ se refiere a una distribución Gaussiana con media m y varianza s^2.

Se asumieron los siguientes valores: $s^2 = 0.2$, $c_1 = 3.0$ and $c_2 = 6.7$.

Basándose en las asunciones de arriba, el valor esperado de cada característica (X) entre AVF y AVNF es:

$$E(X|\text{AVF}) = \mu + \frac{0.1_{\sigma}(c_2 + c_1)}{2}$$

$$E(X|\text{AVNF}) = \mu^* + \frac{0.1_{\sigma}(c_2 + c_1)}{2}$$

Los valores para μ y μ^*se escogen de tal forma que $E(X|\text{AVF}) = 0$ y $E(X|\text{AVNF})$ toma los valores (0, 0.05, 0.10, …,0.50). Basándose en todos los parámetros anteriores la varianza promedio y la asimetría de cada característica fue de 0.80 y 1.50, respectivamente.

Los resultados derivados de los cálculos anteriores demuestran el tamaño de muestra mínimo requerido para determinar si un algoritmo de clasificación funciona significativamente mejor que el azar. Aunque un tamaño de muestra elevado puede resultar en un poder elevado, el clasificador resultante puede no incluir todos los marcadores existentes que realmente discriminan entre AVF y AVNF. Esto también podría resultar en conjuntos de marcadores que tienen poca superposición en experimentos comparables. La estrategia de simulación y los tamaños de muestra resultantes discutidos por Guo y colaboradores [165] son apropiados para estudios realizados durante la etapa inicial del descubrimiento de marcadores, como entornos clínicos en los que se desconoce la existencia de un conjunto de marcadores con capacidad de discriminación clínicamente útil entre clases de resultados. Cuando las fases iniciales de investigación han demostrado ser exitosas, los estudios de validación posteriores pueden diseñarse con condiciones más estrictas destinadas a garantizar el descubrimiento de todos los marcadores relevantes, así como a establecer un alto grado de reproducibilidad.[166]

4.6.1.4 Criterios de inclusión y exclusión

4.6.1.4.1 Criterios de inclusión

- Sujetos adultos
- Diagnosticados con por lo menos una de las siguientes enfermedades crónicas no transmisibles:
    - Hipertensión arterial sistémica

- Diabetes Mellitus
- Dislipidemia

- Cuenten con angiotomografía de vasos supra aórticos

4.6.1.4.2 Criterios de no inclusión

- Cáncer
- Enfermedades Respiratorias Crónica p.ej. neumopatía obstructiva crónica, asma.
- Demencia

4.6.1.4.3 Criterios de exclusión

- Expediente incompleto
- Imágenes de angiotomografía de vasos supra aórticos no disponibles o sin interpretación radiológica

4.6.1.5 Variables independientes

4.6.1.5.1 Demográficas:

Variable	Definición Operacional	Definición Conceptual	Instrumento	Tipo	Codificación Unidades
Edad	Cálculo de tiempo basado en interrogatorio la historia clínica	Tiempo transcurrido desde el nacimiento del paciente	Historia Clínica	Continua	Años
Sexo	Sexo del paciente basado en interrogatorio la historia clínica	Fenotipo Sexual del paciente	Historia Clínica	Categórica Nominal Dicotómica	1 Mujer 2 Hombre

4.6.1.5.2 Antropométricas

Variable	Definición Operacional	Definición Conceptual	Instrumento	Tipo	Codificación Unidades
Talla	Medición en centímetros de la longitud del cuerpo del sujeto desde el	Tamaño del individuo desde la coronilla de la cabeza hasta los pies	Historia Clínica	Continua	centímetros

	vértice craneal hasta la planta de los pies en contacto con el piso				
Peso	Magnitud que mide la fuerza con la que la gravedad presiona o atrae la masa del sujeto hacia la tierra	Medida de la masa corporal	Historia Clínica	Continua	kilogramos

4.6.1.5.3 Clínicas

Variable	Definición Operacional	Definición Conceptual	Instrumento	Tipo	Codificación Unidades
Hipertensión Arterial (HTAS)	Registro en el expediente de presión sistólica por encima de 139mm$_{Hg}$ o presión diastólica mayor de 89mm$_{Hg}$ O registro de estar tomando medicamentos antihipertensivos	Enfermedad crónica caracterizada por un incremento continuo de las cifras de presión sanguínea en las arterias	Historia Clínica	Categórica Nominal Dicotómica	0 Sin HTAS 1 Con HTAS
Tiempo con HTAS	Tiempo transcurrido en años en el paciente cumple la definición operacional de HTAS	Duración en el tiempo en que el sujeto ha padecido HTAS	Historia Clínica	Continua	años
Diabetes mellitus (DM)	Registro en el expediente de por lo menos un valor de glucosa sérica en ayuno >125mg/dL O registro de actualmente	Conjunto de trastornos metabólicos, que afectan a diferentes órganos y tejidos, caracterizados por un aumento de los niveles de glucosa en la sangre	Historia Clínica	Categórica Nominal Dicotómica	0 Sin Diabetes 1Con Diabetes

	tomando medicamentos hipoglucemiantes				
Tiempo con DM	Tiempo transcurrido en años en el paciente cumple la definición operacional de DM	Duración en el tiempo en que el sujeto ha padecido DM	Historia Clínica	Continua	años
Tabaquismo	Registro en el expediente de consumo de por lo menos 1 cigarrillo semanal por lo menos 2 semanas de cada mes durante los últimos 6meses O registro de hábito de fumar.	El hábito de fumar (inhalar y exhalar los humos producidos al quemar el tabaco	Historia Clínica	Categórica Nominal Dicotómica	0 No Fumador 1 Fumador
Tiempo con tabaquismo	Tiempo transcurrido en años en el paciente cumple la definición operacional de tabaquismo	Duración en el tiempo en que el sujeto incurre en tabaquismo	Historia Clínica	Continua	años
Hiper colesterolemia	Nivel de colesterol total sérico >200mg/dL en ayuno O actualmente tomando medicamentos hipolipemiantes	Aumento de los niveles séricos de colesterol en la sangre	Historia Clínica	Categórica Nominal Dicotómica	0 Colesterol Normal 1 Colesterol Alto
Tiempo con Hiper colesterolemia	Tiempo transcurrido en años en el paciente cumple la	Duración en el tiempo en que el sujeto ha	Historia Clínica	Continua	años

	definición operacional de hipercolesterolemia	padecido hipercolesterolemia			
Hiper trigliceridemia	Nivel de triglicéridos séricos >175mg/dL en ayuno O actualmente tomando medicamentos para bajar los triglicéridos	Aumento de los niveles séricos de triglicéridos en la sangre	Historia Clínica	Categórica Nominal Dicotómica	0 Colesterol Normal 1 Colesterol Alto
Tiempo con Hiper trigliceridemia	Tiempo transcurrido en años en el paciente cumple la definición operacional de hipertrigliceridemia	Duración en el tiempo en que el sujeto ha padecido hipertrigliceridemia	Historia Clínica	Continua	años

4.6.1.5.4 De laboratorio

Variable	Definición Operacional	Definición Conceptual	Instrumento	Tipo	Codificación Unidades
Nivel sérico de glucosa	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición del nivel de glucosa sérica por medio de un ensayo de laboratorio estandarizado	Cantidad de glucosa circulante en sangre en el momento de la toma de la muestra	Quimioluminiscencia	Continua	(mg/dL)
Nivel sérico de creatinina	Resultado con fecha más cercana a la realización del estudio de angiotomografía de	Cantidad de creatinina circulante en sangre en el momento de	Quimioluminiscencia	Continua	(mg/dL)

	vasos supra aórticos de la medición del nivel de creatinina sérica por medio de un ensayo de laboratorio estandarizado	la toma de la muestra			
Nivel sérico de urea	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición del nivel de urea sérica por medio de un ensayo de laboratorio estandarizado	Cantidad de urea circulante en sangre en el momento de la toma de la muestra	Quimioluminiscencia	Continua	(mg/dL)
Nivel sérico de nitrógeno ureico	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición del nivel de nitrógeno ureico sérico por medio de un ensayo de laboratorio estandarizado	Cantidad de nitrógeno ureico circulante en sangre en el momento de la toma de la muestra	Quimioluminiscencia	Continua	(mg/dL)
Nivel sérico de ácido úrico	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición del nivel de ácido úrico sérico por medio de un ensayo de	Cantidad de glucosa circulante en sangre en el momento de la toma de la muestra	Quimioluminiscencia	Continua	(mg/dL)

	laboratorio estandarizado				
Leucocitos	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición sérica del número de leucocitos sanguíneos por medio de un ensayo de laboratorio estandarizado	Cantidad de leucocitos circulantes en sangre en el momento de la toma de la muestra	Auto analizador automático	Continua	10^9/L
Eritrocitos	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición sérica del número de eritrocitos sanguíneos por medio de un ensayo de laboratorio estandarizado	Cantidad de eritrocitos circulantes en sangre en el momento de la toma de la muestra	Auto analizador automático	Continua	10^{12}/L
Plaquetas	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición sérica del número de plaquetas sanguíneas por medio de un ensayo de laboratorio estandarizado	Cantidad de plaquetas circulantes en sangre en el momento de la toma de la muestra	Auto analizador automático	Continua	10^9/L

Hemoglobina	Resultado con fecha más cercana a la realización del estudio de angiotomografía de vasos supra aórticos de la medición sérica de la hemoglobina sanguínea por medio de un ensayo de laboratorio estandarizado	Cantidad de hemoglobina circulante en sangre en el momento de la toma de la muestra	Auto analizador automático	Continua	g/L
INR	Valor numérico registrado en el expediente clínico derivado de la prueba de coagulación.	Cociente internacional normalizado del tiempo de protrombina, que estandariza la evaluación de la coagulación.	Coagulometría	Continua	Unidades INR
Tiempo de protrombina	Tiempo en segundos anotado en el expediente clínico correspondiente a la prueba de TP.	Tiempo que tarda el plasma en coagular mediante la vía extrínseca y común de la coagulación.	Coagulometría	Continua	Segundos
Tiempo parcial de tromboplastina (TPT)	Tiempo en segundos registrado tras la prueba del TPT o TTPa.	Tiempo requerido para formar un coágulo mediante la vía intrínseca y común de la coagulación.	Coagulometría	Continua	Segundos

4.6.1.6 Variable dependiente

Clasificación del tipo de arco aórtico observado en el estudio de angiotomografía de vasos supra aórticos por medio del estándar de oro, que es la evaluación por parte del médico

especialista (Médico Radiólogo/Neurólogo), en Tipo 1 = favorable o Tipo 2 = desfavorable, tal y como se muestra en la Figura 14.

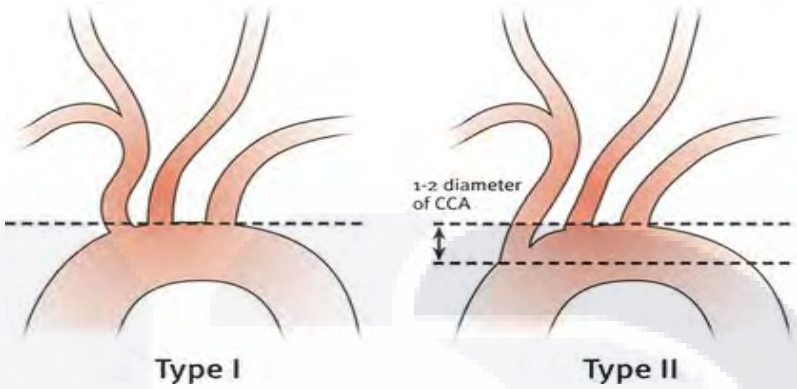


Figura 14 Clasificación del tipo de arco aórtico.[164]

Las imágenes de TC se utilizaron para crear una reconstrucción tridimensional de los vasos arco aórtico y supraaórticos. A continuación, las reconstrucciones tridimensionales se limpiaron manualmente del tejido extravascular. Una vez finalizado el proceso de limpieza, el arco aórtico se clasificó siguiendo la definición propuesta por Casserly y colaboradores [167], que establece que un arco aórtico tiene una anatomía favorable si la arteria innominada, la arteria carótida común izquierda y la arteria subclavia emergen del arco aórtico a una distancia igual o menor que el grosor de la arteria carótida común izquierda medido desde una línea tangente hasta el punto más alto del borde superior del arco aórtico en una vista anterior.

Todos los arcos aórticos que no cumplieron con estos requisitos fueron considerados desfavorables. La Figura 15 muestra imágenes de muestra de la reconstrucción tridimensional de un arco aórtico con anatomía vascular cervical desfavorable (panel A) y un arco aórtico con anatomía vascular cervical favorable (panel B).

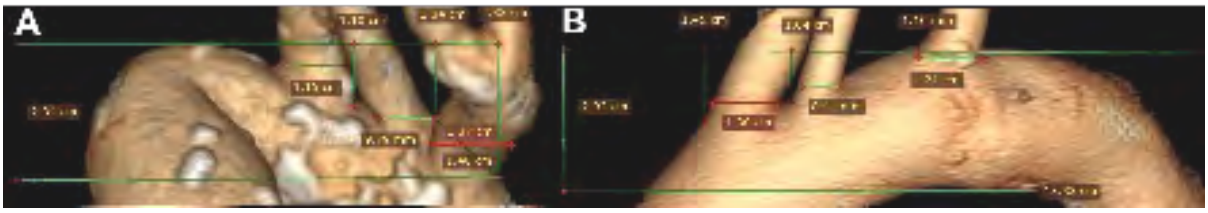


Figura 15 Ejemplos de imágenes de angiotomografía.

4.6.2 Preprocesamiento de los datos

Los expedientes electrónicos revisados para el estudio contenían algunos de los atributos como datos nominales y otros como datos numéricos. Se utilizó un preprocesamiento para transformar los datos sin procesar en un formato adecuado para alimentarlos a los algoritmos de aprendizaje automático.

Todos los datos nominales se capturaron como datos numéricos digitales. Por ejemplo, la anatomía favorable se convirtió en 0, y la desfavorable en 1.

En este estudio, se evitaron los datos faltantes consultando entradas adicionales a los registros electrónicos de salud si había valores faltantes en la entrada más cercana a la fecha de la angiotomografía.

Los registros en el expediente electrónico se consultaron retrospectivamente hasta que no hubo ningún dato faltante.

También se llevó a cabo un proceso de normalización de los datos, restando la media de cada columna en el conjunto de datos a cada valor continuo en la misma columna. Los valores generados se dividieron luego entre su desviación estándar.

Este proceso, estandarizó los datos y los hizo más manejables para el análisis.

Posteriormente, todo el conjunto de datos se dividió aleatoriamente en diez subconjuntos que contenían clases equilibradas (mismo número de casos con anatomía vascular cervical favorable y desfavorable en cada subconjunto), asegurando la equidad en la división del conjunto de datos.

4.6.3 Segunda fase

4.6.3.1 Análisis del conjunto de datos

4.6.3.1.1 Lenguaje R

Para la implantación de todos los experimentos en esta tesis se utilizó el lenguaje de programación estadística R, [168] una herramienta de código abierto para estadísticas y programación que se desarrolló como una extensión del lenguaje S. El software R cuenta con el respaldo de una gran comunidad de usuarios activos y alberga varios paquetes excelentes para aprendizaje automático que son flexibles y fáciles de usar. R es un lenguaje

computacionalmente eficiente que es fácilmente comprensible sin una formación especial en informática. El lenguaje R es similar a muchos otros lenguajes de programación estadística, incluidos MATLAB, SAS y STATA. Los paquetes para R se organizan en diferentes vistas de tareas en la red integral de archivos de R. La vista de tareas de aprendizaje automático cuenta actualmente con casi 100 paquetes dedicados a aprendizaje automático. El programa de acceso a R más común es RStudio®, [169] un entorno de desarrollo integrado de código abierto que está diseñado para facilitar el trabajo en R. Tanto R como RStudio® son de uso gratuito y están disponibles para su uso bajo una licencia de código abierto.

4.6.3.1.2 Preprocesamiento de los datos

Paso 1. Importar y preparar el conjunto de datos. A partir del archivo generado en SPSS se exporto en formato .xlsx y posteriormente se importó a RStudio®.

Paso 2. Creación de los subconjuntos de datos de entrenamiento y prueba utilizando funciones programadas *exprofeso*. La función acepta tres parámetros:

1. f es la fórmula para crear el modelo de aprendizaje automático. La fórmula tiene la siguiente sintaxis:
    a. VARIABLE DEPENDIENTE ~ VARIABLES INDEPENDIENTES (vector de variables concatenadas con el operador "+", o usando el operador comodín "." Para la utilización de todas las variables contenidas en el set de datos. Ejemplo:

```
f = ARC ~ .  # Crea un modelo con todas las variables disponibles
f = ARC~ Edad + Sexo # Crea un modelo solo con Edad y Sexo.
```

2. k es el número de subconjuntos a crear. Diez en el presente trabajo.
3. Split es la proporción de cada subconjunto a utilizar como datos de entrenamiento. Setenta por ciento en este trabajo. Ejemplo:

```
train_ds(f = "ARC ~ .", k = 10, split = 70)
```

El proceso de preprocesamiento se hace explícito en el seudocódigo siguiente:

Seudocódigo para el preprocesamiento de los datos.

1. Cargar el paquete necesario para leer archivos de Excel.
   a. Utilizar una función de lectura de archivos para cargar un archivo desde una ruta de acceso especificada.
2. Convertir los datos a un formato de tabla (*data frame*).
3. Escalar ciertas columnas del conjunto de datos:
   a. Seleccionar un rango de columnas y aplicar una función de escalado a esas columnas.
   b. Combinar las columnas escaladas con el resto de las columnas no escaladas.
4. Convertir un conjunto específico de columnas en factores:
   a. Aplicar una transformación para convertir los valores de estas columnas en factores.
5. Definir una función para dividir el conjunto de datos en subconjuntos de entrenamiento:
   a. Establecer una semilla aleatoria para garantizar la reproducibilidad de los resultados.
   b. Convertir la entrada recibida (una fórmula) en una fórmula válida.
6. Crear particiones para el conjunto de entrenamiento:
   a. Utilizar una función que particione los datos basándose en la variable objetivo.
   b. Especificar el número de particiones (k) y la proporción del conjunto de datos que se usará para entrenar.
   c. Configurar los parámetros para determinar cuántos grupos se usarán en la partición, asegurando que haya al menos un mínimo de grupos.
7. Retornar las particiones de datos resultantes desde la función.
8. La función devuelta puede ser utilizada posteriormente para crear subconjuntos de datos de entrenamiento.

4.6.3.1.3 Clasificación

Los algoritmos se implementaron para construir modelos de aprendizaje automático que pudieran analizar los datos de entrada y luego, se utilizaron los modelos creados para clasificar la anatomía vascular cervical como favorable o desfavorable.

Cada algoritmo de aprendizaje automático se entrenó utilizando el conjunto de entrenamiento, para encontrar patrones que mapeaban los atributos de los datos de entrada a la clase objetivo. El modelo de aprendizaje automático generado después del entrenamiento captura estos patrones para futuras predicciones. Luego, el modelo generado se utilizó para predecir la clase de los datos en el conjunto de prueba.

El proceso de entrenamiento/prueba se realizó a través de una validación cruzada de diez iteraciones utilizando recursivamente cada subconjunto generado durante el preprocesamiento como un conjunto de prueba y los subconjuntos restantes como conjuntos de entrenamiento. Por lo tanto, cada iteración de la validación cruzada se realizó con conjuntos de entrenamiento y prueba en proporción 70%-30%. Ambos conjuntos de

datos se utilizaron para computar la clasificación de cada sujeto de acuerdo con su anatomía vascular cervical, como se muestra en la Figura 16, siendo el proceso a) la fase de entrenamiento y el proceso b) la fase de prueba.

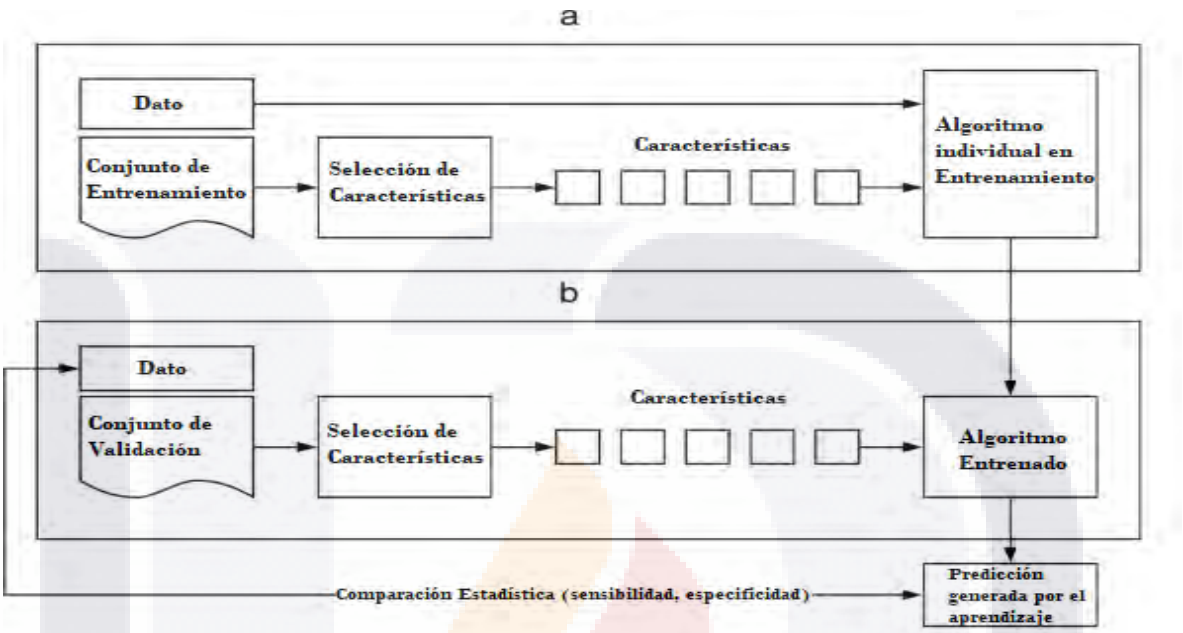


Figura 16 Esquema general de la Fase 2 del estudio.

En resumen, los conjuntos de entrenamiento se utilizaron para entrenar a los algoritmos, y los conjuntos de prueba se utilizaron para evaluar su rendimiento en cada iteración. Aunque todas las divisiones de los conjuntos de entrenamiento y prueba estuvieron compuestas por 108 instancias, cada división contenía una combinación diferente de las instancias. Los algoritmos fueron computados utilizando el entorno de programación de código abierto R. Entrenamiento de los algoritmos de aprendizaje automático. Una vez organizado el conjunto de datos en un formato adecuado, se entrenaron los siguientes algoritmos:

1. Regresión logística (RL) utilizando modelos lineales generalizados, se implementó en R usando el paquete *glmnet*. [39]
2. Clasificador ingenuo de Bayes (CIB) que se implementó en R con el paquete *naivebayes*.[170]
3. k-vecinos más próximos (k-NN), se implementó en R usando el paquete *class*.[171]
4. Árboles de decisión (ArD) que se implementaron en R con el paquete *rpart*.[172]

5. Análisis discriminante (AnD) que se implementó en su variante regularizada en R con el paquete *MASS*.[173]
6. Red neuronal (NNet) que se implementó en R con el paquete *nnet*.[174]
7. Máquinas de vectores de soporte (SVM) con un núcleo de función de base radial, se implementó en R usando el paquete *e1071*.[175]
8. Cópula Gaussiana (CG) que se implementara en R con el paquete *MLCOPULA*. [176]

4.6.4 Tercera fase

4.6.4.1 Comparación del desempeño

4.6.4.1.1 Métricas de desempeño

Una vez que se tengan los resultados para cada uno de los métodos de clasificación estudiados, estos serán incorporados a una tabla 2x2 para cada uno de los métodos, Tabla 8.

Tabla 8 Matriz de confusión para el cálculo de los resultados finales

		Clasificador	
		AVF	AVNF
Referencia	AVF	a	c
	AVNF	b	d

n= algoritmo de aprendizaje automático. a = Verdaderos positivos. b = Falsos positivos. c = Falsos negativos. d = Verdaderos negativos. AVF = Anatomía vascular cervical favorable. AVNF = Anatomía vascular cervical no favorable.

Posteriormente se calcularon siete métricas de desempeño con las siguientes formulas:

$$Ex = \frac{a + d}{a + b + c + d}$$

Ecuación 9 Exactitud

$$Se = \frac{a}{a + c}$$

Ecuación 10 Sensibilidad/*Recall*

$$Sp = \frac{d}{b + d}$$

Ecuación 11 Especificidad

$$VPP = \frac{a}{a + b}$$

Ecuación 12 Valor predictivo positivo/Precisión

$$VPN = \frac{d}{c + d}$$

Ecuación 13 Valor predictivo negativo

$$F1 = \frac{2(VPP \cdot Se)}{VPP + Se}$$

Ecuación 14 F1

$$Ex_b = \frac{Se + Sp}{2}$$

Ecuación 15 Exactitud Balanceada

Además de las métricas anteriores, se calcularon dos métricas adicionales:

- **Área bajo la curva ROC.** La curva ROC (*Receiver Operating Characteristic*) es una representación gráfica del rendimiento de un modelo binario a diferentes umbrales de clasificación. La curva se construye calculando la tasa de verdaderos positivos (Ecuación 10 Sensibilidad/*Recall*) y la tasa de falsos positivos a diferentes umbrales de probabilidad. Para calcular los puntos de una curva ROC, se siguen los siguientes pasos:
  1. Ordenar las predicciones del modelo por la probabilidad predicha de la clase positiva, de mayor a menor
  2. Calcular las tasas (verdaderos y falsos positivos) para cada umbral de decisión. La fórmula para el cálculo de la tasa de falsos positivos se muestra en la Ecuación 16.

$$Tasa\ de\ Falsos\ Positivos = \frac{c}{c+d} = 1 - Sp$$

Ecuación 16 Tasa de Falsos Positivos

3. Graficar la curva ROC. El eje x representa la tasa de falsos positivos. El eje y representa la tasa de verdaderos positivos. Cada punto de la curva ROC corresponde a un umbral diferente de clasificación.
4. Calcular el área bajo la curva dividiendo la curva ROC en pequeños segmentos trapezoidales y sumando sus áreas.

- **Tiempo de computación.** Se refiere al tiempo total que tarda un modelo en completar tanto el proceso de entrenamiento (ajustar el modelo a los datos de entrenamiento) como el proceso de evaluación o prueba (aplicar el modelo entrenado a un conjunto de prueba para generar predicciones). Este tiempo incluye todas las operaciones computacionales involucradas, como el ajuste de parámetros, la optimización interna, y la generación de predicciones, y es un factor crítico para evaluar la eficiencia y escalabilidad de un algoritmo de aprendizaje automático.

4.6.4.1.2 Aclaración acerca de la codificación y cálculo de métricas

Aunque AVF fue codificada como 0 y AVNF como 1 para su procesamiento digital, en la evaluación del desempeño del clasificador se consideró a AVF como la clase positiva, dado su mayor interés clínico. Por ello, en la matriz de confusión (Tabla 8), los verdaderos positivos (a) corresponden a casos de anatomía favorable correctamente identificados.

4.6.4.1.3 Análisis Estadístico

Se compararon las métricas de rendimiento obtenidas de cada modelo de aprendizaje automático. Antes de realizar el análisis comparativo, se utilizó la prueba de Shapiro-Wilk para evaluar la normalidad de los valores obtenidos en cada iteración del proceso de validación cruzada. Luego, se utilizó un ANOVA de una vía para evaluar las diferencias entre los clasificadores. Después de comparar los resultados de los algoritmos, se seleccionó el modelo más adecuado con las mejores métricas de rendimiento. Finalmente se compararon las métricas de desempeño de cada clasificador para obtener los resultados finales.

5. Resultados

5.1 Características del conjunto de datos

El conjunto de datos suministrado a los clasificadores consto de un *data frame* con 108 instancias, cada instancia con 28 atributos, más una variable adicional con la clase. El conjunto de datos contiene dos clases posibles (anatomía vascular cervical favorable y no favorable).

5.2 Balance de clases en el conjunto de datos

El 75% de las instancias (n=81) pertenecieron a la clase de anatomía vascular cervical no favorable, dando como resultado un conjunto de datos desbalanceado en proporción 1:3.

5.3 Atributos

Se revisaron las historias clínicas electrónicas coincidentes de cada paciente utilizando la lista descrita anteriormente como referencia. Los datos clínicos de interés se recogieron a partir de la entrada de datos más cercana a la fecha de la angiografía por tomografía computada. En la Tabla 9 se muestran los resultados de la caracterización epidemiológica de los pacientes cuyos datos constituyeron los atributos alimentados a los algoritmos de aprendizaje automático.

Tabla 9 Atributos y valores incluidos en el conjunto de datos.

Atributos	Todas las instancias	AVF	AVNF	p
Sexo	57 (56.8)	12 (44.4)	45 (55.6)	0.436*
Edad (años)	68.1 (13.8)	66.6 (12.7)	68.6 (14.1)	0.479**
Peso (kg)	45 (41.7)	13 (48.1)	32 (39.5)	0.573*
Altura (m)	76 (70.4)	15 (55.6)	61 (75.3)	0.089*
Tabaquismo	44 (40.7)	9 (33.3)	35 (43.2)	0.498*
Diabetes	14 (13.0)	4 (14.8)	10 (12.3)	0.746*
Hipertensión arterial	9 (8.33)	3 (11.1)	6 (7.41)	0.688*
Hipercolesterolemia	9 (8.33)	3 (11.1)	6 (7.41)	0.688*
Hipertrigliceridemia	1.61 (0.12)	1.61 (0.13)	1.61 (0.12)	0.758**
Obesidad	69.2 (19.5)	71.1 (19.0)	68.5 (19.8)	0.538**
Años con tabaquismo	0.38 (1.85)	0.96 (3.22)	0.19 (1.04)	0.227**

Años con diabetes	7.64 (6.17)	5.78 (6.30)	8.26 (6.04)	0.080**
Años con hipertensión arterial	4.61 (6.11)	4.04 (5.58)	4.80 (6.30)	0.553**
Años con hipercolesterolemia	0.95 (3.13)	0.41 (1.53)	1.14 (3.50)	0.138**
Años con hipertrigliceridemia	0.73 (2.52)	0.81 (2.42)	0.70 (2.57)	0.840**
Glucosa en sangre aleatoria (mg/dL)	128 (59.8)	116 (57.9)	131 (60.3)	0.252**
Creatinina sérica (mg/dL)	0.89 (0.17)	0.81 (0.13)	0.91 (0.18)	0.002**
Urea en sangre (mg/dL)	37.2 (9.01)	35.9 (10.6)	37.7 (8.44)	0.413**
Nitrógeno ureico en sangre (mg/dL)	18.6 (4.88)	17.8 (4.61)	18.8 (4.97)	0.369**
Ácido úrico (mg/dL)	6.10 (1.86)	5.91 (1.75)	6.16 (1.89)	0.523**
Leucocitos (×10^9/L)	8.08 (1.95)	8.47 (2.03)	7.95 (1.92)	0.248**
Eritrocitos (×10^9/L)	5.32 (0.86)	5.31 (0.93)	5.32 (0.84)	0.976**
Plaquetas (×10^9/L)	234 (51.1)	238 (51.4)	232 (51.3)	0.626**
Hemoglobina (g/dL)	16.0 (1.44)	15.7 (1.35)	16.0 (1.47)	0.293**
INR	1.03 (0.10)	1.03 (0.10)	1.03 (0.10)	0.956**
Tiempo de protrombina	13.3 (1.42)	13.4 (1.37)	13.2 (1.44)	0.505**
Tiempo parcial de tromboplastina	28.5 (4.24)	28.9 (4.91)	28.4 (4.02)	0.631**

Todos los valores n (%) a menos que se especifique. * Prueba χ^2. ** Prueba t para muestras independientes.

5.4 Identificación de predictores de la anatomía vascular favorable

Para determinar los factores predictores que permiten clasificar a pacientes con hipertensión, diabetes o dislipidemia de acuerdo con su anatomía vascular cervical, se llevó a cabo un análisis de regresión logística, por medio del cual se determinó el subconjunto óptimo de predictores para la anatomía vascular cervical desfavorable. En primer lugar, se seleccionaron los posibles factores de confusión utilizando el criterio de causa disyuntiva de todas las covariables medidas. [177] A continuación, se utilizó la regresión logística escalonada con eliminación hacia atrás para eliminar las variables de una en una en función del criterio de información de Akaike (AIC). En cada iteración, si el nuevo modelo tenía más de 2 unidades AIC más bajas que el anterior, se consideraba significativamente mejor y se mantenía. De lo contrario, se eliminó.

Después de seleccionar los factores de confusión, el modelo de regresión logística inicial incluyó las siguientes variables: sexo, edad, IMC, tabaquismo, creatinina sérica, diabetes mellitus, hipertensión e hipercolesterolemia. El modelo de regresión logística final identificó tres variables asociadas a una anatomía vascular cervical favorable con un AIC de 110.9.

Los predictores identificados fueron la duración de la hipertensión arterial y de la hipertrigliceridemia y los niveles de creatinina sérica. En la Tabla 10 se muestra el modelo final con las razones de momios con intervalos de confianza del 95% para cada predictor.

Tabla 10 Modelo final, predictores asociados a anatomía vascular cervical favorable

Predictor	Razón de Momios	IC del 95%	p
Años con hipertensión arterial	-0.31	-0.62 – -0.07	0.021
Años con hipercolesterolemia	0.11	0.03 – 0.20	0.012
Creatinina Sérica (mg/dL)	5.0	1.9 – 8.4	0.002
IC = Intervalo de Confianza			

5.5 Desempeño de los Clasificadores

La Tabla 11 muestra la media y la desviación estándar (entre paréntesis) de cada métrica de desempeño obtenida por los ocho clasificadores después de pruebas de validación cruzada de 5 iteraciones. La última columna muestra el resultado de una prueba de hipótesis (ANOVA de un factor) comparando el valor medio obtenido por cada clasificador.

Tabla 11 Métricas de desempeño obtenidas después de la validación cruzada de 5 iteraciones.

	RL	CIB	k-NN	ArD	AnD	NNet	SVM	CG	p
Exactitud	0.58 (0.08)	0.69 (0.02)	0.69 (0.08)	0.58 (0.11)	0.73 (0.03)	0.58 (0.02)	0.65 (0.07)	0.75 (0.02)	<0.001
Sensibilidad/ *Recall*	0.34 (0.19)	0.13 (0.14)	0.20 (0.11)	0.22 (0.14)	0.11 (0.12)	0.33 (0.11)	0.41 (0.07)	0.04 (0.06)	<0.001
Especificidad	0.66 (0.17)	0.88 (0.05)	0.85 (0.12)	0.70 (0.18)	0.93 (0.07)	0.67 (0.02)	0.73 (0.11)	0.98 (0.01)	<0.001
Valor Predictivo Positivo/ Precisión	0.24 (0.05)	0.21 (0.15)	0.36 (0.12)	0.16 (0.09)	0.24 (0.23)	0.24 (0.07)	0.35 (0.07)	0.22 (0.33)	0.513

Valor Predictivo Negativo	0.75 (0.02)	0.75 (0.02)	0.76 (0.03)	0.73 (0.02)	0.76 (0.01)	0.75 (0.03)	0.79 (0.02)	0.75 (0.02)	0.031
F1	0.27 (0.09)	0.15 (0.15)	0.23 (0.10)	0.18 (0.11)	0.14 (0.14)	0.28 (0.09)	0.37 (0.04)	0.06 (0.10)	0.003
Exactitud Balanceada	0.50 (0.03)	0.51 (0.05)	0.53 (0.05)	0.46 (0.03)	0.52 (0.02)	0.50 (0.05)	0.57 (0.04)	0.51 (0.03)	0.021
ABCROC	0.49 (0.05)	0.49 (0.09)	0.53 (0.06)	0.46 (0.03)	0.55 (0.09)	0.52 (0.09)	0.54 (0.10)	0.50 (0.02)	0.542

RL = Regresión Logística. CIB = Clasificador Ingenuo de Bayes. K-NN = k-Vecinos más cercanos. ArD = Arboles de Decisión. AnD = Análisis Discriminante. NNet = Red Neuronal. SVM = Maquinas de Vectores de Soporte. CG = Copula Gaussiana. ABCROC = Área Bajo la Curva ROC

La Figura 17 muestra la exactitud media de cada uno de los ocho clasificadores en una línea continua. La cadena punteada de triángulos representa la exactitud lograda en cada pliegue del proceso de validación cruzada.

La Figura 18 muestra la exactitud balanceada de cada uno de los ocho clasificadores en una línea continua. La cadena punteada de triángulos representa la exactitud balanceada lograda en cada pliegue del proceso de validación cruzada.

La Figura 19 muestra las curvas ROC promedio posterior a validación cruzada de 5 iteraciones para cada clasificador entrenado con líneas punteadas de colores; la leyenda muestra los valores del área bajo la curva ROC.

La Figura 20 muestra la puntuación media F1 de cada uno de los ocho clasificadores en una línea continua. La cadena punteada de triángulos representa la puntuación F1 alcanzada en cada pliegue del proceso de validación cruzada.

En la Figura 21 se muestra el desempeño ordenado de los clasificadores en todas las métricas de desempeño.

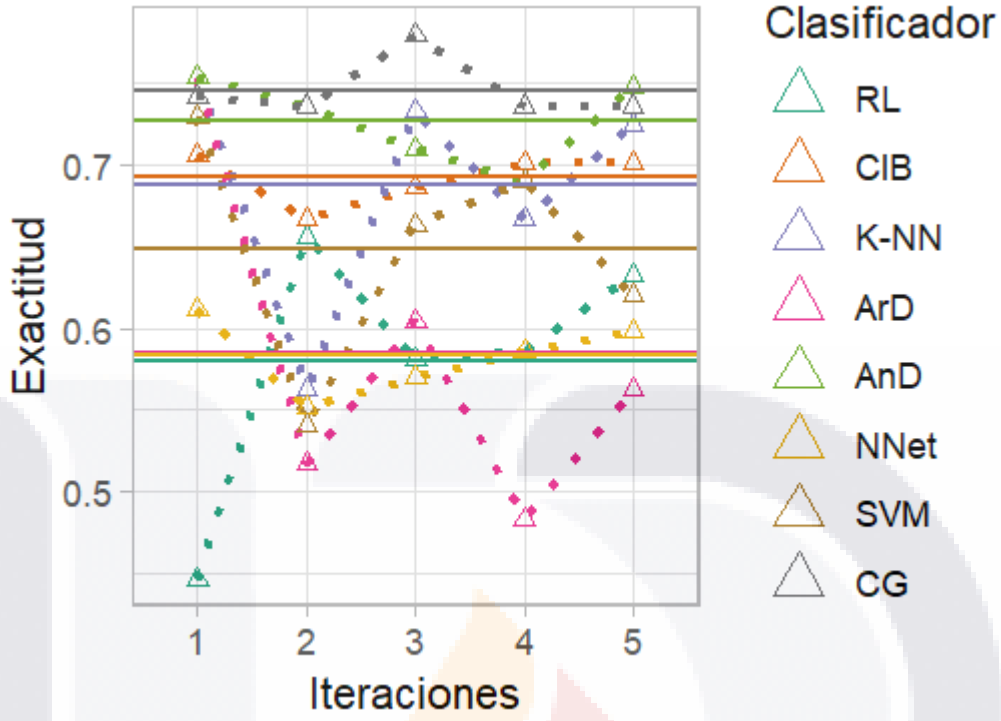


Figura 17 Exactitud obtenida por los clasificadores estudiados.

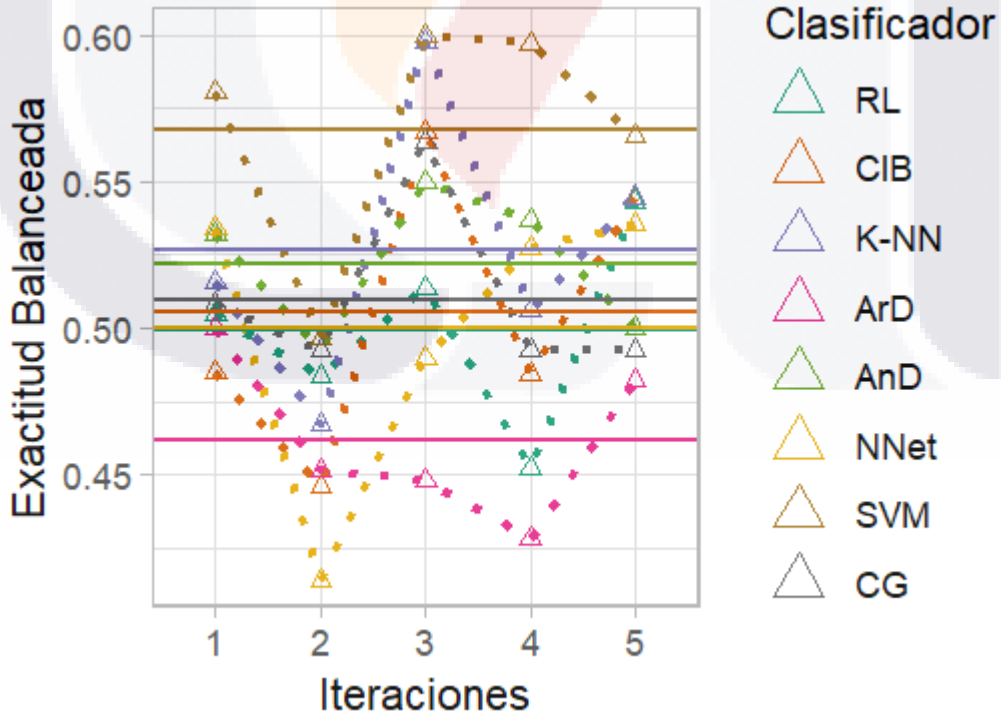


Figura 18 Exactitud balanceada obtenida por los clasificadores estudiados.

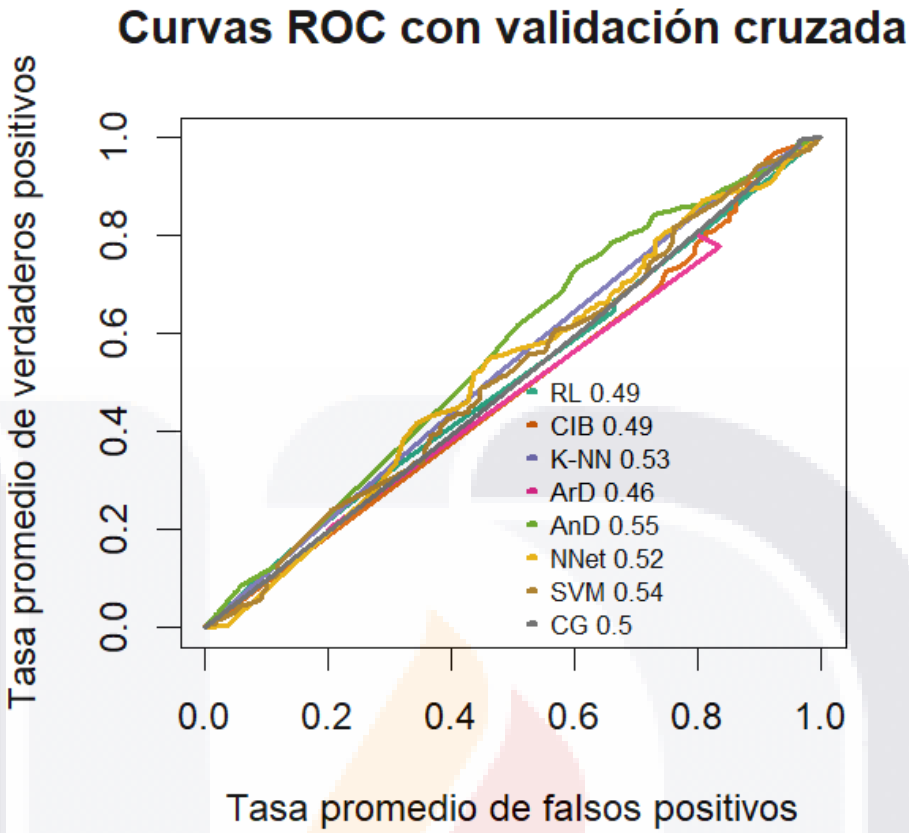


Figura 19 Área bajo la curva ROC obtenida por los clasificadores.

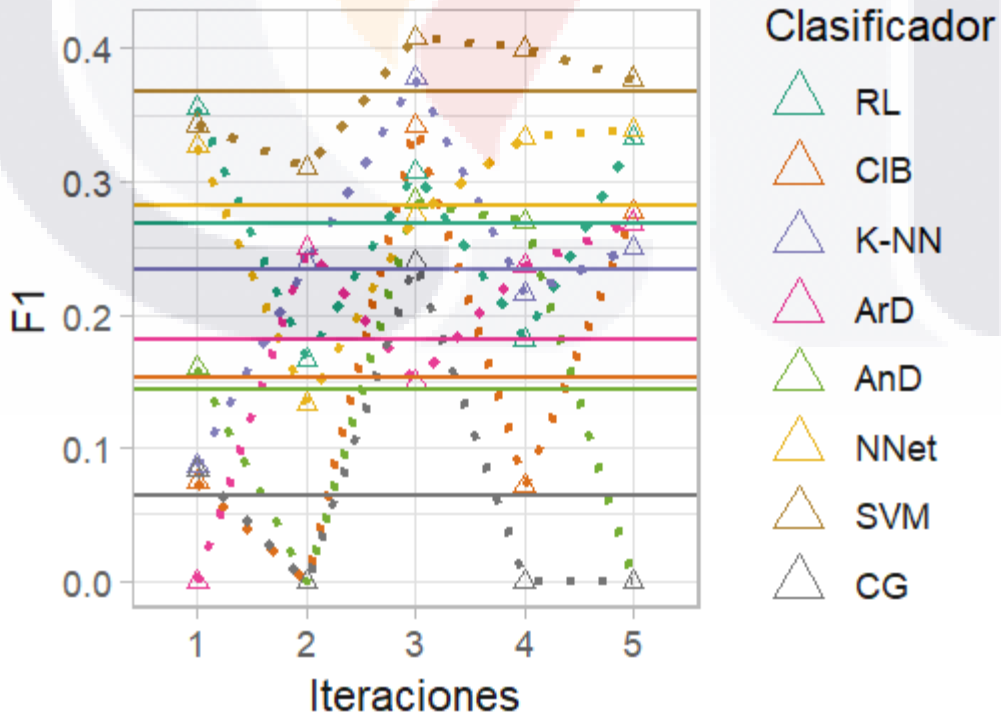


Figura 20 Índice F1 obtenido por los clasificadores.

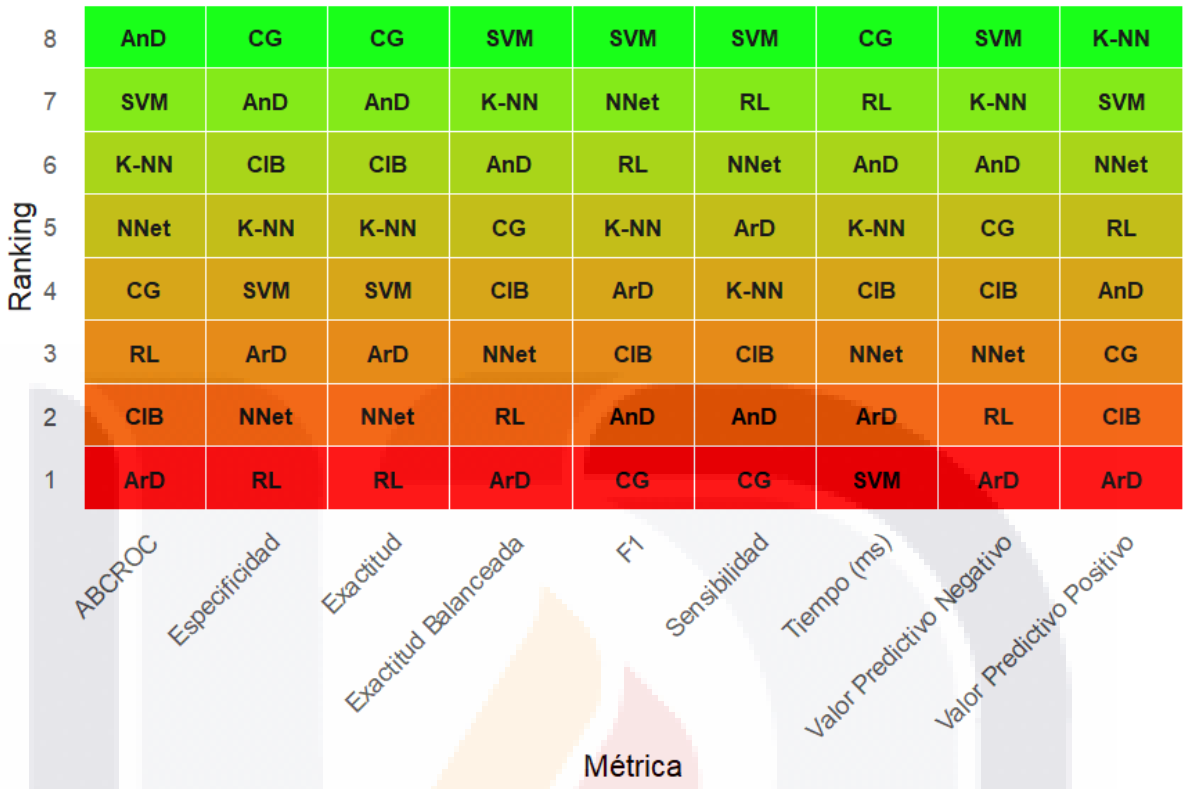


Figura 21 Orden del desempeño de los clasificadores.

La Tabla 12 muestra la comparación por prueba de hipótesis del tiempo en segundos que tomo a cada clasificador el procesamiento de 10 subconjuntos de datos.

Tabla 12 Comparación del tiempo de computación por clasificador.

	RL	CIB	k-NN	ArD	AnD	NNet	SVM	CG	p
Tiempo (ms)	1.38 (0.16)	0.05 (0.01)	0.09 (0.01)	0.02 (0.01)	0.69 (0.87)	0.04 (0.02)	0.01 (0.00)	7.74 (0.18)	<0.001

RL = Regresión Logística. CIB = Clasificador Ingenuo de Bayes. K-NN = k-Vecinos más cercanos. ArD = Arboles de Decisión. AnD = Análisis Discriminante. NNet = Red Neuronal. SVM = Maquinas de Vectores de Soporte. CG = Copula Gaussiana.

5.6 Desempeño individual

En el análisis del desempeño medio de los clasificadores, se observó que el clasificador CG ocupó el primer lugar en las métricas de Especificidad y Tiempo de procesamiento, mientras que SVM se ubicó en la primera posición para Sensibilidad, F1 y Exactitud Balanceada. El

clasificador K-NN encabezó el ranking de Valor Predictivo Positivo, mientras que SVM y CG estuvieron entre los tres primeros puestos en múltiples métricas.

El algoritmo AnD alcanzó la primera posición en la métrica ABCROC y se posicionó dentro de los tres primeros lugares en Especificidad, Valor Predictivo Negativo y Exactitud. En contraste, RL y ArD se ubicaron predominantemente en las posiciones intermedias a bajas en varias métricas, con RL en los últimos lugares en Especificidad, Valor Predictivo Positivo y ABCROC.

NNet mostró una ubicación consistente en rangos medios, destacando con el segundo lugar en Sensibilidad y el cuarto lugar en ABCROC. El clasificador CIB mostró un mejor posicionamiento en las métricas de Especificidad y Valor Predictivo Positivo, pero se ubicó en posiciones inferiores en ABCROC y F1.

En cuanto a los algoritmos que con mayor frecuencia ocuparon las posiciones más altas, se identificaron SVM, CG, y AnD como los que aparecen más veces dentro de los tres primeros lugares a lo largo de las distintas métricas evaluadas. Por otro lado, ArD y RL fueron los clasificadores con menor presencia en las posiciones superiores.

5.7 Desempeño general

Los resultados generales obtenidos en las métricas evaluadas muestran un desempeño limitado por parte de todos los clasificadores. La mayoría de los valores se ubicaron en torno al 50%, lo cual representa un rendimiento apenas superior al azar en un problema binario. Este patrón fue consistente en múltiples indicadores, incluyendo Exactitud, Sensibilidad, Valor Predictivo Positivo, F1 y Exactitud Balanceada.

Una posible causa de estos resultados es el marcado desbalance en el conjunto de datos: el 75% de las instancias correspondieron a pacientes con anatomía vascular cervical no favorable, lo que genera una proporción de 1:3 entre la clase positiva (anatomía favorable) y la clase negativa. Este tipo de distribución tiende a sesgar los algoritmos hacia la clase mayoritaria, reduciendo su capacidad de identificar correctamente los casos menos frecuentes. Esta limitación es especialmente evidente en métricas sensibles al desempeño

sobre la clase positiva, como la Sensibilidad y el Valor Predictivo Positivo, cuyos valores se mantuvieron consistentemente bajos.

Adicionalmente, el conjunto de datos incluyó una amplia variedad de atributos clínicos y de laboratorio, cuya distribución fue en general homogénea entre los grupos, sin diferencias estadísticamente significativas en la mayoría de las variables. Esta falta de discriminación entre clases en los predictores disponibles puede haber reducido la capacidad de los algoritmos para aprender patrones distintivos. Esto es más evidente al tomar en cuenta análisis de predictores por medio de RL, en donde el modelo final conservó únicamente tres predictores con asociación significativa, lo que sugiere una débil relación entre las variables y el desenlace.

5.8 Abordaje del desbalance de clases

Dado el desequilibrio observado en la proporción de clases, la aplicación de técnicas para abordar el desbalance es fundamental antes de continuar con cualquier intento de optimización o comparación entre clasificadores. Existen tres enfoques principales para tratar este tipo de problema: submuestreo de la clase mayoritaria (*undersampling*), sobremuestreo de la clase minoritaria (*oversampling*) y el uso de algoritmos generativos de datos sintéticos, como la Técnica de Sobremuestreo de Minorias Sintéticas (SMOTE: *Synthetic Minority Over-sampling Technique*). [178]

El submuestreo reduce el número de instancias de la clase mayoritaria para balancear el conjunto de datos, pero esta estrategia implica la pérdida de información potencialmente útil, lo cual puede ser crítico en estudios clínicos con muestras limitadas.[179] Por su parte, el sobremuestreo tradicional consiste en duplicar instancias de la clase minoritaria, pero tiene el riesgo de sobreajuste, ya que no introduce nueva información, sino que repite observaciones.[180]

En contraste, SMOTE genera nuevas instancias sintéticas de la clase minoritaria mediante interpolación entre ejemplos existentes, lo que permite enriquecer el espacio de representación sin perder información ni inflar el conjunto de entrenamiento con copias exactas.[181]

Esta técnica resulta especialmente adecuada para el presente conjunto de datos, ya que mantiene todas las observaciones clínicas disponibles y, al mismo tiempo, proporciona una representación más equilibrada para el aprendizaje de los modelos. Dada la proporción 1:3 entre clases y la relativa homogeneidad de los atributos originales, el uso de SMOTE representa una estrategia apropiada para aumentar la sensibilidad del modelo y mejorar la discriminación entre clases sin comprometer la integridad de los datos observacionales. [182-184]

5.9 Desempeño posterior al abordaje del desbalance de clases

Posterior al a aplicación de SMOTE al conjunto de datos original, la Tabla 13 muestra la media y la desviación estándar (entre paréntesis) de cada métrica de desempeño obtenida por los ocho clasificadores después de pruebas de validación cruzada de 5 iteraciones. La última columna muestra el resultado de una prueba de hipótesis (ANOVA de un factor) comparando el valor medio obtenido por cada clasificador.

Tabla 13 Métricas de desempeño obtenidas por los clasificadores después de validación cruzada de 5 iteraciones y corrección del desbalance.

	RL	CIB	k-NN	ArD	AnD	NNet	SVM	CG	p
Exactitud	0.57 (0.06)	0.69 (0.03)	0.55 (0.03)	0.53 (0.04)	0.58 (0.05)	0.60 (0.03)	0.58 (0.03)	0.73 (0.04)	<0.001
Sensibilidad/ *Recall*	0.58 (0.08)	0.69 (0.10)	0.62 (0.08)	0.57 (0.14)	0.69 (0.10)	0.71 (0.12)	0.72 (0.15)	0.55 (0.09)	0.080
Especificidad	0.56 (0.08)	0.69 (0.14)	0.47 (0.08)	0.48 (0.12)	0.46 (0.12)	0.50 (0.12)	0.45 (0.10)	0.90 (0.06)	<0.001
Valor Predictivo Positivo/ Precisión	0.57 (0.06)	0.71 (0.08)	0.54 (0.03)	0.52 (0.04)	0.57 (0.04)	0.59 (0.02)	0.56 (0.01)	0.86 (0.07)	<0.001
Valor Predictivo Negativo	0.57 (0.07)	0.70 (0.03)	0.56 (0.04)	0.53 (0.05)	0.60 (0.05)	0.64 (0.05)	0.64 (0.10)	0.67 (0.04)	0.001
F1	0.57 (0.06)	0.69 (0.03)	0.58 (0.04)	0.54 (0.08)	0.62 (0.05)	0.64 (0.04)	0.63 (0.06)	0.66 (0.06)	0.003
Exactitud Balanceada	0.57 (0.06)	0.69 (0.03)	0.55 (0.03)	0.53 (0.04)	0.58 (0.04)	0.60 (0.03)	0.58 (0.03)	0.73 (0.04)	<0.001

ABCROC	0.57 (0.06)	0.77 (0.04)	0.56 (0.03)	0.54 (0.06)	0.58 (0.03)	0.65 (0.02)	0.49 (0.13)	0.80 (0.04)	<0.001

RL = Regresión Logística. CIB = Clasificador Ingenuo de Bayes. K-NN = k-Vecinos más cercanos. ArD = Arboles de Decisión. AnD = Análisis Discriminante. NNet = Red Neuronal. SVM = Maquinas de Vectores de Soporte. CG = Copula Gaussiana. ABCROC = Área Bajo la Curva ROC

Tras la aplicación de SMOTE, se observaron mejoras generalizadas, especialmente en sensibilidad y F1. Los clasificadores con mayores incrementos fueron CIB y CG. En CIB, la sensibilidad aumentó de 0.13 a 0.69, la F1 de 0.15 a 0.69 y el AUROC de 0.49 a 0.77. CG mostró un patrón similar, con aumentos en sensibilidad (0.04 a 0.55), precisión (0.22 a 0.86), F1 (0.06 a 0.66) y AUROC (0.50 a 0.80).

NNet y SVM también registraron mejoras importantes, con sensibilidades superiores a 0.70 y F1 mayores a 0.60. En ambos casos, las ganancias se concentraron en métricas asociadas al desempeño sobre la clase minoritaria.

Por el contrario, RL y k-NN mostraron cambios más limitados. En RL, aunque la sensibilidad y F1 aumentaron, la exactitud y el AUROC apenas se modificaron. En k-NN, la F1 mejoró, pero otras métricas como especificidad y AUROC mostraron descensos o permanecieron estables. ArD y AnD también reflejaron aumentos en sensibilidad y F1, pero con reducciones marcadas en especificidad y valor predictivo negativo.

En conjunto, los efectos de SMOTE variaron entre algoritmos, beneficiando sobre todo a aquellos con bajo desempeño inicial en sensibilidad, sin que hubiera una mejoría uniforme en todas las métricas.

El AUROC mostró una tendencia variable entre clasificadores tras la aplicación de SMOTE. CG y CIB fueron los únicos algoritmos que alcanzaron valores altos de AUROC, con 0.80 y 0.77, respectivamente, lo que representa un aumento considerable en comparación con sus valores previos (0.50 en ambos casos). NNet también mejoró su AUROC, pasando de 0.52 a 0.65. En contraste, SVM mostró una disminución importante, de 0.54 a 0.49, mientras que k-NN, ArD, AnD y RL presentaron solo aumentos modestos, manteniéndose en el rango de 0.54 a 0.58. Algunos clasificadores, como SVM, mejoraron considerablemente en

sensibilidad, pero a costa de una menor especificidad, lo que impactó negativamente su área bajo la curva.

La Figura 22 muestra las curvas ROC de todos los clasificadores después del ajuste por desbalanceo.

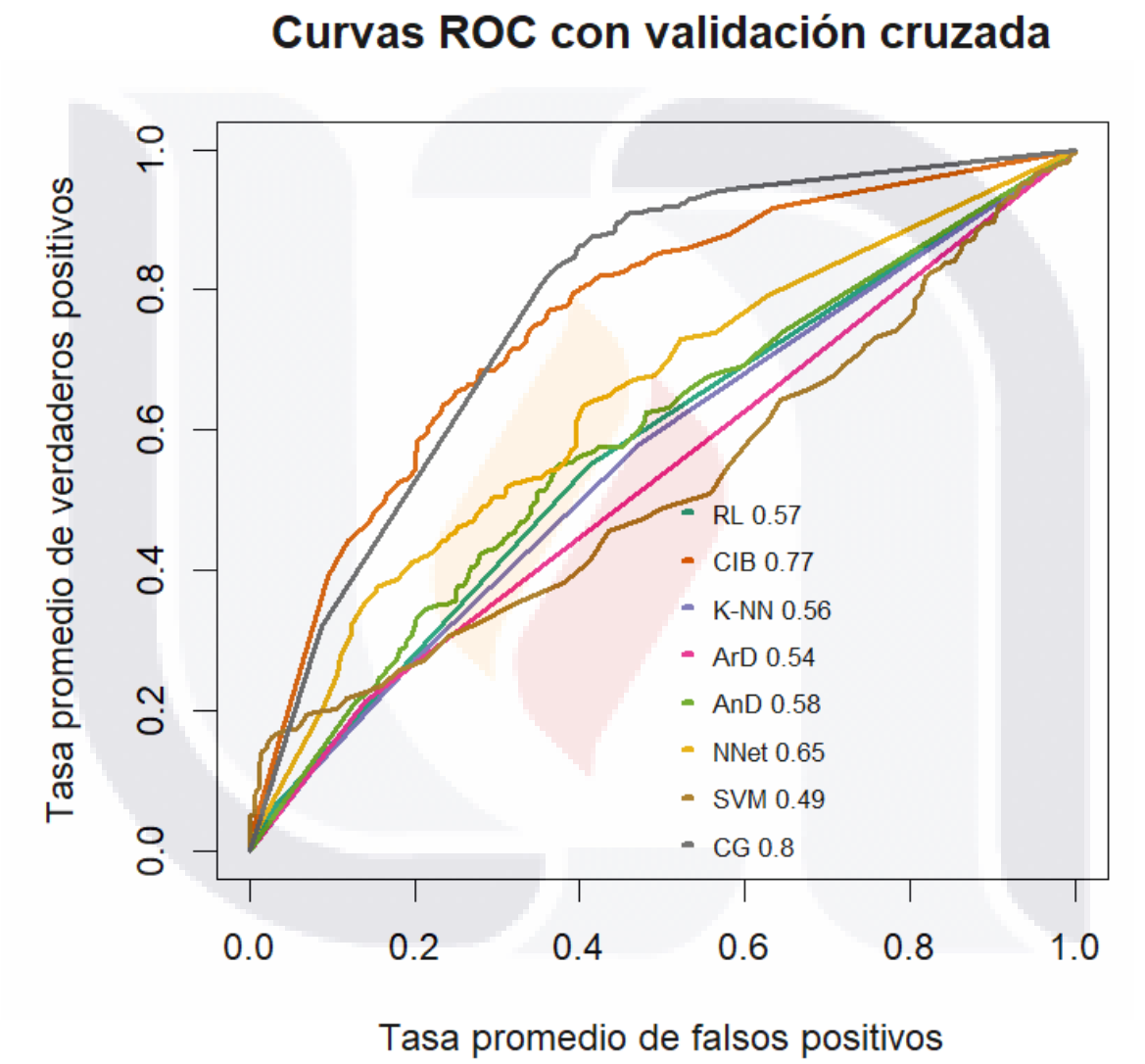


Figura 22 Área bajo la curva ROC posterior al ajuste del desbalance.

El análisis del ranking por métrica permite identificar patrones complementarios al análisis basado en valores absolutos. RL ocupa la primera posición en exactitud y aparece consistentemente entre los tres primeros lugares en varias métricas clave, como valor

predictivo negativo y exactitud balanceada. Sin embargo, no figura en los primeros puestos de sensibilidad ni de F1. CG se posiciona como el mejor clasificador en sensibilidad y también obtiene el primer lugar en valor predictivo positivo.

ArD destaca en especificidad, valor predictivo negativo y AUROC, a pesar de ocupar posiciones más bajas en sensibilidad y F1. CIB muestra un perfil balanceado, posicionándose entre los primeros tres lugares en sensibilidad, precisión, F1 y AUROC. NNet mantiene un buen desempeño en exactitud, F1 y balance de métricas, mientras que AnD obtiene su mejor lugar en sensibilidad y F1, aunque ocupa el último puesto en AUROC. SVM presenta un desempeño intermedio, sin liderar ninguna métrica, pero manteniéndose en posiciones medias en la mayoría.

Por último, k-NN se ubica predominantemente en posiciones bajas, excepto en sensibilidad. La Figura 23 ilustra este ordenamiento visualmente, facilitando la comparación del rendimiento relativo de los clasificadores en cada métrica.

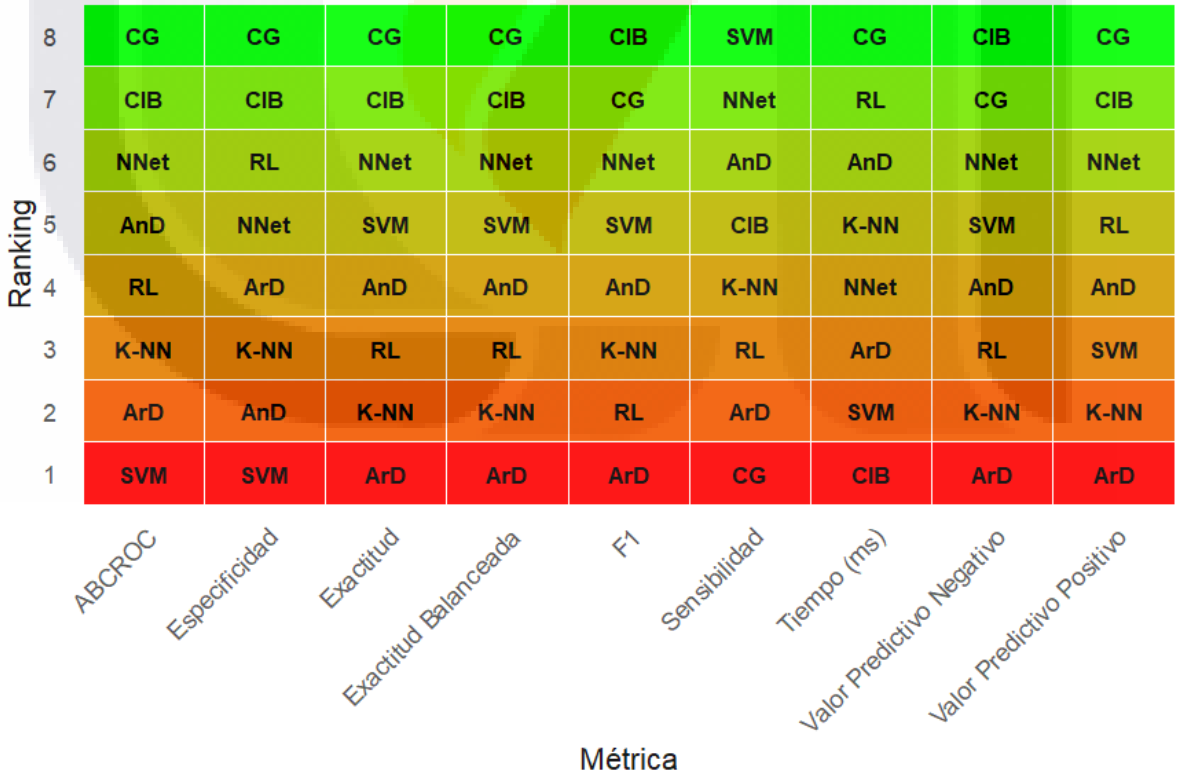


Figura 23 Orden del desempeño posterior al ajuste del desbalance.

5.10 Desempeño individual

5.10.1 Regresión Logística

En la nueva evaluación, la RL presentó una ligera disminución en la exactitud, pasando de 0.58 a 0.57. La sensibilidad mostró un cambio más notorio, con un aumento desde 0.34 hasta 0.58. Por otro lado, la especificidad disminuyó de 0.66 a 0.56.

El valor predictivo positivo o precisión aumentó notablemente, de 0.24 a 0.57, mientras que el valor predictivo negativo mostró un descenso, pasando de 0.75 a 0.57. En cuanto a la métrica F1, se observó una mejora significativa, con un incremento de 0.27 a 0.57.

La exactitud balanceada pasó de 0.50 a 0.57, evidenciando un aumento moderado. Finalmente, el área bajo la curva ROC (ABCROC) también mostró una mejora, al pasar de 0.49 a 0.57.

Este conjunto de cambios en el desempeño de la RL antes y después del ajuste por desbalanceo se ilustra en la Figura 24, que muestra de forma comparativa cada métrica evaluada para este algoritmo.

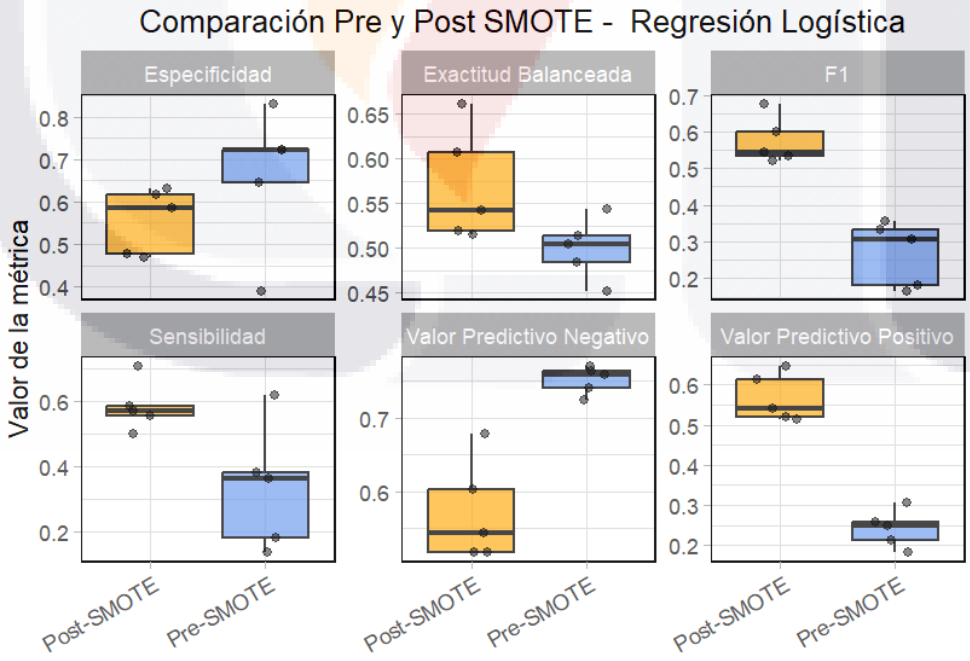


Figura 24 Desempeño Final de la Regresión Logística

5.10.2 Clasificador Ingenuo de Bayes

después del balanceo de clases, el CIB conservó su nivel de exactitud en 0.69. Mostró un incremento sustancial en sensibilidad, que pasó de 0.13 a 0.69, mientras que su especificidad descendió de 0.88 a 0.69. La precisión también mejoró, subiendo de 0.21 a 0.71. El valor predictivo negativo tuvo una ligera reducción de 0.75 a 0.70. La puntuación F1 reflejó un avance notable, al pasar de 0.15 a 0.69. La exactitud balanceada aumentó de 0.51 a 0.69, y el área bajo la curva ROC se elevó de 0.49 a 0.77.

La Figura 25 muestra visualmente esta transición en el desempeño del CIB a través de cada una de las métricas utilizadas.

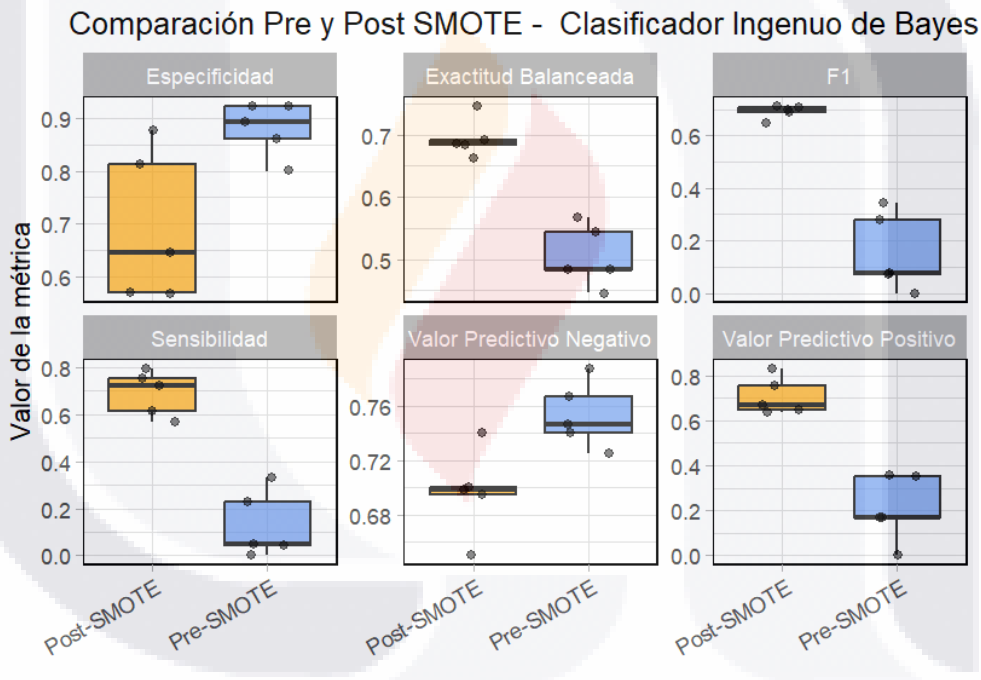


Figura 25 Desempeño Final del Clasificador Ingenuo de Bayes

5.10.3 K-vecinos más cercanos

Para k-NN, la exactitud disminuyó ligeramente de 0.69 a 0.55. La sensibilidad, en cambio, mostró un aumento considerable, de 0.20 a 0.62, mientras que la especificidad descendió de 0.85 a 0.47. La precisión avanzó de 0.36 a 0.54, y el valor predictivo negativo bajó de 0.76 a 0.56. La métrica F1 mejoró, pasando de 0.23 a 0.58. También se observó una reducción

en la exactitud balanceada, de 0.53 a 0.55. El área bajo la curva ROC se mantuvo relativamente estable, con un valor final de 0.56 frente al anterior de 0.53.

Estos resultados están en la Figura 26, donde se muestra el desempeño de k-NN en cada una de las métricas tras el ajuste por desbalanceo.

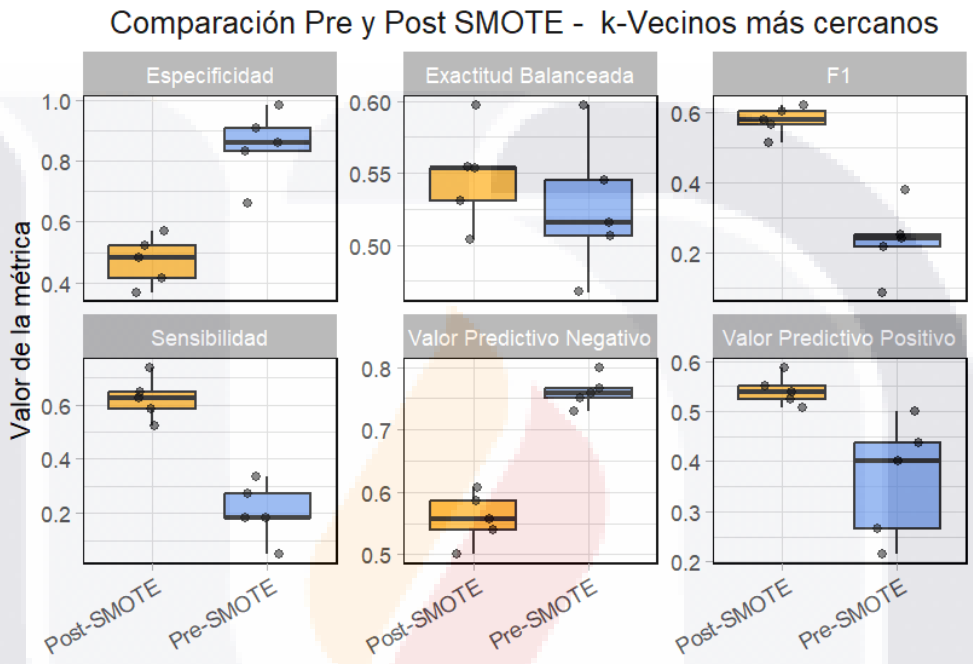


Figura 26 Desempeño Final de k vecinos más próximos

5.10.4 Árboles de decisión

En el caso de los Árboles de Decisión, la exactitud disminuyó de 0.58 a 0.53. La sensibilidad pasó de 0.22 a 0.57, mientras que la especificidad se redujo de 0.70 a 0.48. La precisión aumentó levemente, de 0.16 a 0.52. El valor predictivo negativo también mostró una ligera baja, de 0.73 a 0.53. La puntuación F1 subió de 0.18 a 0.54. La exactitud balanceada avanzó de 0.46 a 0.53. Por último, el área bajo la curva ROC se incrementó de 0.46 a 0.54.

El cambio en el desempeño de este clasificador se detalla en la Figura 27, que muestra los valores obtenidos en cada métrica después del balanceo de clases.

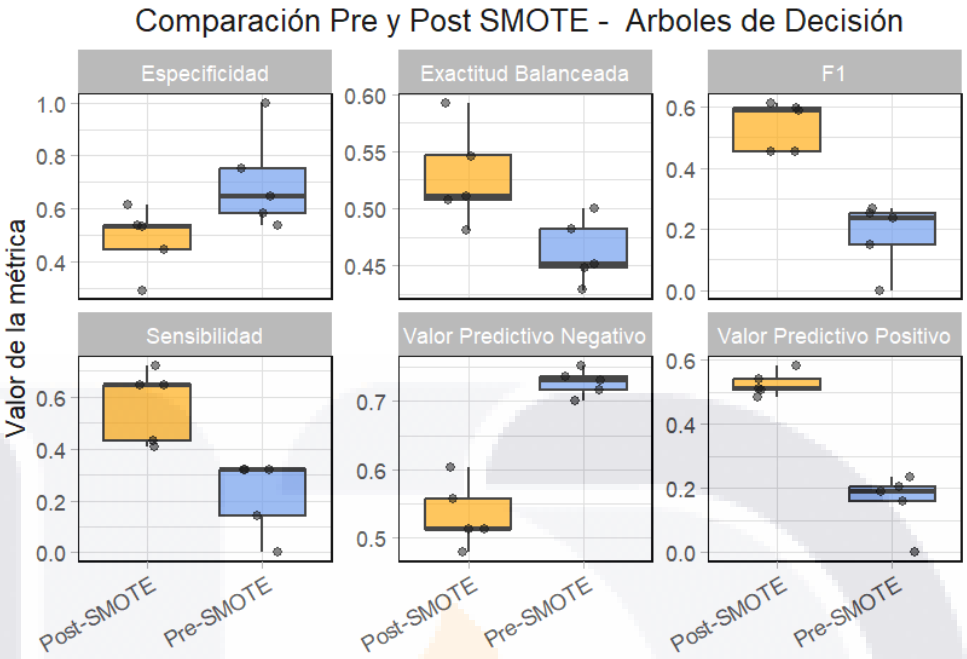


Figura 27 Desempeño Final de los Árboles de Decisión

5.10.5 Análisis Discriminante

El Análisis Discriminante mostró un leve descenso en la exactitud, que pasó de 0.73 a 0.58. La sensibilidad aumentó de 0.11 a 0.69, mientras que la especificidad se redujo de 0.93 a 0.46. La precisión se mantuvo estable en torno a 0.57, en comparación con el valor anterior de 0.24.

El valor predictivo negativo disminuyó ligeramente, de 0.76 a 0.60. En cuanto a la métrica F1, se registró una mejora considerable, subiendo de 0.14 a 0.62. La exactitud balanceada mostró un ascenso de 0.52 a 0.58. Finalmente, el área bajo la curva ROC pasó de 0.55 a 0.58.

Todos estos cambios tras la aplicación del sobreajuste sintético se muestran en la Figura 28.

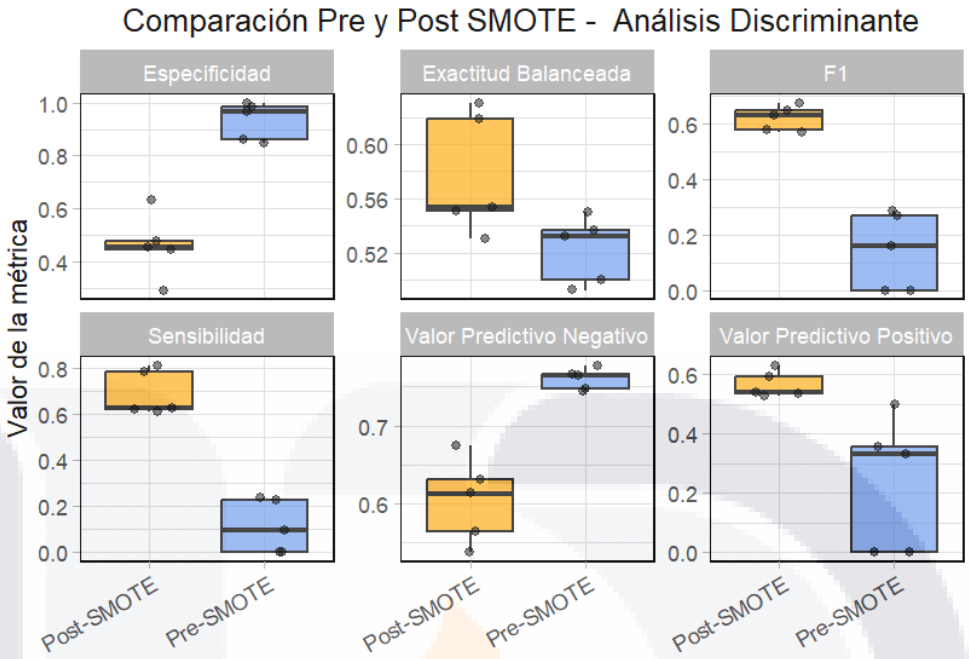


Figura 28 Desempeño Final del Análisis Discriminante

5.10.6 Red Neuronal

Para NNet, la exactitud se mantuvo prácticamente constante, con un valor de 0.60 frente al previo de 0.58. La sensibilidad mostró una mejora notable, pasando de 0.33 a 0.71, mientras que la especificidad se redujo ligeramente, de 0.67 a 0.50. La precisión se incrementó de 0.24 a 0.59, y el valor predictivo negativo disminuyó levemente, de 0.75 a 0.64. La puntuación F1 mostró una ganancia significativa, pasando de 0.28 a 0.64. La exactitud balanceada mejoró de 0.50 a 0.60. Finalmente, el área bajo la curva ROC aumentó de 0.52 a 0.65.

La Figura 29 muestra estos cambios en el desempeño del modelo, mostrando su comportamiento tras el balanceo de clases.

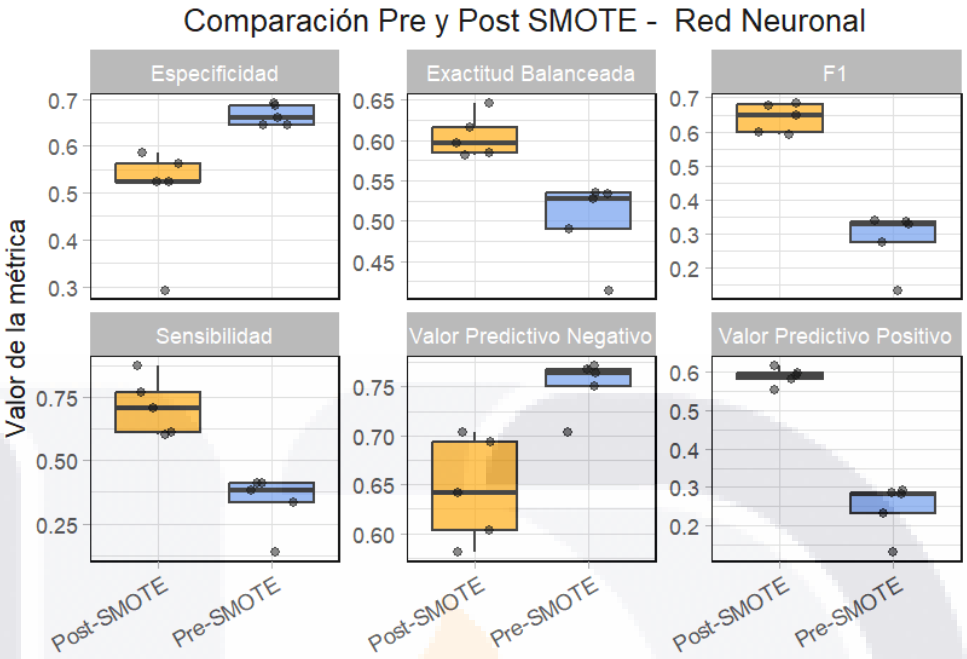


Figura 29 Desempeño Final de la Red Neuronal

5.10.7 Máquinas de Vectores de Soporte

Para SVM se observó una ligera reducción en la exactitud, que pasó de 0.65 a 0.58. La sensibilidad experimentó una mejora importante, subiendo de 0.41 a 0.72, mientras que la especificidad disminuyó de 0.73 a 0.45. La precisión aumentó de 0.35 a 0.56, y el valor predictivo negativo descendió de 0.79 a 0.64. La métrica F1 presentó una mejora clara, de 0.37 a 0.63. La exactitud balanceada mostró un incremento leve, de 0.57 a 0.58. En contraste, el área bajo la curva ROC disminuyó, pasando de 0.54 a 0.49.

Estos cambios en el desempeño del clasificador SVM se presentan en la Figura 30.

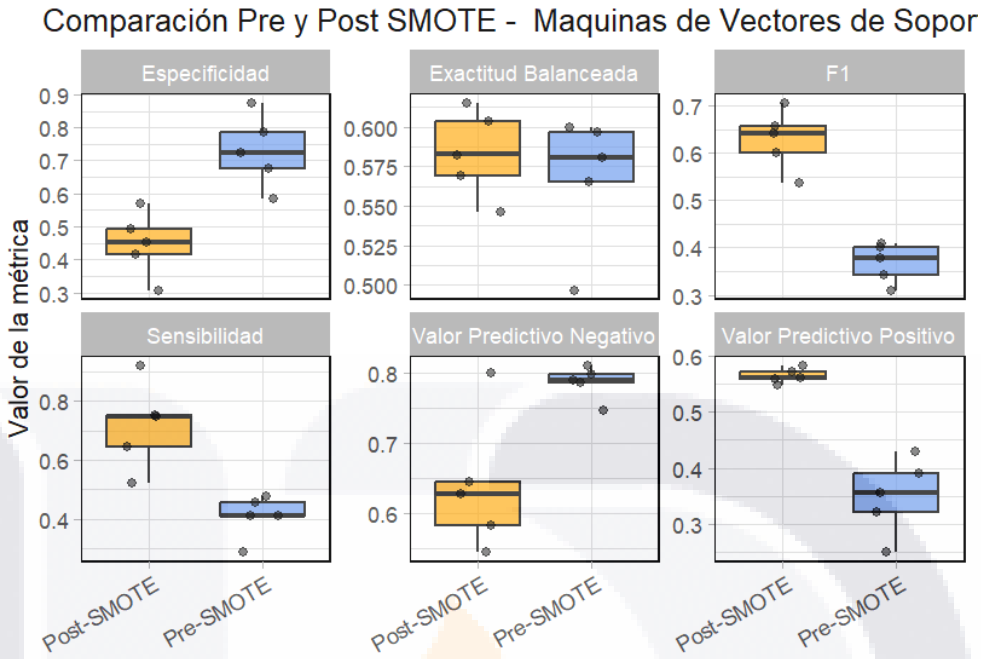


Figura 30 Desempeño Final de las Máquinas de Vectores de Soporte

5.10.8 Cópula Gaussiana

En la evaluación posterior al balanceo, la CG mantuvo una alta exactitud, con un valor final de 0.73 frente al previo de 0.75. La sensibilidad mostró un incremento, al pasar de 0.04 a 0.55, mientras que la especificidad descendió de 0.98 a 0.90. La precisión mejoró de 0.22 a 0.86, y el valor predictivo negativo se redujo de 0.75 a 0.67. La métrica F1 evidenció un aumento notable, de 0.06 a 0.66. La exactitud balanceada también se elevó, de 0.51 a 0.73. Finalmente, el área bajo la curva ROC alcanzó 0.80, superando el valor previo de 0.50. Todos estos cambios pueden observarse en la Figura 31.

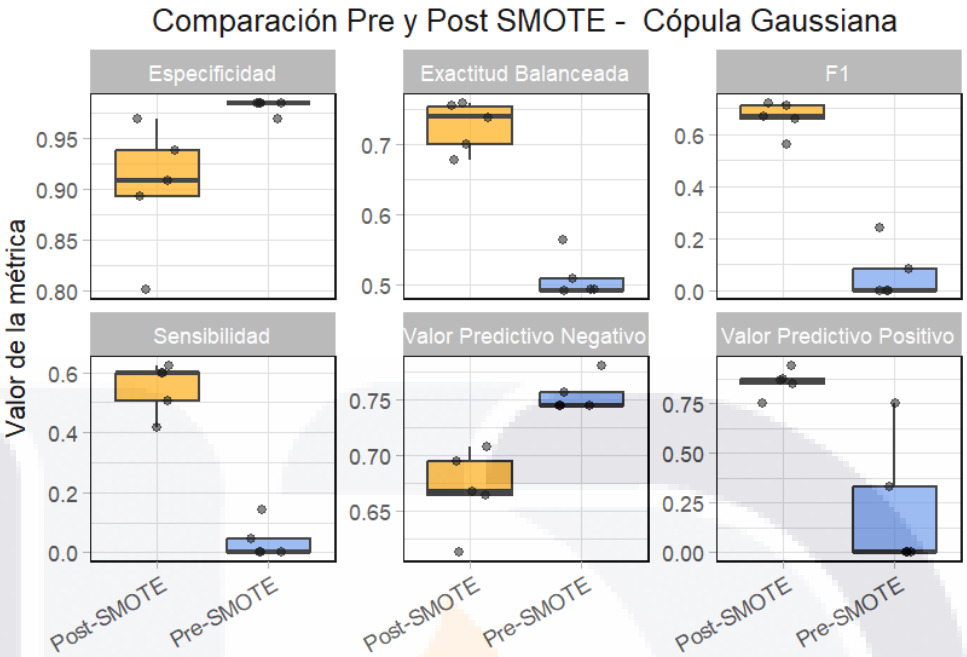


Figura 31 Desempeño Final de la Cópula Gaussiana

6. Discusión

La presente tesis evaluó el desempeño de ocho algoritmos de aprendizaje automático para predecir la anatomía vascular cervical en pacientes con hipertensión, diabetes y dislipidemia. Los hallazgos iniciales mostraron que todos los clasificadores tienen un desempeño limitado, con métricas apenas superiores al azar, debido al marcado desbalance en la proporción de clases (1:3). No obstante, tras aplicar la técnica SMOTE para corregir este desbalance, se observaron mejoras sustanciales, especialmente en sensibilidad y puntuación F1. Los clasificadores que mostraron mayores beneficios fueron la Cópula Gaussiana, con un AUROC final de 0.80, y el Clasificador Ingenuo de Bayes, con un AUROC de 0.77. Sin embargo, el rendimiento general se mantuvo moderado, lo que sugiere que las variables utilizadas tienen una capacidad limitada para discriminar entre los pacientes con anatomía vascular favorable y desfavorable.

Estudios recientes han explorado el uso de algoritmos de aprendizaje automático en el análisis de factores de riesgo clínicos y su relación con estructuras anatómicas de interés a la neurología vascular. En una muestra de la clínica de memoria de Singapur, Tan y colaboradores utilizaron resonancia magnética y aprendizaje automático para derivar un índice de edad cerebral, encontrando que este actuaba como mediador entre factores clínico-vasculares y deterioro cognitivo, especialmente en individuos con carga vascular elevada. [185] Esta aproximación refuerza la viabilidad de utilizar variables clínicas para inferir características internas mediante modelos de predicción multivariada, como se propuso en la presente tesis. Aunque el objetivo de esta tesis se enfocó en anatomía vascular y no en función cognitiva, ambos estudios comparten la premisa de que la información obtenida de expedientes clínicos y registros de pruebas de laboratorio puede capturar parcialmente propiedades estructurales subyacentes, especialmente cuando se incorpora el aprendizaje automático como herramienta de inferencia.

En otro estudio realizado en Japón, Ishida y colaboradores [186] desarrollaron un modelo predictivo de eventos cardiovasculares posteriores a endarterectomía carotídea mediante un enfoque en dos pasos: primero generaron un puntaje de riesgo a partir de imágenes patológicas de placas carotídeas mediante detección de anomalías, y luego combinaron ese

puntaje con datos clínicos para entrenar un modelo XGBoost con resultados sobresalientes en AUC y sensibilidad. Estudios como el anterior muestran al igual que esta tesis, la capacidad de modelos los algoritmos de aprendizaje automático que combinan características clínicas y estructurales para mejorar la predicción de condiciones relacionadas con el sistema vascular.

Además, Selitser y colaboradores, en un metaanálisis con datos de más de 100,000 sujetos, mostraron que algoritmos de aprendizaje automático relacionan consistentemente la diabetes con el envejecimiento cerebral aparente, con un tamaño de efecto más del doble que el de la hipertensión. [187] Este hallazgo es relevante para los modelos de predicción clínica de anatomía aórtica de esta tesis, ya que refuerza la noción de que las enfermedades crónico-degenerativas pueden influir no solo en procesos funcionales, sino también en estructuras internas de los sistemas cardiovascular y nervioso. Lo anterior apoya el uso de variables clínicas como variables aproximadas de las alteraciones estructurales/anatómicas, especialmente cuando no se dispone de imagenología directa.

También, el uso de técnicas de aprendizaje profundo ha permitido recientemente identificar patrones de alteración cerebral específicos de patologías como demencia vascular, Alzheimer y cuerpos de Lewy a partir de neuroimagen estructural. En Estados Unidos, Wang y colaboradores, desarrollaron una arquitectura multietiqueta explicable basada en T1-MRI que alcanzó alta precisión para discriminar entre estas condiciones en vida, vinculando sus predicciones a hallazgos neuropatológicos validados. [188] Sus hallazgos respaldan la posibilidad de identificar patrones anatómicos estructurales específicos mediante algoritmos de aprendizaje automático entrenados con fuentes indirectas. Esto sugiere que la predicción de la anatomía vascular, como la planteada en esta tesis, puede beneficiarse en futuras etapas de modelos más complejos.

Con respecto al desempeño, en la presente tesis los algoritmos KNN, NNet y, en menor medida, AnD demostraron un mejor desempeño. Estos clasificadores fueron los únicos que alcanzaron valores superiores a 0.60 en sensibilidad tras la aplicación de SMOTE, destacando especialmente KNN, con registros por encima de 0.80 en algunas iteraciones. Este comportamiento puede ser especialmente útil en el contexto clínico (tratamiento

endovascular del infarto cerebral) donde la omisión de un caso positivo tiene consecuencias significativas para el tratamiento o pronóstico del paciente. En estudios previos, algoritmos como KNN y redes neuronales han sido asociados con alta sensibilidad en tareas clínicas de clasificación similares, como se ha observado en trabajos de Xu y colaboradores y Liu y colaboradores realizados en China. [123, 189, 190]

Este aumento en sensibilidad se asoció, sin embargo, a una disminución en la especificidad. Este comportamiento opuesto entre sensibilidad y especificidad ha sido ampliamente discutida en el ámbito del aprendizaje automático aplicado a medicina, y se ha propuesto el uso de métricas compuestas como la exactitud balanceada para apoyar la toma de decisiones.[191] En medicina, existen múltiples escenarios clínicos donde es preferible utilizar pruebas diagnósticas con alta sensibilidad, aun si esto implica sacrificar especificidad. [192] Esto se justifica cuando el costo de omitir un caso verdadero (falso negativo) es clínicamente más grave que el de realizar estudios o intervenciones innecesarias (falsos positivos). Un ejemplo clásico es el uso de pruebas de tamizaje para cáncer en etapas tempranas, como la mamografía para cáncer de mama o la prueba de sangre oculta en heces para cáncer colorrectal. [193] En estos casos, detectar todos los casos posibles es prioritario, aun si algunos pacientes requieren estudios confirmatorios posteriores que finalmente pueden resultar negativos. Otro ejemplo se observa en pruebas rápidas para enfermedades infecciosas transmisibles, como el VIH o la tuberculosis, donde una alta sensibilidad permite identificar a casi todos los pacientes infectados, incluso si se incluyen algunos no infectados que luego se descartarán con estudios confirmatorios más específicos. [194]

Este principio también es aplicable a la predicción de la anatomía vascular aórtica en el contexto de la terapia endovascular para infarto cerebral. En estos casos, contar con una anatomía aórtica favorable (como un arco tipo I) facilita el acceso rápido al sitio de oclusión y reduce las complicaciones del procedimiento. Un modelo predictivo con alta sensibilidad permitiría identificar la mayoría de los pacientes con esta anatomía favorable, lo que es crucial para planificar con anticipación el abordaje, movilizar recursos especializados o tomar decisiones sobre la derivación a centros con capacidad para intervención urgente. [9,

13] Aunque esta estrategia pueda etiquetar erróneamente a algunos pacientes sin anatomía favorable (falsos positivos), lo cual podría reflejar cierto grado de sobreajuste del modelo a las características de la clase minoritaria, el riesgo clínico de esta sobre inclusión es bajo en comparación con el riesgo de omitir a un paciente que sí presenta un acceso vascular ideal y que podría beneficiarse de un tratamiento más rápido y eficaz.

En cuanto a esta posibilidad de sobreajuste en los hallazgos de la tesis, aunque se emplearon estrategias robustas de validación cruzada estratificada y se mantuvieron configuraciones uniformes durante el ajuste de hiperparámetros, algunos algoritmos como NNet y SVM mostraron variabilidad significativa entre iteraciones en métricas como sensibilidad y AUROC. Esta inestabilidad puede indicar que estos modelos están capturando patrones específicos del conjunto de entrenamiento, en lugar de generalizar a patrones clínicamente relevantes. Tal comportamiento ha sido descrito anteriormente en escenarios de bajo volumen muestral, donde modelos altamente flexibles como las redes neuronales pueden ajustar ruido en lugar de estructura informativa. [195]

Esta posible tendencia al sobreajuste, aun cuando se aplicaron estrategias de validación cruzada, resalta una de las limitaciones inherentes del presente trabajo: la disponibilidad restringida de datos y la dificultad para obtener conjuntos suficientemente representativos y diversos. Antes de considerar la aplicabilidad clínica de los modelos desarrollados, resulta indispensable evaluar su capacidad de generalización en escenarios distintos, con poblaciones externas que compartan características relevantes pero que no hayan sido utilizadas durante el entrenamiento. [196] Solo mediante esta validación independiente es posible confirmar la utilidad real de la propuesta metodológica.

Para validar los resultados obtenidos en una base de datos externa sería necesario contar con un conjunto independiente que incluya pacientes con características clínicas similares, y en el cual la variable objetivo se haya definido de forma equivalente a través de angio-TAC. El nuevo conjunto debería permitir una partición estratificada similar, con una distribución comparable entre clases, e idealmente provenir de una fuente institucional diferente para evaluar la posibilidad de generalización. Adicionalmente, sería recomendable repetir el mismo protocolo de preprocesamiento, ajuste y validación

cruzada, y comparar el desempeño en métricas principales. El proceso de validación externa es una condición fundamental para considerar la aplicabilidad clínica real de modelos de aprendizaje automático en medicina, como señalan recientes guías sobre desarrollo de modelos predictivos clínicos. [197, 198]

Esta validación externa, además de ser metodológicamente indispensable, permitiría posicionar los hallazgos de esta tesis en el contexto más amplio de la investigación actual. En efecto, el uso de algoritmos de aprendizaje automático en el campo de la neurología vascular es un área de investigación de punta actualmente. La mayoría de los estudios se enfocan en diversas aplicaciones del aprendizaje automático para mejorar el pronóstico y la predicción de complicaciones en pacientes con infarto cerebral y otras afecciones relacionadas. Por ejemplo, algunos trabajos abordan la predicción personalizada de la mortalidad y el deterioro neurológico temprano. [189, 190] Otros estudios subrayan el uso de enfoques de aprendizaje automático para la predicción del riesgo de infarto cerebral, alineándose con los esfuerzos para desarrollar modelos más precisos que puedan mejorar la toma de decisiones clínicas.[199] En general, los artículos publicados se centran en mejorar el manejo de pacientes con EVC mediante herramientas avanzadas de análisis predictivo. Sin embargo, ninguno de los estudios identificados hasta el momento aborda el uso de algoritmos de aprendizaje automatizado para aplicaciones en terapia endovascular para el tratamiento del infarto cerebral.

En comparación con los estudios antes mencionados, se observan algunas similitudes y diferencias en cuanto al desempeño de los algoritmos y el número de métricas reportadas. En el estudio de Xu y colaboradores de China,[189] el algoritmo k-NN alcanzó una exactitud de 0.80 y un AUROC de 0.75, mientras que las redes neuronales también mostraron un desempeño elevado. En términos de sensibilidad, Xu y colaboradores [189] reportan que estos mismos algoritmos fueron los más eficaces para identificar correctamente los casos positivos, lo que coincide con los resultados observados en esta tesis.

Por otro lado, los algoritmos tradicionales como la RL y las SVM mostraron un desempeño más limitado, tanto en este estudio como en los reportes de Vu y colaboradores [199] y Liu y colaboradores [190] En estos estudios, la RL y las SVM mostraron valores de exactitud de

0.64 y 0.68 respectivamente, mientras que su sensibilidad fue considerablemente baja, en especial en Liu y colaboradores,[190] donde la regresión logística tuvo una sensibilidad de 0.0256. Estos resultados refuerzan la idea de que estos algoritmos pueden no ser ideales cuando el objetivo es minimizar los falsos negativos en contextos clínicos.

En cuanto al número de métricas de desempeño reportadas, esta tesis evaluó ocho métricas principales. Este enfoque es mejor que el de estudios como Vu y colaboradores [199] y Liu y colaboradores, [190] que reportaron cinco métricas. Por otro lado, Xu y colaboradores [189] también incluyeron ocho métricas, incorporando correlación además de las métricas estándar. Mientras que Caliandro y colaboradores de Italia [200] optaron por un conjunto más reducido de métricas, incluyendo el error cuadrático medio (RMSE).

En términos del número de algoritmos evaluados, esta tesis utilizó ocho algoritmos de aprendizaje automático, Xu y colaboradores [189] evaluaron seis algoritmos, Liu y colaboradores [190] incluyeron siete, y estudios como los de Vu y colaboradores [199] y Caliandro y colaboradores [200] reportaron cinco. Finalmente, Li y colaboradores [201] evaluaron cuatro algoritmos, siendo los más comúnmente utilizados en todos ellos, k-NN, NNet, SVM y RL, lo que concuerda con la presente tesis y refleja la tendencia generalizada a utilizar tanto algoritmos tradicionales como avanzados en estudios comparativos.

Otro aspecto para considerar es el número de subconjuntos de validación cruzada. En esta tesis se utilizaron cinco subconjuntos, ya que este número representa un equilibrio adecuado entre variabilidad, sesgo y estabilidad de la estimación del desempeño.

Primero, al dividir el conjunto de datos en cinco partes aproximadamente iguales (de alrededor de 21-22 instancias por iteración), cada modelo es entrenado con cerca del 80% de los datos y validado con el 20% restante. Este porcentaje permite que cada modelo cuente con una cantidad suficiente de datos para aprender, sin sacrificar la diversidad del conjunto de prueba en cada iteración. A diferencia de una partición única entrenamiento-prueba, esta estrategia reduce el riesgo de que los resultados dependan de una sola división aleatoria.

Además, un número mayor de subconjuntos, como 10 o uno contra todos, aunque más común en estudios con grandes volúmenes de datos, podría ser inapropiado en este

contexto por dos razones: (1) reduciría aún más el número de muestras por iteración, generando estimaciones más inestables en conjuntos pequeños; y (2) aumentaría significativamente el costo computacional, especialmente al ajustar hiperparámetros múltiples en varios algoritmos, como se hizo en esta tesis.

Finalmente, la validación cruzada de cinco iteraciones ha sido recomendada en literatura metodológica para conjuntos de datos pequeños y medianos como un compromiso razonable entre sesgo y varianza de la estimación del error, [202] y ha sido empleada en estudios similares en el ámbito clínico predictivo. [203] Algunos estudios, como Liu y colaboradores [190] no usaron validación cruzada, y estudios como Xu y colaboradores [189] y Vu y colaboradores [199] utilizaron menos o igual cantidad.

Una diferencia con los estudios antes mencionados es el uso del algoritmo Boosting, que fue reportado en estudios como Xu y colaboradores [189] y Liu y colaboradores [190] el cual demostró un excelente desempeño. Otra diferencia, es la evaluación de los tiempos de procesamiento de los algoritmos, esta métrica no fue reportada en ninguno de los estudios previos ni se reporta rutinariamente. [204] Si bien esto representa una ventaja en términos de análisis de eficacia, la falta de comparativos con otros estudios limita la capacidad de evaluar el rendimiento temporal de los clasificadores en un contexto más amplio.

En conjunto, tanto los resultados de la presente tesis como los obtenidos de la literatura publicada previamente, subrayan la importancia de, uno, elegir el algoritmo adecuado para el contexto específico del problema clínico, considerando no solo la exactitud como métrica de desempeño principal sino también otras métricas menos sensibles al desbalance de clases [205], y dos, distanciarse de enfoques más simples que evalúan solo una métrica o una única partición entrenamiento/prueba, [91, 189, 190, 199-201, 206-209].

6.1 Limitaciones de la tesis

Esta tesis presenta varias limitaciones que deben considerarse al interpretar los resultados. En primer lugar, el tamaño del conjunto de datos fue limitado, con solo 108 instancias, lo cual restringe la capacidad de generalización de los modelos y aumenta el riesgo de sobreajuste, especialmente en algoritmos con alta flexibilidad como las redes neuronales y

las máquinas de vectores de soporte. Aunque se implementaron estrategias de validación cruzada estratificada para mitigar este efecto, la variabilidad observada entre iteraciones sugiere que algunos modelos podrían estar capturando patrones específicos del conjunto de entrenamiento más que características clínicas generalizables.

En segundo lugar, el conjunto de datos presentó un marcado desbalance de clases (proporción 1:3), lo cual afectó negativamente métricas como sensibilidad y valor predictivo positivo en la evaluación inicial. Si bien la técnica SMOTE permitió mejorar significativamente estas métricas, este tipo de sobremuestreo sintético no sustituye a la información real y podría introducir artefactos que no reflejan fielmente la distribución clínica original.

Otra limitación importante es la homogeneidad observada entre las variables clínicas y de laboratorio, muchas de las cuales no mostraron diferencias estadísticamente significativas entre los grupos. Esto indica que los atributos disponibles tienen un poder discriminativo limitado, lo que se evidencio en el análisis aislado de regresión logística.

Finalmente, ninguno de los modelos generados ha sido validado en una cohorte externa. La falta de validación independiente impide asegurar su aplicabilidad clínica en pacientes distintos a los incluidos en el set de datos original, por lo que los hallazgos deben considerarse preliminares hasta que se confirmen en otros contextos y con poblaciones más amplias.

7. Conclusiones

Aunque los resultados de desempeño obtenidos en este estudio fueron, en general, modestos, la experiencia adquirida a lo largo del proceso representó una oportunidad valiosa para el aprendizaje práctico y estructurado de técnicas de aprendizaje automático aplicadas a un problema clínico real. La tarea planteada: predecir la anatomía aortica a partir de variables clínicas, implicó retos significativos, ya que se identificó una desconexión natural entre los atributos disponibles y la variable objetivo, lo que se tradujo en una limitada capacidad predictiva de los algoritmos evaluados, incluso tras un preprocesamiento cuidadoso y la implementación de estrategias avanzadas de optimización y manejo del desbalance.

A lo largo del desarrollo del proyecto se llevaron a cabo múltiples actividades de alto valor metodológico: la obtención, procesamiento y limpieza de imágenes de angiotomografía, el pareamiento con los registros clínicos provenientes de expedientes electrónicos correspondientes a cada paciente, la curación y limpieza del conjunto de datos, la codificación y transformación de variables, la implementación de validación cruzada estratificada, el ajuste de hiperparámetros específicos para cada algoritmo, y la aplicación de técnicas para corregir el desbalance de clases, específicamente SMOTE. Se exploraron diversos algoritmos de aprendizaje automático supervisado, y se calcularon múltiples métricas que permitieron una evaluación robusta del desempeño de cada clasificador.

Más allá de los resultados numéricos, el valor principal del trabajo radica en haber diseñado, implementado y evaluado un flujo completo de aprendizaje automático, desde la formulación del problema hasta el análisis crítico de las salidas. Esta experiencia no solo fortaleció las competencias técnicas en programación, estadística y modelado predictivo, sino que permitió desarrollar una visión crítica sobre los límites y posibilidades del aprendizaje automático en entornos clínicos. Asimismo, evidenció la importancia de considerar la calidad y relevancia de las variables disponibles para tareas de clasificación, así como la necesidad de integrar dominios complementarios de información cuando se busca predecir fenómenos complejos, como los anatómicos.

7.1 Trabajo Futuro

El trabajo futuro puede incluir el enfoque combinado en lugar del competitivo. Esto significa usar varios algoritmos de aprendizaje automático para una tarea de clasificación (aprendizaje de conjuntos). El aprendizaje por conjuntos combina las predicciones de varios modelos para mejorar el desempeño general, reducir el sobreajuste y aumentar la solidez. Varias técnicas son adecuadas para ser estudiadas en estudios futuros, por ejemplo, el *Bagging* (*Bootstrap Aggregating*), [210] que consiste en entrenar múltiples modelos (generalmente del mismo tipo) en diferentes subconjuntos de los datos de entrenamiento (obtenidos a través de *bootstrapping*) y promediar sus predicciones. [211]

El *boosting,* que consiste en entrenar varios modelos de forma secuencial, donde cada modelo intenta corregir los errores de su predecesor. [212] El apilamiento (generalización apilada) implica entrenar varios modelos y, a continuación, usar otro modelo (llamado meta aprendiz) para combinar sus predicciones. [213] Los modelos base se entrenan con los datos originales y el meta aprendiz se entrena con sus predicciones. Y *voting*, entrenando múltiples modelos y combinando sus predicciones por votación mayoritaria. Puede ser una votación dura (en la que cada modelo vota por una clase) o una votación blanda (en la que se promedian las probabilidades previstas). [214] Estas técnicas pueden ayudar a aprovechar las fortalezas de diferentes algoritmos de aprendizaje automático para lograr un mejor desempeño que cualquier modelo por sí solo.

Asimismo, dada la variabilidad observada en el desempeño de los clasificadores individuales aplicados en esta tesis, y considerando que ningún algoritmo logró dominar todas las métricas de forma consistente, el uso de enfoques combinados podría resultar especialmente prometedor en este contexto. La base de datos empleada, caracterizada por un tamaño moderado, desbalance entre clases y un reto inherente en la predicción de una variable estructural interna a partir de datos clínicos, ofrece un escenario ideal para explorar técnicas de aprendizaje por conjuntos. Estas podrían mitigar el riesgo de sobreajuste observado en algunos modelos altamente flexibles y mejorar la generalización mediante la integración de sus diferentes fortalezas. Futuras implementaciones podrían

probar el desempeño de combinaciones específicas de algoritmos que, en este estudio, destacaron por su sensibilidad, exactitud balanceada o eficiencia computacional.

Con base en los resultados obtenidos en esta tesis, se pueden proponer combinaciones específicas de algoritmos para su evaluación futura en esquemas de aprendizaje por conjuntos. Estas combinaciones podrían aprovechar las fortalezas observadas en diferentes clasificadores, optimizando la sensibilidad, la robustez o el equilibrio general del sistema. A continuación, se presentan algunos ejemplos concretos:

- **CIB + NNet + CG**: Esta combinación es prometedora para priorizar la sensibilidad sin sacrificar totalmente la especificidad o la exactitud balanceada. En el estudio, CIB mostró los valores más altos de sensibilidad, NNet tuvo un desempeño competitivo en múltiples métricas incluyendo F1 y exactitud balanceada, y CG destacó por su especificidad. Un votador blando entre estos tres modelos podría equilibrar sus sesgos individuales.
- **ArD + RL + NNet**: ArD mostró un rendimiento robusto en varias métricas y el mejor tiempo de procesamiento, RL tuvo una exactitud aceptable y una estructura simple que puede actuar como base estable, mientras que NNet proporcionó buen AUROC. Esta combinación podría ser adecuada para un esquema de apilamiento, donde RL podría usarse como modelo base y NNet como meta-aprendiz.
- **k-NN + NNet + CG**: Para lograr mayor exactitud balanceada y una buena relación sensibilidad/F1. k-NN destacó por su equilibrio, NNet por su F1, y CG por su especificidad. La heterogeneidad de estos algoritmos los hace buenos candidatos para enfoques tipo *boosting*, donde se espera que los errores de uno puedan ser corregidos por otro.

Estas combinaciones podrían implementarse en futuros estudios, para validar si efectivamente logran un mejor rendimiento agregado que los clasificadores individuales, especialmente al aplicarse en una base de datos externa o expandida.

Bibliografía

1. Amigo, J.M. and M. Small, *Mathematical methods in medicine: neuroscience, cardiology and pathology.* Philos Trans A Math Phys Eng Sci, 2017. **375**(2096).
2. Boateng, E.Y. and D.A. Abaye, *A review of the logistic regression model with emphasis on medical research.* Journal of data analysis and information processing, 2019. **7**(4): p. 190-207.
3. Fiuza Pérez, M.D. and J.C. Rodríguez Pérez, *La regresión logística: una herramienta versátil.* Nefrología, 2000. **20**(6): p. 495-500.
4. Shehab, M., et al., *Machine learning in medical applications: A review of state-of-the-art methods.* Computers in Biology and Medicine, 2022. **145**: p. 105458.
5. Elyan, E., et al., *Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward.* Artificial Intelligence Surgery, 2022. **2**.
6. Tu, J.V., *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.* Journal of clinical epidemiology, 1996. **49**(11): p. 1225-1231.
7. Deo, R.C., *Machine learning in medicine.* Circulation, 2015. **132**(20): p. 1920-1930.
8. Berkhemer, O.A., et al., *A randomized trial of intraarterial treatment for acute ischemic stroke.* N Engl J Med, 2015. **372**(1): p. 11-20.
9. Goyal, M., et al., *Randomized assessment of rapid endovascular treatment of ischemic stroke.* N Engl J Med, 2015. **372**(11): p. 1019-30.
10. Saver, J.L., et al., *Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke.* N Engl J Med, 2015. **372**(24): p. 2285-95.
11. Campbell, B.C., et al., *Endovascular therapy for ischemic stroke with perfusion-imaging selection.* N Engl J Med, 2015. **372**(11): p. 1009-18.
12. Jovin, T.G., et al., *Thrombectomy within 8 hours after symptom onset in ischemic stroke.* N Engl J Med, 2015. **372**(24): p. 2296-306.
13. Goyal, M., et al., *Endovascular Therapy in Acute Ischemic Stroke: Challenges and Transition From Trials to Bedside.* Stroke, 2016. **47**(2): p. 548-53.

14. Sarraj, A., et al., *Direct to Angiography vs Repeated Imaging Approaches in Transferred Patients Undergoing Endovascular Thrombectomy.* JAMA Neurol, 2021. **78**(8): p. 916-926.
15. Braga-Neto, U., *Fundamentals of pattern recognition and machine learning*. 2020: Springer.
16. Ozdemir, O., et al., *Copula based classifier fusion under statistical dependence.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. **40**(11): p. 2740-2748.
17. Peres, M.V.d.O., *Bivariate distributions based on copulas functions: developments and applications in medical studies*. 2021, Universidade de São Paulo.
18. Grazian, C., L. Dalla Valle, and B. Liseo, *Approximate Bayesian conditional copulas.* Computational Statistics & Data Analysis, 2022. **169**: p. 107417.
19. Duque-Molina, C., *Protocolos de Atención Integral, una estrategia para las enfermedades crónicas.* Rev Med Inst Mex Seguro Soc., 2022. **60**: p. S1-3.
20. Cornelio Presenda, A.d.C., *Nivel de discapacidad y factores de riesgo asociados en los pacientes con accidente cerebrovascular atendidos en el Hospital de Especialidades de Puebla*. 2019, Benemérita Universidad Autónoma de Puebla: México.
21. Legg, S. and M. Hutter, *A collection of definitions of intelligence.* Frontiers in Artificial Intelligence and applications, 2007. **157**: p. 17.
22. Real Academia Española, *Diccionario.* Madrid: Espasa, 2001.
23. McPherson, S.S., *Artificial intelligence: building smarter machines*. 2017: Twenty-First Century Books™.
24. Mogali, S. *Artificial Intelligence and its applications in Libraries*. in *Conference: Bilingual International Conference on Information Technology: Yesterday, Today and Tomorrow, At Defence Scientific Information and Documentation Centre, Ministry of Defence Delhi*. 2014.
25. Hulick, K., *Artificial Intelligence*. 2015: ABDO.

26. Tegmark, M., *Benefits and risks of artificial intelligence.* Future of life, 2016: p. 29-31.
27. Berryman, S., *Ancient automata and mechanical explanation.* Phronesis, 2003. **48**(4): p. 344-369.
28. Thomas, P., *Artificial Intelligence*. 2005: Lucent Books.
29. McCorduck, P., et al. *History of artificial intelligence*. in *IJCAI*. 1977.
30. Ramesh, A., et al., *Artificial intelligence in medicine.* Annals of the Royal College of Surgeons of England, 2004. **86**(5): p. 334.
31. Greenhill, A., *B. A Primer of AI in Medicine.* Techniques in Gastrointestinal Endoscopy. Published online, 2019.
32. Amisha, P.M., M. Pathania, and V.K. Rathaur, *Overview of artificial intelligence in medicine.* Journal of family medicine and primary care, 2019. **8**(7): p. 2328.
33. Hamet, P. and J. Tremblay, *Artificial intelligence in medicine.* Metabolism, 2017. **69**: p. S36-S40.
34. Hill, R.K., *What an algorithm is.* Philosophy & Technology, 2016. **29**(1): p. 35-59.
35. Rapaport, W.J., *Semiotic systems, computers, and the mind: How cognition could be computing.* International Journal of Signs and Semiotic Systems (IJSSS), 2012. **2**(1): p. 32-71.
36. Rajkomar, A., J. Dean, and I. Kohane, *Machine learning in medicine.* New England Journal of Medicine, 2019. **380**(14): p. 1347-1358.
37. Simonite, T., *The recipe for the perfect robot surgeon*. 2016, MIT Technology Review.
38. Kaul, V., S. Enslin, and S.A. Gross, *History of artificial intelligence in medicine.* Gastrointestinal endoscopy, 2020. **92**(4): p. 807-812.
39. Hastie, T., et al., *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. 2009: Springer.
40. Ng, A. and M. Jordan, *On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes.* Advances in neural information processing systems, 2001. **14**.

41. Bishop, C.M. and N.M. Nasrabadi, *Pattern recognition and machine learning*. Vol. 4. 2006: Springer.
42. Vapnik, V., *Statistical learning theory.* John Wiley & Sons google schola, 1998. **2**: p. 831-842.
43. Murphy, K.P., *Machine learning: a probabilistic perspective*. 2012: MIT press.
44. Saberioon, M., et al., *Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (Oncorhynchus mykiss) classification using image-based features.* Sensors, 2018. **18**(4): p. 1027.
45. Zhang, H., *The optimality of naive Bayes.* Aa, 2004. **1**(2): p. 3.
46. Kumar, R., et al., *Naive bayes in focus: a thorough examination of its algorithmic foundations and use cases.* Int. J. Innov. Sci. Res. Technol, 2024. **9**(5): p. 2078-2081.
47. Sarang, P., *Naive Bayes: A Supervised Learning Algorithm for Classification*, in *Thinking data science: A data science practitioner's guide*. 2023, Springer. p. 143-152.
48. Mohan, P. and I. Paramasivam, *A study on impact of dimensionality reduction on Naive Bayes classifier.* Indian Journal of Science and Technology, 2017. **10**(20).
49. Wang, J.-W.D., *Naïve Bayes is an interpretable and predictive machine learning algorithm in predicting osteoporotic hip fracture in-hospital mortality compared to other machine learning algorithms.* PLOS Digital Health, 2025. **4**(1): p. e0000529.
50. Pajila, P.B., et al. *A comprehensive survey on naive bayes algorithm: Advantages, limitations and applications*. in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*. 2023. IEEE.
51. Fukunaga, K. and D.M. Hummels, *Bayes error estimation using Parzen and k-NN procedures.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987(5): p. 634-643.
52. Noto, A.P. and D.R. Saputro. *Classification data mining with Laplacian Smoothing on Naïve Bayes method*. in *AIP Conference Proceedings*. 2022. AIP Publishing.

53. Sutton, O., *Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction.* University lectures, University of Leicester, 2012. **1**.
54. Cunningham, P. and S.J. Delany, *k-Nearest neighbour classifiers-A Tutorial.* ACM Computing Surveys (CSUR), 2021. **54**(6): p. 1-25.
55. Kataria, A. and M. Singh, *A review of data classification using k-nearest neighbour algorithm.* International Journal of Emerging Technology and Advanced Engineering, 2013. **3**(6): p. 354-360.
56. Suthaharan, S. and S. Suthaharan, *Decision tree learning.* Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 2016: p. 237-269.
57. Magee, J.F., *Decision trees for decision making*. 1964: Harvard Business Review Brighton, MA, USA.
58. Dougherty, G., *Pattern recognition and classification: an introduction*. 2012: Springer Science & Business Media.
59. Schober, P. and T.R. Vetter, *Logistic Regression in Medical Research.* Anesth Analg, 2021. **132**(2): p. 365-366.
60. Bender, R. and U. Grouven, *Ordinal logistic regression in medical research.* Journal of the Royal College of physicians of London, 1997. **31**(5): p. 546.
61. Hilbe, J.M., *Logistic regression models*. 2009: Chapman and hall/CRC.
62. Maalouf, M., *Logistic regression in data analysis: an overview.* International Journal of Data Analysis Techniques and Strategies, 2011. **3**(3): p. 281-299.
63. Cokluk, O., *Logistic Regression: Concept and Application.* Educational Sciences: Theory and Practice, 2010. **10**(3): p. 1397-1407.
64. Noble, W.S., *What is a support vector machine?* Nature biotechnology, 2006. **24**(12): p. 1565-1567.
65. Suthaharan, S., *Support vector machine*, in *Machine learning models and algorithms for big data classification*. 2016, Springer. p. 207-235.
66. Pisner, D.A. and D.M. Schnyer, *Support vector machine*, in *Machine learning*. 2020, Elsevier. p. 101-121.

67. Byvatov, E. and G. Schneider, *Support vector machine applications in bioinformatics.* Applied bioinformatics, 2003. **2**(2): p. 67-77.

68. Wang, S.-C., *Artificial neural network*, in *Interdisciplinary computing in java programming*. 2003, Springer. p. 81-100.

69. Guresen, E. and G. Kayakutlu, *Definition of artificial neural networks with comparison to other networks.* Procedia Computer Science, 2011. **3**: p. 426-433.

70. Gallant, S.I., *Perceptron-based learning algorithms.* IEEE Transactions on neural networks, 1990. **1**(2): p. 179-191.

71. Gurney, K., *An introduction to neural networks*. 2018: CRC press.

72. Narayan, S., *The generalized sigmoid activation function: Competitive supervised learning.* Information sciences, 1997. **99**(1-2): p. 69-82.

73. Schluchter, M.D., *Mean square error.* Encyclopedia of Biostatistics, 2005. **5**.

74. Kohan, A., E.A. Rietman, and H.T. Siegelmann, *Signal propagation: The framework for learning and inference in a forward pass.* IEEE Transactions on Neural Networks and Learning Systems, 2023.

75. Cilimkovic, M., *Neural networks and back propagation algorithm.* Institute of Technology Blanchardstown, Blanchardstown Road North Dublin, 2015. **15**(1): p. 18.

76. Mijwel, M.M., *Artificial neural networks advantages and disadvantages.* Retrieved from LinkedIn https//www. linkedin. com/pulse/artificial-neuralnet Work, 2018.

77. Becker, S., *Unsupervised learning procedures for neural networks.* International Journal of Neural Systems, 1991. **2**(01n02): p. 17-33.

78. Worth, A.P. and M.T. Cronin, *The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects.* Journal of Molecular Structure: THEOCHEM, 2003. **622**(1-2): p. 97-111.

79. Sachin, D., *Dimensionality reduction and classification through PCA and LDA.* International journal of computer Applications, 2015. **122**(17).

80. Uzhga-Rebrov, O. and P. Grabusts. *Comparative Evaluation of Four Methods for Exploratory Data Analysis*. in *2021 62nd International Scientific Conference on*

Information Technology and Management Science of Riga Technical University (ITMS). 2021. IEEE.

81. Zhao, S., et al., *Linear discriminant analysis.* Nature Reviews Methods Primers, 2024. **4**(1): p. 70.
82. Kachigan, S.K., *Multivariate statistical analysis: A conceptual introduction*. 1991: Radius Press.
83. Pohar, M., M. Blas, and S. Turk, *Comparison of logistic regression and linear discriminant analysis: a simulation study.* Metodoloski zvezki, 2004. **1**(1): p. 143.
84. Salinas-Gutiérrez, R., A. Hernández-Aguirre, and E.R. Villa-Diharce, *Copula selection for graphical models in continuous Estimation of Distribution Algorithms.* Computational Statistics, 2014. **29**(3): p. 685-713.
85. Koch, I. and A. De Schepper, *The comonotonicity coefficient: a new measure of positive dependence in a multivariate setting.* 2006: p. https://repository.uantwerpen.be/desktop/irua.
86. Hua, L. and H. Joe, *Tail order and intermediate tail dependence of multivariate copulas.* Journal of Multivariate Analysis, 2011. **102**(10): p. 1454-1471.
87. Frees, E.W. and E.A. Valdez, *Understanding relationships using copulas.* North American actuarial journal, 1998. **2**(1): p. 1-25.
88. Lang, W. and Q. Jin, *Copula Models*, in *Applied Multivariate Statistical Analysis and Related Topics with R*. 2021, EDP Sciences. p. 97-110.
89. Morgenstern, L.B., et al., *Excess stroke in Mexican Americans compared with non-Hispanic Whites: the Brain Attack Surveillance in Corpus Christi Project.* Am J Epidemiol, 2004. **160**(4): p. 376-83.
90. Zahuranec, D.B., et al., *Differences in intracerebral hemorrhage between Mexican Americans and non-Hispanic whites.* Neurology, 2006. **66**(1): p. 30-4.
91. Siqueira, C. and D.L.B. de Souza, *Reduction of mortality and predictions for acute myocardial infarction, stroke, and heart failure in Brazil until 2030.* Sci Rep, 2020. **10**(1): p. 17856.

92. Schargrodsky, H., M.C. Escobar, and E. Escobar, *Cardiovascular disease prevention: a challenge for Latin America.* Circulation, 1998. **98**(20): p. 2103-4.

93. Kissela, B., et al., *Stroke in a biracial population: the excess burden of stroke among blacks.* Stroke, 2004. **35**(2): p. 426-31.

94. Broderick, J.P., et al., *Intracerebral hemorrhage more than twice as common as subarachnoid hemorrhage.* J Neurosurg, 1993. **78**(2): p. 188-91.

95. Qureshi, A.I., et al., *Changes in cost and outcome among US patients with stroke hospitalized in 1990 to 1991 and those hospitalized in 2000 to 2001.* Stroke, 2007. **38**(7): p. 2180-4.

96. Feigin, V.L., et al., *Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century.* Lancet Neurol, 2003. **2**(1): p. 43-53.

97. Ruiz-Sandoval, J.L., et al., *[Intracerebral haemorrhage in a referral hospital in the central-western region of Mexico].* Rev Neurol, 2005. **40**(11): p. 656-60.

98. Labovitz, D.L., et al., *The incidence of deep and lobar intracerebral hemorrhage in whites, blacks, and Hispanics.* Neurology, 2005. **65**(4): p. 518-22.

99. Jiang, B., et al., *Incidence and trends of stroke and its subtypes in China: results from three large cities.* Stroke, 2006. **37**(1): p. 63-8.

100. Kubo, M., et al., *Trends in the incidence, mortality, and survival rate of cardiovascular disease in a Japanese community: the Hisayama study.* Stroke, 2003. **34**(10): p. 2349-54.

101. Morgenstern, L.B. and W.D. Spears, *A triethnic comparison of intracerebral hemorrhage mortality in Texas.* Ann Neurol, 1997. **42**(6): p. 919-23.

102. Inagawa, T., et al., *Primary intracerebral hemorrhage in Izumo City, Japan: incidence rates and outcome in relation to the site of hemorrhage.* Neurosurgery, 2003. **53**(6): p. 1283-97; discussion 1297-8.

103. Flaherty, M.L., et al., *Racial variations in location and risk of intracerebral hemorrhage.* Stroke, 2005. **36**(5): p. 934-7.

104. Ayala, C., et al., *Sex differences in US mortality rates for stroke and stroke subtypes by race/ethnicity and age, 1995-1998.* Stroke, 2002. **33**(5): p. 1197-201.

105. Broderick, J.P., et al., *The risk of subarachnoid and intracerebral hemorrhages in blacks as compared with whites.* N Engl J Med, 1992. **326**(11): p. 733-6.

106. Woo, D., et al., *Effect of untreated hypertension on hemorrhagic stroke.* Stroke, 2004. **35**(7): p. 1703-8.

107. Thrift, A.G., et al., *Three important subgroups of hypertensive persons at greater risk of intracerebral hemorrhage. Melbourne Risk Factor Study Group.* Hypertension, 1998. **31**(6): p. 1223-9.

108. Ariesen, M.J., et al., *Risk factors for intracerebral hemorrhage in the general population: a systematic review.* Stroke, 2003. **34**(8): p. 2060-5.

109. Tirschwell, D.L., et al., *Association of cholesterol with stroke risk varies in stroke subtypes and patient subgroups.* Neurology, 2004. **63**(10): p. 1868-75.

110. Woo, D., et al., *Genetic and environmental risk factors for intracerebral hemorrhage: preliminary results of a population-based study.* Stroke, 2002. **33**(5): p. 1190-5.

111. Amarenco, P., et al., *High-dose atorvastatin after stroke or transient ischemic attack.* N Engl J Med, 2006. **355**(6): p. 549-59.

112. Woo, D., et al., *Hypercholesterolemia, HMG-CoA reductase inhibitors, and risk of intracerebral hemorrhage: a case-control study.* Stroke, 2004. **35**(6): p. 1360-4.

113. Leker, R.R., et al., *Prior use of statins improves outcome in patients with intracerebral hemorrhage: prospective data from the National Acute Stroke Israeli Surveys (NASIS).* Stroke, 2009. **40**(7): p. 2581-4.

114. Fujii, Y., et al., *Multivariate analysis of predictors of hematoma enlargement in spontaneous intracerebral hemorrhage.* Stroke, 1998. **29**(6): p. 1160-6.

115. Thrift, A.G., G.A. Donnan, and J.J. McNeil, *Heavy drinking, but not moderate or intermediate drinking, increases the risk of intracerebral hemorrhage.* Epidemiology, 1999. **10**(3): p. 307-12.

116. Feldmann, E., et al., *Major risk factors for intracerebral hemorrhage in the young are modifiable.* Stroke, 2005. **36**(9): p. 1881-5.

117. Kurth, T., et al., *Smoking and risk of hemorrhagic stroke in women.* Stroke, 2003. **34**(12): p. 2792-5.

118. Ruiz-Sandoval, J.L., et al., *Hypertensive intracerebral hemorrhage in young people: previously unnoticed age-related clinical differences.* Stroke, 2006. **37**(12): p. 2946-50.

119. Rotondi, M., F. Magri, and L. Chiovato, *Risk of coronary heart disease and mortality for adults with subclinical hypothyroidism.* JAMA, 2010. **304**(22): p. 2481; author reply 2482.

120. Alizargar, J., et al., *Use of the triglyceride-glucose index (TyG) in cardiovascular disease patients.* Cardiovasc Diabetol, 2020. **19**(1): p. 8.

121. Azizi, F., et al., *Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II.* Trials, 2009. **10**: p. 5.

122. Dzubur, A., et al., *The serum triglyceride to high-density lipoprotein (HDL) ratio in patients with acute coronary syndrome with and without renal dysfunction.* Med Glas (Zenica), 2019. **16**(1): p. 28-34.

123. Liu, W., et al., *Expression of Hcy and blood lipid levels in serum of CHD patients and analysis of risk factors for CHD.* Exp Ther Med, 2019. **17**(3): p. 1756-1760.

124. Lozano, J.V., et al., *Serum lipid profiles and their relationship to cardiovascular disease in the elderly: the PREV-ICTUS study.* Curr Med Res Opin, 2008. **24**(3): p. 659-70.

125. Mayer, O., Jr., et al., *Fibrate treatment and prevalence risk of mild hyperhomocysteinaemia in clinical coronary heart disease patients.* Eur J Cardiovasc Prev Rehabil, 2004. **11**(3): p. 244-9.

126. Pandya, D., A.K. Nagrajappa, and K.S. Ravi, *Assessment and Correlation of Urea and Creatinine Levels in Saliva and Serum of Patients with Chronic Kidney Disease, Diabetes and Hypertension- A Research Study.* J Clin Diagn Res, 2016. **10**(10): p. ZC58-ZC62.

127. Yao, H., et al., *Associations of multiple serum biomarkers and the risk of cardiovascular disease in China.* BMC Cardiovasc Disord, 2020. **20**(1): p. 426.

128. Rubinstein, S.M., et al., *A systematic review of the risk factors for cervical artery dissection.* Stroke, 2005. **36**(7): p. 1575-80.

129. Loeffen, R., H.M. Spronk, and H. ten Cate, *The impact of blood coagulability on atherosclerosis and cardiovascular disease.* J Thromb Haemost, 2012. **10**(7): p. 1207-16.

130. Girolami, A., L. Sambado, and A.M. Lombardi, *The impact of blood coagulability on atherosclerosis and cardiovascular disease: a rebuttal.* J Thromb Haemost, 2013. **11**(1): p. 213-4; discussion 215-6.

131. Lee, C.D., et al., *White blood cell count and incidence of coronary heart disease and ischemic stroke and mortality from cardiovascular disease in African-American and White men and women: atherosclerosis risk in communities study.* Am J Epidemiol, 2001. **154**(8): p. 758-64.

132. Qiu, F., et al., *Changes of coagulation function and risk of stroke in patients with COVID-19.* Brain Behav, 2021. **11**(6): p. e02185.

133. Miller, J.D. and R.M. Forte, *Mastering Predictive Analytics with R*. 2017: Packt Publishing Ltd.

134. Schöner, H., *Working with real world datasets: preprocessing and prediction with large incomplete and heterogeneous datasets*. 2005, Berlin, Techn. Univ., Diss., 2004.

135. Lesmeister, C. and S.K. Chinnamgari, *Advanced machine learning with R: tackle data analytics and machine learning challenges and build complex applications with R 3.5*. 2019: Packt Publishing Ltd.

136. Palmer, A., R. Jiménez, and E. Gervilla, *Data mining: Machine learning and statistical techniques.* Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), 2011: p. 373-396.

137. Ai, J., P.L. Brockett, and L.L. Golden, *Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables.* North American Actuarial Journal, 2009. **13**(4): p. 438-458.

138. Hahn, L.D., K. Baeumler, and A. Hsiao, *Artificial intelligence and machine learning in aortic disease.* Curr Opin Cardiol, 2021. **36**(6): p. 695-703.

139. Ullah, N., et al., *Machine learning algorithms for the prognostication of abdominal aortic aneurysm progression: a systematic review.* Minerva Surg, 2023.

140. Huo, D., et al., *A machine learning model to classify aortic dissection patients in the early diagnosis phase.* Sci Rep, 2019. **9**(1): p. 2701.

141. Liang, L., W. Mao, and W. Sun, *A feasibility study of deep learning for predicting hemodynamics of human thoracic aorta.* J Biomech, 2020. **99**: p. 109544.

142. Zaidat, O.O., et al., *First Pass Effect: A New Measure for Stroke Thrombectomy Devices.* Stroke, 2018. **49**(3): p. 660-666.

143. Baek, J.H., et al., *Number of Stent Retriever Passes Associated With Futile Recanalization in Acute Stroke.* Stroke, 2018. **49**(9): p. 2088-2095.

144. Bourcier, R., et al., *More than three passes of stent retriever is an independent predictor of parenchymal hematoma in acute ischemic stroke.* J Neurointerv Surg, 2019. **11**(7): p. 625-629.

145. Huang, X., et al., *Influence of procedure time on outcome and hemorrhagic transformation in stroke patients undergoing thrombectomy.* J Neurol, 2019. **266**(10): p. 2560-2570.

146. Ribo, M., et al., *Difficult catheter access to the occluded vessel during endovascular treatment of acute ischemic stroke is associated with worse clinical outcome.* J Neurointerv Surg, 2013. **5 Suppl 1**: p. i70-3.

147. Saver, J.L., et al., *Time to Treatment With Endovascular Thrombectomy and Outcomes From Ischemic Stroke: A Meta-analysis.* JAMA, 2016. **316**(12): p. 1279-88.

148. Ribo, M., et al., *Association Between Time to Reperfusion and Outcome Is Primarily Driven by the Time From Imaging to Reperfusion.* Stroke, 2016. **47**(4): p. 999-1004.

149. Alawieh, A., et al., *Impact of Procedure Time on Outcomes of Thrombectomy for Stroke.* J Am Coll Cardiol, 2019. **73**(8): p. 879-890.

150. Spiotta, A.M., et al., *The golden hour of stroke intervention: effect of thrombectomy procedural time in acute ischemic stroke on outcome.* J Neurointerv Surg, 2014. **6**(7): p. 511-6.

151. Snelling, B.M., et al., *Unfavorable Vascular Anatomy Is Associated with Increased Revascularization Time and Worse Outcome in Anterior Circulation Thrombectomy.* World Neurosurg, 2018. **120**: p. e976-e983.

152. Mazighi, M., et al., *Impact of onset-to-reperfusion time on stroke mortality: a collaborative pooled analysis.* Circulation, 2013. **127**(19): p. 1980-5.

153. Hassan, A.E., et al., *Impact of procedural time on clinical and angiographic outcomes in patients with acute ischemic stroke receiving endovascular treatment.* J Neurointerv Surg, 2019. **11**(10): p. 984-988.

154. Bourcier, R., et al., *Susceptibility vessel sign on MRI predicts better clinical outcome in patients with anterior circulation acute stroke treated with stent retriever as first-line strategy.* J Neurointerv Surg, 2019. **11**(4): p. 328-333.

155. Kaesmacher, J., et al., *Reasons for Reperfusion Failures in Stent-Retriever-Based Thrombectomy: Registry Analysis and Proposal of a Classification System.* AJNR Am J Neuroradiol, 2018. **39**(10): p. 1848-1853.

156. Lee, J.S., J.M. Hong, and J.S. Kim, *Diagnostic and Therapeutic Strategies for Acute Intracranial Atherosclerosis-related Occlusions.* J Stroke, 2017. **19**(2): p. 143-151.

157. Jeong, D.E., et al., *Impact of Balloon-Guiding Catheter Location on Recanalization in Patients with Acute Stroke Treated by Mechanical Thrombectomy.* AJNR Am J Neuroradiol, 2019. **40**(5): p. 840-844.

158. Nagata, T., et al., *Three-dimensional computed tomographic analysis of variations of the carotid artery.* J Craniomaxillofac Surg, 2016. **44**(6): p. 734-42.

159. La Barbera, G., et al., *Kinking, coiling, and tortuosity of extracranial internal carotid artery: is it the effect of a metaplasia?* Surg Radiol Anat, 2006. **28**(6): p. 573-80.

160. Association, G.A.o.t.W.M., *World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects.* The Journal of the American College of Dentists, 2014. **81**(3): p. 14-18.

161. Reglamento de la Ley General de Salud en Materia de Investigación para la Salud. En Diario Oficial de la Federación. México.6 de enero de 1987. (Capítulo III, a.

162. *Engagement of Institutions in Human Subjects Research*. 2008, U.S. Department of Health & Human Services.

163. Tawfik, A.M., et al., *Prevalence and Types of Aortic Arch Variants and Anomalies in Congenital Heart Diseases.* Acad Radiol, 2019. **26**(7): p. 930-936.

164. Kotelis, D., et al., *Morphological risk factors of stroke during thoracic endovascular aortic repair.* Langenbecks Arch Surg, 2012. **397**(8): p. 1267-73.

165. Guo, Y., et al., *Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms.* BMC Bioinformatics, 2010. **11**: p. 447.

166. Ein-Dor, L., O. Zuk, and E. Domany, *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5923-8.

167. Casserly, I.P. and S.R. Kapadia, *Advances in percutaneous valvular intervention.* Expert Rev Cardiovasc Ther, 2005. **3**(1): p. 143-58.

168. R Core Team, R., *R: A language and environment for statistical computing.* 2013.

169. *Posit team (2025). RStudio: Integrated Development Environment for R.*

Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

170. Majka, M. and M.M. Majka, *Package 'naivebayes'*. 2020, CRAN.

171. Zhang, Z., *Introduction to machine learning: k-nearest neighbors.* Annals of translational medicine, 2016. **4**(11).

172. Therneau, T., et al., *Package 'rpart'.* Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016), 2015.

173. Tharwat, A., et al., *Linear discriminant analysis: A detailed tutorial.* AI communications, 2017. **30**(2): p. 169-190.

174. Mahdi, S., et al., *A Review of R Neural Network Packages (with NNbenchmark): Accuracy and Ease of Use.*

175. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.

176. Hofert, M., et al., *Multivariate dependence with copulas.* 2013.

177. VanderWeele, T.J. and I. Shpitser, *A new criterion for confounder selection.* Biometrics, 2011. **67**(4): p. 1406-13.

178. Ramyachitra, D. and P. Manikandan, *Imbalanced dataset classification and solutions: a review.* International Journal of Computing and Business Research (IJCBR), 2014. **5**(4): p. 1-29.

179. Mohammed, R., J. Rawashdeh, and M. Abdullah. *Machine learning with oversampling and undersampling techniques: overview study and experimental results*. in *2020 11th international conference on information and communication systems (ICICS)*. 2020. IEEE.

180. Mujahid, M., et al., *Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering.* Journal of Big Data, 2024. **11**(1): p. 87.

181. Elreedy, D., A.F. Atiya, and F. Kamalov, *A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning.* Machine Learning, 2024. **113**(7): p. 4903-4923.

182. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.

183. Elreedy, D. and A.F. Atiya, *A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance.* Information Sciences, 2019. **505**: p. 32-64.

184. Wang, A.X., et al., *Addressing imbalance in health data: Synthetic minority oversampling using deep learning.* Computers in Biology and Medicine, 2025. **188**: p. 109830.

185. Tan, W.Y., et al., *Role of Brain Age Gap as a Mediator in the Relationship Between Cognitive Impairment Risk Factors and Cognition.* Neurology, 2025. **105**(2): p. e213815.

186. Ishida, S., et al., *Prediction of cardiovascular events after carotid endarterectomy using pathological images and clinical data.* International Journal of Computer Assisted Radiology and Surgery, 2025. **20**(4): p. 643-652.

187. Selitser, M., et al., *Cardiometabolic risk factors and brain age: a meta-analysis to quantify brain structural differences related to diabetes, hypertension, and obesity.* Journal of Psychiatry and Neuroscience, 2025. **50**(2): p. E102-E111.

188. Wang, D., et al., *Deep learning reveals pathology-confirmed neuroimaging signatures in Alzheimer's, vascular and Lewy body dementias.* Brain, 2025. **148**(6): p. 1963-1977.

189. Xu, L., et al., *Personalized prediction of mortality in patients with acute ischemic stroke using explainable artificial intelligence.* European Journal of Medical Research, 2024. **29**(1): p. 341.

190. Liu, W., et al., *Prediction of early neurologic deterioration in patients with perforating artery territory infarction using machine learning: a retrospective study.* Frontiers in Neurology, 2024. **15**: p. 1368902.

191. Bekkar, M., H.K. Djemaa, and T.A. Alitouche, *Evaluation measures for models assessment over imbalanced data sets.* J Inf Eng Appl, 2013. **3**(10).

192. Zavon, M.R., *Principles and Practice of Screening for Disease.* Archives of internal medicine, 1969. **123**(3): p. 349-349.

193. Hara, A.K., et al., *Detection of colorectal polyps with CT colography: initial assessment of sensitivity and specificity.* Radiology, 1997. **205**(1): p. 59-65.

194. Gallagher, E.J., *The problem with sensitivity and specificity….* Annals of emergency medicine, 2003. **42**(2): p. 298-303.

195. Li, D.-C. and C.-W. Liu, *A neural network weight determination model designed uniquely for small data set learning.* Expert systems with applications, 2009. **36**(6): p. 9853-9858.

196. Li, D.-C., C.-W. Liu, and S.C. Hu, *A learning method for the class imbalance problem with medical data sets.* Computers in biology and medicine, 2010. **40**(5): p. 509-518.

197. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement.* Circulation, 2015. **131**(2): p. 211-219.

198. Moons, K.G., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration.* Annals of internal medicine, 2015. **162**(1): p. W1-W73.

199. Vu, T., et al., *Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study.* Journal of Cardiovascular Development and Disease, 2024. **11**(7): p. 207.

200. Caliandro, P., et al., *Artificial intelligence to predict individualized outcome of acute ischemic stroke patients: The SIBILLA project.* Eur Stroke J, 2024: p. 23969873241253366.

201. Li, X., et al., *Development and performance assessment of novel machine learning models for predicting postoperative pneumonia in aneurysmal subarachnoid hemorrhage patients: external validation in MIMIC-IV.* Front Neurol, 2024. **15**: p. 1341252.

202. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*. 1995. Montreal, Canada.

203. Varoquaux, G., *Cross-validation failure: Small sample sizes lead to large error bars.* Neuroimage, 2018. **180**: p. 68-77.

204. Coleman, C., et al., *Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark.* ACM SIGOPS Operating Systems Review, 2019. **53**(1): p. 14-25.

205. Thumpati, A. and Y. Zhang. *Towards Optimizing Performance of Machine Learning Algorithms on Unbalanced Dataset*. in *CS & IT Conference Proceedings*. 2023. CS & IT Conference Proceedings.

206. Hadanny, A., et al., *Machine learning-based prediction of 1-year mortality for acute coronary syndrome().* J Cardiol, 2022. **79**(3): p. 342-351.

207. Wu, C.T., T.W. Chu, and J.R. Jang, *Current-Visit and Next-Visit Prediction for Fatty Liver Disease With a Large-Scale Dataset: Model Development and Performance Comparison.* JMIR Med Inform, 2021. **9**(8): p. e26398.

208. Zachariah, F.J., et al., *Prospective Comparison of Medical Oncologists and a Machine Learning Model to Predict 3-Month Mortality in Patients With Metastatic Solid Tumors.* JAMA Netw Open, 2022. **5**(5): p. e2214514.

209. Zheng, B., et al., *An Interpretable Model-Based Prediction of Severity and Crucial Factors in Patients with COVID-19.* Biomed Res Int, 2021. **2021**: p. 8840835.

210. Bulut, F. *Heart attack risk detection using Bagging classifier*. in *2016 24th Signal Processing and Communication Application Conference (SIU)*. 2016. IEEE.

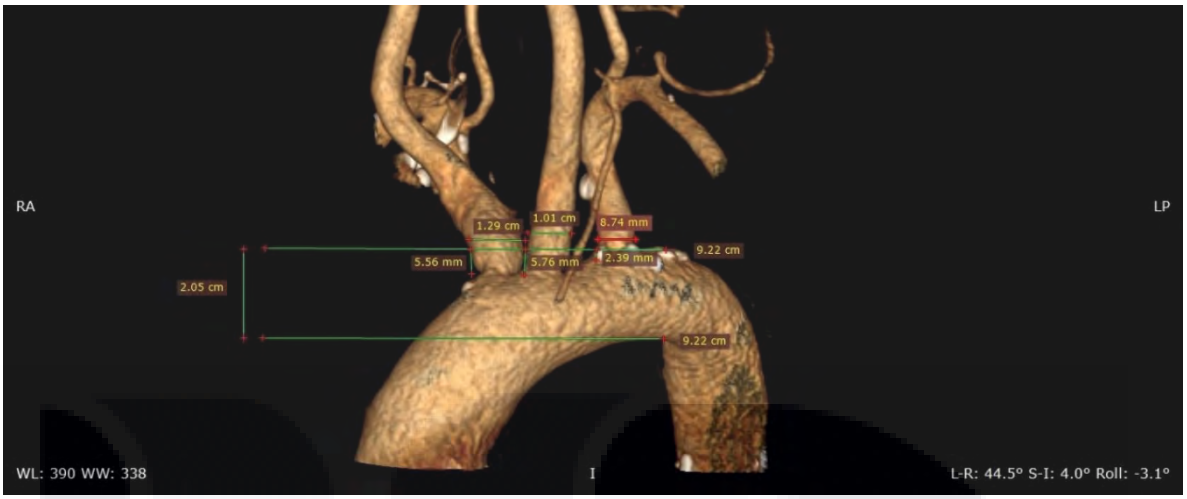
211. Farid, D.M., M.Z. Rahman, and C.M. Rahman, *An ensemble approach to classifier construction based on bootstrap aggregation.* International Journal of Computer Applications, 2011. **25**(5): p. 30-34.

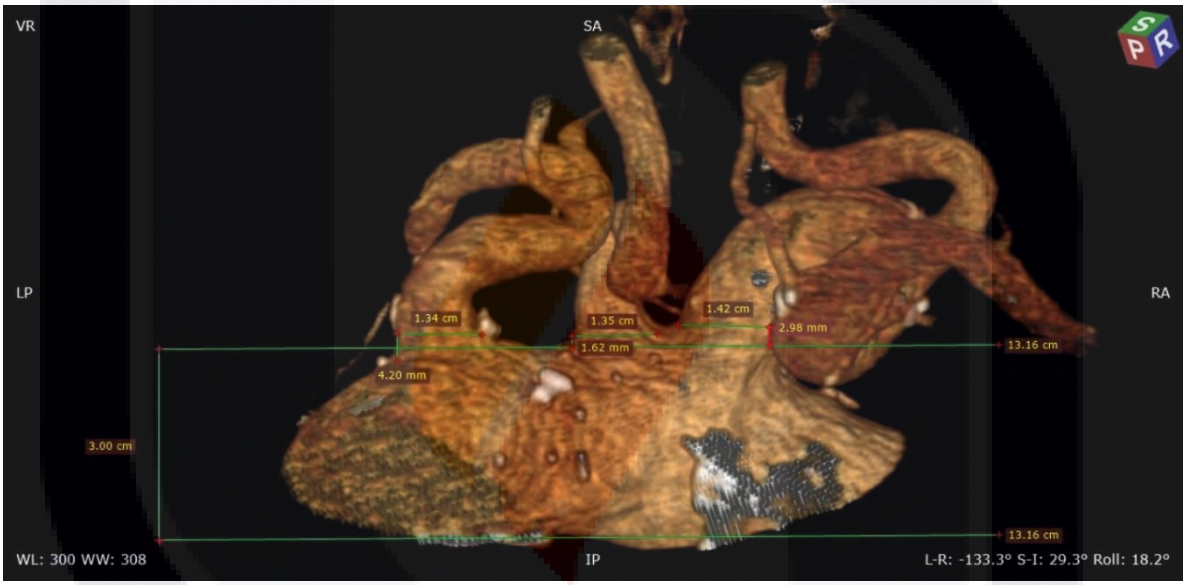
212. Ferreira, A.J. and M.A. Figueiredo, *Boosting algorithms: A review of methods, theory, and applications.* Ensemble machine learning: Methods and applications, 2012: p. 35-85.

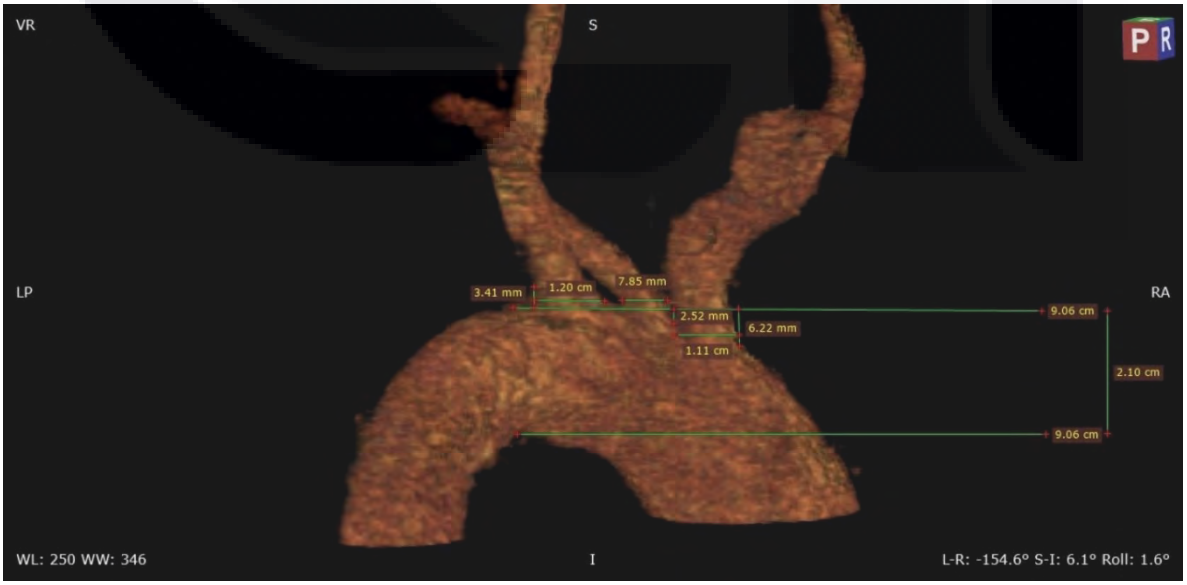
213. Pavlyshenko, B. *Using stacking approaches for machine learning models*. in *2018 IEEE second international conference on data stream mining & processing (DSMP)*. 2018. IEEE.

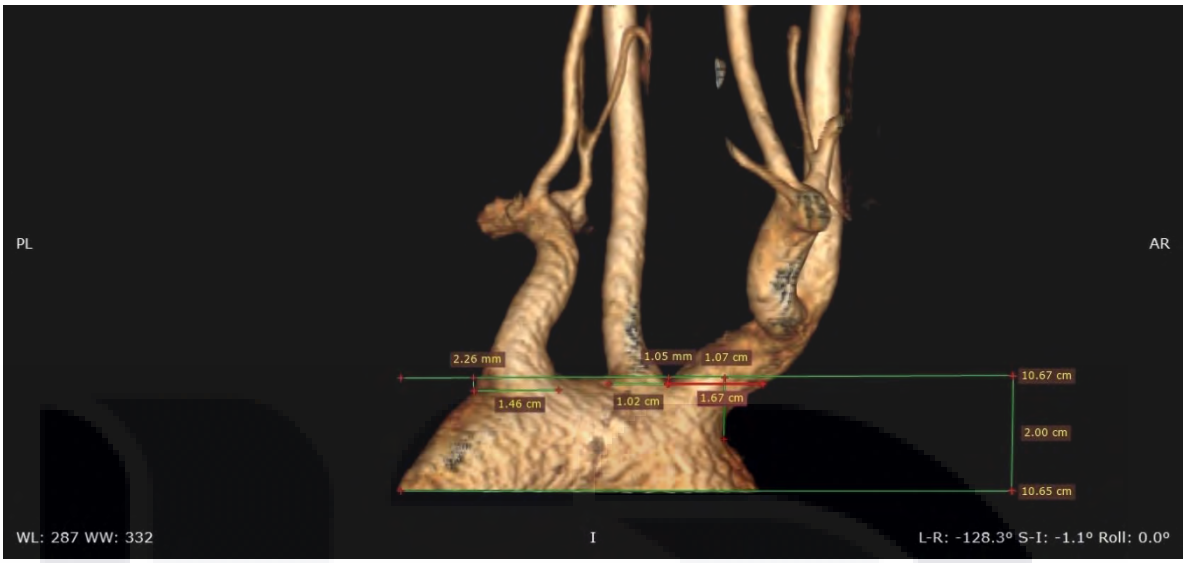
214. Burka, D., et al., *Voting: A machine learning approach.* European Journal of Operational Research, 2022. **299**(3): p. 1003-1017.

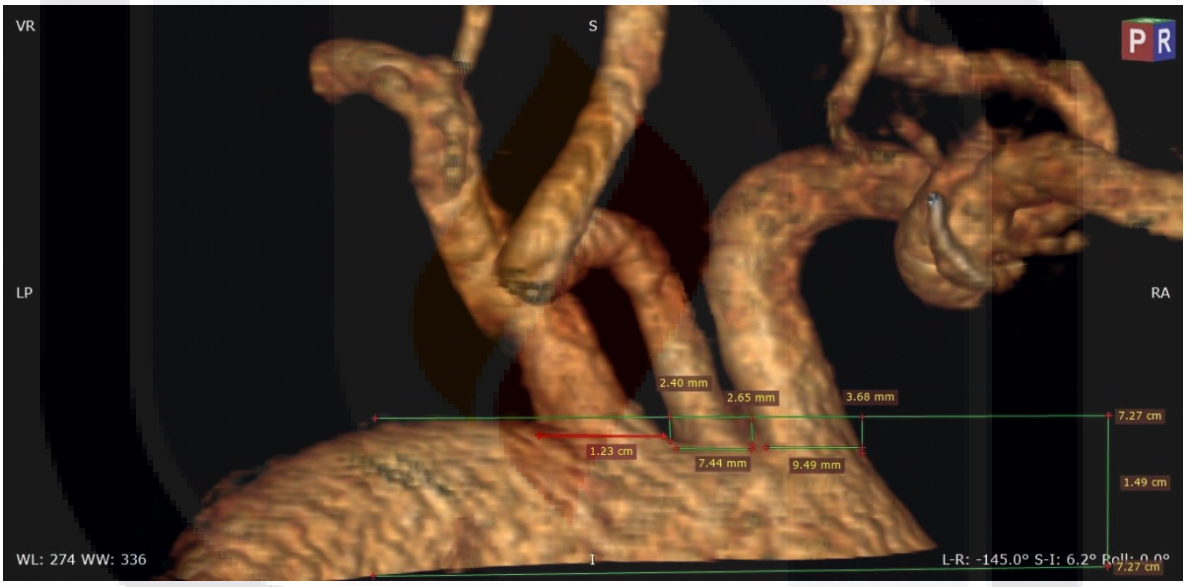
Anexo A. Galería de imágenes de Angiotomografía de vasos supraórticos de los pacientes estudiados

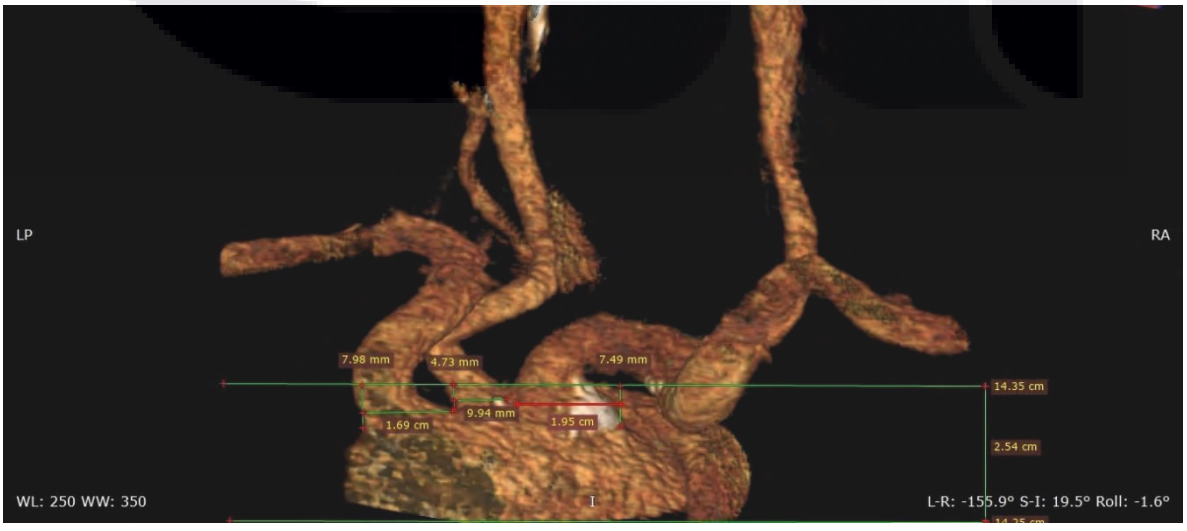


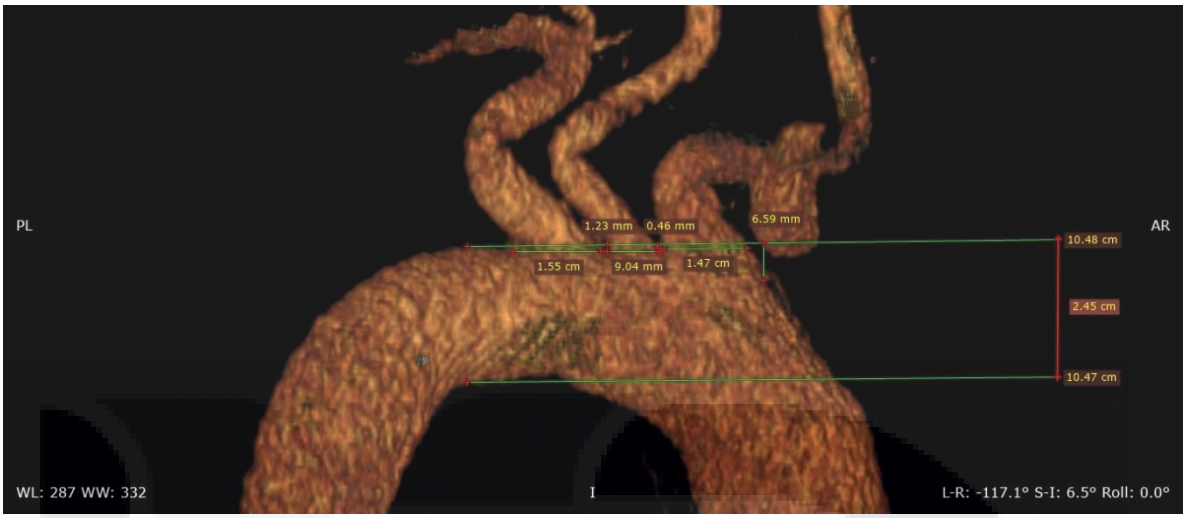


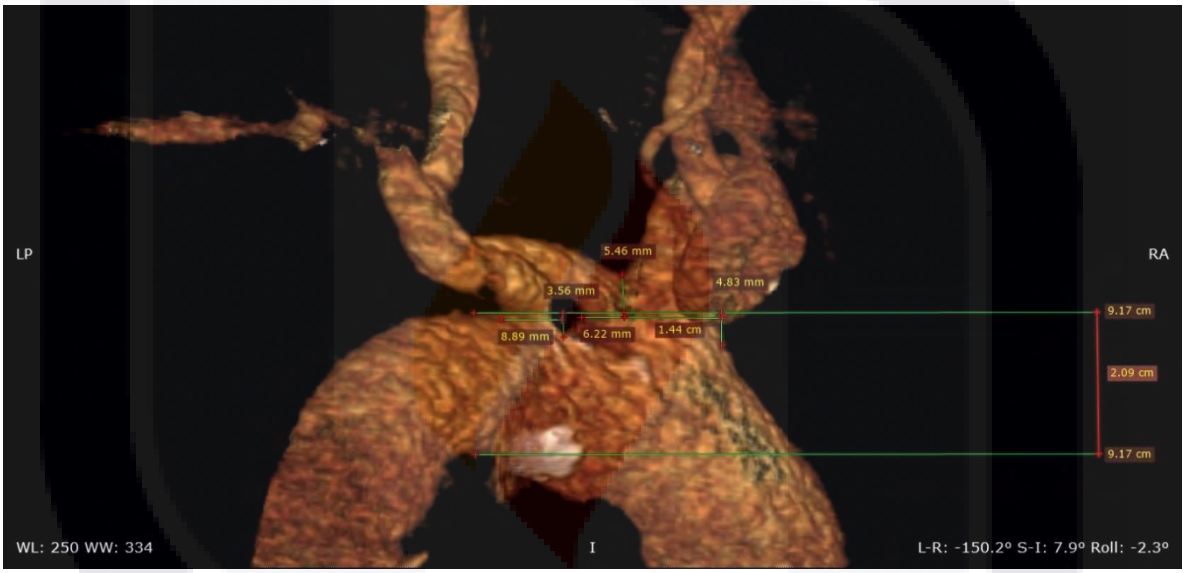


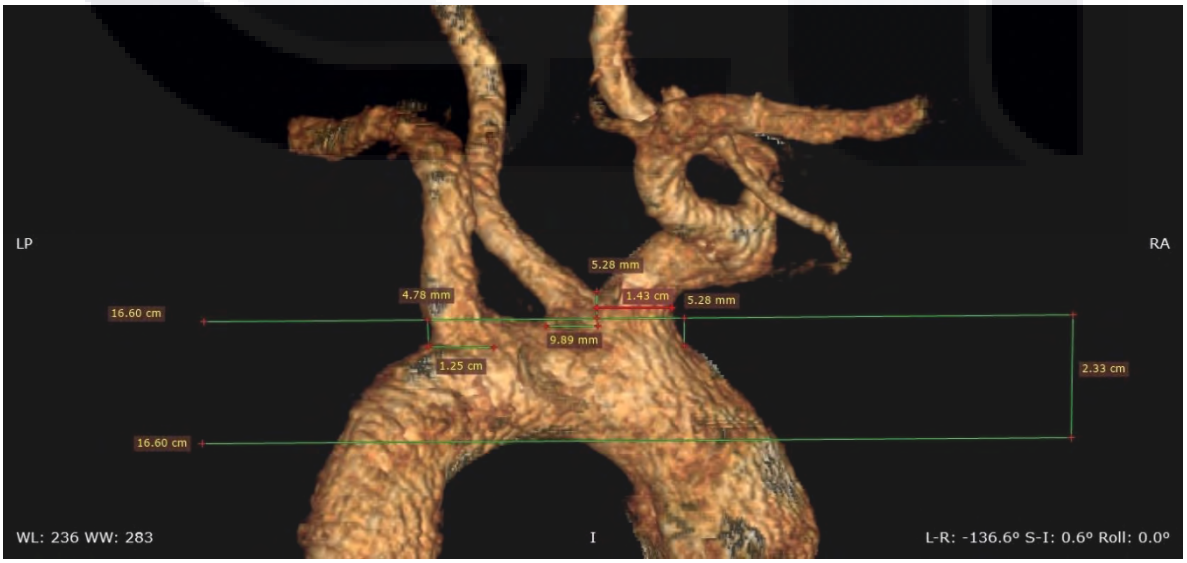


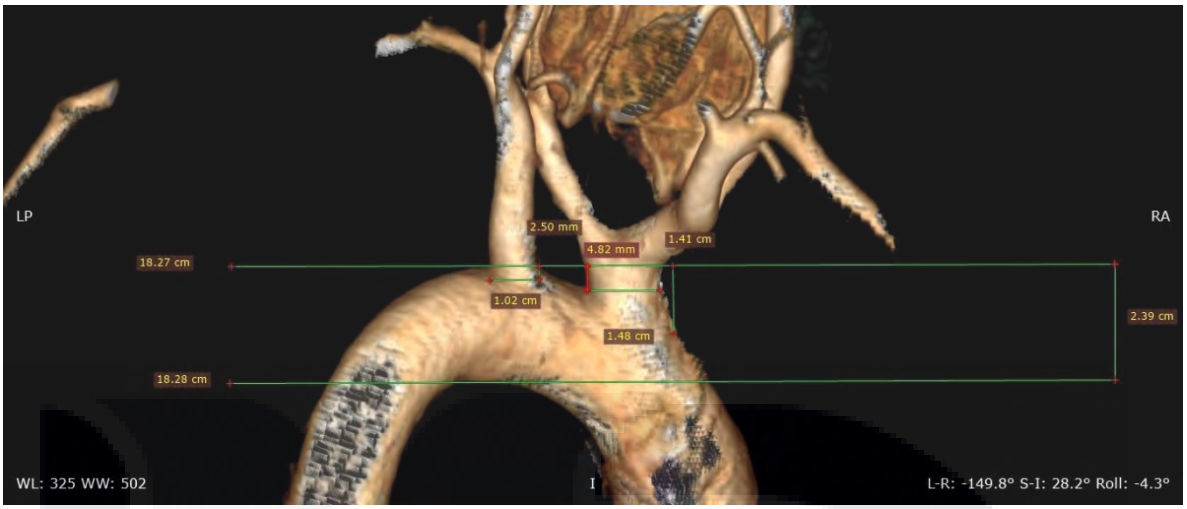


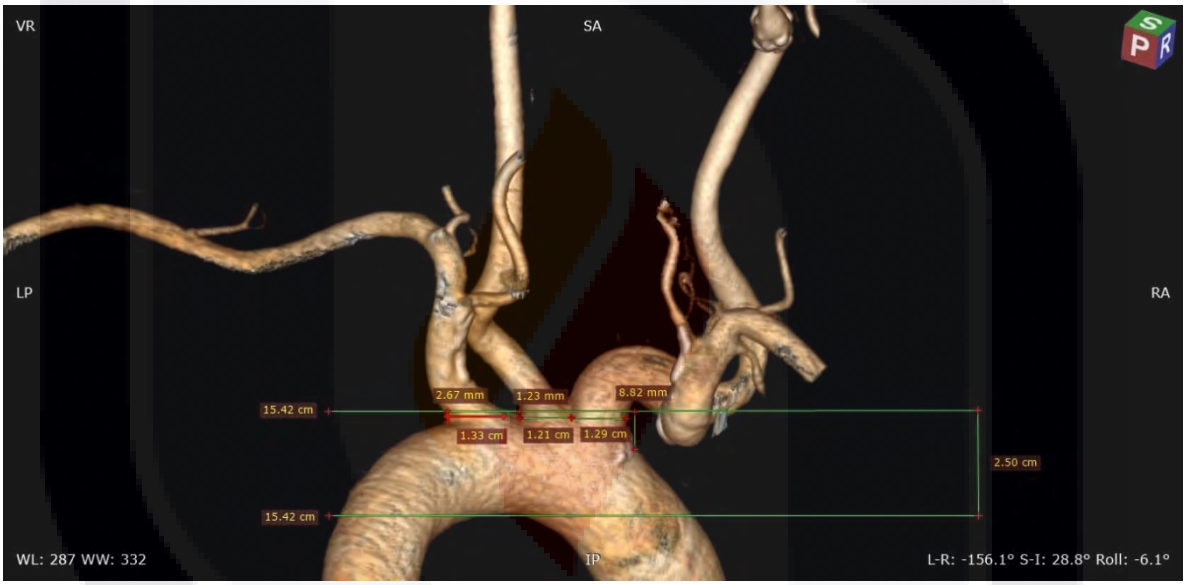


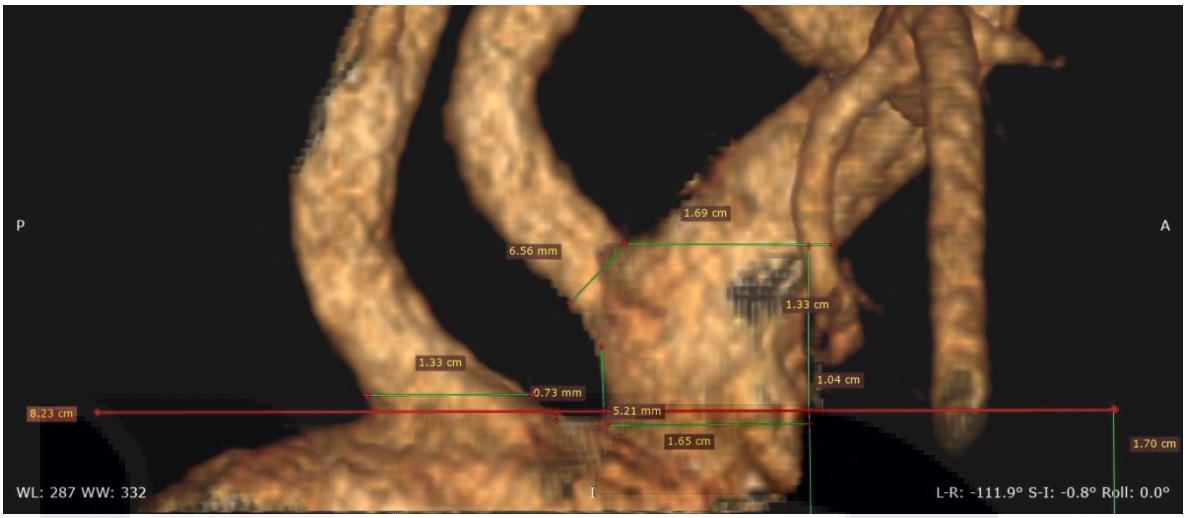


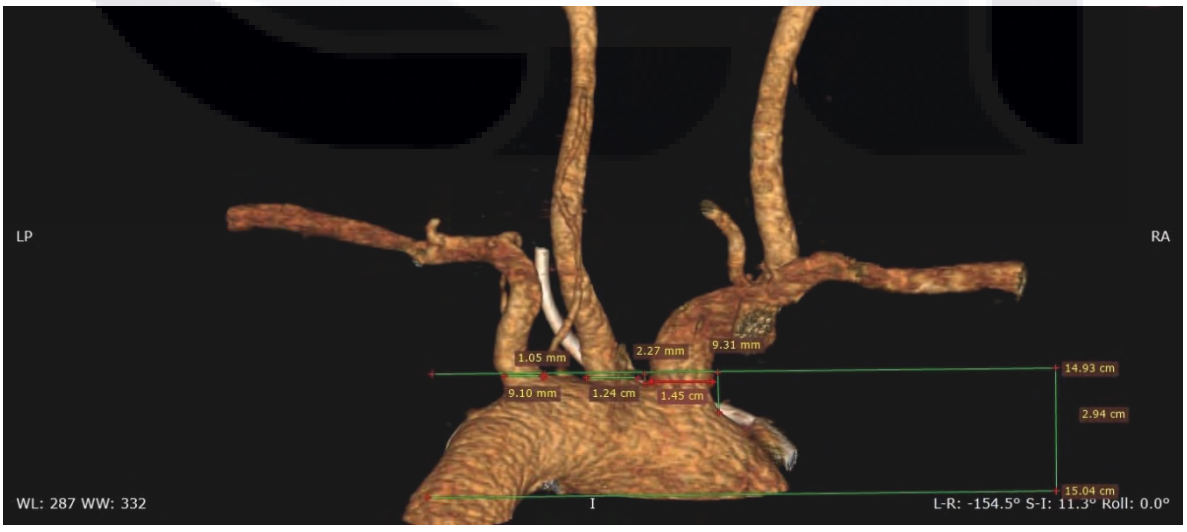


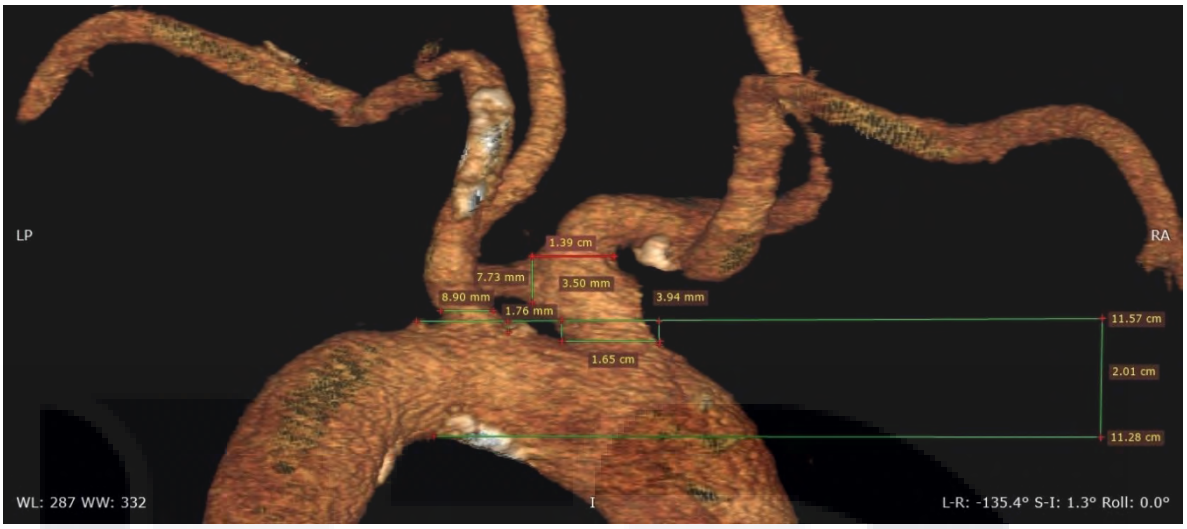


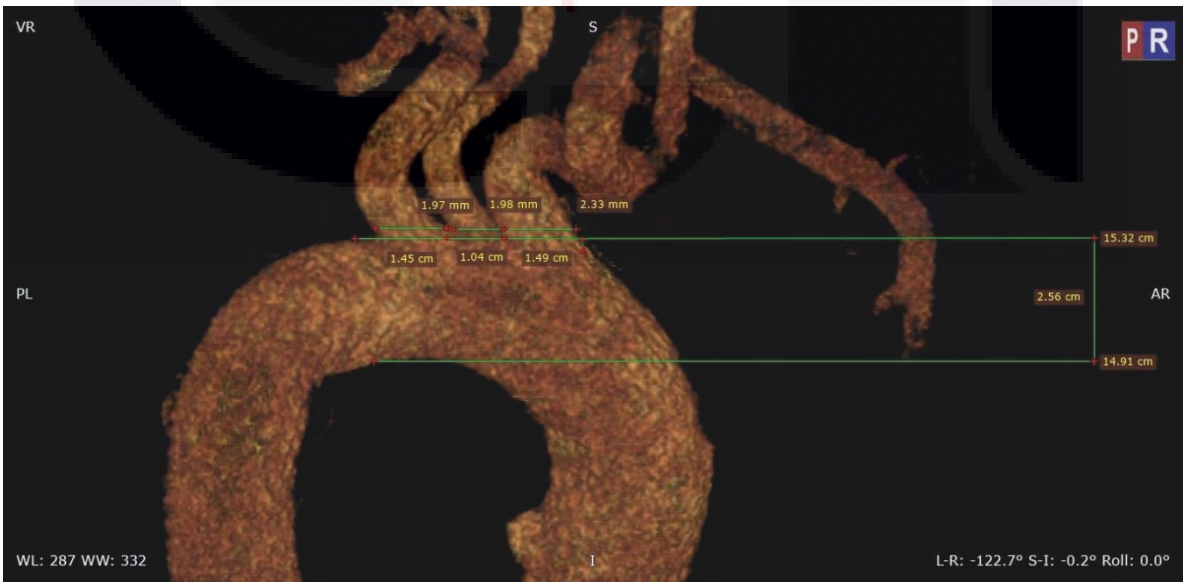


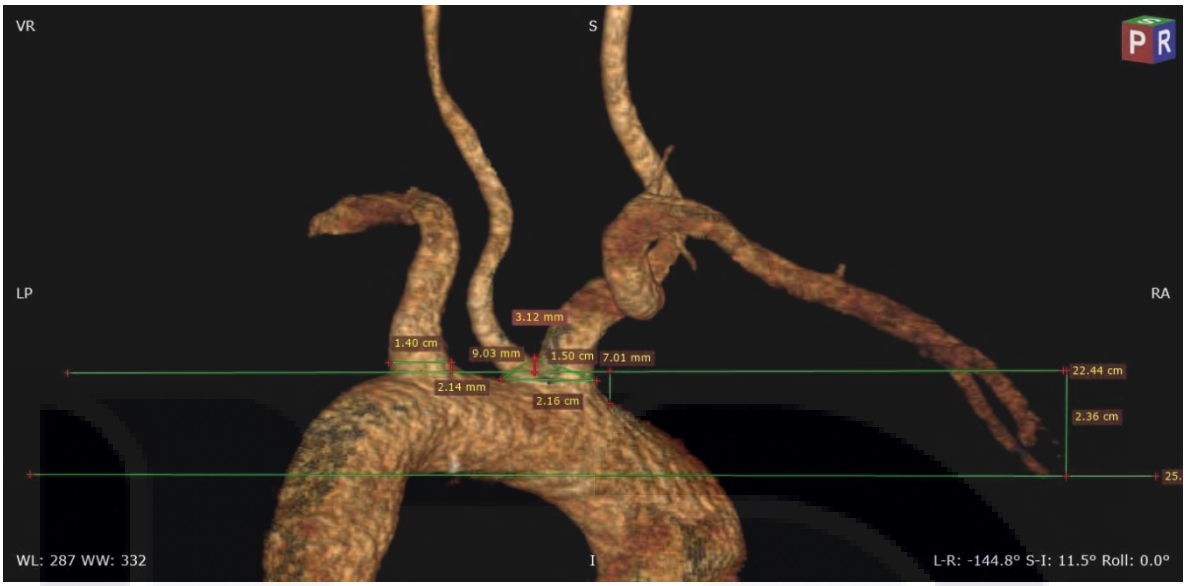


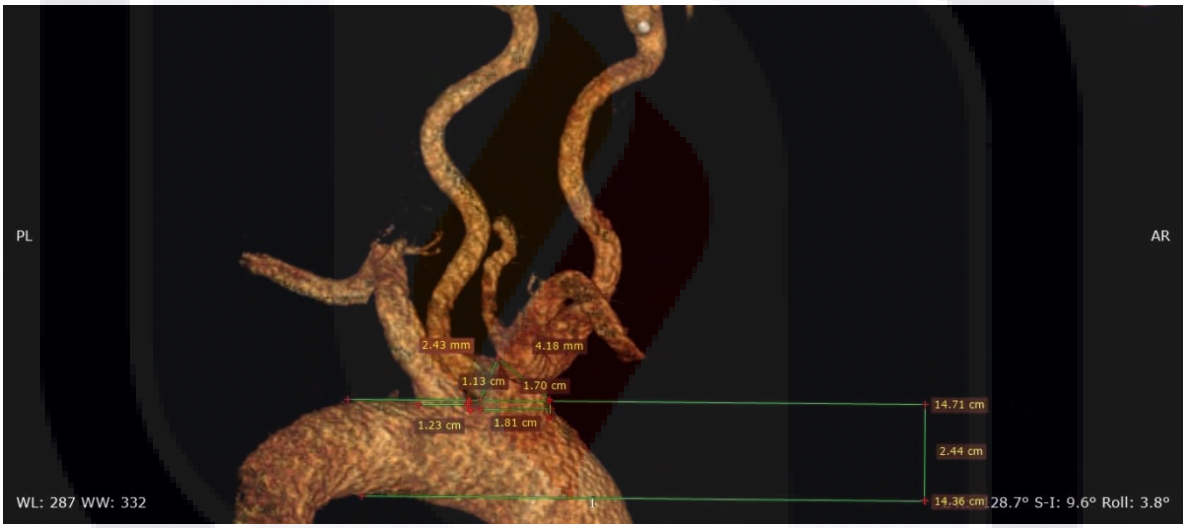


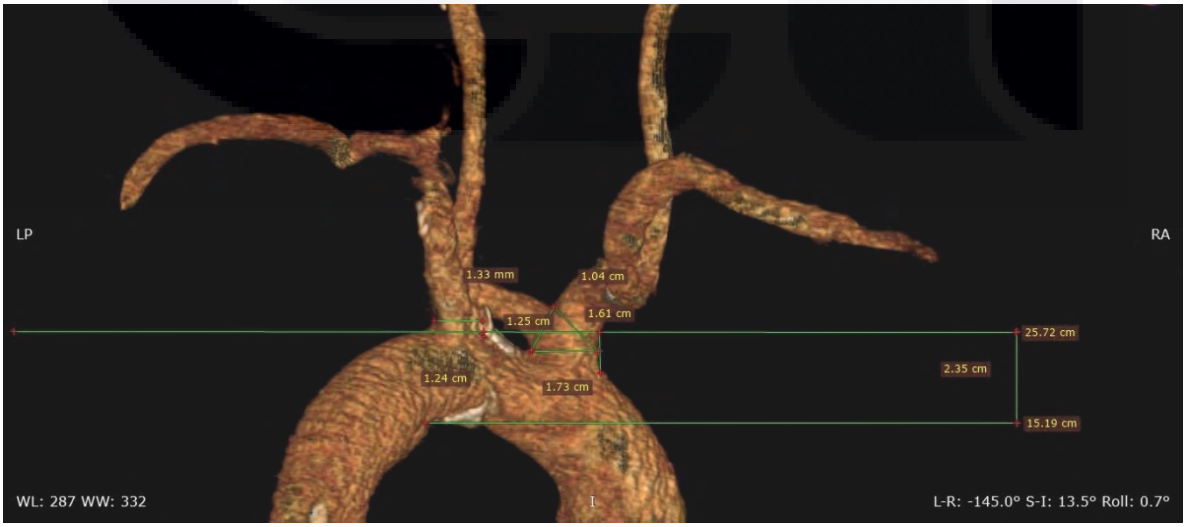


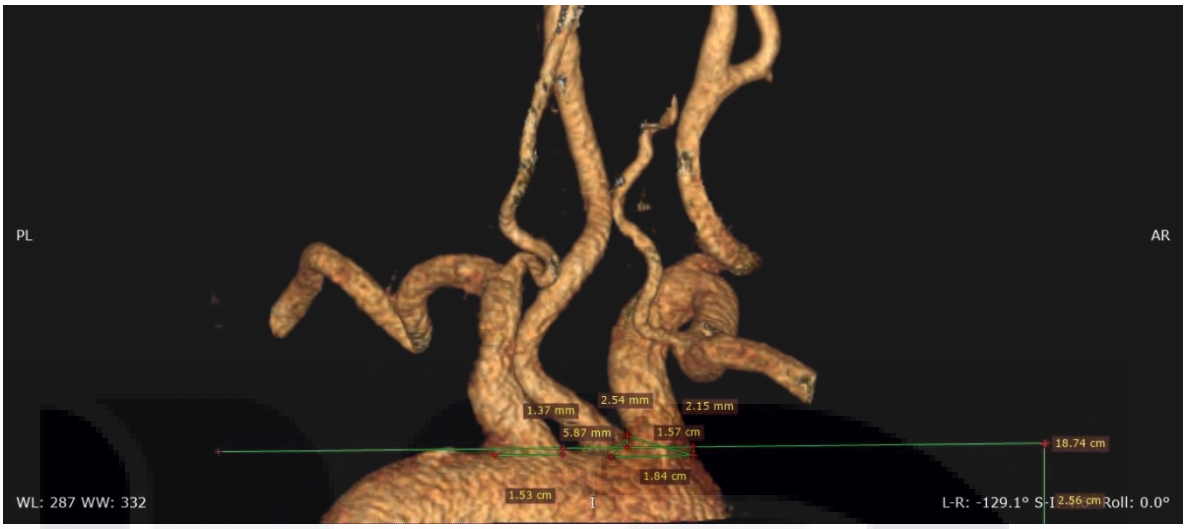


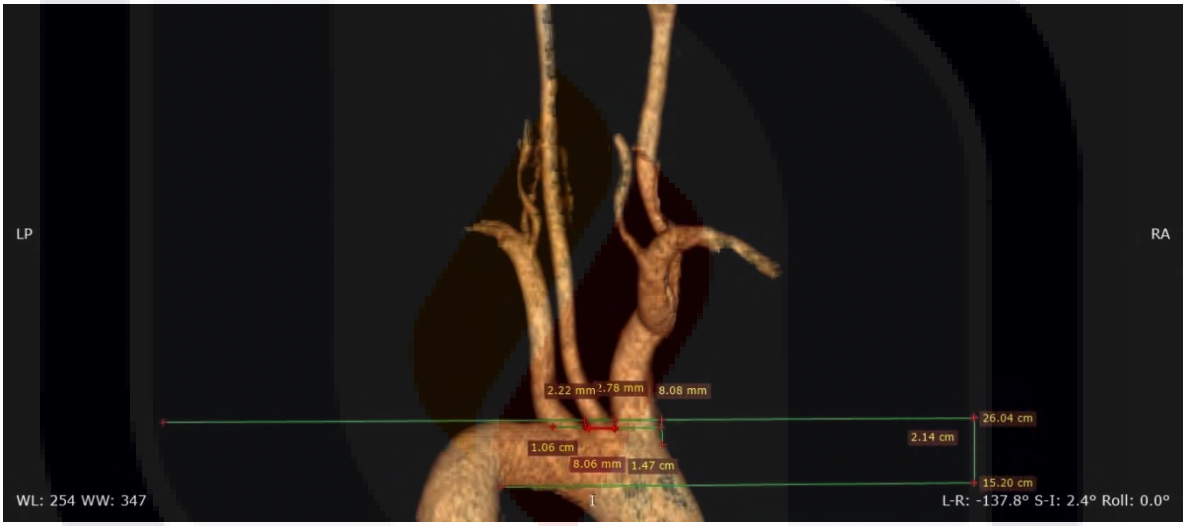


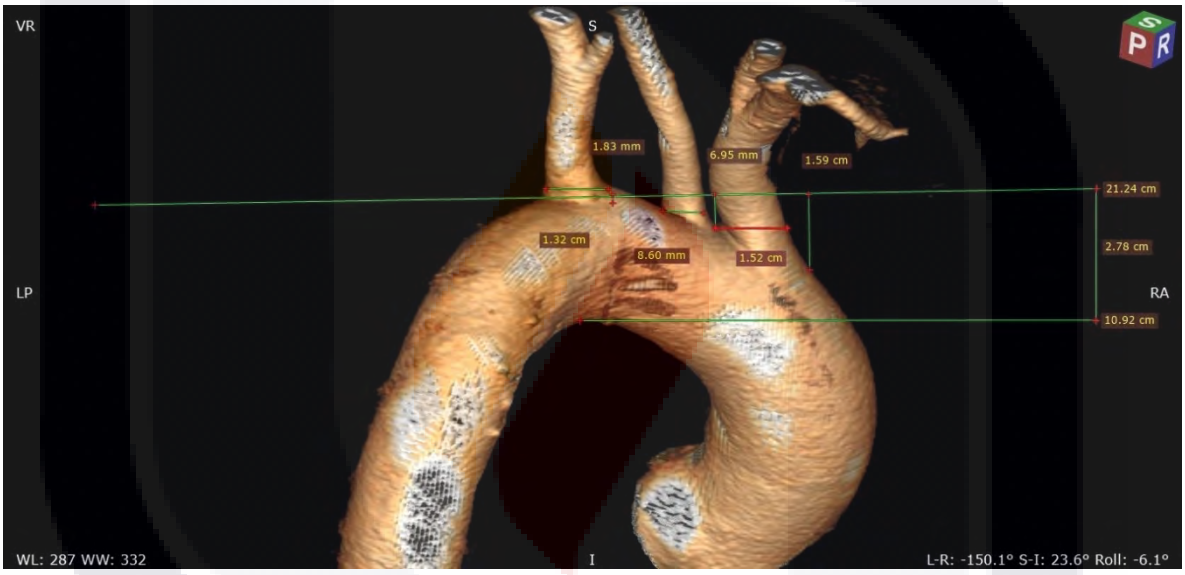


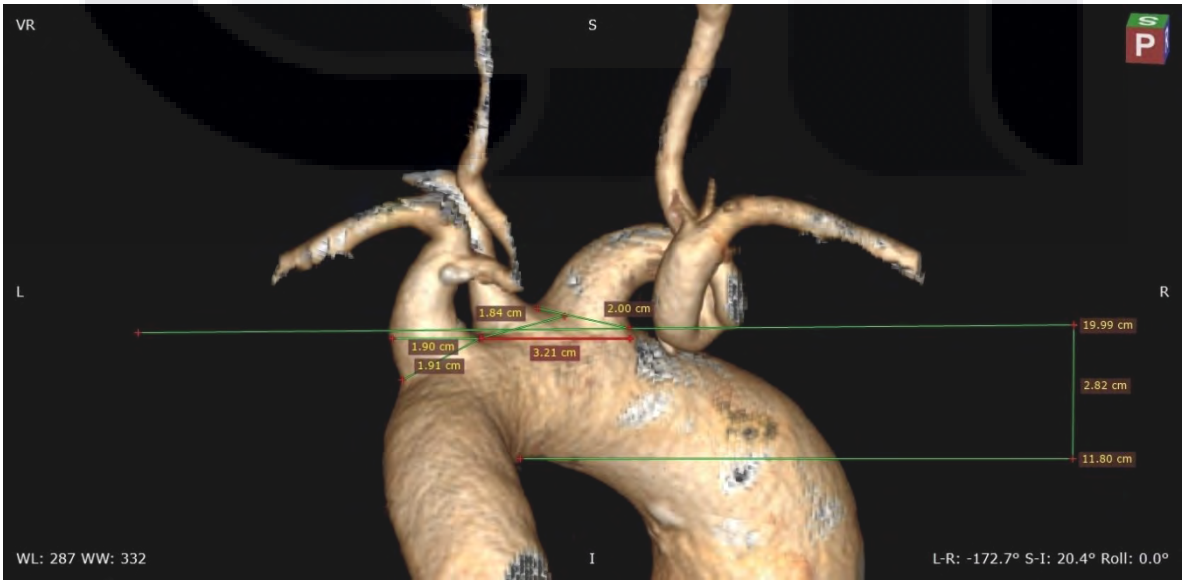


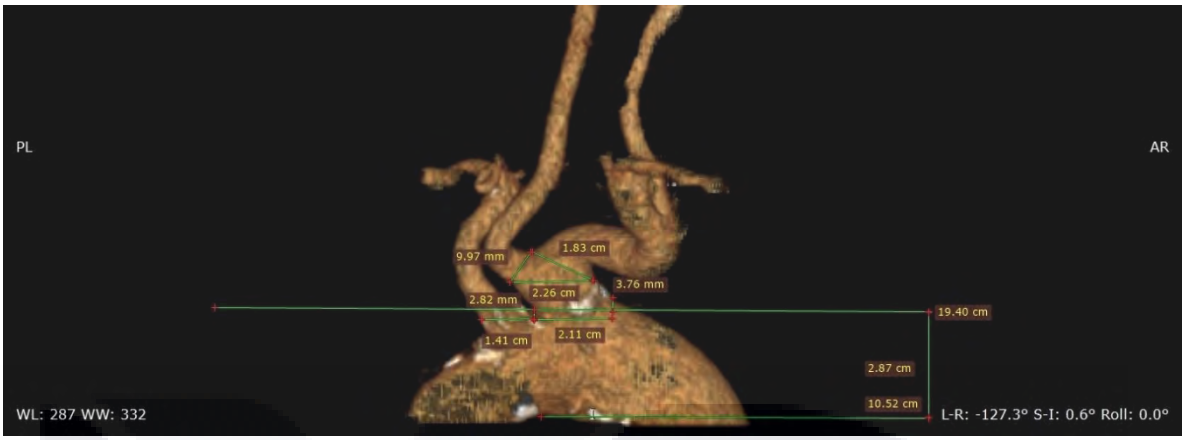


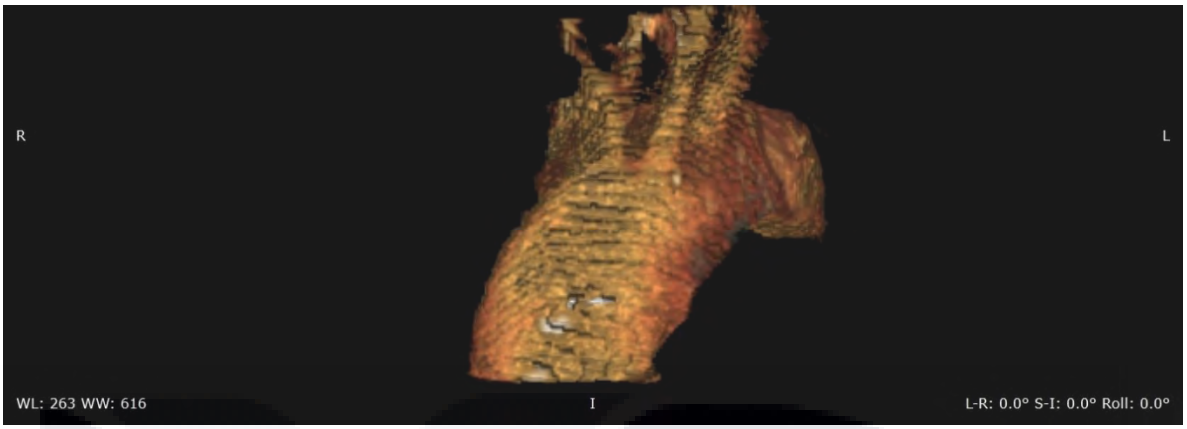


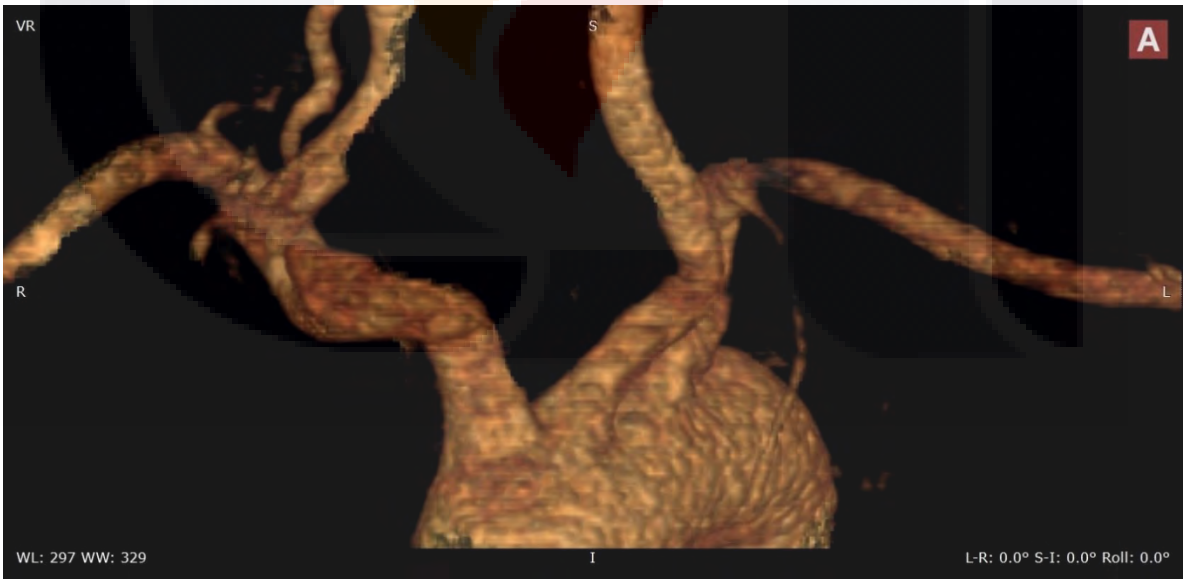


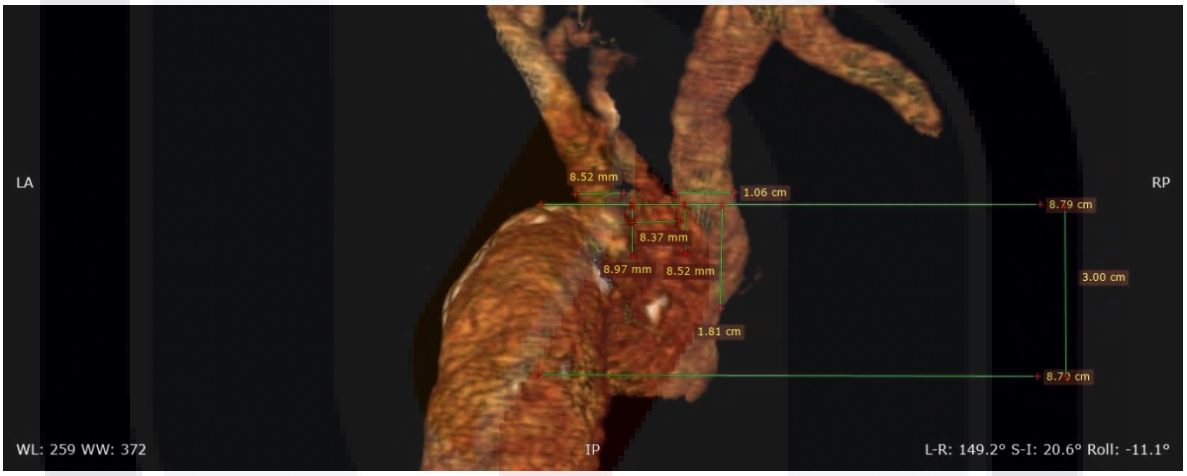


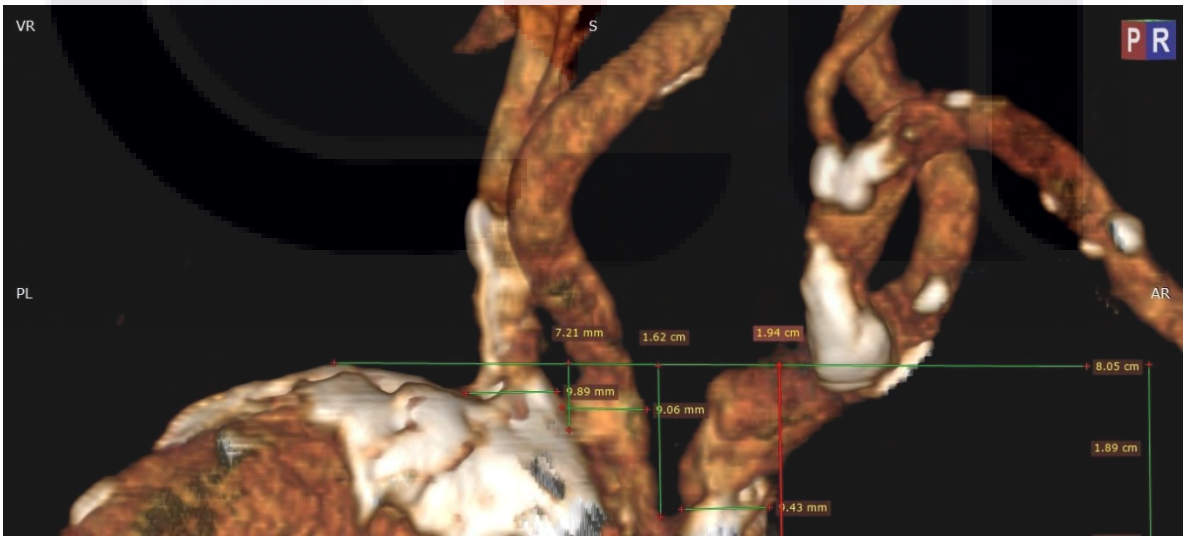


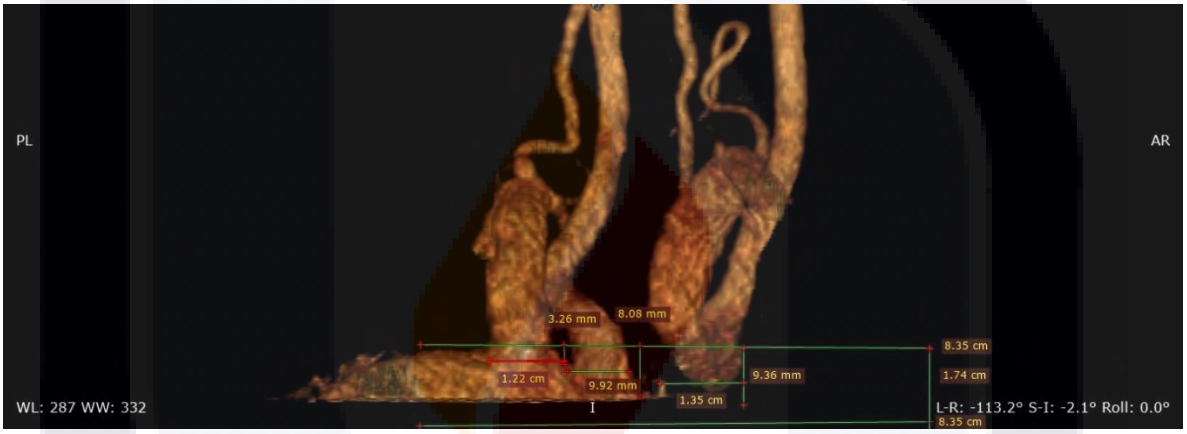


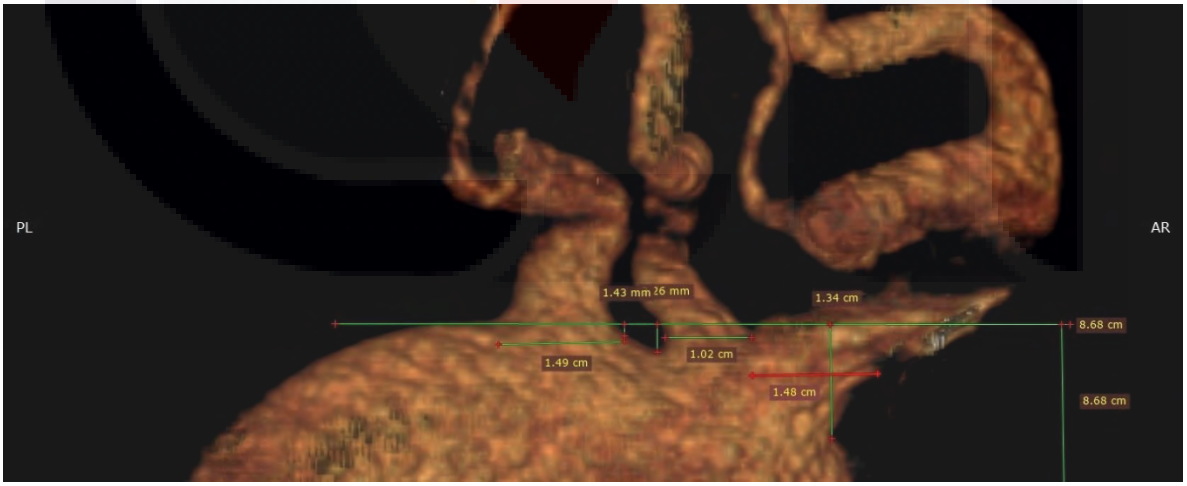


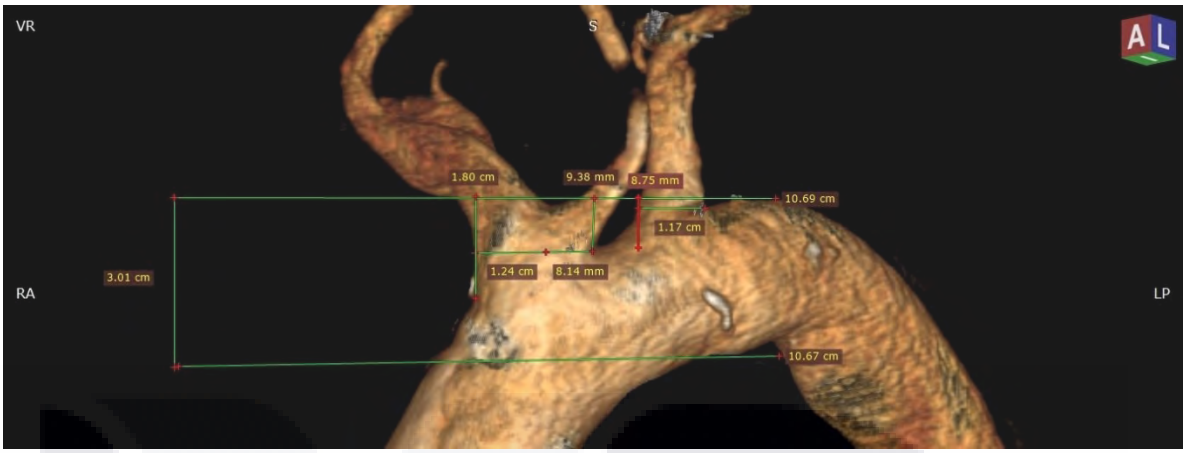


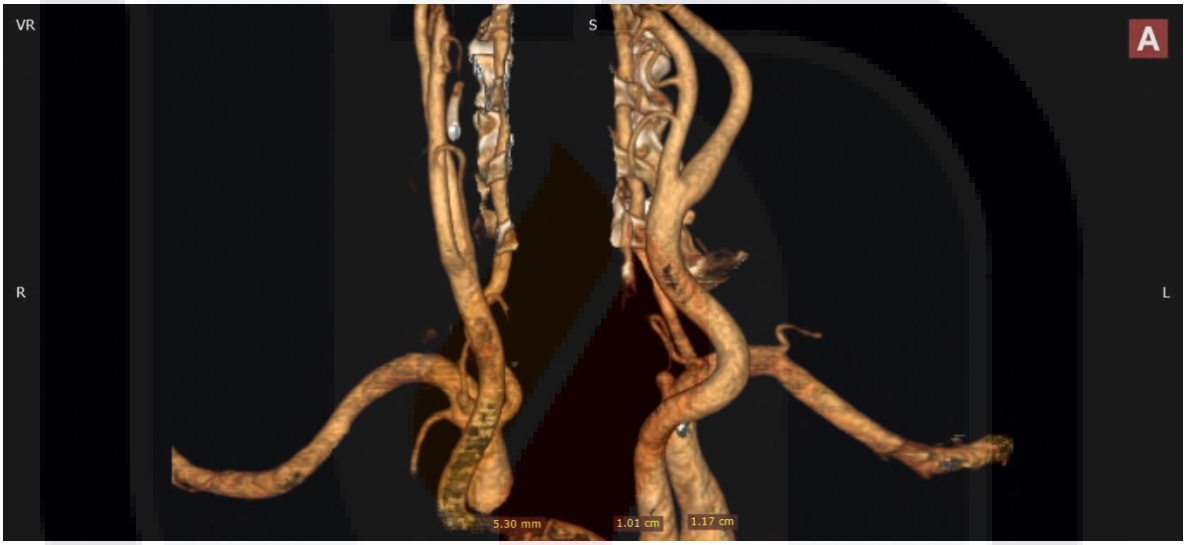


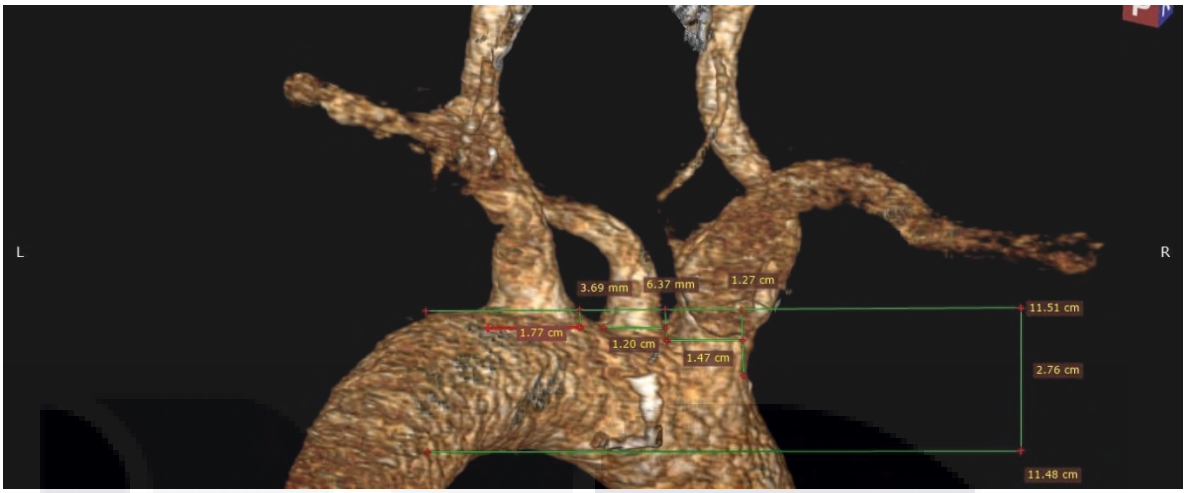


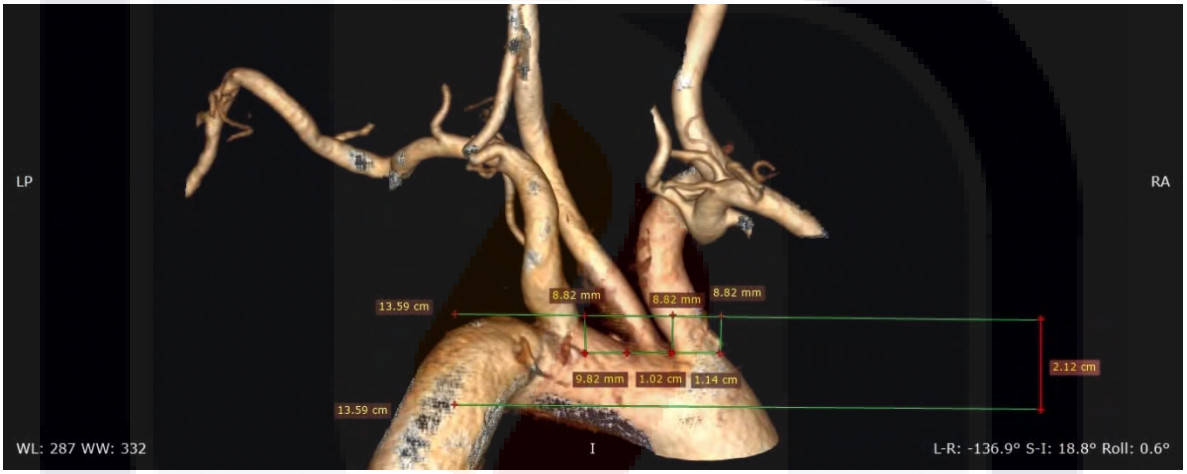


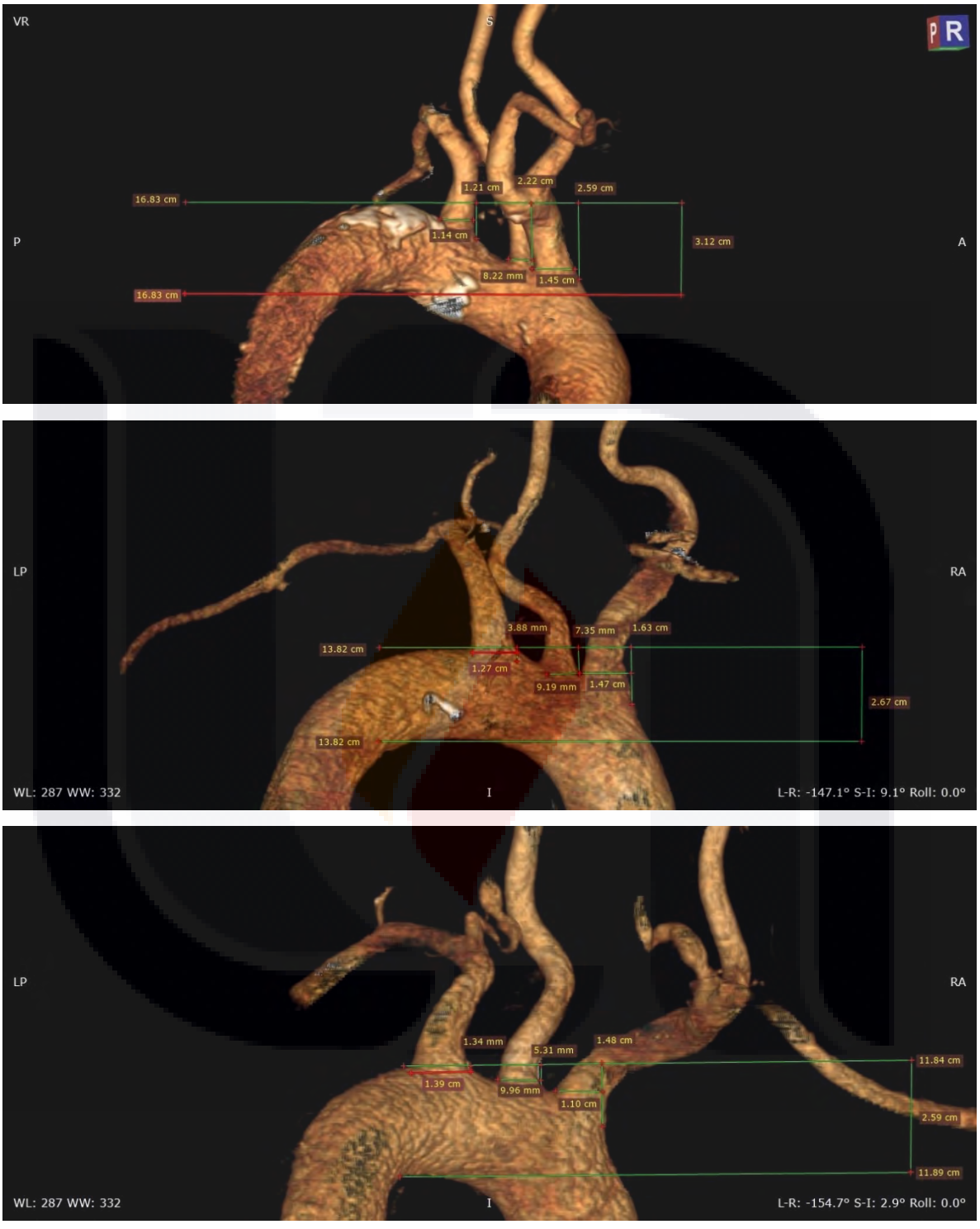


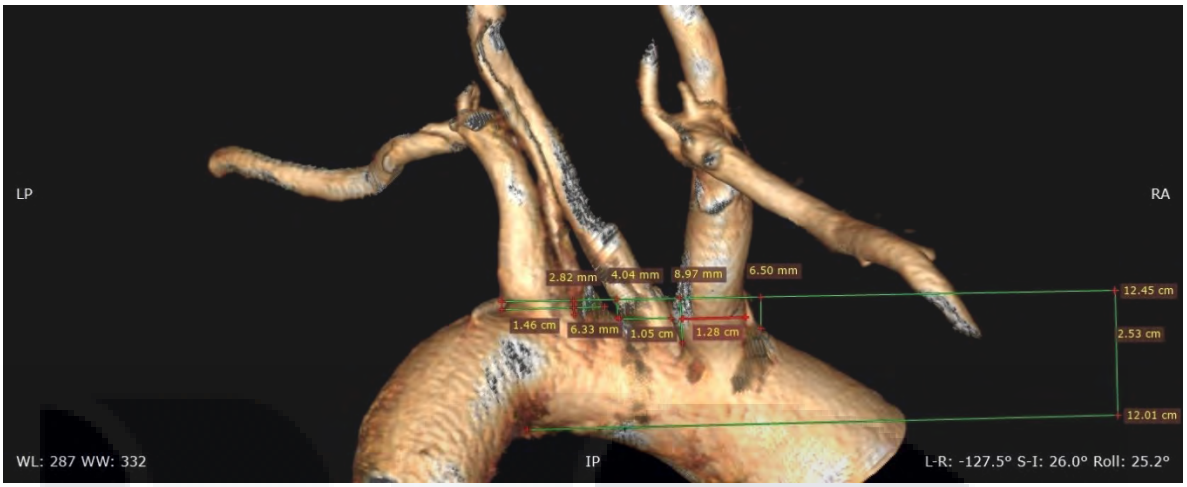


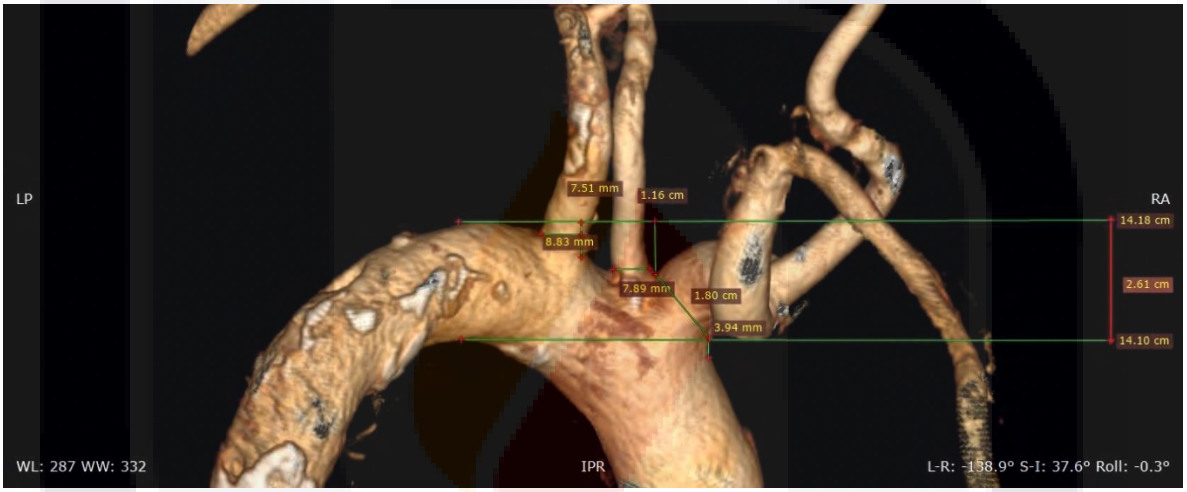


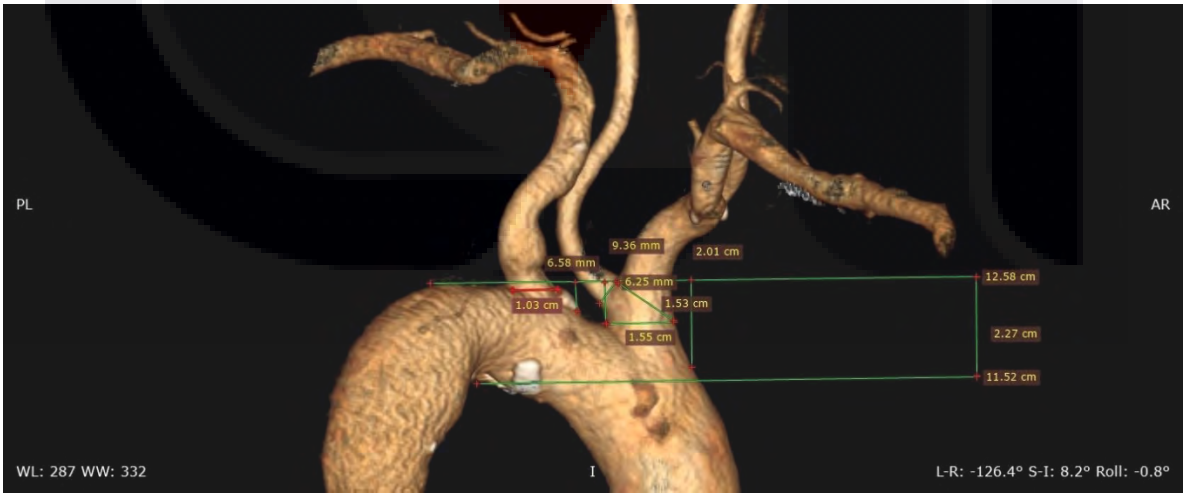


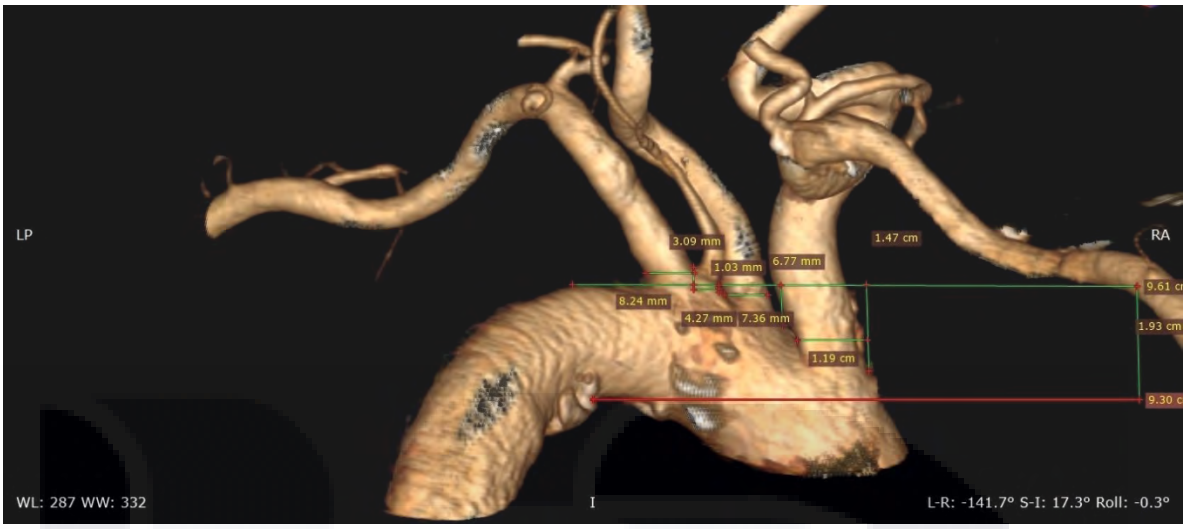


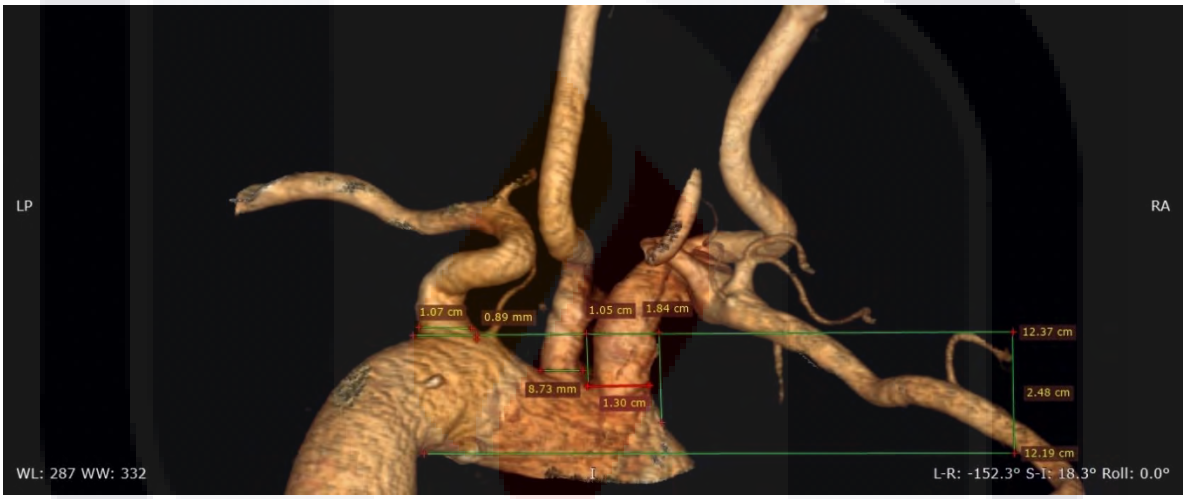


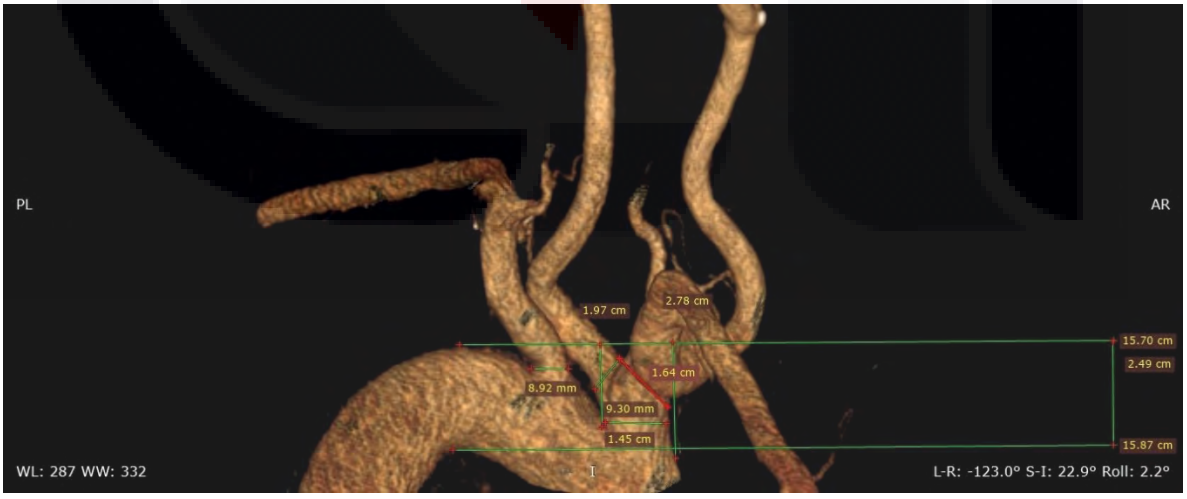


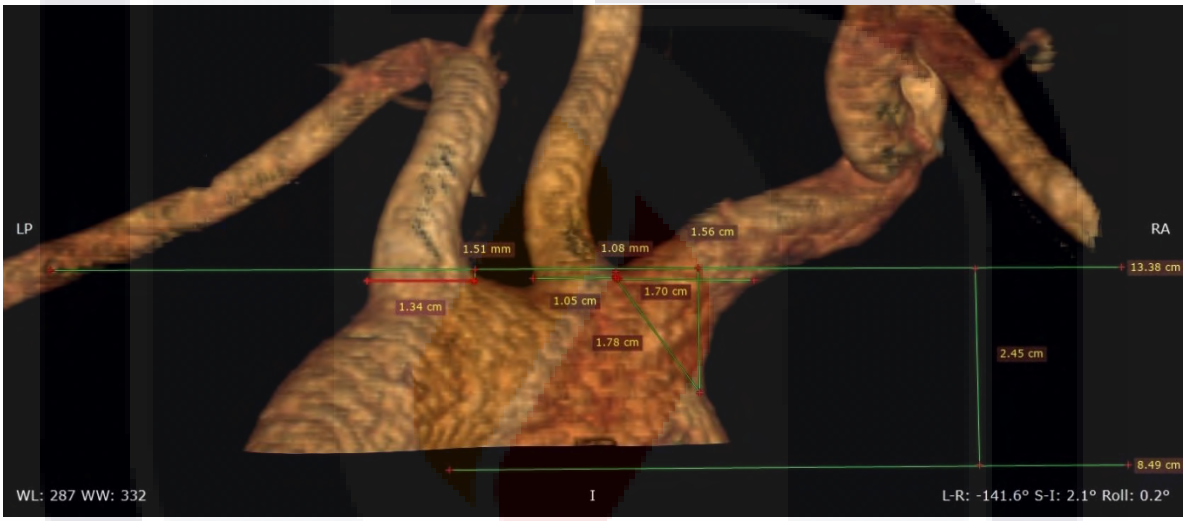


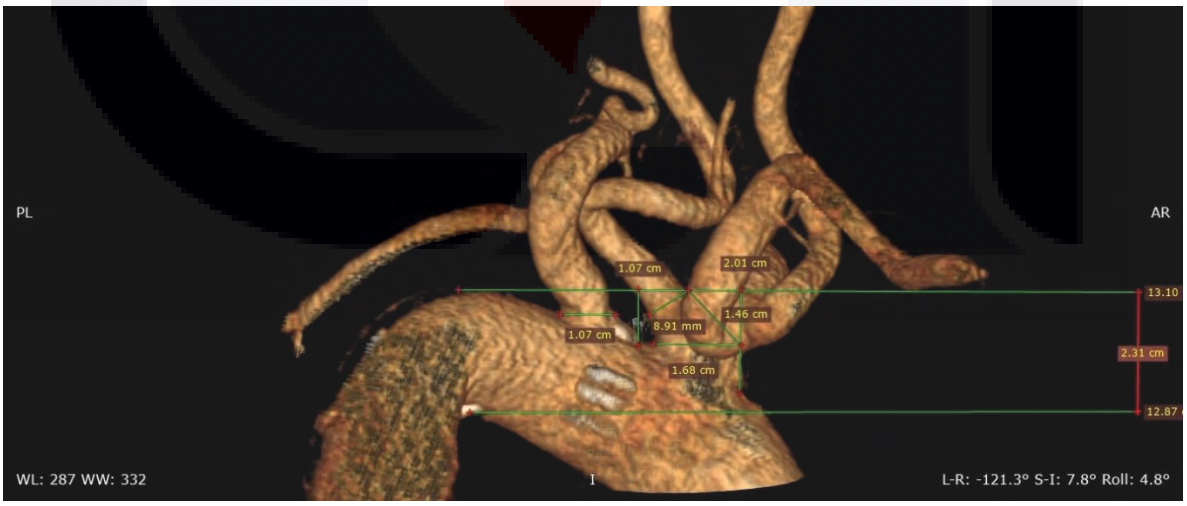


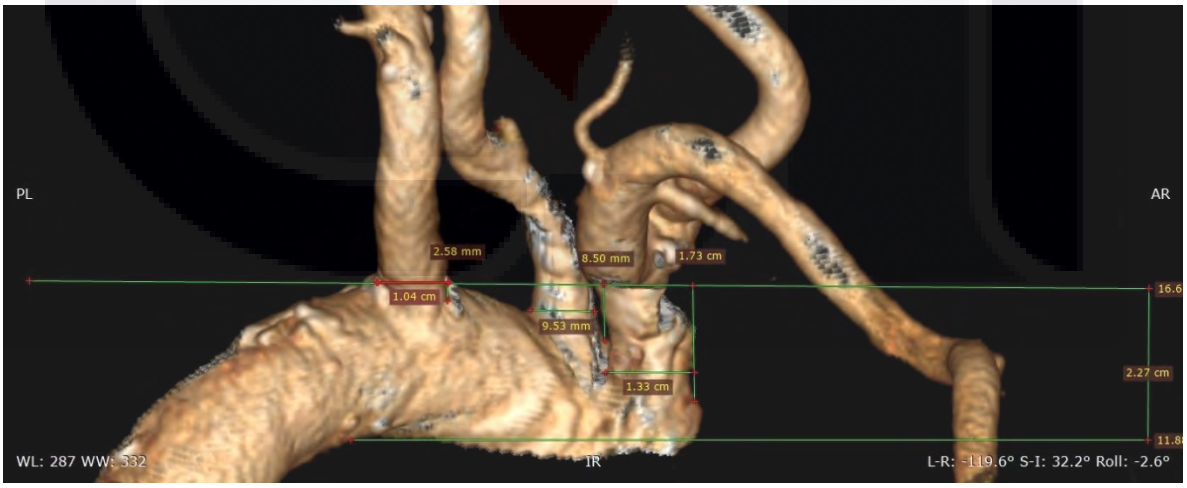


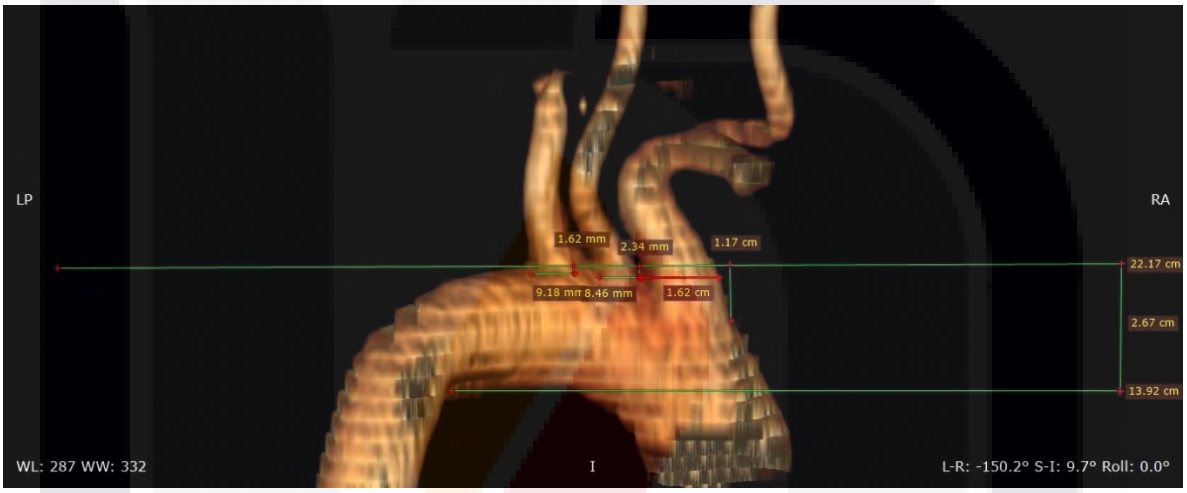


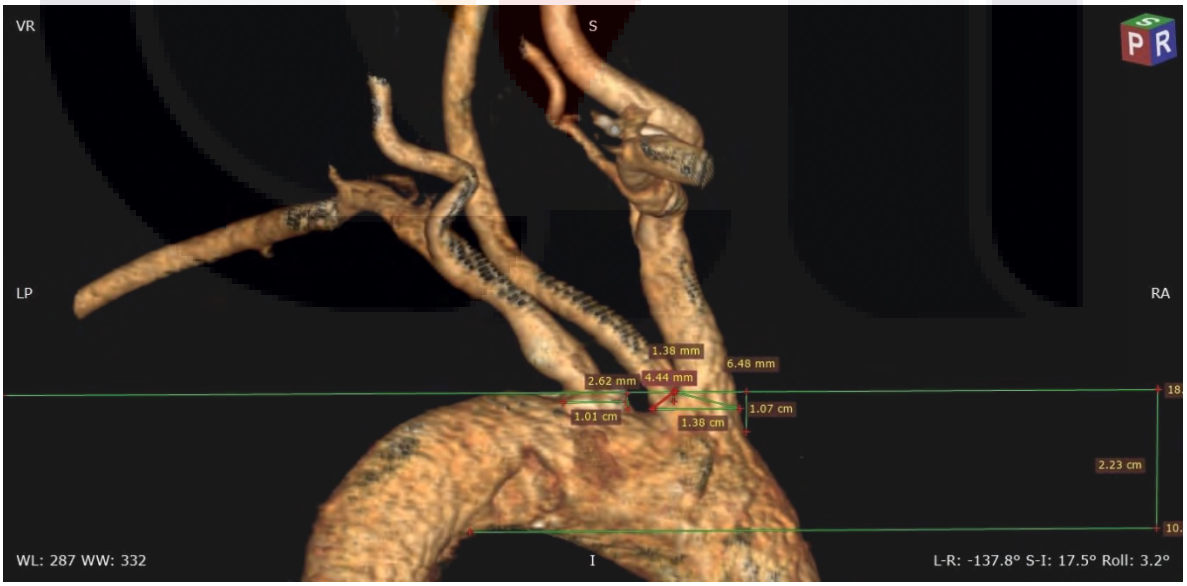


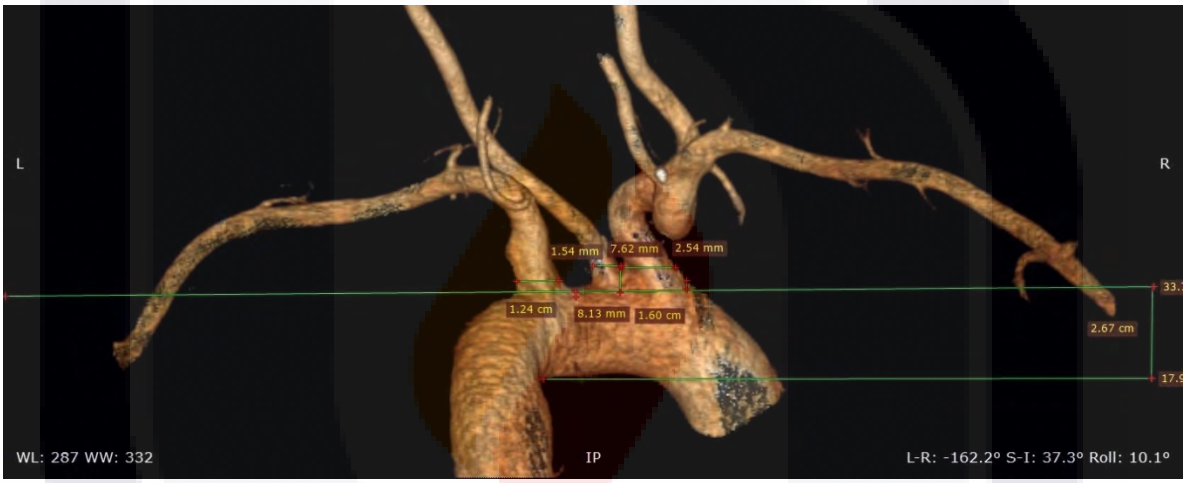


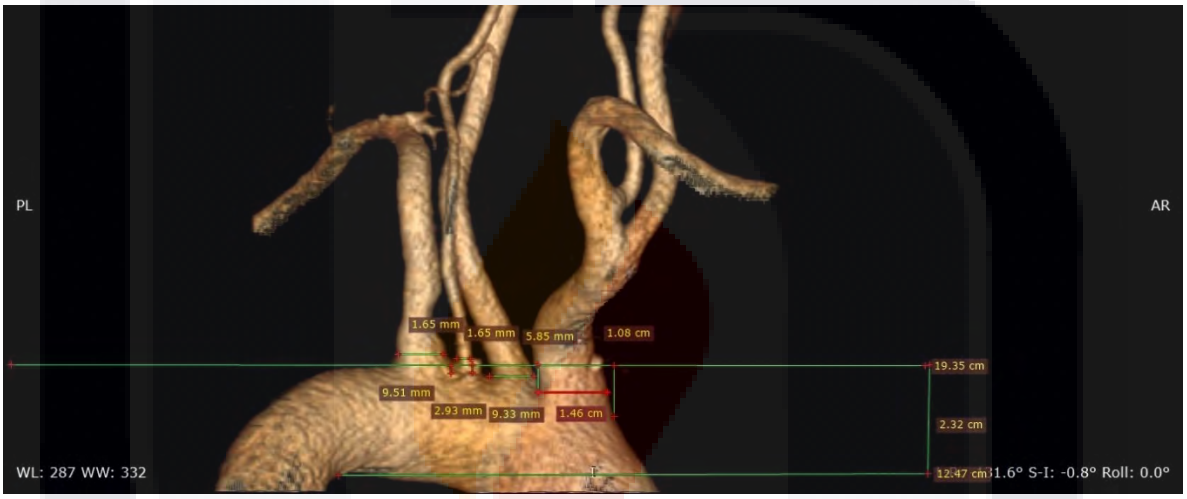


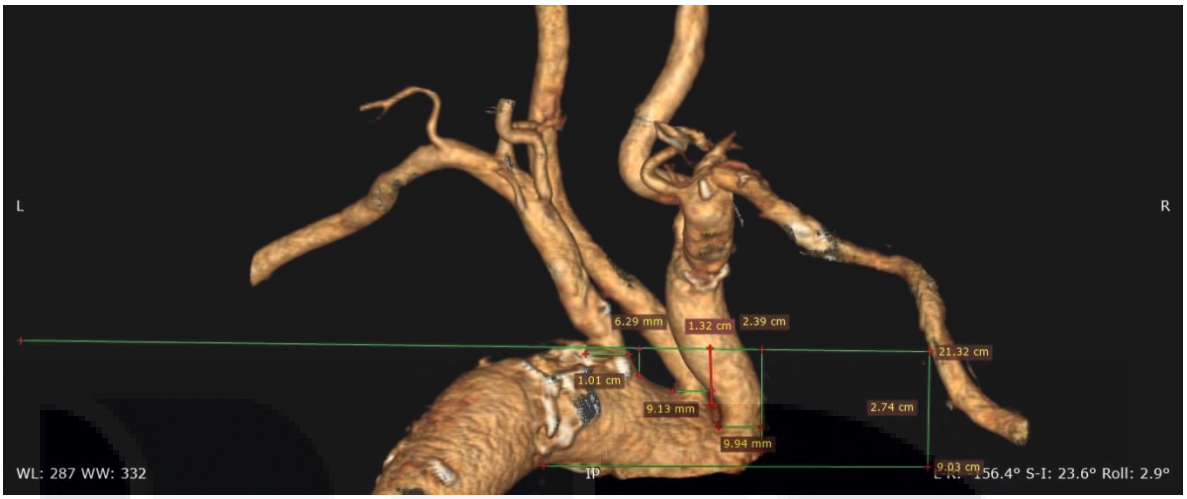


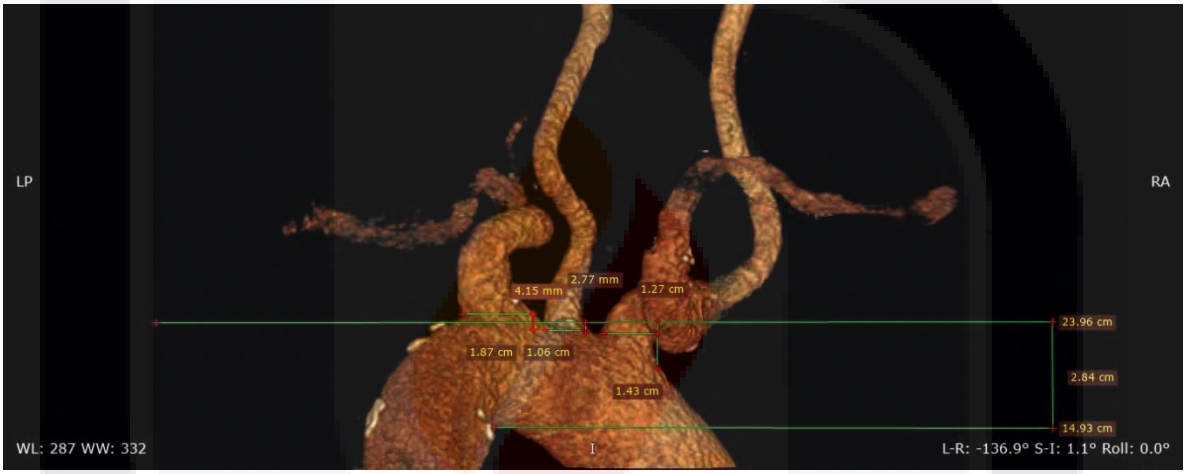


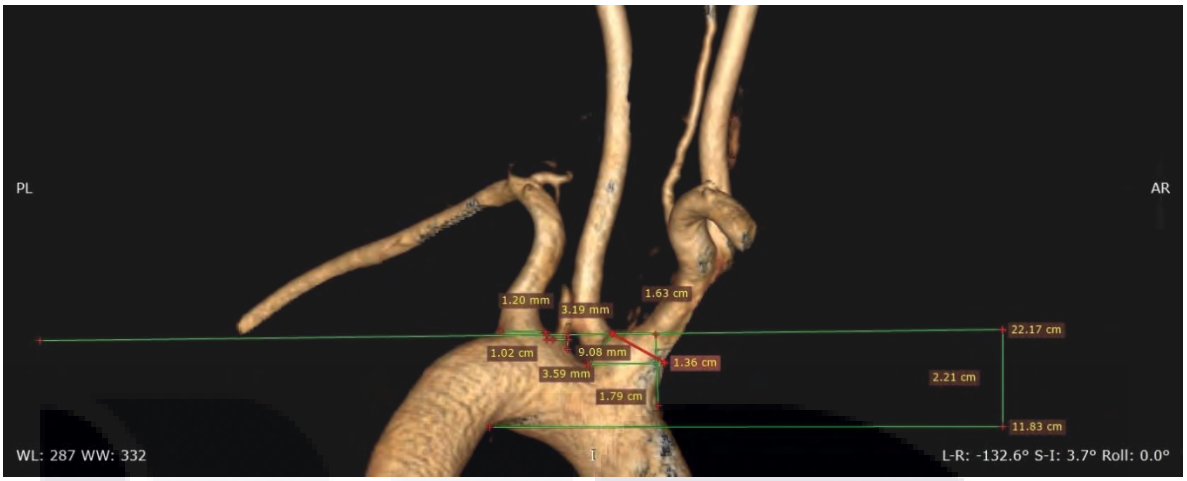


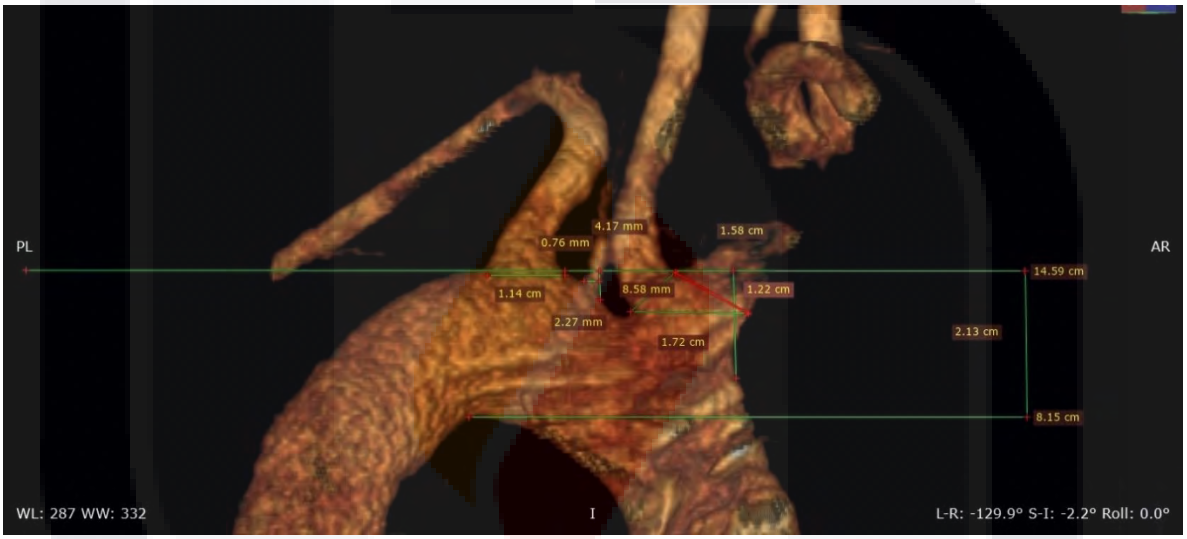


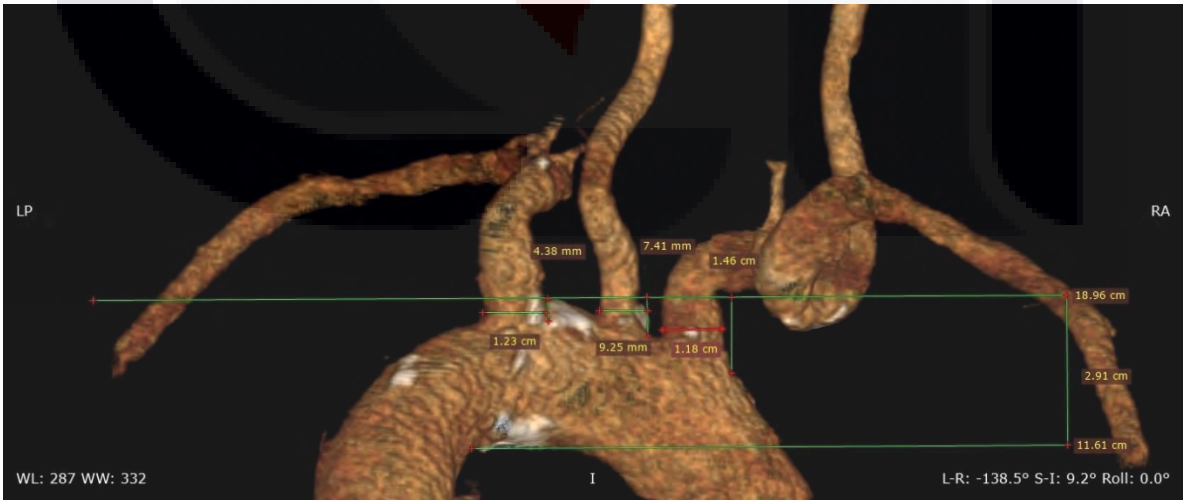


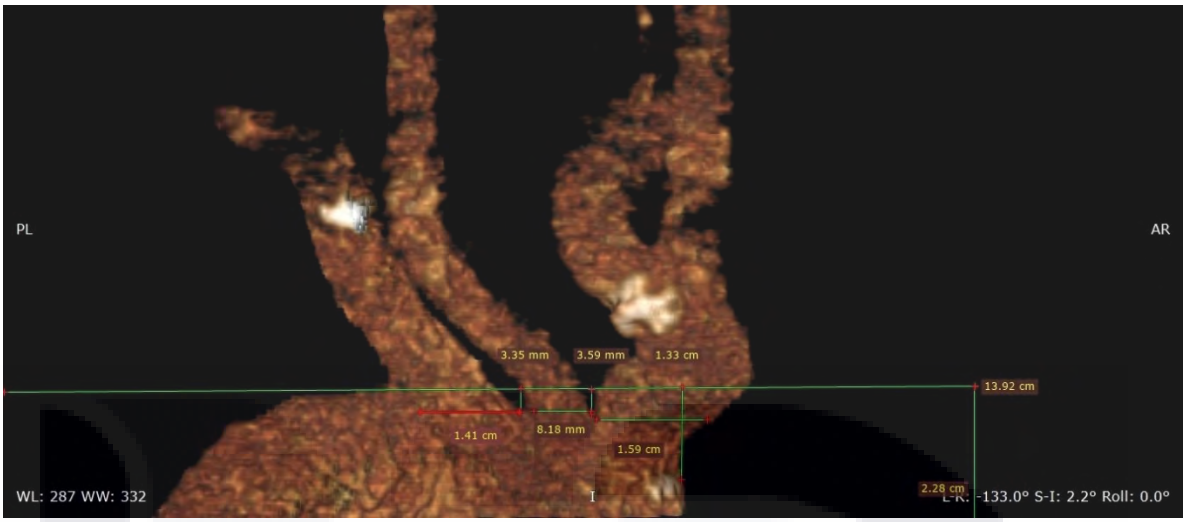


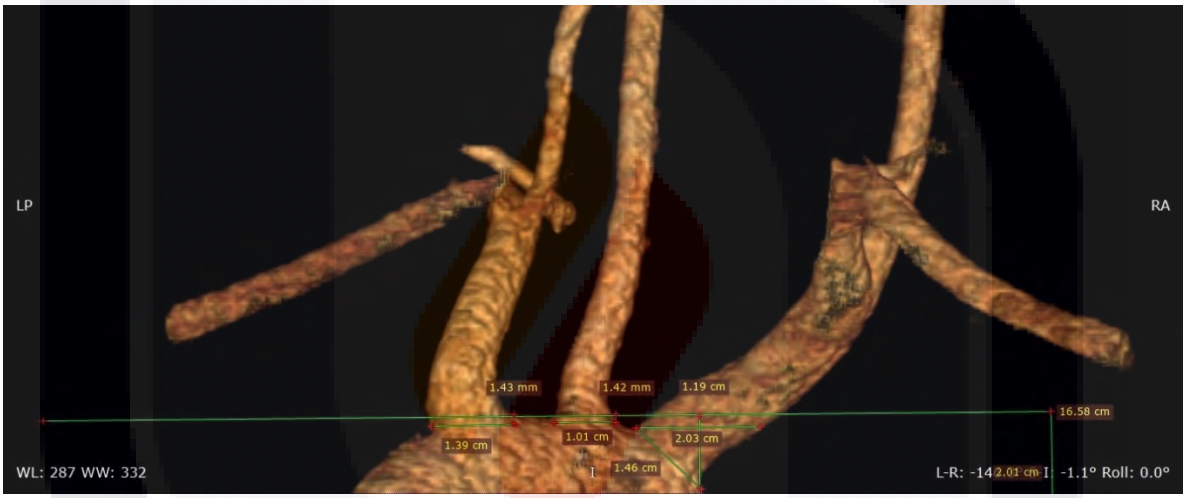


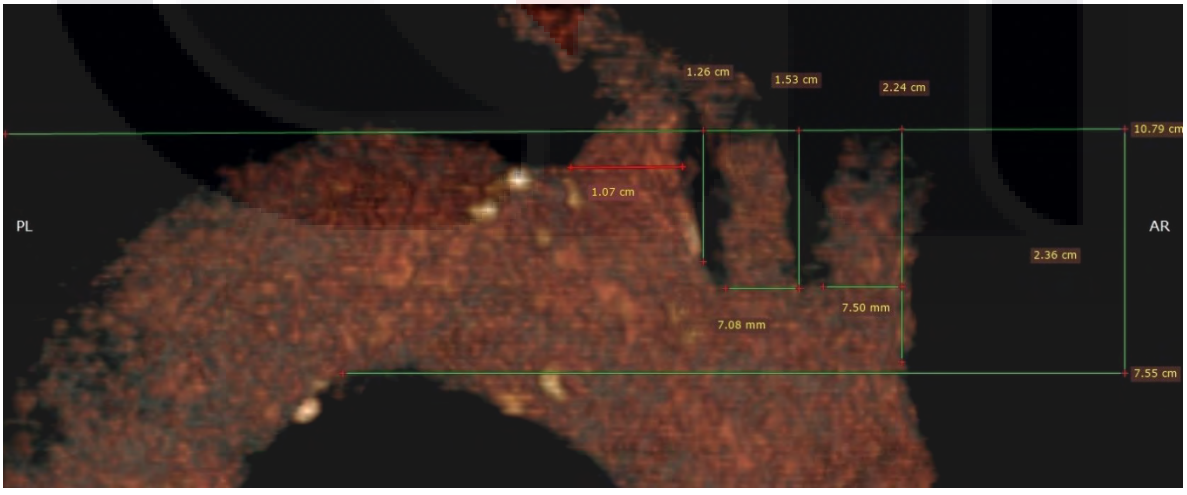


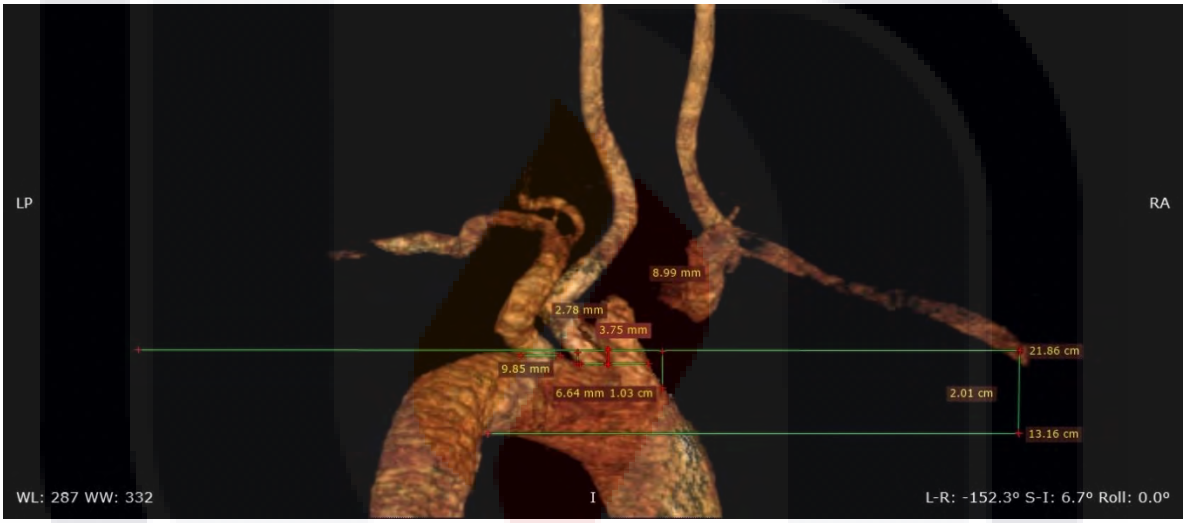


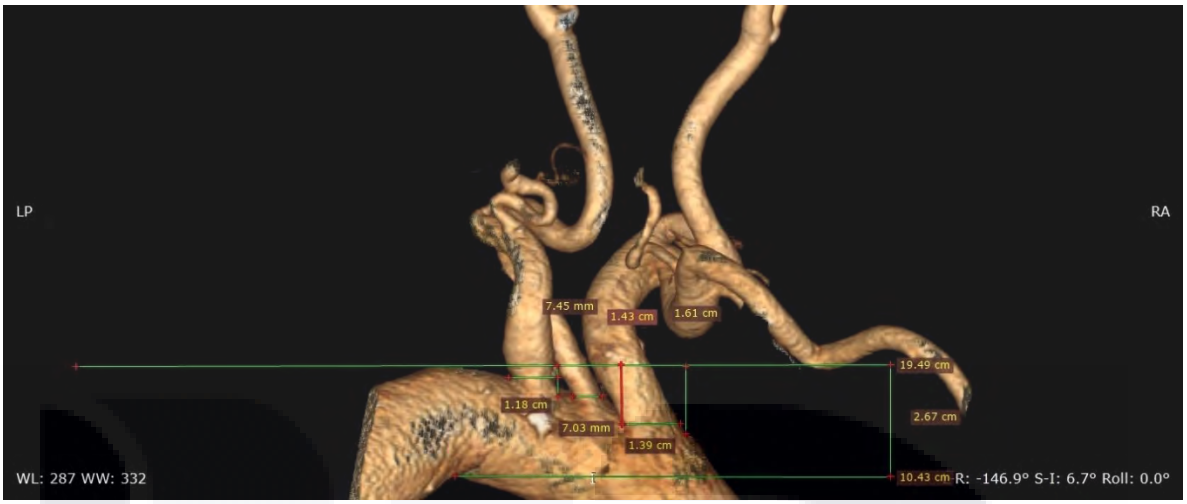


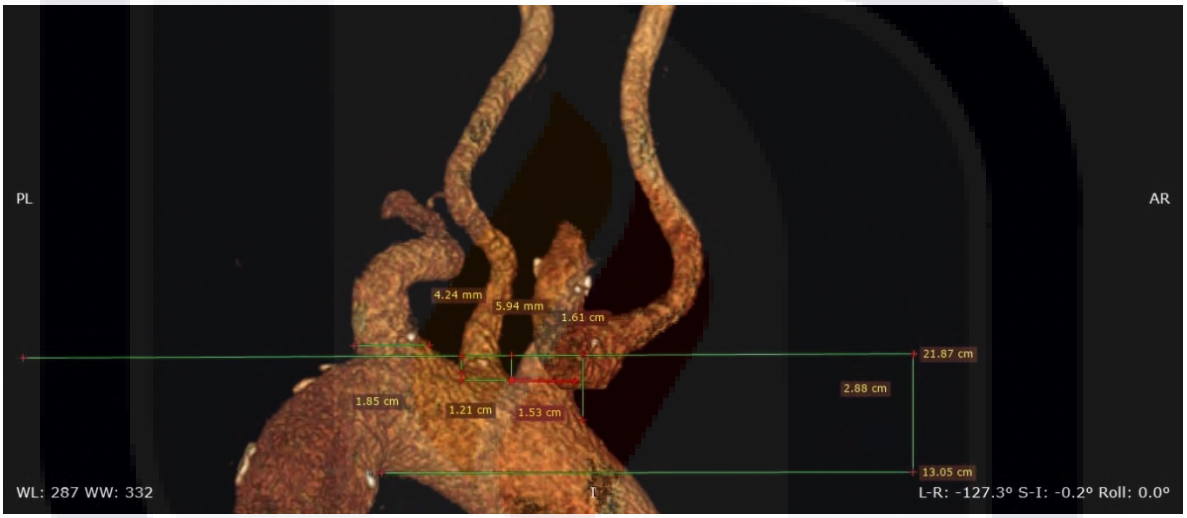


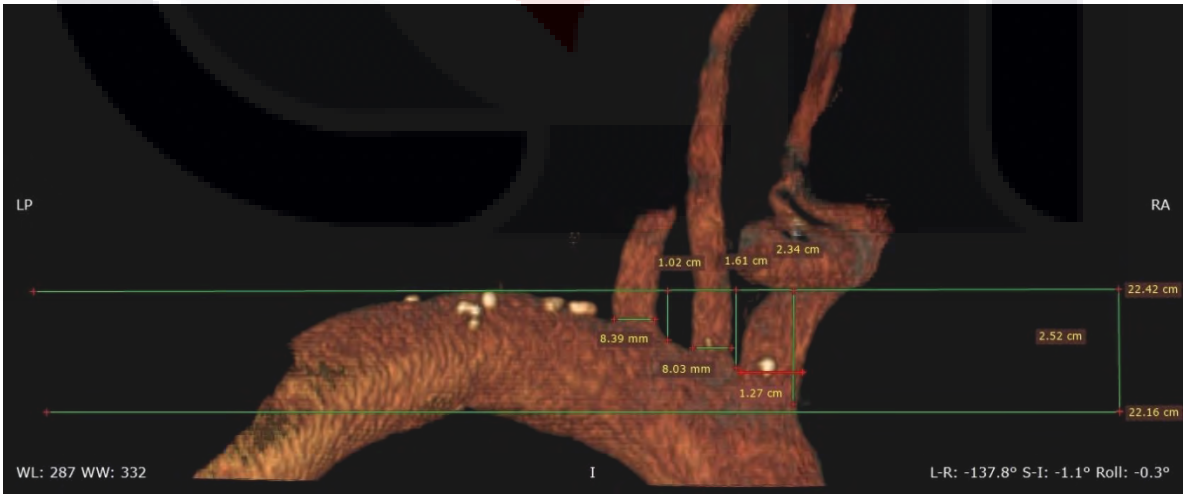


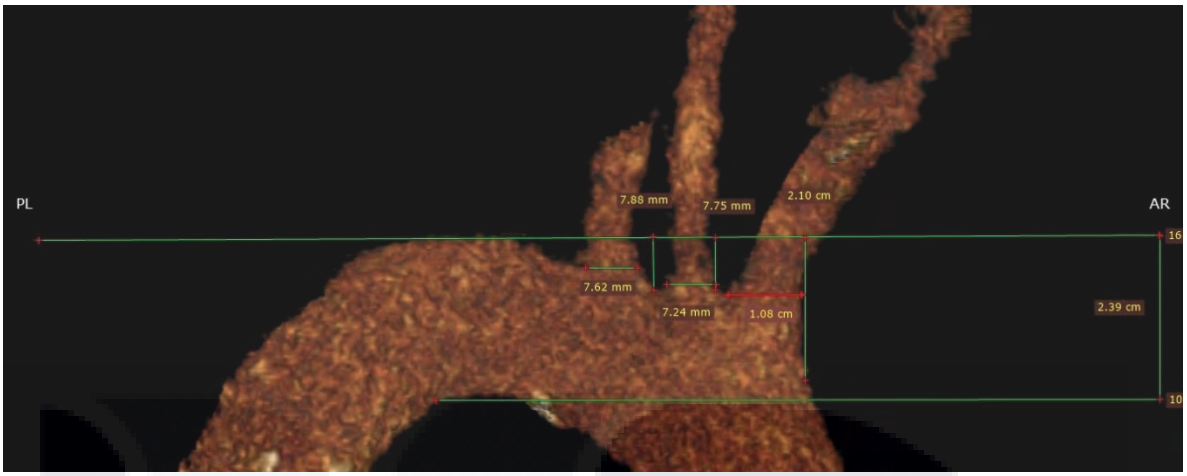


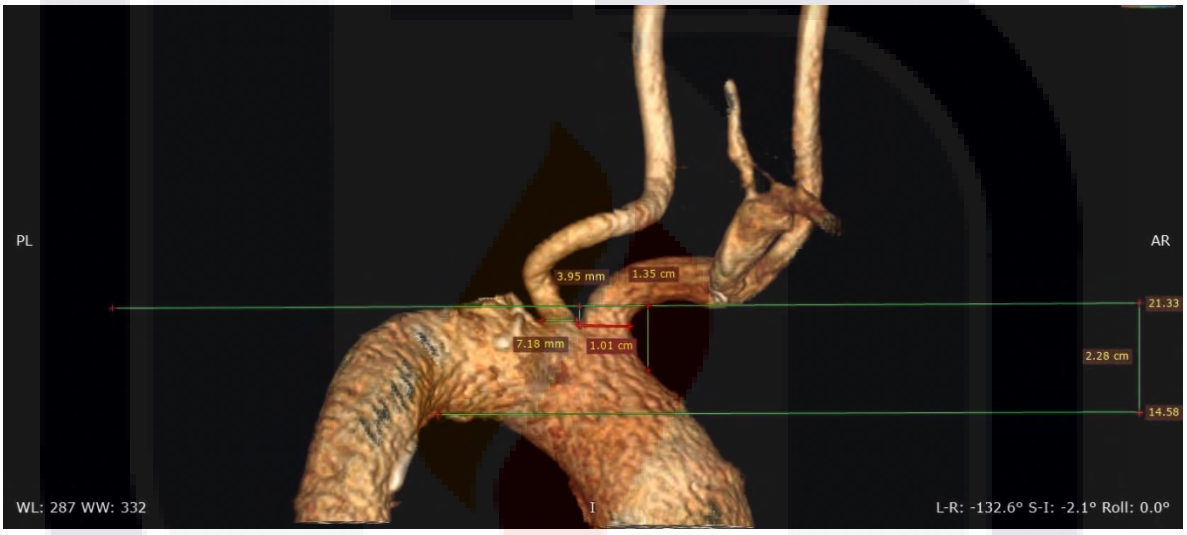


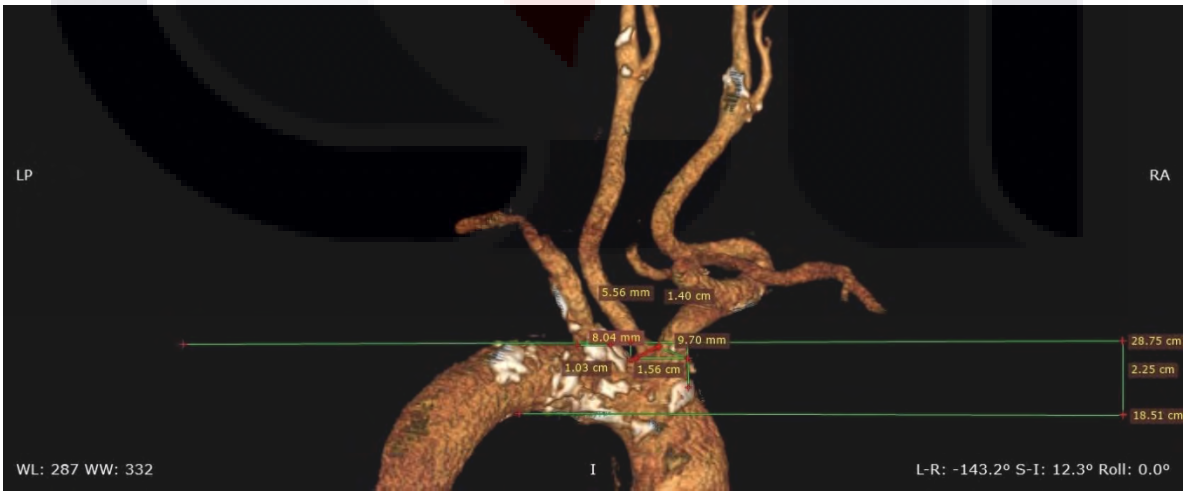


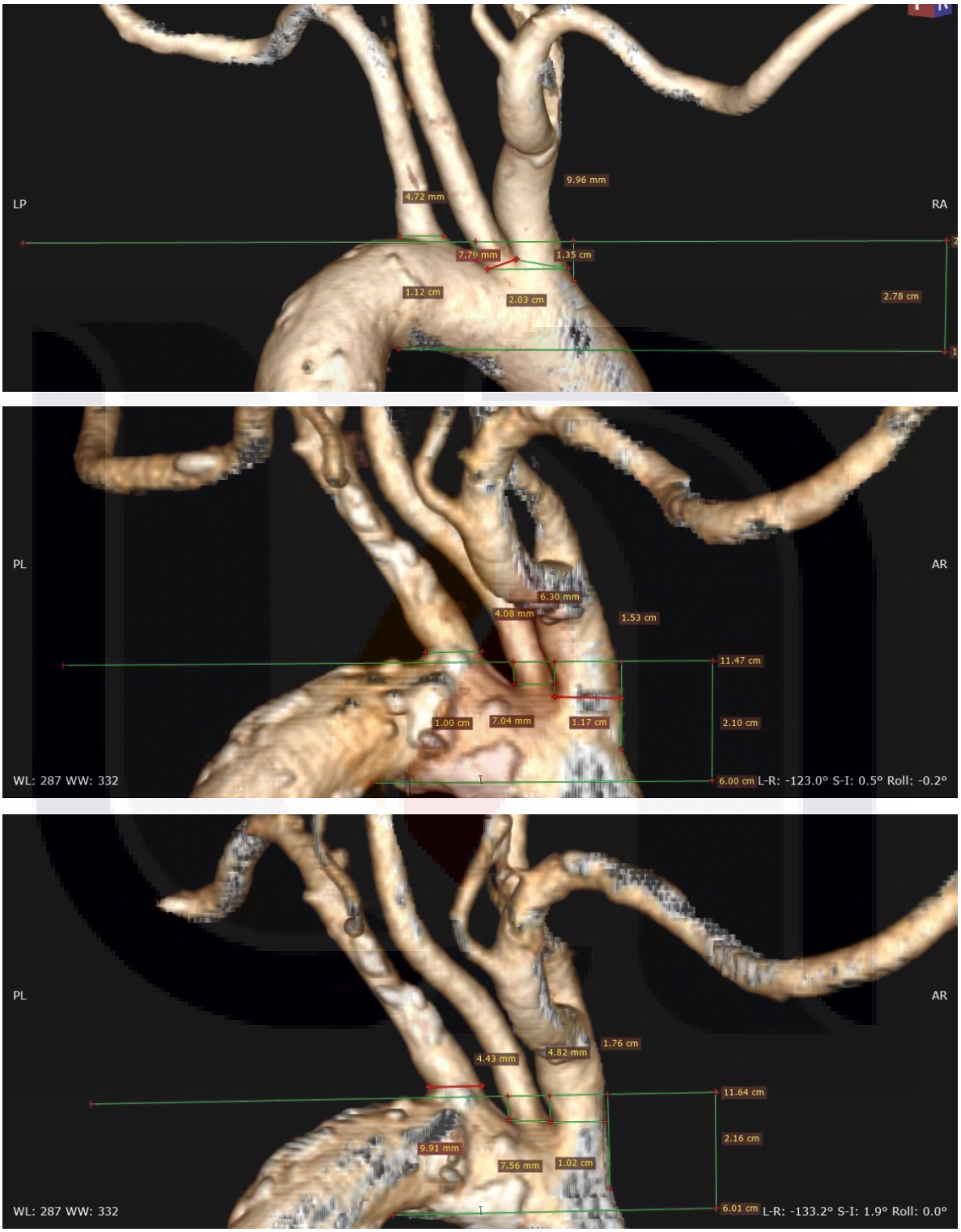


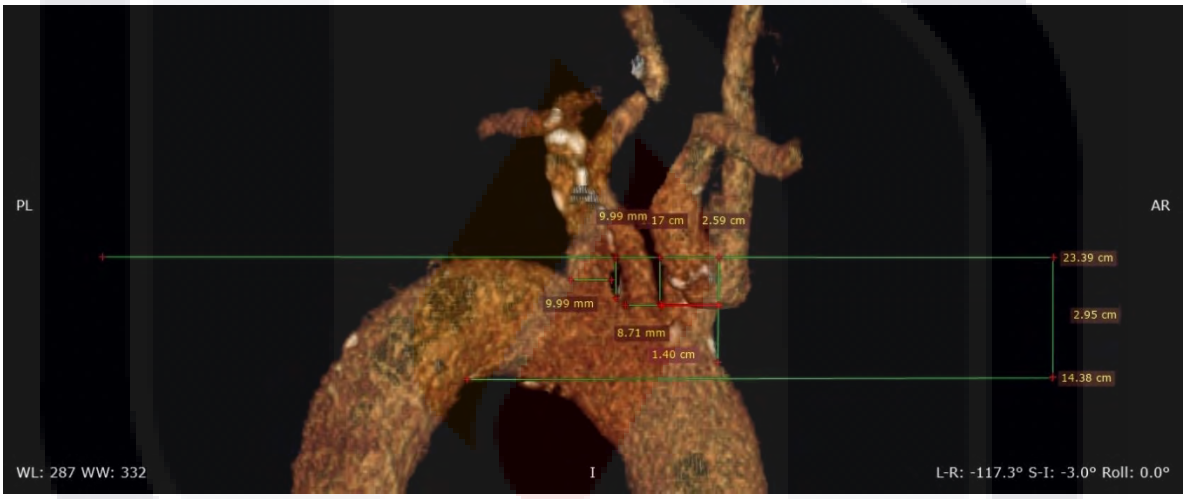


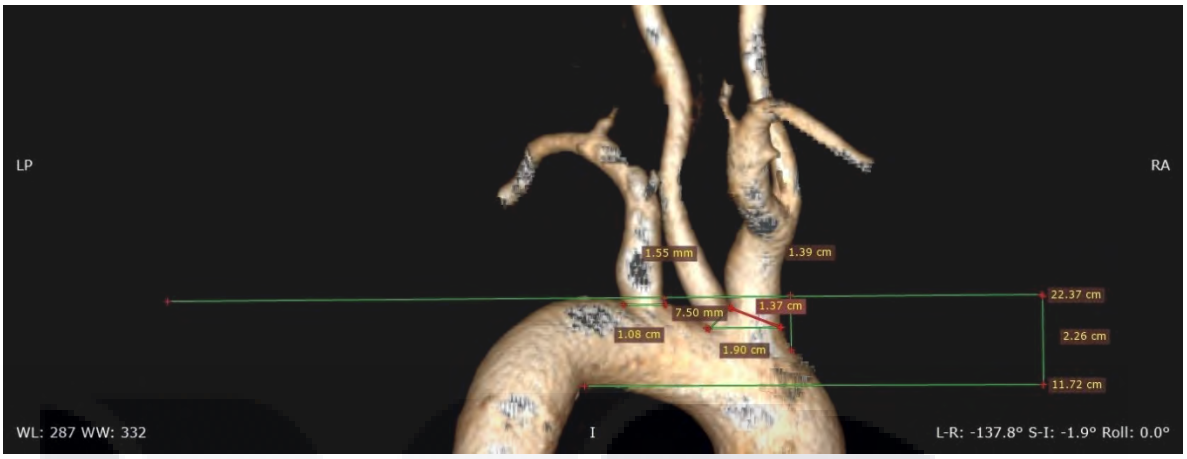


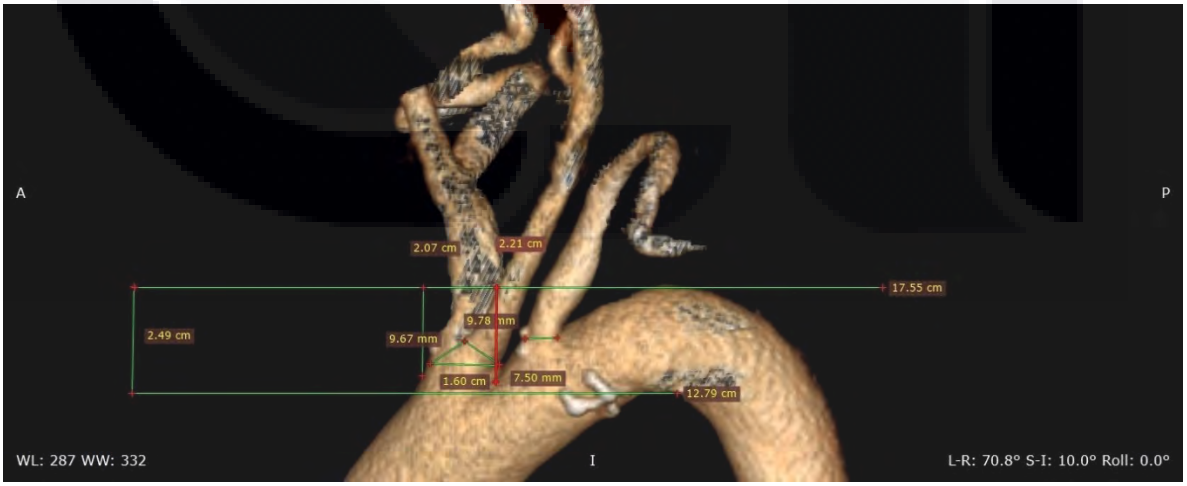


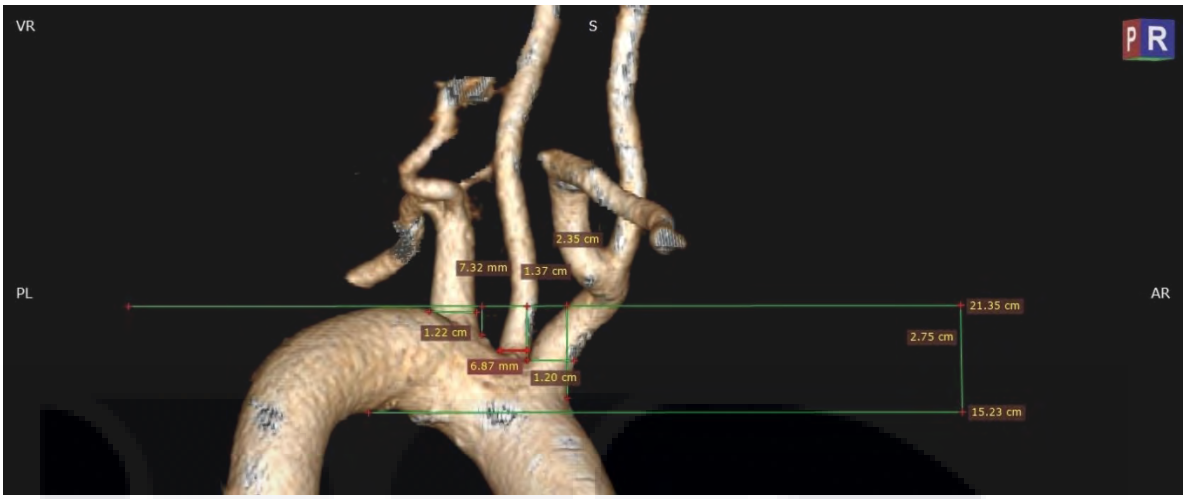


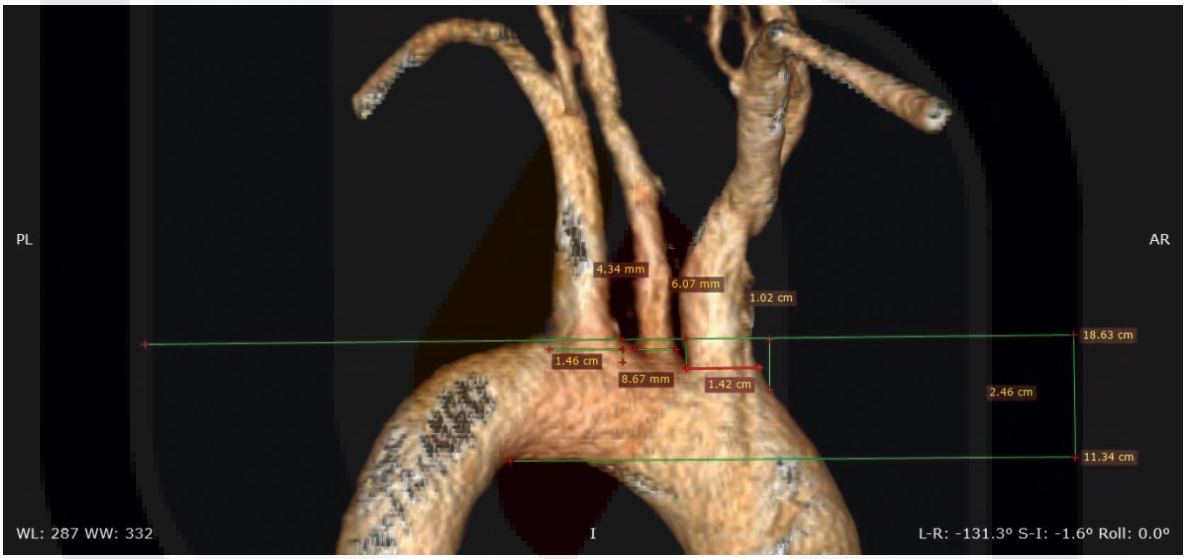


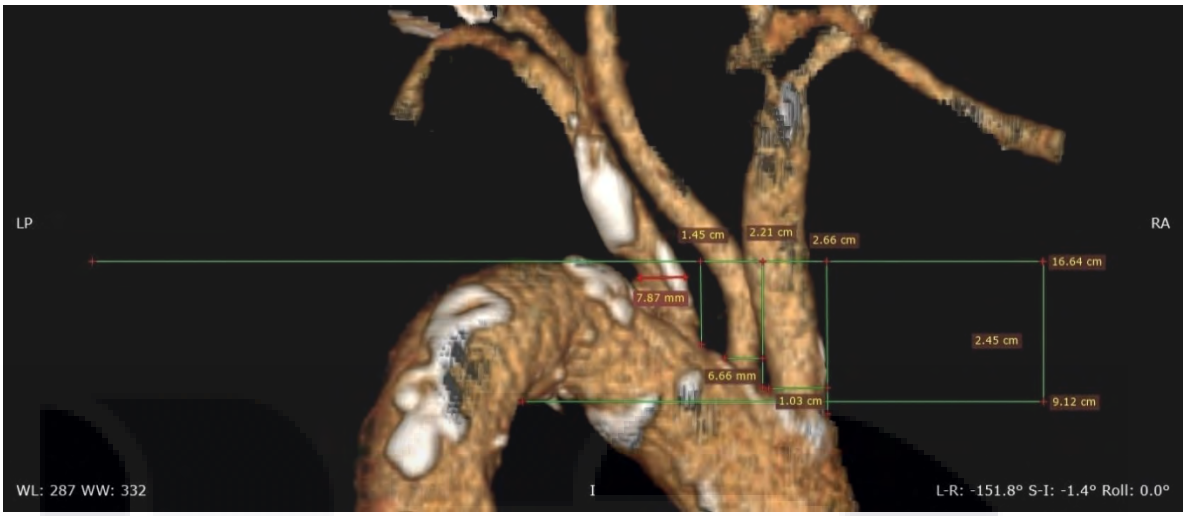


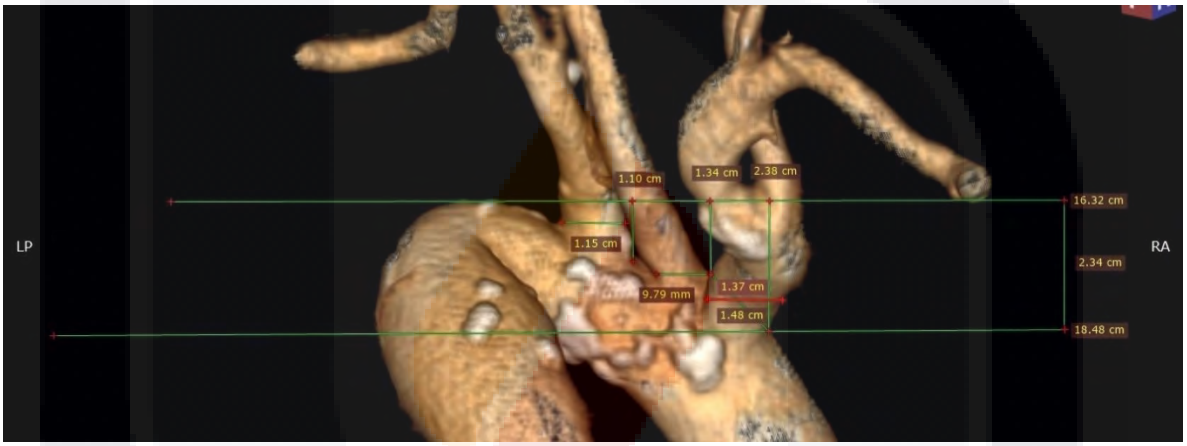


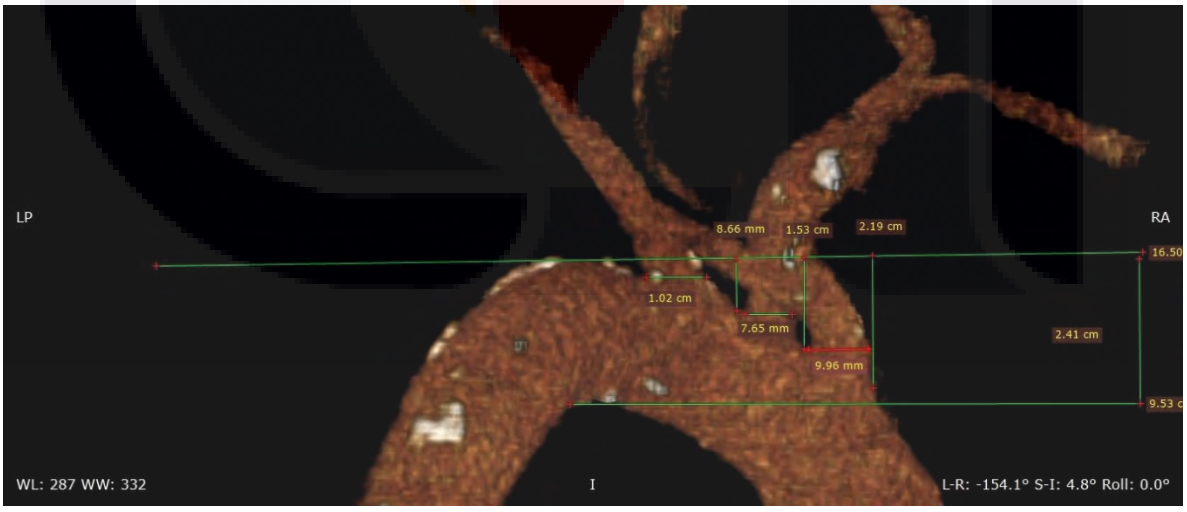


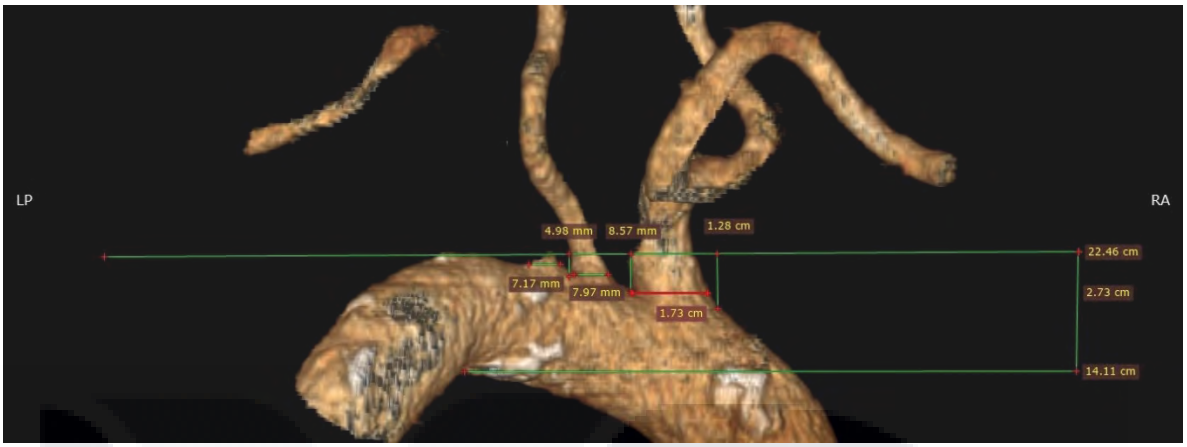


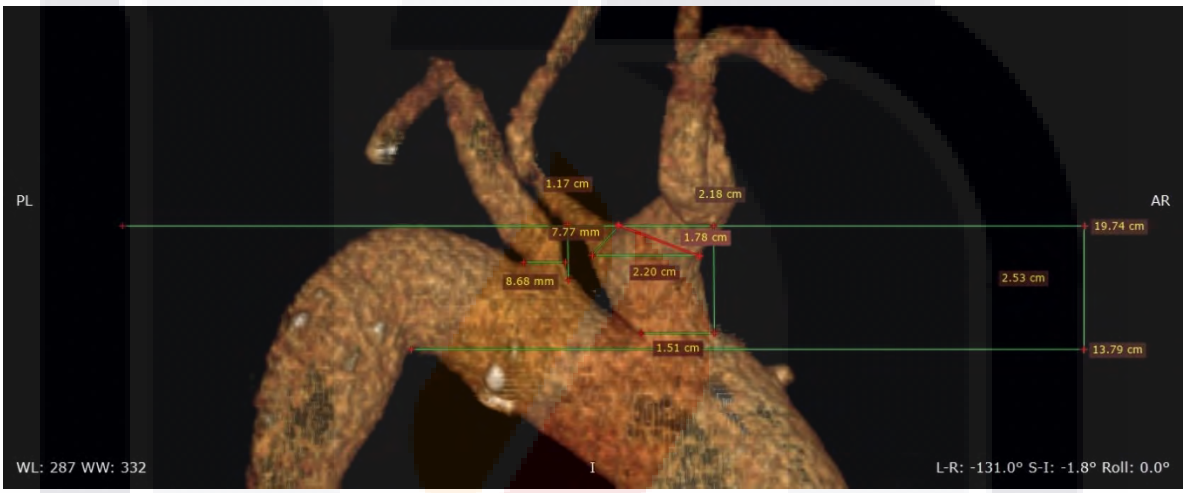


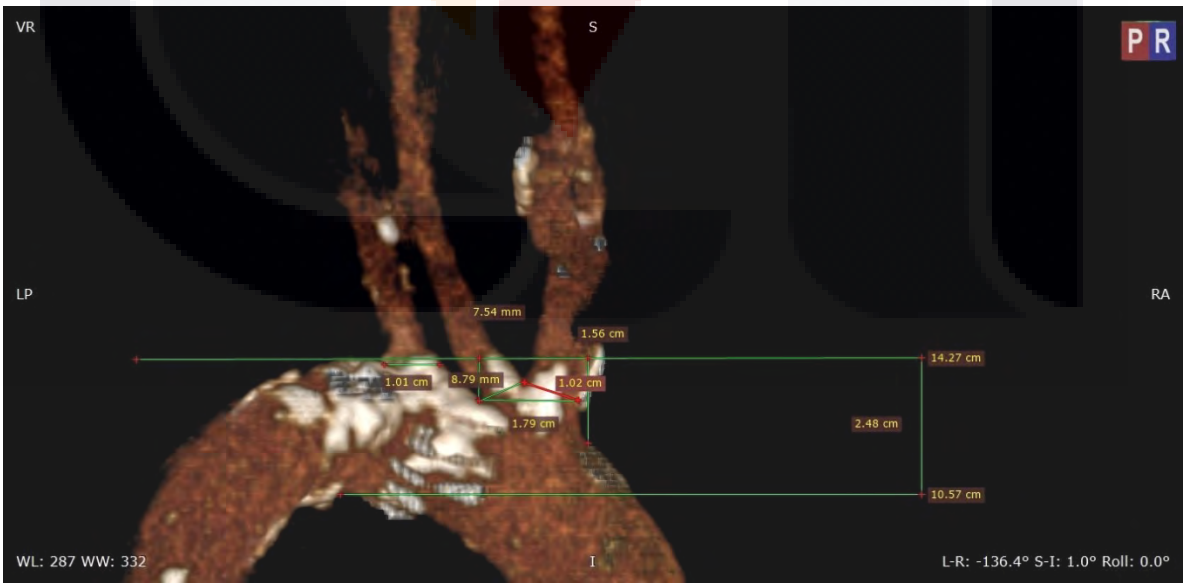


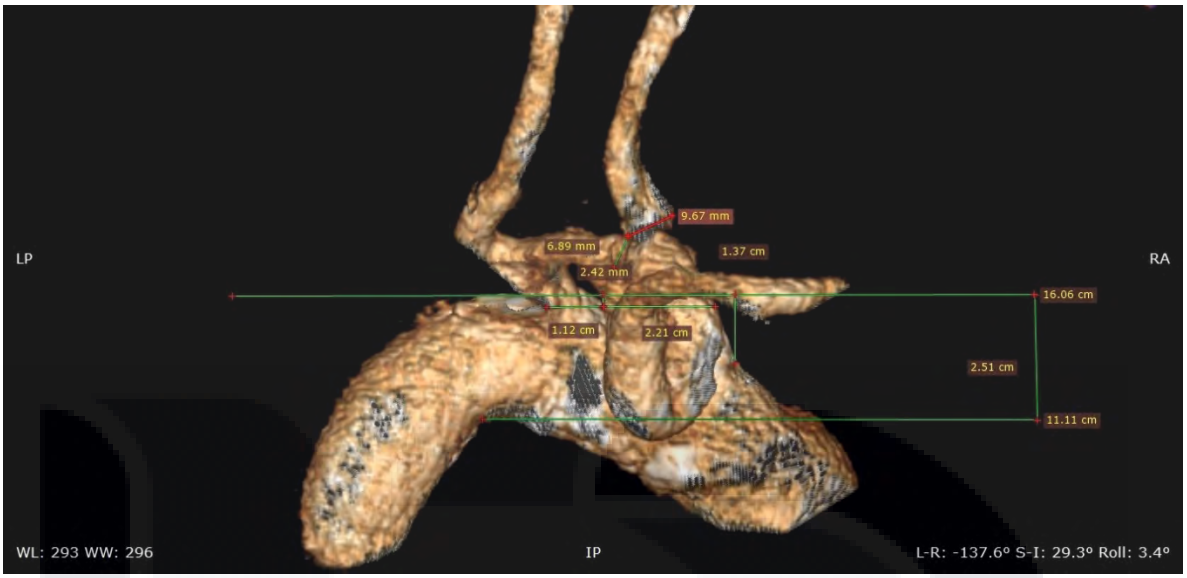


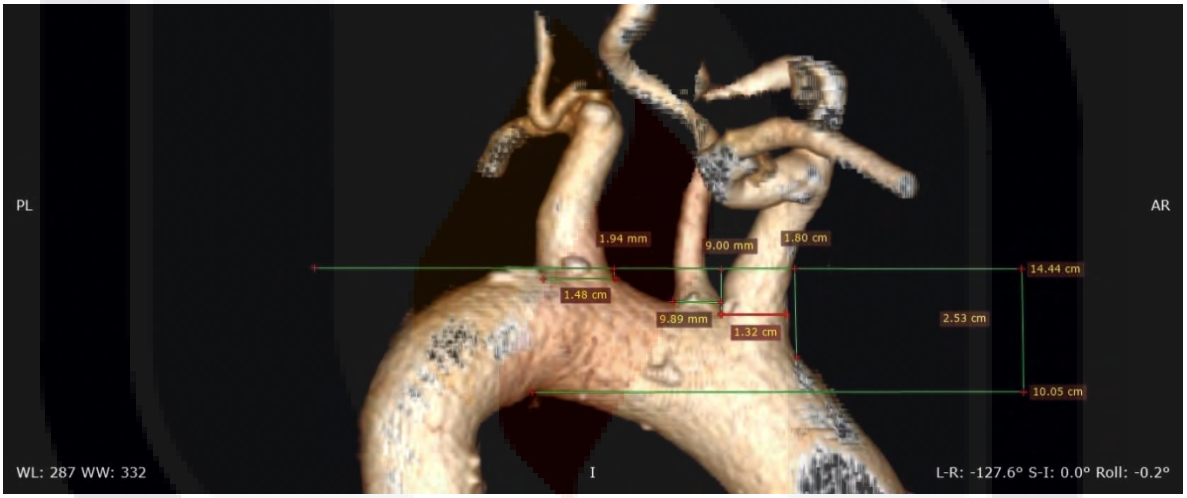












Anexo B. Código en lenguaje R utilizado para los análisis

LLAMADA FINAL

```
source('classify_folds_smote.r')
library(readxl)

data = as.data.frame(read_excel("C:/Dropbox/Proyectos y Protocolos Activos/Doctorado UAA/Datos_Angio_TAC_Only_Predictors.xlsx"))
data[,1] = as.factor(data[,1])
data[,9:28] = scale(data[,9:28])
# f = Formula, k = Numero de folds, split = vector con la proporción
resultados = classify(data = data,
                      algorithms = c("LR", "NBay", "DT", "RDA","NNet", "kNN", "SVM", "GC"),
                      f = ARC ~ ., k =5)
```

classify_folds_smote.r

```
classify = function(data, algorithms, f, k, split) {
 source('train_f.r')
 library(smotefamily)

 f = formula(f)
 resultados = list()
 Train = train_f(data, f, k)
 unique_items = unique(algorithms)
 for (i in 1:length(Train)) {
  train.data = as.data.frame(data[-Train[[i]],])
  test.data = as.data.frame(data[Train[[i]],])

  smote_result = SMOTE(X = train.data[,-1], target = train.data$ARC,
                       K = 4, dup_size = 2)
  data_c = smote_result$data[,1:7]
  for (e in 1:7) {
   data_c[,e] = ifelse(data_c[,e]> 0.5, 1, 0)
  }
  train.data = cbind(ARC = smote_result$data$class, data_c, smote_result$data[,8:27])
  train.data[1:8] = lapply(train.data[1:8], as.factor)
  test.data[1:8] = lapply(test.data[1:8], as.factor)

  if ("LR" %in% unique_items){
   source('LR.r')
   cat(i, "Modeling LR \n")
   single_level_vars = sapply(train.data, function(x) is.factor(x) && length(unique(x)) < 2)
```

```
  train.data_lr = train.data[, !single_level_vars]
  test.data_lr = test.data[, colnames(train.data)]
  resultados[[paste("Fold ",i)]]["LR"] =
  clasificar_lr(train = train.data_lr, test = test.data_lr, f = f)
 }
 if ("NBay" %in% unique_items){
  source('Nbay.r')
  cat(i, "Modeling NBay \n")
  resultados[[paste("Fold ",i)]]["NBay"] =
   clasificar_nbay(train = train.data, test = test.data, f = f)
  }
 if ("kNN" %in% unique_items) {
  source('kNN.r')
  cat(i, "Modeling k-NN \n")
  resultados[[paste("Fold ",i)]]["kNN"] =
   clasificar_knn(train = train.data, test = test.data, f = f)
  }
 if ("DT" %in% unique_items) {
  source('DT.r')
  cat(i, "Modeling DT \n")
  resultados[[paste("Fold ",i)]]["DT"] =
   clasificar_dt(train = train.data, test = test.data, f = f)
  }
 if ("RDA" %in% unique_items) {
  source('RDA.r')
  cat(i, "Modeling RDA \n")
  resultados[[paste("Fold ",i)]]["RDA"] =
   clasificar_rda(train = train.data, test = test.data, f = f)
   }
 if ("NNet" %in% algorithms) {
  source('NNet_1.r')
  cat(i, "Modeling NNet \n")
  train.data_nnet = cbind(train.data[which(colnames(data)==f[[2]])],
                train.data[, sapply(train.data, function(x) is.numeric(x))])
  test.data_nnet = cbind(test.data[which(colnames(data)==f[[2]])],
                test.data[, sapply(test.data, function(x) is.numeric(x))])
  resultados[[paste("Fold ",i)]]["NNet"] =
   clasificar_nnet(train = train.data_nnet, test = test.data_nnet, f = f)
  }
 if ("SVM" %in% algorithms) {
  source('SVM_f.r')
```

```
    cat(i, "Modeling SVM \n")
    resultados[[paste("Fold ",i)]]["SVM"] =
     clasificar_svm(train = train.data, test = test.data, f = f, data = data)
   }
   if ("GC" %in% algorithms) {
    source('Copula_f.r')
    cat(i, "Modeling GC \n")
    resultados[[paste("Fold ",i)]]["GC"] =
     clasificar_cop(train = train.data, test = test.data, f = f)
    }
  }
   return(resultados)
}
```

train_f.r

```
train_f = function(data, f, k){
 set.seed(202523)
 f = formula(f)
 Train = caret::createFolds(
  data[,which(colnames(data)==f[[2]])],
  k = k,
  list = TRUE
 )
 return(Train)
}
```

LR.r

```
clasificar_lr = function(train, test, f){
 library(caret)
 library(tidyverse)
 metricas=NULL
 f = formula(f)
 s.time = Sys.time()
 model = glm(f, data = train, family = binomial) %>%
  MASS::stepAIC(trace = FALSE)
 e.time = Sys.time()
 probabilities = model %>% predict(test, type = "response")
 predicted.classes = ifelse(probabilities > 0.5, 1, 0)
 predicted.classes = as.factor(predicted.classes)
```

```
  levels(predicted.classes)=levels(test[,which(colnames(data)==f[[2]])])
  metrica.log = confusionMatrix(data=predicted.classes, reference =
test[,which(colnames(data)==f[[2]])])
  tt = round(e.time - s.time,2)
  metrica.log[paste("Time")] = list(as.numeric(tt))
  metrica.log[[paste("rocr")]][[paste("predictions")]] = probabilities
  metrica.log[[paste("rocr")]][[paste("labels")]] = predicted.classes
  metrica.log[[paste("rocr")]][[paste("true_labels")]] = test[,which(colnames(data)==f[[2]])]
  metricas["Logistic Regression"] = list(metrica.log)
  return(metricas)
}
```

Nbay.r

```
clasificar_nbay = function(train, test, f){
  metricas=NULL
  library(caret)
  f = formula(f)

  if(!require('naivebayes')) {
    install.packages('naivebayes')
    library('naivebayes')
  }
  s.time = Sys.time()
  model.B = naive_bayes(f, data = train, usekernel = T, laplace = 1)
  e.time = Sys.time()
  b.pred = predict(model.B, test[,-which(colnames(data)==f[[2]])])
  pred.b = predict(model.B, test[,-which(colnames(data)==f[[2]])], type = "prob")
  probabilities.b = pred.b[,2]
  probabilities.b[is.na(probabilities.b)] = 0
  metrica.b = confusionMatrix(data=b.pred, reference =
as.factor(test[,which(colnames(data)==f[[2]])]))
  tt = round(e.time - s.time,2)
  metrica.b[paste("Time")] = list(as.numeric(tt))
  metrica.b[[paste("rocr")]][[paste("predictions")]] = list(probabilities.b)
  metrica.b[[paste("rocr")]][[paste("labels")]] = list(b.pred)
  metrica.b[[paste("rocr")]][[paste("true_labels")]] =
as.factor(test[,which(colnames(data)==f[[2]])])
  metricas["NBayes"] = list(metrica.b)

  return(metricas)
```

```
}
```

kNN.r

```
clasificar_knn = function(train, test, f){
  library(caret)
  metricas=NULL
  f = formula(f)
  tags = which(colnames(train)==f[[2]])
  kas=round(sqrt(length(train[,tags])), digits = 0)
  acuracies = rep(0,kas)
  s.time = Sys.time()
  for (e in 1:kas){
    knn.mod = class::knn(train=train[,-tags], test=test[,-tags], cl=train[,tags], k=e)
    metrica.knn = confusionMatrix(data=knn.mod, reference = as.factor(test[,tags]))
    acuracies[e] = as.numeric(metrica.knn$byClass[11])
  }
  acuracies[1] = 0
  knn.opt.mod = class::knn(train=train[,-tags], test=test[,-tags],
                cl=train[,tags], k=which.max(acuracies), prob = TRUE)
  e.time = Sys.time()
  metrica.knn.opt = confusionMatrix(data=knn.opt.mod, reference =
as.factor(test[,which(colnames(data)==f[[2]])]))
  tt = round(e.time - s.time,2)
  probabilities.knn = ifelse(knn.opt.mod == "1", attr(knn.opt.mod, "prob"), 1 - attr(knn.opt.mod,
"prob"))
  probabilities.knn[is.na(probabilities.knn)] = 0
  metrica.knn.opt[paste("Time")] = list(as.numeric(tt))
  knn.classes = as.factor(as.vector(knn.opt.mod))
  levels(knn.classes)= c(0,1)
  metrica.knn.opt[[paste("rocr")]][[paste("predictions")]] = list(probabilities.knn)
  metrica.knn.opt[[paste("rocr")]][[paste("labels")]] = list(knn.classes)
  metrica.knn.opt[[paste("rocr")]][[paste("true_labels")]] =
as.factor(test[,which(colnames(data)==f[[2]])])
  metricas[paste("kNN, k=",which.max(acuracies))] = list(metrica.knn.opt)

  return(metricas)
}
```

DT.r

```
clasificar_dt = function(train, test, f){
  metricas=NULL
  f = formula(f)
  library(caret)

  if(!require('parsnip')) {
    install.packages('parsnip')
    library('parsnip')
  }
  s.time = Sys.time()
  tree_spec = decision_tree() %>%
    set_engine("rpart") %>%
    set_mode("classification")
  tree_fit = fit(tree_spec, f, data = train)
  e.time = Sys.time()
  tree_pred = predict(tree_fit, as.data.frame(test))
  tree_pred = as.vector(tree_pred)
  tree.pred = predict(tree_fit, test, type = "prob")
  probabilities.tree = as.data.frame(tree.pred[,2],give.attr=FALSE)
  probabilities.tree[is.na(probabilities.tree)] = 0
  metrica.tree.mod = confusionMatrix(data=tree_pred$.pred_class,
                                     reference = as.factor(test[,which(colnames(data)==f[[2]])]))
  tt = round(e.time - s.time,2)
  metrica.tree.mod[paste("Time")] = list(as.numeric(tt))
  metrica.tree.mod[[paste("rocr")]][[paste("predictions")]] = list(probabilities.tree)
  metrica.tree.mod[[paste("rocr")]][[paste("labels")]] = list(tree_pred)
  metrica.tree.mod[[paste("rocr")]][[paste("true_labels")]]                                    =
as.factor(test[,which(colnames(data)==f[[2]])])
  metricas["Decision Tree"] = list(metrica.tree.mod)

  return(metricas)
}
```

RDA.r

```
clasificar_rda = function(train, test, f){
  metricas=NULL
  library(caret)
  f = formula(f)

  if(!require('klaR')) {
```

```
  install.packages('klaR')
  library('klaR')
 }
 s.time = Sys.time()
 lda.mod = rda(f,data = train, CV = TRUE)
 e.time = Sys.time()
 lda.pred = predict(lda.mod, test)
 metrica.lda.mod = confusionMatrix(data=lda.pred$class,
                          reference = as.factor(test[,which(colnames(data)==f[[2]])]))
 tt = round(e.time - s.time,2)
 metrica.lda.mod[paste("Time")] = list(as.numeric(tt))
 probabilities.lda = lda.pred$posterior[,2]
 probabilities.lda[is.na(probabilities.lda)] = 0
 metrica.lda.mod[[paste("rocr")]][[paste("predictions")]] = list(probabilities.lda)
 metrica.lda.mod[[paste("rocr")]][[paste("labels")]] = list(lda.pred$class)
 metrica.lda.mod[[paste("rocr")]][[paste("true_labels")]]                    =
as.factor(test[,which(colnames(data)==f[[2]])])
 metricas["LDA"] = list(metrica.lda.mod)

 return(metricas)
}
```

NNet_1.r

```
clasificar_nnet = function(train, test, f){
 metricas=NULL
 library(caret)
 f = formula(f)

 if(!require('neuralnet')) {
  install.packages('neuralnet')
  library('neuralnet')
 }
 tryCatch({
  tags = which(colnames(train)==f[[2]])
  s.time = Sys.time()
  seed = 2024
  model_w = neuralnet(
   f,
   data = train,
   hidden=c(8),
```

```
    linear.output = FALSE,
    stepmax = 1e6,
    threshold = 0.01,
    rep = 5
   )
   e.time = Sys.time()
   arq = "LR0.0001/20:8:10:1"
   pred_w = predict(model_w, test[,-tags])

   if (ncol(pred_w) == 1) {
    cual = ifelse(pred_w > 0.5, levels(test[, tags])[2], levels(test[, tags])[1])
   } else {
    cual = ifelse(pred_w[, 1] > pred_w[, 2], levels(test[, tags])[1], levels(test[, tags])[2])
   }
   metrica_w.mod = confusionMatrix(data=as.factor(cual),
                           reference = as.factor(test[,tags]))
   tt = round(e.time - s.time,2)
   metrica_w.mod[paste("Time")] = list(as.numeric(tt))
   probabilities.w = pred_w[,2]
   probabilities.w[is.na(probabilities.w)] = 0
   metrica_w.mod[[paste("rocr")]][[paste("predictions")]] = probabilities.w
   metrica_w.mod[[paste("rocr")]][[paste("labels")]] = factor(cual, levels = c(0,1))
   metrica_w.mod[[paste("rocr")]][[paste("true_labels")]] = list(as.factor(test[,tags]))
   metricas['20:11:10:1 NNet'] = list(metrica_w.mod)
   errors[e,1] = arq
  }, error=function(e){
   cat("ERROR :",conditionMessage(e), "\n")})

  return(metricas)
}
```

SVM_f.r

```
clasificar_svm = function(train, test, f, data) {
  metricas = NULL
  library(caret)
  library(e1071)

  f = formula(f)
  train = data.frame(ARC = train$ARC, train[, !sapply(train, is.factor), drop = FALSE])
  test = data.frame(ARC = test$ARC, test[, !sapply(test, is.factor), drop = FALSE])
```

```
  s.time = Sys.time()
  svmfit = svm(f, data = train, kernel = "linear", cost = 1, probability = TRUE, scale = FALSE)
  e.time = Sys.time()
  pred_svm = predict(svmfit, test)
  pred.svm = predict(svmfit, test, probability = TRUE)
  probabilities.svm = attr(pred.svm, "probabilities")[, 2]
  probabilities.svm[is.na(probabilities.svm)] = 0
  reference = as.factor(test[[which(colnames(data) == f[[2]])]])
  metrics.svmfit = confusionMatrix(data = as.factor(pred_svm), reference = reference)
  tt = round(e.time - s.time, 2)
  metrics.svmfit$Time = as.numeric(tt)
  metrics.svmfit$rocr = list(
    predictions = probabilities.svm,
    labels = pred_svm,
    true_labels = reference
  )

  metricas["SVM"] = list(metrics.svmfit)
  return(metricas)
}
```

Copula_f.r

```
clasificar_cop = function(train, test, f, data) {
  metricas = NULL
  library(caret)
  library(MASS)
  library(Matrix)

  if (!require("MLCOPULA")) {
    install.packages("MLCOPULA")
    library("MLCOPULA")
  }

  f = formula(f)
  tags = which(colnames(train) == f[[2]])
  numeric_cols = sapply(train, is.numeric)
  numeric_cols[tags] = TRUE
  train = train[, numeric_cols, drop = FALSE]
  test = test[, numeric_cols, drop = FALSE]
```

```
print("Numeric columns selected:")
print(colnames(train))
near_zero_var = caret::nearZeroVar(train[,-tags, drop = FALSE], saveMetrics = TRUE)
print("Near-zero variance predictors:")
print(near_zero_var)
valid_columns = unique(c(tags, which(!near_zero_var$nzv)))
train = train[, valid_columns, drop = FALSE]
test = test[, colnames(train), drop = FALSE]
print("Processed training data:")
print(str(train))
print("Processed test data:")
print(str(test))
X_train = train[,-tags, drop = FALSE]
y_train = train[,tags]

print("Dimensions of X (predictors):")
print(dim(X_train))
sigma = as.matrix(nearPD(cov(X_train))$mat)
colnames(sigma) = colnames(X_train)
rownames(sigma) = colnames(X_train)
print("Covariance matrix dimensions:")
print(dim(sigma))
print("Covariance matrix:")
print(sigma)
if (!all(colnames(X_train) == colnames(sigma))) {
  stop("Mismatch between predictor columns and covariance matrix.")
}
print("Training the copula-based classifier...")
s.time = Sys.time()
model_cop = copulaClassifier(
  X = X_train,
  y = y_train,
  copula = "frank",
  distribution = "kernel",
  graph_model = "chain"
)

print("Trained Copula Classifier:")
print(model_cop)
print("Aligning test predictors with training covariance matrix...")
X_test = test[,-tags, drop = FALSE]
```

```
  X_test = X_test[, colnames(sigma), drop = FALSE]
  if (!all(colnames(X_test) == colnames(sigma))) {
    stop("Mismatch between test predictor columns and covariance matrix.")
  }
  print("Aligned test predictors for prediction:")
  print(dim(X_test))
  print(head(X_test))
  print("Predicting on the test set...")
  pred_cop = copulaPredict(X = X_test, model = model_cop)
  print("Prediction output:")
  print(pred_cop)

  probabilities.cop = pred_cop$prob[,2]
  probabilities.cop[is.na(probabilities.cop)] = 0
  e.time = Sys.time()

  reference = as.factor(test[,tags])
  metrica_cop = confusionMatrix(
    data = as.factor(pred_cop$class),
    reference = reference
  )
  print("Confusion Matrix:")
  print(metrica_cop)

  tt = round(e.time - s.time, 2)
  metrica_cop$Time = as.numeric(tt)
  metrica_cop$rocr = list(
    predictions = probabilities.cop,
    labels = factor(pred_cop$class, levels = c(0, 1)),
    true_labels = reference
  )

  metricas["Copula"] = list(metrica_cop)
  return(metricas)
}
```