UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ALGORITMO DE APRENDIZAJE AUTOMÁTICO PARA EL ESTUDIO DE LA ASOCIACIÓN ENTRE ENFERMEDAD CARDIOVASCULAR Y DEPRESIÓN.

POR

Rodrigo Ignacio Navarro Araneda

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero Civil Biomédico

Profesor Guía:

Dra. Rosa L. Figueroa I.

Comisión:

Dr. Pablo E. Aqueveque N.

Dr. Sergio K. Sobarzo G.

Septiembre 2023

Concepción (Chile)

© 2023 Rodrigo Ignacio Navarro Araneda

Agradecimientos

A mi familia, por su comprensión y apoyo.

A mi profesora guía, por su dedicación y compromiso.

A mis amigos, por enseñarme y crecer juntos.

A la Capoeira, por brindarme hermandad, desafíos y bienestar.

Resumen

Desde hace 20 años las enfermedades cardiovasculares son la principal causa de muerte a nivel global y en Chile son las responsables de un 25.6% de defunciones. Por su parte, la depresión es la principal causa de baja laboral, afectado a más de 450 millones de personas a nivel global, y en Chile se estima que un 15.8% de la población la padece. La base de datos CHS posee datos de 20 años, más de 300 variables y un aproximado de 5000 pacientes adultos mayores con enfermedades cardiovasculares, siendo al año 2000 el estudio longitudinal más extenso realizado nunca. Se utilizaron los algoritmos de aprendizaje automático Random Forest y Red Neuronal Artificial para realizar un estudio de las variables que predicen la fatalidad por enfermedades cardiovasculares y aquellas que predicen la depresión. El ensamble de ambos algoritmos permitió encontrar variables predictoras de depresión cuyo origen es cardiovascular, sin encontrar a la depresión como un predictor directo de las enfermedades cardiovasculares.

Summary

For the past 20 years, cardiovascular diseases have been the leading cause of death globally, and in Chile, they are responsible for 25.6% of deaths. Meanwhile, depression is the leading cause of work-related absenteeism, affecting over 450 million people worldwide, with an estimated 15.8% of the population in Chile being affected. The CHS database contains 20 years' worth of data, comprising over 300 variables and approximately 5000 elderly patients with cardiovascular diseases, making it the most extensive longitudinal study conducted up to the year 2000. Machine learning algorithms, specifically Random Forest and Artificial Neural Network, were employed to study the variables predicting fatality from cardiovascular diseases and those predicting depression. The ensemble of both algorithms allowed for the identification of predictor variables for depression with a cardiovascular origin but did not find depression to be a direct predictor of cardiovascular diseases.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Abreviaciones

Mayúsculas

E.C. V : Enfermedad cardiovascular

I.A.M : Infarto agudo al miocardio

C.H.S : Del inglés cardiovascular health study

S.V.M : Del inglés support vector machine

C.A.D : Arteriopatía coronaria

S.C.D : Del inglés sudden cardiac death

E.H.R : Del inglés electronic health record

R.F : Del inglés Random Forest

I.D.E : Del ingles Integrated Development Environment

R.N.A : Red Neuronal Artificial

A.C.V : Accidente cerebrovascular

I.C.C : Insuficiencia cardíaca congestiva

Capítulo 1. Introducción

1.1 Introducción General

A nivel mundial existen 10 causas de muerte que ocasionan el 55% de las defunciones, dentro de estas las enfermedades cardiovasculares ocupan el primer lugar. De acuerdo con la Organización de las Naciones Unidas (ONU), el 2019 alrededor de 19 millones de personas murieron a causa de una enfermedad cardiovascular (ECV). [1, 2]. En el contexto chileno se puede observar una tendencia similar. De acuerdo con datos del anuario de estadísticas vitales del Instituto Nacional de Estadística (INE), durante el año 2019 las enfermedades del sistema circulatorio fueron la segunda causa de muerte, con un 25.6% de defunciones (ver anexo B, Fig. B.9).

Por otro lado, la depresión afecta a más de 450 millones de personas en todo el mundo. Se estima que una de cada cuatro personas sufrirá en algún momento de su vida algún episodio depresivo, sin importar su edad ni condición social [4]. A nivel local, en Chile, un 6.2% de la población sufre depresión y se estima que un 15.8% se encontraría en sospecha según la encuesta nacional de salud 2016-2017 del Ministerio de Salud (MINSAL) [5].

Anteriormente, no se tenía conciencia de la posible asociación que existe entre depresión y la ECV, sin embargo, en los últimos años se ha reportado en numerosos artículos científicos la sinergia que producen ambas patologías debido a la complejidad de sus efectos fisiológicos y psicológicos. Si bien se está estudiando su posible asociación, aún no existe conocimiento claro de los factores que las vinculan, ni de medidas clínicas eficaces para el tratamiento antidepresivo en pacientes con *ECV*, así como tampoco medidas estandarizadas para su diagnóstico [6, 7].

Actualmente existe un auge del uso de Aprendizaje Automático para el estudio de bases de datos médicas debido a la mejora de las capacidades computacionales vividas durante los últimos años, esto ha permitido el desarrollo de algoritmos que predicen y averiguan tendencias en grandes bases de datos. El presente trabajo de la base de datos del estudio cardiovascular o Cardiovascular Health Study (CHS), que cuenta con el historial clínico de cerca de 5000 pacientes adultos mayores que padecían enfermedades del sistema cardiovascular [8].

1.2 Trabajos Previos

En esta sección se hará revisión del material bibliográfico referente a los tópicos de la investigación. Analizando artículos científicos en 6 tópicos principales, los cuales son: i) Incidencia de E.C.V y depresión a nivel global y local; ii) Relación bidireccional existente entre ambas patologías y afecciones sinérgicas para producir E.C.V; iii) Estudio de la relación entre depresión y ECV para adultos mayores; iv) Estudio de los mecanismos que rigen en la relación depresión-ECV; v) Algoritmos de aprendizaje utilizados para el estudio de bases de datos médicas y vi) Revisión de estudios de la base de datos Cardiovascular Health Study.

1.2.1 Base de Datos cardiovascular health study: design and rationale.

- Fried, L. P., Borhani, N. O., Enright, P. et al. "The Cardiovascular Health Study: design and rationale," *Ann Epidemiol*, vol. 2, issue 3, pp.263-276, Feb. 1991

El presente artículo aborda un análisis de los objetivos, el diseño y estructura de la base de datos *Cardiovascular Health Study (CHS)*. Este corresponde a un estudio longitudinal de las enfermedades cardiovasculares en pacientes mayores de 65 años, con un seguimiento en la recopilación de datos de hasta 24 años. La cantidad de participantes del estudio en un comienzo fue de 1200 hombres y mujeres procedentes de las siguientes 4 comunidades: Carolina del Norte, Condado de Forsyth; Sacramento, California; Maryland, Condado de Washington; y Pittsburgh, Pennsylvania. Sin embargo, al final del estudio fueron monitoreados cerca de 5000 sujetos.

El objetivo del estudio fue cuantificar las asociaciones y plantear los factores de riesgo para accidentes cardíacos y cerebrovasculares, evaluar la asociación de enfermedades subclínicas medibles por exámenes con la incidencia de cardiopatía coronaria y accidente cerebrovascular. Cuantificar la asociación de factores de riesgo generados por enfermedades subclínicas, caracterizando el historial de cardiopatía coronaria y accidente cerebrovascular, encontrando sus factores asociados. Finalmente, se buscó describir la prevalencia y distribución de los siguientes factores de riesgo: enfermedades subclínicas, cardiopatía coronaria y accidente cerebrovascular.

Los exámenes realizados en el estudio son amplios, contemplando factores psicosociales como calidad de vida percibida, apoyo social, redes sociales, eventos de la vida y depresión; también contempla la atención médica, como medicación, historial médico y hospitalizaciones; junto a esto, considera también exámenes clínicos varios, como presión sanguínea, exámenes físicos, actividad

física, función neurológica, hábitos dietéticos, exámenes de laboratorio, ultrasonido y electrocardiografía. La periodicidad de los datos obtenidos en estos exámenes varía desde semestral, anual y trianual.

1.2.2 Corazones y mentes: estrés, ansiedad y depresión

♣ Silverman, A. L., Herzog, A. A., & Silverman, D. I. (2019). Hearts and Minds: Stress, Anxiety, and Depression: Unsung Risk Factors for Cardiovascular Disease. *Cardiology in Review*, *27*(4), 202-207. https://doi.org/10.1097/crd.0000000000000228

Mediante la revisión de diversos artículos científicos, el presente artículo busca establecer la relación causal entre la depresión, ansiedad y estrés con las enfermedades cardiovasculares (ECV). Se destaca que la tasa de depresión en pacientes con ECV es de aproximadamente un 20%, además se cuantifican los efectos adversos que produce en ellos. El artículo indica que es más predictivo el padecer depresión para el desarrollo de infartos agudos del miocardio que el tabaquismo, obesidad, diabetes o hipertensión como causa de estos eventos. Además, destaca que la depresión disminuye la supervivencia luego de estos eventos en un 30%. Mediante un metaanálisis, el artículo establece que la relación entre depresión y el riesgo de muerte cardiaca súbita es de un 1.6% de probabilidad.

Por otro lado, el artículo detalla cómo se establece la relación entre el estrés o la ansiedad y sus efectos en la salud, a través de la sinergia de sus patogénesis. Se destaca que el estrés y la ansiedad pueden generar respuestas corporales sin una amenaza real, lo que lleva a la vasoconstricción de arterias y puede provocar arritmias o isquemias. Asimismo, se señala que los pacientes con depresión suelen presentar altos niveles de cortisol. Por otro lado, el artículo resalta que los pacientes con antecedentes de ataques de pánico tienen un 75% de probabilidad de sufrir una isquemia miocárdica cuando hiperventilan, en comparación con el 6.7% de probabilidad en aquellos sin antecedentes de ataques de pánico.

Los pacientes depresivos y con *arteriopatía coronaria (C.A.D)* presentan marcadores inflamatorios, como la proteína c reactiva, interleucina-6 y fibrinógenos, elevados en relación con los no depresivos. Además, se destaca que las diferencias fisiológicas en las respuestas hormonales varían entre los géneros masculino y femenino. Las mujeres presentan un mayor volumen y actividad de receptores hormonales en el cerebro que los hombres, lo que conlleva una mayor prevalencia de *ECV* y depresión. Asimismo, existen diferencias en la agregación plaquetaria entre los no deprimidos y los

deprimidos con CAD: los pacientes deprimidos muestran una mayor activación plaquetaria, lo cual aumenta el estrés agudo y los marcadores de este tipo. Estos cambios no se observan en pacientes no deprimidos con angina de pecho. También es importante señalar que la disminución en la variabilidad cardíaca en pacientes deprimidos se considera un factor de riesgo de *muerte cardiaca súbita (S.C.D).* Además de los factores biológicos que establecen esta compleja relación, se evidencia que factores psicosociales contribuyen a aumentar la incidencia y letalidad de la ECV. Estos factores incluyen el bajo apoyo social percibido, la presencia de pocas redes interpersonales, el estado civil de estar divorciado, una baja integración social y una reducida participación en actividades sociales. Por otra parte, los pacientes deprimidos no siguen los comportamientos recomendados según sus diagnósticos y tienden a un consumo excesivo de alcohol o cigarrillos.

Después de establecer la relación entre la depresión y las ECV, el artículo se enfoca en examinar el diagnóstico y tratamiento de la depresión en estos pacientes. Para ello, se analizaron diversas técnicas de tratamiento para el manejo del estrés, la ansiedad y la depresión. Los resultados demostraron impactos positivos en pacientes con ECV. En lo que concierne a la reducción del estrés, las técnicas de meditación trascendental y relajación muscular progresiva presentaron efectos positivos. Por otro lado, la meditación de atención plena mostró resultados favorables en la modulación neurofisiológica del miedo, la ansiedad y la hipertensión. Respecto a la regulación de pensamientos, comportamientos y emociones, se observó que la terapia cognitivo-conductual resultó altamente eficaz. Esto condujo a una reducción del 41% en eventos de *síndrome coronario agudo (S.C.D)* y un 45% en *infartos agudos al miocardio (I.A.M).* En el ámbito farmacológico, se resalta que la sertralina no demostró mejoras significativas en la depresión ni en los resultados cardiovasculares. Por el contrario, el antidepresivo Bupropion mostró ser eficaz para ambas afecciones. Finalmente, se concluye que los medicamentos tricíclicos no son apropiados y deben evitarse en pacientes con ECV.

1.2.3 La depresión y su impacto en la salud pública

- ♣ M. T. Corea Del Cid Depression and its impact in public health. REV MÉD HONDUR 2021;89 (Supl. No. 1): S1-68. DOI: 10.5377/rmh.v89iSupl.1.12047

Mediante una revisión bibliográfica, el presente estudio enfatiza la depresión como un serio problema de salud pública, resultado de su alta incidencia a nivel global, con un diagnóstico que supera los 400 millones de personas. Además, se subraya la notable repercusión económica de la

depresión a nivel mundial. El estudio se inicia al abordar la compleja etiología de esta afección, la cual involucra factores de índole psicosocial, genética y biológica. La revisión revela que los factores genéticos que describen la heredabilidad de los trastornos de depresión mayor representan el 37%. El restante 63% se atribuye a factores ambientales y sociales. Es importante destacar que, a pesar de la marcada influencia ambiental y social en el desarrollo de los trastornos depresivos, existe una notoria carencia en la atención de estos trastornos, especialmente en la región de América Latina y el Caribe, donde la tasa de atención presenta un preocupante déficit del 73.9%.

Se realizó un análisis de la depresión para diferentes rangos etarios, cuantificando la prevalencia en la infancia de 2.8%, porcentaje que no difiere entre niños y niñas. En la adolescencia la prevalencia de depresión fue de un 5.6% y afectó el doble de veces a mujeres que a hombres. La depresión en los dos rangos etarios mencionados anteriormente fue asociada con condiciones médicas crónicas, necesidades especiales y niños procedentes de minorías indígenas. En el rango etario de los adultos mayores, la depresión se presenta como la patología más prevalente, siendo crónica y recurrente, como causa de su aparición se identificaron las problemáticas familiares, las enfermedades y los tratamientos farmacológicos como las causas predisponentes para depresión en los ancianos, se destacó que los factores genéticos tienen menos relevancia para este rango etario, y se hizo énfasis en la comorbilidad de la depresión geriátrica con numerosas enfermedades, como *infarto agudo al miocardio (IAM)*, párkinson, alzhéimer, cardiopatías, diabetes, hipotiroidismo y cáncer. Cabe subrayar que el artículo sitúa a las enfermedades como las causantes de depresión secundaria, destacando una menor probabilidad de poseer depresión en ausencia de enfermedades crónicas.

El último aspecto abordado en el artículo se enfoca en la discapacidad originada por la depresión en adultos, proporcionando datos significativos. Se revela que aproximadamente el 80% de los individuos con depresión enfrenta dificultades en su desempeño laboral, tareas domésticas y habilidades sociales. Este impacto es más pronunciado que en numerosas enfermedades crónicas como angina de pecho, diabetes mellitus, artritis o asma. Para culminar, se establece la posición de la depresión unipolar como la tercera causa más importante de morbilidad, contribuyendo con el 4.3% de la carga global de enfermedades. Los problemas derivados de esta condición representan la principal razón de la incapacidad laboral, tanto temporal como permanente, en naciones desarrolladas. Conscientes del significativo impacto de la depresión en la salud mundial, países como Estados Unidos, Canadá y el Reino Unido destinan recursos a la creación de centros especializados para la

prevención y tratamiento de esta afección.

1.2.4 Depresión y enfermedad cardiovascular en ancianos: conocimiento actual

- Zhang, Y., Chen, Y., & Ma, L. (2018). Depression and cardiovascular disease in elderly: Current understanding. *Journal of Clinical Neuroscience, 47,* 1-5. https://doi.org/10.1016/j.jocn.2017.09.022

El artículo caracteriza a la población mundial como una sociedad en proceso de envejecimiento y clasifica como psicosomáticas a las enfermedades que suelen acompañar la vejez, como la hipertensión, enfermedad coronaria, diabetes y E.C.V. Estas condiciones, al coexistir con la depresión, contribuyen al deterioro tanto fisiológico como psicológico de los pacientes. Se resalta una alta incidencia de depresión geriátrica en las naciones desarrolladas, con tasas que varían entre un 14.6% y un 33.5%. El estudio pone énfasis en la relevancia y la frecuencia global de la depresión en la tercera edad. Además, se establece la relación cuantitativa entre la depresión y las enfermedades propias de la vejez. Se encuentra una prevalencia del 15% al 20% de pacientes con ECV que también padecen depresión, así como un 19.8% en pacientes que han sufrido un I.A.M. Se subraya que el 31.1% de los pacientes con I.A.M tienen antecedentes de depresión significativa. Además, se destaca una prevalencia del 40.1% de depresión en pacientes diagnosticados con hipertensión arterial.

A continuación, se llevó a cabo un análisis epidemiológico más profundo con el objetivo de medir la relación entre estas patologías. En primer lugar, se estableció una conexión entre la depresión y las *E.C.V*, subrayando que las *E.C.V* generan disfunciones somáticas, una carga financiera significativa y una mayor dependencia en el cuidado de otras personas. Estos factores pueden contribuir a la aparición y el desarrollo de la depresión geriátrica. De manera recíproca, la depresión puede originar o agravar enfermedades crónicas, además de aumentar el riesgo de enfermedades coronarias en un 60% a un 80%. En segundo lugar, un metaanálisis reveló una correlación entre la depresión y la hipertensión, así como con la morbilidad y mortalidad relacionadas con el *I.A.M*, accidentes cerebrovasculares, muerte súbita y otros eventos cardiovasculares graves, que se incrementan en pacientes con depresión. A continuación, se evaluó de manera cuantitativa la asociación entre la depresión y las enfermedades coronarias, demostrando que la depresión aumenta el riesgo de *I.A.M* entre 1.5 y 4.5 veces en comparación con la población sana. Por otro lado, se destacó que experimentar emociones positivas reduce la incidencia de *I.A.M* durante un período de 10 años.

Finalmente, se cuantificó la relación entre la depresión y la diabetes, mostrando una incidencia del 27.3% en pacientes diabéticos. En estos casos, los trastornos psicológicos derivados de ambas enfermedades tienden a exacerbarse, generando un impacto conjunto en la salud mental y física de los individuos afectados. Luego del análisis epidemiológico, el artículo establece las bases biológicas que gobiernan el complejo impacto de la depresión en las enfermedades cardiovasculares, destacando los 4 factores predominantes: la alteración de la variabilidad de la frecuencia cardiaca, la inflamación crónica sistémica, la disfunción del eje hipotalámico-pituitario-adrenal y la disfunción endotelial. A modo de conclusión, el articulo presenta que el tratamiento clínico antidepresivo en pacientes geriátricos con ECV es una medida fundamental con el objetivo de disminuir la mortalidad y la morbilidad, fomentando un modelo biopsicosocial que mejore la calidad vida, aumente la esperanza de vida y reduzca los costos médicos.

1.2.5 Depresión en pacientes cardiovasculares en las poblaciones de Medio Oriente

- ♣ Donnelly, T. T., Al Suwaidi, J. M., Alqahtani, A., Assad, N., Qader, N. A., Singh, R., Fung, T. S., Mueed, I., El Banna, N., & Sharara, S. (2013). Depression in cardiovascular patients in the Middle Eastern populations [Letter]. *International Journal of Cardiology*, *168*(5), 5110-5111. https://doi.org/10.1016/j.ijcard.2013.07.242

Este estudio se llevó a cabo mediante una revisión bibliográfica de 60 artículos, estableciendo un equilibrio en la evaluación de la prevalencia de la depresión en pacientes con E.C.V en las poblaciones de Medio Oriente. Se subraya que, en las últimas tres décadas, la tasa de mortalidad por cardiopatía isquémica ha experimentado un incremento notable del 146% en mujeres y del 174% en hombres. Además de esta patología, los *I.A.M* y el síndrome coronario agudo afectan a una población más joven en Medio Oriente en comparación con las poblaciones de América del Norte y Europa. En este contexto, el artículo revela que la depresión ha demostrado ser un fuerte predictor de eventos cardiovasculares y que contribuye al aumento de la morbilidad y mortalidad en estos pacientes.

De los 60 estudios sometidos a revisión, 30 se enfocaron en investigar la coexistencia de enfermedades cardiovasculares y depresión en la región de Medio Oriente. Por su parte, 8 estudios indagaron en esta misma relación en poblaciones de Medio Oriente que residen en el extranjero, mientras que 14 estudios se centraron en examinar los mecanismos biológicos que incrementan el riesgo cardiovascular en pacientes con depresión. Los hallazgos revelaron que la incidencia de

depresión entre pacientes con enfermedades cardiovasculares y condiciones crónicas osciló entre un 4.5% y un 66%. Se enfatizó que los pacientes de edad avanzada presentaron una prevalencia considerable de enfermedades cerebrovasculares y diabetes en conjunto con la depresión, alcanzando un 56% y un 59%, respectivamente. Asimismo, se destacó en diez de los estudios que las mujeres mostraron una mayor prevalencia y gravedad de la depresión en comparación con los hombres. En referencia a los pacientes originarios de Medio Oriente que residían en el extranjero, se identificó una mayor incidencia de depresión en comparación con los pacientes de ascendencia caucásica. Un estudio que involucró a 321 pacientes árabes con E.C.V residentes en Estados Unidos reveló que un 29% de estos presentaban síntomas de depresión. Otro caso, que incluyó a 55 pacientes de origen turco y persa con E.C.V residentes en Australia, señaló que la mitad de las mujeres presentaban síntomas de depresión. En relación con los estudios que pusieron énfasis en los factores que influyen en la relación entre la depresión y las E.C.V en esta población, se identificaron cambios fisiológicos y biológicos compartidos entre ambas enfermedades que establecen dicha conexión. Asimismo, se observó que factores como la edad, el género, niveles educativos más bajos, ingresos limitados, una disminución en las redes de apoyo, pertenecer a grupos minoritarios, tener habilidades lingüísticas limitadas, acceso insuficiente a servicios médicos y enfrentar estigmatización social en relación con los problemas de salud mental también desempeñan un papel influyente en esta relación.

Se concluye este artículo destacando la importancia de un manejo no farmacológico de la depresión para los pacientes con ECV, que contemple un enfoque multidisciplinario y enfocado en la cultura de medio oriente, controlando la dieta, entrenamiento físico, asesoramiento educativo y psicosocial, enfatizando que con este tratamiento se reduce la mortalidad cardíaca en más de un 25%.

1.2.6 Fusión de minería de datos, aprendizaje automático y estadísticas tradicionales para detectar biomarcadores asociados con la depresión.

♣ Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., & Meyer, D. (2016). Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression. *PLOS ONE*, *11*(2), e0148195. https://doi.org/10.1371/journal.pone.0148195

El presente artículo, destaca la importancia que han tenido las técnicas de aprendizaje automático aplicadas a grandes bases de datos, en un comienzo aplicadas principalmente en

marketing, debido a que presentan modelos altamente predictivos que han permitido ahorrar costos a las empresas. Sin embargo, actualmente se destaca el rol relevante que ha tomado en neurociencia, biomedicina y recientemente en psiquiatría. Es en este último campo donde centra su atención este estudio y propone un método hibrido que permite identificar con precisión los biomarcadores relevantes que se asocien con la depresión para un estudio con aproximadamente 10000 civiles estadounidenses, con edades comprendidas entre 18 y 80 años, procedentes de diferentes géneros y razas.

El método híbrido, consta de técnicas de minería de datos y técnicas estadísticas tradicionales que permitan reconocer los biomarcadores más relevantes para estudiar la asociación entre la depresión y un total de 67 biomarcadores. Con el objetivo de mejorar la predicción del modelo planteado, se eliminaron los participantes que tuvieran datos faltantes para los 67 biomarcadores incluidos, dejando una muestra final de 5227 participantes para la investigación. El trabajo realizado, consta de mejorar la selección de variables para utilizar en el modelo, esto por medio de 3 pasos. El primer paso corresponde a la imputación múltiple de datos faltantes, es aplicada cuando los datos faltantes no son aleatorios, y puede llegar a generar 50 conjuntos datos imputados con el objetivo de obtener resultados estables, aunque se destaca que, si los datos faltantes no son demasiados, se recomienda imputar entre 5 a 20 conjuntos de datos. El segundo paso consta de la técnica de selección inicial de predictores, el enfoque de este análisis es seleccionar predictores en función de su importancia relativa en el conjunto de datos original, se incluyen los biomarcadores que tengan una importancia relativa superior o una importancia promedio en los datos imputados del 2 %, se concluye esta etapa con 21 biomarcadores a considerar en el siguiente paso. El tercer paso consta de la regresión estadística tradicional, que divide los datos en entrenamiento y prueba, para así evitar el sobreajuste en el modelo, obteniendo luego de este paso tan sólo 6 biomarcadores con una relación significativa con la depresión, estos son: Hemoglobina, ancho de distribución de los glóbulos rojos, cadmio en sangre, cotonina, bilirrubina total y glucosa.

Los resultados del estudio concluyen en una mejora en la precisión para la selección de variables, reduciendo desde 67 a 6 los biomarcadores que se relacionan con depresión, y destacando principalmente la distribución de los glóbulos rojos, glucosa sérica y bilirrubina total, resultados que son ampliamente concordantes con la literatura actual.

1.2.7 Técnicas de Machine Learning en medicina cardiovascular

- De la Hoz Manotas, A. K., Martínez-Palacio, U. J., & Mendoza-Palechor, F. E. (2013). Técnicas de ML en medicina cardiovascular. *Memorias*, *11*(20), 41-46.

El objetivo principal de este artículo fue llevar a cabo una comparación entre diversas técnicas avanzadas de aprendizaje supervisado en términos de su precisión en la predicción de enfermedades cardiovasculares. Esta evaluación se realizó utilizando una base de datos de Cleveland, la cual está disponible en el repositorio de aprendizaje automático UCI. Esta base de datos consta de 14 atributos y un total de 303 registros. En primer lugar, el trabajo define las técnicas a utilizar, partiendo por el árbol de decisión, este es un algoritmo que en forma gráfica se representa por nodos, ramas y hojas, las que permiten una descripción narrativa de un problema mediante la observación de la homogeneidad entre sus variables. Consta de dos etapas, la primera consiste en su construcción mediante datos de entrenamientos repartidos en los nodos de acuerdo con los valores de los atributos y la segunda etapa es la de clasificación, aquí los atributos nuevos son clasificados mediante un recorrido desde el nodo raíz hasta la hoja. El segundo algoritmo es *Maquina de soporte vectorial (SVM)*, el cual aprende la frontera de decisión de dos clases distintas, lo que le permite resolver problemas de regresión y clasificación de manera muy eficiente, hallando una muy buena separación entre clases, además se afirma que este algoritmo es especialmente útil cuando hay pocos datos de entrenamiento. El tercer algoritmo corresponde al modelo de regresión logística, el cual permite conocer la relación entre variables dependientes cualitativas, binarias, con más de dos valores; una o más variables independientes, ya sean estás, cualitativas o cuantitativas. Sus objetivos primordiales son modelar la influencia de las variables en la probabilidad de ocurrencia de un suceso particular y describir la relación entre las variables de respuesta y las variables regresoras.

El trabajo de comparación consistió en aplicar los modelos a la base de datos de pacientes cardiovasculares, en los tres modelos se distribuyeron los datos en 60% para entrenamiento, 20% para validación y un 20% de los datos para prueba. Los resultados se cuantificaron mediante la aplicación de la matriz de confusión y la exactitud obtenida por cada modelo. La regresión lineal consiguió una precisión del 85.15% para los problemas de clasificación, el algoritmo SVM por su parte, obtuvo un valor de precisión del 82.17%, y en último lugar, los árboles de decisión que obtuvieron una precisión del 76.56%. Estos datos comparativos son ilustrados en la **TABLA A.1** y Fig. B.1.

1.2.8 Asociación entre depresión y mortalidad en adultos mayores

- ♣ R. Schulz, S. R. Beach, D. G. Ives, L. M. Martire, A. A. Ariyo and W. J. Kop Archives of Internal Medicine 2000 Vol. 160 Issue 12 Pages 1761 DOI: 10.1001/archinte.160.12.1761

El presente estudio es del año 2000, está basado en la base de datos CHS, posicionada en esa época como la más amplia gama de variables de control físicas y demográficas jamás estudiadas, esta base de datos cuenta con historial médico para 5201 hombres y mujeres mayores de 65 años. El estudio planteó como objetivo cuantificar la relación entre los síntomas depresivos con la mortalidad al cabo de 6 años, de esta forma, aspiró a resolver las inconsistencias respecto a la asociación entre depresión y riesgo de muerte para adultos mayores, poco dilucidada hasta aquel momento.

Se presenta una completa descripción estadística de la base de datos, mostrada en la **TABLA A.2**. Para analizar el vínculo entre la depresión y la mortalidad durante 6 años, se utilizó regresión de Cox, aplicada a 5 modelos : (1) modelo sociodemográfico, controlando la edad, sexo, raza, nivel educativo, estado civil y los eventos estresantes de la vida; (2) Modelo de enfermedad prevalente, que controla las enfermedades clínicas medibles; (3) Modelo de enfermedad subclínica, que controla las variables como claudicación, estenosis carotidea o anomalías cardiográficas; (4) Modelo de factores de riesgo, este modelo contempla la biología del paciente y sus conductas, tales como el tabaquismo, índice de masa corporal o niveles de glucosa en ayunas; (5) modelo combinado, este modelo contempla todas las variables presentes en los primeros cuatro modelos excepto la depresión. Cabe destacar que, en la realización de estos modelos vario el número de participantes, producto que solo se utilizaron pacientes con un registro completo de las variables a utilizar en el modelo, variando de 5173 participantes en el modelo sociodemográfico a 4710 en el modelo combinado. Es importante notar que los distintos modelos planteados dotan al estudio de la capacidad de relacionar la depresión con la mortalidad para diferentes conjuntos de variables.

En relación con los resultados obtenidos, se observa que, tras un período de 6 años, el 18.9% de la muestra experimentó fallecimientos. De acuerdo con el modelo sociodemográfico (1), se evidencia que los participantes de género masculino, con niveles educativos más bajos y que se encuentran viudos o separados, presentan un mayor riesgo de mortalidad. Además, al considerar una puntuación elevada en la prueba de depresión, se identifica un aumento del 43% en el riesgo de mortalidad en comparación con aquellos cuya puntuación en la prueba de depresión es baja. La incorporación del modelo de enfermedad prevalente (2) junto con las variables sociodemográficas

muestra que las personas con síntomas depresivos significativos tienen un 25% más de probabilidad de fallecer en comparación con aquellas que presentan síntomas depresivos mínimos. Este aumento del riesgo de mortalidad en relación con síntomas depresivos altos también se confirma en los modelos (3) y (4), así como en las combinaciones de modelos (5). En consecuencia, a través de estos modelos se concluye que una puntuación inicial elevada en la prueba de depresión constituye un factor independiente de riesgo de mortalidad durante los próximos 6 años, con un rango de aumento en el riesgo que va del 25% al 43% en comparación con individuos con puntuaciones bajas en la prueba de depresión. Vale la pena resaltar los resultados de este estudio en contraste con aquellos que no encuentran una asociación entre la depresión y la mortalidad. Esto se atribuye al uso de una medida objetiva de depresión en la prueba, en contraposición con muchas investigaciones que se basan en pruebas de depresión auto reportadas por los pacientes.

1.2.9 Los modelos de aprendizaje automático en los registros de salud electrónicos pueden superar a los modelos de supervivencia convencionales para predecir la mortalidad de los pacientes con enfermedad de las arterias coronarias.

♣ Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLOS ONE, 13*(8), e0202344. https://doi.org/10.1371/journal.pone.0202344

Este articulo evalúa la eficacia en el desarrollo de distintos modelos de aprendizaje automático para estudiar la supervivencia luego de 5 años en pacientes con CAD, desarrollados en base a un *registro de salud electrónico (EHR)* con una cohorte de 80.000 pacientes procedentes del programa CALIBER, programa que vincula 4 fuentes de datos de salud electrónicos de Inglaterra.

El estudio afirma, mediante un metaanálisis realizado, que la mediana de variables seleccionadas por expertos para el desarrollo convencional de modelos de aprendizaje automático es de 27 variables. Esta cantidad se utiliza para desarrollar el modelo convencional en el presente estudio y fue comparada con una cantidad de 586 variables, para el denominado modelo de variables extendidas. Se utilizó el índice de concordancia de Harrel (índice-C) y la calibración para comparar el modelo convencional con el modelo de variables extendidas para 3 modelos, estos son: el modelo de Cox, *Random Forest (R.F)* Y Red elástica, alterando la cantidad de variables y su preprocesamiento

para los modelos y dando por resultado la información entregada por la Fig. B.2.

Los resultados muestran que la predicción tiene un buen desempeño en todos los modelos, lo que sugiere que no es necesario una imputación exhaustiva de los datos en EHR, ahorrando así tiempo al investigador, que podría utilizarse para examinar los efectos de las variables omitidas en el modelo. Por otra parte, la transformación de datos continuos a categóricos no altero los resultados para el modelo con RF, por lo que no existe necesidad de normalizar y transformar manualmente los datos para su posterior análisis. En cuanto a la comparación de modelos, RF no supero al modelo de Cox, esto se atribuye a que la cantidad de variables no informativas presentes en el modelo extendido pueden empeorar su desempeño, por su parte, las redes elásticas lograron el mejor desempeño en discriminación, dado que dicho modelo selecciona las variables más relevantes. El último punto por destacar es en relación con la selección de variables, donde RF tuvo un desempeño peor en este aspecto, dada la utilización del modelo de variables extendidas que añade ruido al modelo, por otro lado, al utilizar solo las variables sin datos faltantes, el desempeño no es tan bueno como el desempeño obtenido con la selección experta de variables.

1.2.10 Discusión

De manera general, la revisión bibliográfica permitió vislumbrar la importancia e incertidumbre existente respecto a los mecanismos que asocian la ECV con la depresión, situando el conocimiento de predictores de ambas patologías como uno de los puntos centrales para la investigación actual, desde donde se obtenga conocimiento del riesgo asociado al tener ambas patologías y saber cuáles son las variables inherentes a ambas patologías que interactúan entre sí. La literatura revisada destaca que ambas enfermedades están muy presentes en la sociedad actual, siendo las enfermedades cardiovasculares la principal causa de muerte a nivel mundial, y la depresión la principal causa de discapacidad a nivel mundial, representando la depresión como un índice de muerte prematura para todas las causas.

Se destaca que los factores que intervienen en la depresión son en un 63% relacionado con el entorno y la sociedad, mientras que un 37% son de influencia hereditaria. Además, se detalla cómo los aspectos psicosociales potencian tanto la incidencia, como la gravedad de las enfermedades cardíacas. Estos aspectos psicosociales incluyen la escasa red de apoyo social, la limitada interconexión interpersonal, el estado civil y una integración social deficiente. Es necesario destacar

que la presencia de depresión muestra una mayor capacidad predictiva para el desarrollo de IAM en comparación con factores como el tabaquismo, la obesidad, la diabetes o la hipertensión. Además, se sitúa la tasa de prevalencia de la depresión en pacientes con IAM en un 19.8%. Por otro lado, en un artículo basado en el estudio CHS se destaca que la depresión posee un papel autónomo en el riesgo de mortalidad a 6 años, incrementándose este índice de 25% a 43% en comparación con individuos que presentan niveles bajos de depresión. La revisión resalta el uso de modelos predictivos en neurociencias, medicina y psiquiatría dado su destacado rendimiento. Específicamente, se destaca la continua utilización del algoritmo Random Forest en numerosos artículos relacionados con bases de datos médicas. Se subraya cómo la imputación de datos faltantes y selección de variables para el modelo RF mejoró ampliamente su rendimiento, asimismo, se descubrió que en la transformación de datos continuos a categóricos no modifico el rendimiento en RF. Consecuentemente, RF se presenta como un modelo para ser utilizado en predicciones en la base datos CHS. A pesar de que se indique que los mejores resultados se han obtenido con RF, los estudios revisados no exploran otros modelos avanzados de inteligencia artificial, por lo que es necesario en este trabajo analizar el comportamiento de otros algoritmos que permita añadir solidez a los resultados, como, por ejemplo, modelos basados en redes neuronales artificiales.

1.3 Definición del Problema

Se busca desarrollar un modelo de aprendizaje supervisado para predecir eventos cardiovasculares de la base de datos CHS a partir de variables sociodemográficas, clínicas y de depresión. Además, se realizará la predicción de depresión a partir de variables sociodemográficas, clínicas y cardiovasculares. Finalmente, se buscará esclarecer la relación entre la depresión y eventos cardiovasculares mediante un análisis de importancia de características de los modelos.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un modelo de aprendizaje supervisado que permita predecir eventos cardiovasculares utilizando variables sociodemográficas, clínicas y de depresión contenidas en la base de datos CHS, incluyendo un análisis sobre la influencia de la depresión en los eventos cardiovasculares predichos.

1.4.2 Objetivos específicos

- Identificar las variables sociodemográficas, clínicas y psicológicas asociadas a enfermedad cardiovascular, así como las variables sociodemográficas, clínicas y cardiovasculares asociadas a depresión en los adultos mayores de la base de datos CHS.
- Realizar un preprocesamiento de las variables a utilizar en la base de datos.
- Implementar una selección de variables óptima.
- Diseñar y desarrollar un modelo de predicción con los algoritmos Random Forest y Red Neuronal Artificial, que permita estudiar la relación entre las variables asociadas a depresión y ECV.
- Identificar y evaluar el rendimiento de los algoritmos Random Forest y Red Neuronal Artificial en el estudio de la asociación entre las variables depresión y enfermedad cardiovascular.
- Visualizar los resultados obtenidos de las variables predictoras que se relacionan con enfermedades cardiovasculares y la depresión.

1.5 Alcances y Limitaciones

Se espera desarrollar un modelo de aprendizaje automático supervisado, que permita encontrar variables asociadas a la depresión y a enfermedades cardiovasculares, buscando conocer si estás variables inherentes a la depresión, son o no predictoras de enfermedad cardiovascular, y viceversa, si las variables predictoras de ECV, son o no predictoras de depresión. Se busca así, cuantificar su relación mediante el uso de distintos algoritmos de aprendizaje automático, evaluando posteriormente las predicciones de cada algoritmo con la herramienta matriz de confusión y curvas ROC-AUC que permitirán evaluar su desempeño.

1.6 Temario y Metodología

A continuación, se presenta un esquema general de los tópicos incluidos en cada capítulo.

- **Capítulo 1. Introducción:** Los tópicos tratados en este capítulo buscan en primera instancia, establecer el panorama general de las enfermades cardiovasculares y la depresión, basadas en la literatura médica actual referente a estas enfermedades. Adicionalmente, se estudia lo publicado

con respecto a la relación que existe entre ambas patologías. En segunda instancia, se busca ahondar modelos de aprendizaje automático que se hayan utilizado para determinar la relación entre las patologías . El último punto tratado es el estudio de la base de datos CHS, como se conforma, variables que contiene, su descripción estadística y la incidencia de depresión en la población.

- **Capítulo 2. Marco Teórico:** El capítulo dos consta de una revisión del material teórico a utilizar para el desarrollo del modelo de aprendizaje automático. En primer lugar, se estudian las ECV frecuentes en CHS. Luego se estudian los procesos de selección de características, preprocesamiento de las variables, así como también los algoritmos a implementar y los métodos para su evaluación.
- **Capítulo 3. Desarrollo del modelo:** El capítulo número tres denota el desarrollo del modelo. Primero se explora la base de datos y se seleccionan los archivos para trabajar. Luego, se realiza un preprocesamiento de los archivos, con el fin de preparar el set de datos para la selección de características y luego, explica el procedimiento de transformación de variables categóricas a binarias en preparación para el proceso de clasificación. Finalmente, se explica el proceso de entrenamiento de los modelos, para obtener variables predictoras que definan una relación entre la depresión y los eventos cardiacos de IAM, ACV y ICC.
- **Capítulo 4. Resultados:** El capítulo cuatro contiene los resultados del desarrollo del modelo, se evalúa el rendimiento de los algoritmos mediante curva de aprendizaje, curva ROC-AUC y métricas derivadas de la matriz de confusión. Adicionalmente, se estudian las variables predictoras encontradas por los algoritmos individualmente, así como la intersección entre algoritmos.
- **Capítulo 5. Conclusión, Discusión y Trabajos Futuros:** Este capítulo resume los resultados obtenidos y como estos dan respuesta a los objetivos del trabajo. Se finaliza con la propuesta de trabajos futuros.

Capítulo 2. Marco Teórico

2.1 Introducción

Se realizará un marco teórico previo al desarrollo del modelo. En primer lugar, se realizará una caracterización de la base de datos CHS, que incluirá la explicación de las principales enfermedades cardiovasculares en la población bajo estudio. En segundo lugar, se dividirá el modelo de predicción a implementar en sus unidades básicas funcionales, abarcando métodos matemáticos, estadísticos y los modelos de aprendizaje automático junto con los métodos de evaluación utilizados en Python. La subdivisión de temas será; base de datos CHS, ECV más frecuentes en CHS, análisis exploratorio de los datos, preprocesamiento, selección de características, algoritmos de predicción y evaluación de los algoritmos.

2.2 Base de Datos cardiovascular Health Study (CHS)

La base de datos CHS realiza un seguimiento por 24 años a más de 5000 pacientes adultos mayores con ECV, esta base de datos alberga variables sociodemográficas, enfermedades clínicas prevalentes, indicadores de enfermedad subclínica y también factores de riesgo biológicos o conductuales. Existen numerosos estudios en base a CHS, algunos de estos buscan relacionar la depresión con el riesgo de mortalidad para esta población y otorgan un análisis estadístico descriptivo de los datos de CHS, el artículo estudiado en el capítulo 1 nos entrega como conocimiento cuantitativo la información de la **TABLA A.2**, con descripción de las variables y sus estadísticos descriptivos.

Las enfermedades clínicas prevalentes fueron controladas para un total de 5201 pacientes, entre estas enfermedades se encuentra: infarto agudo de miocardio, angina de pecho, insuficiencia cardiaca congestiva, claudicación intermitente, derrame cerebral, ataque isquémico transitorio, diabetes e hipertensión. Mediante los estadísticos descriptivos, se señala como la ECV con más prevalencia a la hipertensión, con un total de 2138 pacientes, lo que es equivalente a un 41.1% del total, seguida por la diabetes con un total de 1173 casos, es decir un 22.6%, la angina de pecho con 814 casos, equivalente a un 15.7% de la cohorte, seguida por el infarto agudo al miocardio con 504 casos, equivalente a un 9.7% , la insuficiencia cardiaca congestiva con un 4.2%, derrame cerebral con un 3.7% y por último, claudicación intermitente y ataque isquémico transitorio, presentes en un 2.6% y 2.5% respectivamente.

2.3 Enfermedades cardiovasculares más frecuentes en CHS

2.3.1 Hipertensión arterial

La ECV con mayor presencia en la cohorte de CHS es la hipertensión, afectando a 2138 participantes, es decir 41,2%. Mientras que otros 774 pacientes, que es un 14.9% del total, se encuentran diagnosticados como hipertensos limítrofes.

La hipertensión arterial es diagnosticada al tener una presión arterial sistólica sobre 130 mmHg y una presión diastólica sobre 80 mmHg, por tiempos prolongados. Esta elevada presión arterial indica que las paredes de las arterias y el corazón están más dilatadas, lo que obstruye el flujo de la sangre a través de este sistema y obliga al corazón a latir con más fuerza. Las arterias de los pacientes sanos son flexibles, fuertes y elásticas, su revestimiento interior es liso y permite con esto que la sangre fluya libremente suministrando los nutrientes y oxígeno a los órganos vitales, en cambio, unas arterias con altas presiones aumentan el daño y estrechamiento de las arterias, dañando su revestimiento interno, que facilita la acumulación de grasas en las arterias dañadas, además, se vuelven menos elásticas por lo que es limitada su función [9].

Una hipertensión arterial no tratada durante años puede acabar generando discapacidad, ataques cardiacos o accidentes cerebrovasculares. Esto, dado que los daños a las arterias son causantes de próximos daños en el corazón, predisponiendo a enfermades de las arterias coronarias y/o agrandamiento del ventrículo izquierdo. También, afecta al cerebro, dado que no otorga una sangre nutritiva a este órgano, ocasionando accidente isquémico transitorio y accidente cerebrovascular. Además, produce daños a los riñones, dado que la función de estos es en grandes rasgos, filtrar los desechos de la sangre, y producto de la alta tensión se pueden dañar sus vasos sanguíneos, produciendo cicatrices o insuficiencia renal. Otros vasos sanguíneos que pueden ser dañados por la hipertensión, son los vasos sanguíneos de los ojos y esto puede producir daños en los nervios, que pueden ocasionar incluso perdida de la visión.

2.3.2 Angina de pecho

La angina de pecho fue la segunda ECV más usual, tuvo 814 casos en la base de datos, lo que equivale a un 15.7% de los participantes. La angina de pecho es un dolor característico en el centro del pecho, descrito como constrictivo, una presión o pesadez. Este dolor se desencadena debido a

estrés psíquico o físico elevado. Los mecanismos biológicos que rigen esta enfermedad hacen referencia a alteraciones en las arterias coronarias que llevan la sangre al músculo cardíaco las cuales se estrechan y no permiten la irrigación de sangre al órgano cardíaco [9].

La angina de pecho es detectada mediante diversos exámenes como electrocardiograma, ecocardiograma, pruebas de esfuerzo, tomografía computarizada, angiografía coronaria o resonancia magnética. Los factores de riesgo para tener este daño en las arterias coronarias son variados y entre estos se encuentra ser mayor de 60 años, tener diabetes, hipertensión, estrés emocional, tabaquismo, abuso de drogas, entre otros. Su tratamiento es en base a medicamentos, procedimientos de angioplastia con colocación de stent, la cual es una cirugía invasiva que busca introducir un globo con una pieza metálica (stent) que al ser expandido fija el stent en la arteria coronaria, expandiéndola, obteniendo por resultado un mejor flujo sanguíneo. Para corregir esta enfermedad, también se realizan cirugías a corazón abierto. En cuanto a la prevención para evitar su aparición se destaca hacer ejercicio, controlar el peso, tener una dieta saludable, aliviar el estrés, evitar fumar y consumir alcohol [11].

2.3.3 Infarto agudo al miocardio

El tercer evento cardiovascular más frecuente en CHS es el infarto agudo al miocardio, afectando en 504 casos, es decir, una prevalencia de 9.7% en la cohorte. Este evento, se produce cuando un coagulo sanguíneo obstruye completamente el paso de la sangre por alguna de las arterias coronarias, ocasionando la muerte de partes del corazón debido a la ausencia de oxígeno. El principal tratamiento para este infarto es realizar rápidamente cateterismo que busque desobstruir la arteria o la administración de fármacos anticoagulantes para disolver el coagulo. Los principales factores de riesgo para sufrir este tipo de eventos es el envejecimiento, una presión arterial alta, alto colesterol, tabaquismo, diabetes, obesidad, infección por COVID-19, malos hábitos alimenticios y poca actividad física [9, 12].

2.3.4 Insuficiencia cardiaca congestiva

El cuarto evento cardiovascular más frecuente es la insuficiencia cardiaca congestiva, que afectó a 217 pacientes, un 4.2% de la población total. Esta afección ocurre, cuando el corazón no es capaz de bombear la sangre con la fuerza necesaria para llegar a todo el organismo y provoca que se acumulen líquidos en los pulmones, esto se denomina congestión. Además, el corazón aumenta su

tamaño debido a la insuficiencia cardiaca. Las causas de este evento suelen ser, haber sufrido un infarto cardiaco previamente y tener hipertensión. Los síntomas que presentan los pacientes son la falta de aliento durante actividades simples como estar recostados, fatiga o debilidad, hinchazón en el tren inferior, arritmias, menor capacidad física y dolor en el pecho. Mientras que su tratamiento requiere control de por vida, puede requerir un trasplante de corazón, reparación de las válvulas cardiacas, control del elevado ritmo cardiaco o un tratamiento de las causas subyacentes [9, 13].

2.4 Modelo de Aprendizaje automático para predicción

El objetivo central de esta sección es definir y describir los métodos que se utilizan en el diseño y desarrollo del modelo de predicción. Primero, se realizará un análisis exploratorio de los datos a utilizar. Este análisis constará de una descripción gráfica de las variables que contienen la información de depresión y enfermedades cardiovasculares. Seguido a esto, se llevará a cabo la selección de características mediante el método del filtro con estadístico Chi-Cuadrado. Posteriormente, se describirá el preprocesamiento del dataframe. A continuación, se estudiarán los algoritmos Random Forest y Red Neuronal Artificial que realizarán la predicción. También se analizarán los mecanismos de evaluación del rendimiento y generalización que poseen los algoritmos. Esto incluye la utilización de la matriz de confusión y la curva ROC-AUC.

Es importante destacar que todo el modelo de aprendizaje automático a realizar será efectuado en el *entorno de desarrollo integrado (I.D.E)* Spyder en el lenguaje de programación Python, este es un lenguaje interpretado y orientado a objetos, ideal para el desarrollo de proyectos del reconocimiento de patrones, dado que permite recolectar y limpiar datos además de explorarlos, modelarlos y visualizarlos. Para esto, Python posee numerosas librerías especializadas para aprendizaje automático, dentro de las cuales serán utilizadas Scikit-Learn, Scipy, Tensor Flow, Keras, Pandas y Seaborn, cada una con diversas utilidades, desde gráficas, estadísticas y matemáticas [15].

2.4.1 Análisis exploratorio de los datos y selección de características

El análisis exploratorio de los datos es fundamental para reconocer patrones, dado que sentará las bases del conocimiento del conjunto de datos a investigar, esto se logrará resumiendo sus principales características, usualmente usando métodos de visualización de los datos tales como gráficos de barras e histogramas. Esta idea de una comprensión de los datos y sus relaciones previa al procesamiento permite determinar si las técnicas a implementar son apropiadas al conjunto de datos,

fue planteada en los años 70 por el matemático estadounidense John Tukey y es utilizada ampliamente en la ciencia de datos actual [16].

Python posee una librería llamada Seaborn la cual está basada en la librería Matplotlib. Seaborn proporciona una visualización de datos estadísticos de alto nivel a través de gráficos atractivos e informativos, que en conjunto con Matplotlib permitirá conocer las variables de la base de datos.

La selección de características es una parte fundamental en el desarrollo de un modelo, en esta etapa se selecciona de una gran cantidad de variables disponibles solo aquellas que se califiquen como las mejores características, buscando de esta manera reducir la dimensionalidad del problema. Sin embargo, se deben tener algunas consideraciones, deben seleccionarse una cantidad de variables lo suficientemente grandes para distinguir las diferencias entre clases y las similitudes entre objetos de la misma clase, y lo suficientemente pequeña para evitar errores de generalización. La calidad del vector de características debe contemplar una buena discriminación entre clases distintas, mientras que para características de la misma clase deben presentar valores similares, en otras palabras, al buscar seleccionar características se tiene que encontrar una separabilidad inter-clase lo más distante posible, y pequeñas varianzas intra-clases, tal como ilustra la Fig. 2.1. La medida de separabilidad hace referencia a la medición de distancias según diversos factores de medición, algunas de estas medidas son el cálculo de la distancia Euclidiana, distancia City Block y distancia Mahalanobis, cada uno utilizando medidas estadísticas diferentes: media, desviación estándar y covarianza, respectivamente [17].

Existen diversos métodos para la selección de características, tal como el análisis univariable, el cual no toma en cuenta la correlación entre variables. El análisis multivariable que examina conjuntamente como vectores de características, y esto le permite testear separabilidad entre clases y se basa en métodos de distancia como los señalados anteriormente. El método wrapper el cual es un proceso costoso computacionalmente que evalúa todos los posibles subsets utilizando un clasificador y ten-fold cross-validation. Por último, se tiene el método del filtrado, el cuál rankea las características individualmente según un criterio univariable, para generar un subset de características con mejores puntajes para el criterio seleccionado, un ejemplo es un filtro de correlación con el estadístico Chi-cuadrado [17].

La prueba Chi-cuadrado es una prueba de independencia entre variables categóricas. El objetivo de la prueba es determinar si dos variables se relacionan o no entre sí, para esto se requiere definir una hipótesis nula, la cual afirma que no existe relación entre las variables, y también una hipótesis alternativa, que afirma que existe una relación significativa entre las variables. La prueba se basa en comparar las frecuencias observadas en los datos con las frecuencias esperadas si las variables fueran independientes. Como resultado, se obtiene un valor chi-cuadrado que indica la discrepancia entre las frecuencias observadas y esperada. Cuanto mayor sea la discrepancia más evidencia habrá en contra de la hipótesis nula, y también se obtiene un valor p, que al ser menor que un umbral fijado previamente, usualmente de 0.05, se concluye que existe una asociación significativa entre las variables [33].

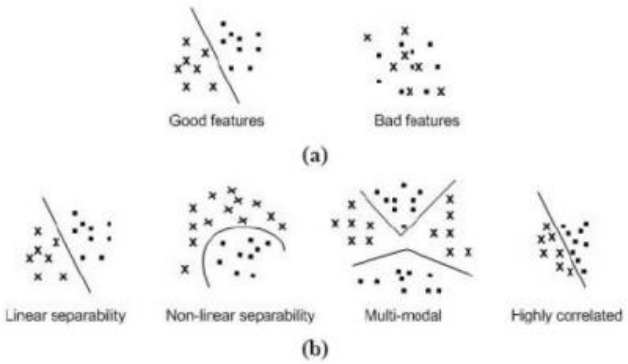

Fig. 2.1 Selección de características a) distinción entre buena y mala separabilidad inter-clase y b) tipos de separabilidad. Fuente: Reconocimiento de patrones y aprendizaje automático (2021).

2.4.2 Preprocesamiento

Luego de realizar el análisis exploratorio de los datos y selección de características para nuestro modelo de aprendizaje, es necesario limpiar las características seleccionadas para aumentar la precisión y coherencia, preparándolas para la aplicación de los algoritmos.

El primer paso es comprender el tipo de variable del que está conformado el vector de características, estas pueden ser variables ordinales discretas, enteras continuas, nominales discretas, variables binarias y variables continuas. Existen diversos problemas a solucionar en este proceso, los cuales son: i) valores faltantes, ii) datos ruidosos o duplicados, iii) datos no coincidentes, iv) un gran volumen de datos. Cada uno de estos problemas, tiene una causa y soluciones diversas, estás son detalladas a continuación:

i) Valores faltantes: es muy común debido al error humano tener problemas en el registro de la información, la base de datos a trabajar consta de 5000 pacientes aproximadamente, con múltiples datos auto informados y exámenes periódicos por 24 años, por lo que es de esperar que existan datos faltantes para algunas de las características a utilizar. Existen dos formas de abordar este problema, la primera consiste en la eliminación de las columnas con datos faltantes según un umbral óptimo, sin embargo, aplicar esto a todos los datos faltantes podría afectar a la clasificación del modelo de aprendizaje. La segunda opción consiste en realizar un remplazo de la información faltante, mediante un proceso denominado imputación, por medio del cual se puede remplazar por un estadístico descriptivo de aquellos datos faltantes. De manera similar, existe un tipo de imputación basada en el algoritmo K-Nearest Neighbors el cual se basa en la idea de calcular las distancias entre un numero K de vecinos más cercanos en el conjunto de datos y luego, con los valores para dicha característica de los vecinos más cercanos se realiza una imputación de los valores faltantes, remplazándose por este nuevo valor calculado [34].

ii) Datos ruidosos: son aquellos valores que pueden ser variables duplicadas, campos no relevantes para el análisis o un desbalanceo entre clases para una misma característica. Existen 3 posibles soluciones, la primera solución consiste en ajustar los datos en función de regresión múltiple o lineal, esta es especialmente útil para grandes conjuntos de datos. La segunda, consiste en eliminación de datos duplicados o no relevantes. La tercera se denomina remuestreo, es utilizada para un ajustar el balance entre clases distintas de una misma característica, permitiendo al algoritmo aprender de manera más precisa y evitando sobreajuste. Además, existe la técnica Clustering, la cual permite agrupar conjuntos de datos similares, tomando en cuenta los extremos de la agrupación.

iii) Datos no coincidentes: Esto suele suceder cuando los datos no se encuentran en el formato adecuado para ser recibidos por el algoritmo de aprendizaje automático, por lo que afectan su capacidad de clasificar o predecir tendencias. Se puede solucionar esto mediante normalización, discretización, casting, eliminación o generalización de los datos.

iv) Un gran volumen de datos: al tener en selección muchas características, existe una ralentización en los procesamientos realizados, necesitando realizar una representación más pequeña de los datos. Algunos de los métodos utilizados para este fin son: seleccionar

una cantidad más acotada de características, reducir la dimensionalidad y reducir la numeración en que se encuentran los datos [18].

2.4.3 Algoritmos por utilizar en el modelo

Existen numerosos algoritmos de clasificación para los modelos de aprendizaje automático, entre estos existen los no probabilísticos, como clasificador de mínima distancia, K-Nearest Neighbords, arboles de decisión y discriminantes lineales. Y también, los modelos probabilísticos, como clasificador bayesiano y regresión, entre una gran cantidad de otros algoritmos. Según la literatura revisada, y consultas con expertos, los modelos apropiados para trabajar con bases de datos médicas son Random Forest y Redes Neuronales Artificiales.

Uno de los algoritmos más utilizados para problemas de clasificación, es RF debido a su alto desempeño, este algoritmo está basado en los árboles de decisión. El árbol de decisión es un tipo de algoritmo supervisado de uso frecuente dada su fácil interpretación, este clasifica de acuerdo con una secuencia de preguntas en las que la siguiente pregunta depende de la respuesta a la pregunta actual. Se compone de nodos, ramas y hojas y es construido de arriba hacia abajo (top-down). Los nodos prueban cada característica, las ramas representan el valor de cada característica y las hojas corresponden a las clases en las que se clasifican los patrones, esto se puede visualizar en la Fig. 2.2. Pese a sus beneficios de representación, los árboles de decisión suelen sobre ajustarse, teniendo un buen desempeño en el set de entrenamiento y un peor desempeño en el set de prueba.

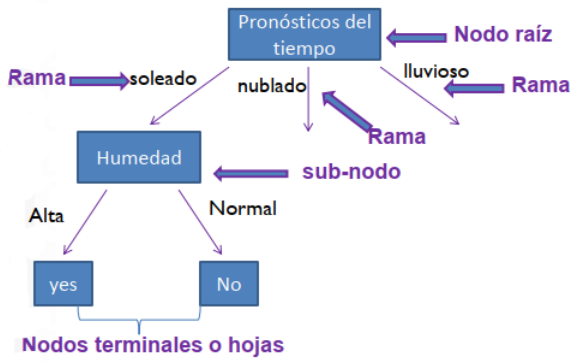

Fig. 2.2 Representación de un árbol de decisión. Fuente: Reconocimiento de patrones y aprendizaje automático (2021).

El clasificador RF, está formado por una gran cantidad de árboles de decisión, los cuales operan entre si como un conjunto. Cada árbol entrega su predicción de la clase, luego la clase mayormente votada por cada árbol individual se selecciona como la variable predicha. La fortaleza de este modelo consiste en que cada árbol de decisión individual tiene una baja correlación con otro, esto dota al modelo de predicciones que en conjunto que son más precisas que la predicción individual de cada árbol. RF se asegura la diversificación de los árboles individuales mediante dos métodos, el primer método, conocido como Bootstrap, aprovecha la sensibilidad de los árboles de decisión a las variaciones en los datos de entrenamiento, y dota a cada árbol para tomar muestras aleatorias de un conjunto de datos con remplazo, como resultado árbol de decisión dará resultados distintos. El segundo método, consiste en la aleatoriedad de características tomadas al momento de dividir un nodo, lo queda representado en la Fig. 2.3 [19].

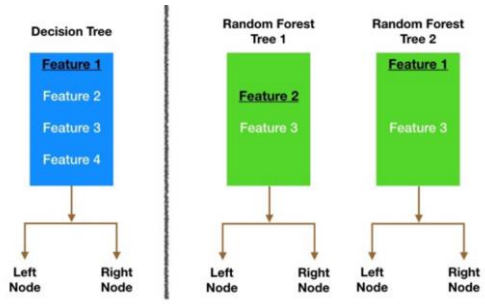

Fig. 2.3 División de nodos en un modelo de RF basada en subconjunto aleatorio. Fuente: Reconocimiento de patrones y aprendizaje automático (2021).

El algoritmo de red neuronal artificial es un modelo de aprendizaje profundo que se basa en simular el modo de funcionamiento del cerebro humano. Este algoritmo se conforma por una unidad básica denominada neurona, que es la unidad básica de aprendizaje. La interconexión de neuronas forma una red neuronal, la cual se conforma por lo tanto como una red de funciones interconectadas entre sí, transmitiendo información de una neurona a otra, dotando al modelo la capacidad de adaptarse a no linealidades. Esta interconexión se representa en el modelo de la Fig. 2.4, donde se observa la capa de entrada en color verde que recibe el valor bruto de las variables, la capa oculta en azul, que recibe los valores de la capa de entrada multiplicada por su peso y la capa de salida, que combina los valores que salen de la capa intermedia para generar la predicción [35].

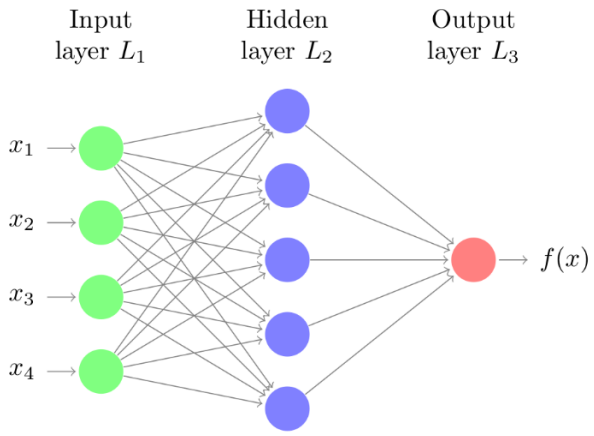

Fig. 2.4 Red Neuronal Artificial con capa de entrada, oculta y de salida. Fuente: Computer Age Statistical Inference 2016.

Dentro de una neurona ocurren dos operaciones: la suma ponderada de las entradas que recibe y la aplicación de la función de activación. De este modo, según se observa en la Fig. 2.5 el valor de entrada a una neurona es la suma de los valores que llegan x_i, multiplicado por el peso de las conexiones ω_i, sumándose el bias b. Esta sumatoria suele ser representada como un producto matricial, es decir, un vector X para valores de entrada y un vector W para pesos, aplicándose luego a la entrada la función de activación, dando por resultado la salida de la neurona, tal como muestra la siguiente ecuación matemática:

$$a = g(entrada) = g\left(\sum_{i=1}^{n} x_i\omega_i + b\right) = g(XW + b) \qquad (2)$$

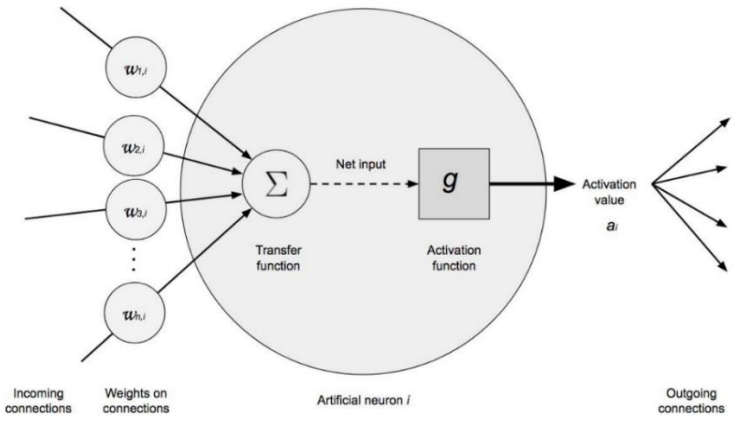

Fig 2.5 Representación de una neurona. Fuente: Deep Learning a Practitioner's Aproach, Josh Patterson and Adam Gibson.

Las funciones de activación se encargan de controlar la información que pasa de una capa a otra, convirtiendo el valor neto de entrada, en un valor nuevo. Es gracias a la combinación de funciones de activación, con múltiples capas en la red neuronal que los modelos pueden ser capaces de aprender relaciones no lineales. De esta forma, las funciones de activación convierten la entrada de una neurona en un rango entre (0,1) o (-1,1). Cuando el valor de activación de una neurona, es decir la salida de su función de activación, es cero, se dice que la neurona esta inactiva y no se propaga información por parte de esa neurona a las siguientes. Existen numerosas funciones de activación, tales como: Sigmoide, Tangente hiperbólica o la más utilizada es la función ReLu, mostrada en la Fig. B.6 y denotada por la ecuación (3), que trabaja asignando valor cero cuando la entrada esta debajo de cero, pero al ser superior a cero, el valor de la salida aumenta de forma lineal con la entrada [35].

$$ReLu(x) = \max(x, 0) \qquad (3)$$

2.4.4 Evaluación de los clasificadores

Comprobar los resultados entregados por los algoritmos utilizados es de vital importancia para la presentación de resultados fiables, debido a que, cada algoritmo presenta respuestas diferentes para el problema planteado. Dado lo anterior, es fundamental su evaluación comparativa y las medidas de desempeño basadas en la matriz de confusión serán implementadas con aquel objetivo. Además de este método, se utilizará también métodos gráficos como el análisis de curva ROC-AUC, y curvas de aprendizaje para el caso de Random Forest.

2.4.4.1 Matriz de confusión

La matriz de confusión es una herramienta utilizada ampliamente en la inteligencia artificial y el aprendizaje automático. Esta matriz permite comprender el desempeño de los algoritmos de aprendizaje supervisado, mediante la evaluación de los aciertos y errores en la clasificación. La configuración que utiliza la matriz es mostrada en la Fig. B.7 y se tiene para la primera columna de predicción, que VP representa el número de predicciones “positivas” que realmente son “positivas” y FP representa el número de predicciones “positivas” que realmente son “negativas”. Para la segunda columna de predicción, FN representa el número de predicciones “negativas” que realmente son “positivas” y VN representa el número de predicciones “negativas” que realmente son “negativas” [14]. La cantidad de VP, FP, FN Y VN es relevante dado que puede dar cuenta de diversas medidas

como exactitud, tasa de error, sensibilidad, especificidad, precisión y valor de predicción negativo, todas estas medidas calculadas como se muestra en el anexo TABLA A.3.

2.4.4.2 Curva de operación (ROC) y área bajo la curva (AUC)

El uso de curva ROC y su área bajo la curva permiten evaluar el rendimiento de un clasificador, mediante un mapeo de la razón entre Falsos Positivos (FPR) ubicada en el horizontal y la razón de Verdadero Positivos (TPR) en el eje vertical, en donde se define:

$$FPR = 1 - Especificidad \;\; y \;\; TPR = Sensibilidad$$

A partir de esto, se obtiene la curva ROC mostrada en Fig. B.8. Esta curva ROC, indica qué tan bueno es el modelo para distinguir entre dos clases. Se toma a todos los valores sobre la recta umbral como verdaderos positivos y los valores debajo del umbral como falsos positivos, es decir, valores predichos incorrectamente. Esta curva permite la comparación entre distintos algoritmos clasificadores para distintos rangos de operación, sin embargo, no permite cuantificar el desempeño general de éstos. Para esto, se utiliza el área bajo la curva ROC, que se denomina AUC y presenta una mayor robustes para comparar la precisión de la clasificación, su valorización se denota en la TABLA 2.1 [32].

Tabla 2.1 Valoración de AUC.

AUC	Desempeño
0.5-0.6	Sin discriminación
0.6-0.7	Malo
0.7-0.8	Regular
0.8-0.9	Bueno
0.9-1.0	Excelente

2.5 Discusión y Conclusiones

Este capítulo, consistió en la investigación del marco teórico necesario para el diseño y desarrollo de un algoritmo de aprendizaje automático que permita realizar predicciones realizando una clasificación de tipo binaria. El primer paso consiste en entender el problema, para esto se recurre a la explicación de los estadísticos descriptivos de ECV más frecuentes en la base de datos, los que fueron hipertensión arterial, angina de pecho, IAM e insuficiencia cardiaca congestiva, se evaluaron sus causas, tratamiento y pronóstico. Respecto a este inciso, cabe destacar la interacción que existe entre una ECV con otra, siendo la hipertensión un riesgo para la aparición de otras enfermedades. El segundo paso, consiste en una preparación de los datos, y se estudia como explorar la información de la base de datos. Se plantea que se debe realizar una exploración para estudiar las correlaciones entre las distintas características. Luego, para la selección de las características, se estudiaron los lineamientos generales que ayudarían a tener una baja dimensionalidad, buena separabilidad entre clases y pequeñas varianzas intra-clases, además de los métodos eficientes de selección de características.

Luego de la selección de características, se estudiaron los métodos adecuados para preprocesar los datos frente a distintas problemáticas, se podría considerar que existen opciones robustas, por lo que estos problemas podrían tener buena solución en el desarrollo del modelo a realizar. Un tercer paso para realizar el modelo es la elección de los algoritmos: Random Forest y Red Neuronal Artificial, los cuales fueron estudiados en cuanto a su funcionamiento. Por último, se estudiaron métodos de evaluación para los algoritmos, estos serán la matriz de confusión entregando medidas cuantitativas del desempeño de los clasificadores y gráficos mediante la curva ROC-AUC.

Capítulo 3. Desarrollo del modelo

3.1 Introducción

Este capítulo comprenderá la sección práctica desarrollada con la base de datos CHS y el lenguaje de programación Python. En primer lugar, se realiza una exploración de la base datos, identificando la información contenida por carpeta y los archivos utilizables para el desarrollo del algoritmo, así como también archivos que contengan información descriptiva de cada variable a estudiar. En segundo lugar, se describe el preprocesamiento aplicado a los datos faltantes de la base de datos. Luego, se detalla el diseño y entrenamiento de los algoritmos Random Forest y Red Neuronal artificial. Finalmente, se explica el proceso de evaluación del rendimiento de los modelos, que nos permitirá analizar la relación entre variables cardiovasculares y depresión a partir del análisis de importancia de características en la predicción. Con el objetivo de mejorar la comprensión del proceso completo, este se detalla en el diagrama de la Figura B.5 del anexo.

3.2 Exploración base de datos y selección de archivos.

La base de datos CHS está contenida en 35 carpetas, la información relativa a cada paciente se distribuye en archivos de tipo de documento de texto, CSV, DOC, PDF, JPG. Dada la cantidad de información, se condensó la exploración realizada de estos archivos en una tabla con el objetivo de conocer el contenido estructural de la base de datos, la tabla aludida anteriormente es la **TABLA A.3**.

La revisión sistemática de la base de datos permitió reconocer aquellos archivos .CSV que fueron seleccionados para el trabajo a realizar, dada la prevalencia de pacientes y gran cantidad de variables médicas, tales como: exámenes físicos, psicológicos, calidad de vida, apoyo social, exámenes de laboratorio y más, se seleccionó ‘basebothfinal.csv’, archivo que contiene información desde el inicio del estudio CHS. Así como también se seleccionó el archivo ‘events.csv’ el cual contiene información cardiovascular y la fatalidad de dichos eventos.

3.3 Preprocesamiento archivos.

3.3.1 Exploración de archivos seleccionados, tratamiento de archivos y su unión en un

nuevo dataframe.

Se procedió a cargar los archivos en el software Spyder utilizando el lenguaje Python 3.8. El archivo por trabajar events.csv se transformó en un dataframe de pandas con dimensiones (18729,32). Durante la exploración de las columnas e índices, se identificó una repetición del número de identificación de cada paciente, que se denominó variable IDNO en el estudio.

Con el objetivo de analizar los factores de riesgo, se decidió enfocarse en los eventos cardíacos que resultaron ser fatales. Para ello, se seleccionaron las columnas "IDNO", "EVTYPE" Y "FATAL" del dataframe mediante la función. drop. Posteriormente, se extrajeron los eventos cardiacos fatales utilizando la condición "FATAL= = 1" con la función. loc, lo que resulto en un nuevo dataframe de dimensiones (5076,3). Este nuevo dataframe, representa a cada paciente de la base de datos una sola vez, lo cual permite realizar un análisis más preciso.

Para garantizar la fiabilidad de los datos, se evaluó los posibles valores faltantes en el dataframe resultante, utilizando la función *"pandas. isnull()sum()sum()"* que devuelve un dataframe booleano con True para las celdas que contienen valores NaN y False para las demás. Luego, se utiliza *.sum().sum()* con el objetivo de obtener la cantidad total de valores NaN en el dataframe. Sin encontrar valores faltantes, lo que indica que este dataframe está completo y listo para el proceso posterior.

El proceso posterior comenzó por realizar la carga del archivo basebothfinal.csv de dimensiones (5888,322), con el objetivo de evaluar la fiabilidad de los datos en este dataframe, se repitió el uso de la función *pandas.isnull()sum()sum()* para evaluar los datos NaN. De este proceso se obtuvo como resultado una cantidad de 128.000 datos NaN, además, se usó la función para evaluar los datos no nulos *pandas.notnull().sum().sum(),* dando una cantidad de datos no nulos 1.767.536.

Estos resultados, indican que existen 128.000 datos NaN en el archivo basebothfinal.csv, mientras que hay 1.767.536 datos no nulos disponibles. A continuación, se procedió a unir los dataframes de events.csv y basebothfinal.csv en un nuevo dataframe denominado "Df_unificado" de dimensiones (5076,325). La unión se realizó con la función *pandas.merge,* con el criterio de coincidencia 'IDNO' y con el método 'inner' que mantiene solo las filas con valores coincidentes en la columna seleccionada, es decir, se agrupan por su número de identificación en el nuevo dataframe.

Posterior a la unión, se evalúa los valores faltantes en el nuevo dataframe mediante la función utilizada en los archivos iniciales del procesamiento, los resultados entregaron un total de 128.400 valores nulos, y la cantidad de valores no nulos resultante fue de 1.561.041.

3.3.2 Eliminación de columnas con exceso de datos nulos.

Se procedió con el preprocesamiento del dataframe unificado, comenzando por eliminar aquellas columnas con un porcentaje de datos faltantes mayor a 30%. Esto, se realizó mediante la función de *pandas.dropna*, en donde mediante una variable auxiliar llamada threshold se fijaron los argumentos de la función para eliminar todos los registros que tuvieran más de 30% de valores NAN. Como resultado de esta eliminación de columnas, tanto la cantidad de datos faltantes se redujo a 51.267 como la dimensionalidad del dataframe de (5076,325) a (5076,316).

3.3.3 Imputación basada en modelo K-Nearest Neighbors

Con el objetivo de optimizar el remplazo que tendrán los valores faltantes, se eligió método de imputación K-nearest Neighbors. Utilizando la clase *KNNImputer* de la biblioteca scikit-learn disponible en Python y utilizando un parámetro n_neighbors de 10. Lo que significa, que se tomaron en cuenta los 10 vecinos más cercanos para la imputación de cada valor faltante. Mediante el método fit_transform (), se aplicó el proceso de imputación en el dataframe, el resultado de esta imputación se almaceno en nuevo dataframe llamado “Df_imputado”, que conservo los índices y columnas del dataframe original.

El resultado de la imputación fue obtenido mediante la función *isnull().sum().sum(),* mostrando que el dataframe no tenía datos faltantes, indicando que todos los datos faltantes fueron remplazados con éxito por los valores estimados según método KNN.

3.3.4 Casting y eliminación de datos duplicados.

Luego de la imputación de datos faltantes, se realizó Casting para los datos a tipo entero utilizando la función *astype(int),* para asegurar que todos los datos imputados fueran de tipo entero y no otro tipo de dato. Posteriormente, se eliminaron las filas duplicadas en la columna correspondiente al número de identificación de los pacientes “IDNO”, esto utilizando la función *drop_duplicates (subset=[“IDNO”], keep= “first”)*. Asegurando con este procesamiento que cada paciente estuviera representado una única vez en el dataframe.

3.4 Selección de características con método del filtro y estadístico Chi-cuadrado

Una vez preprocesado el dataframe, con el objetivo de reducir la dimensionalidad del problema y mejorar la eficiencia computacional asociada, es importante realizar una selección de características efectiva. Existen diversas alternativas para llevar a cabo esta selección, como el análisis univariable que considera a cada característica de manera individual sin evaluar la correlación entre ellas, el análisis Multivariable, el método Wrapper que es costoso computacionalmente y el método del filtro. En el caso de la base de datos CHS, se hace necesario utilizar el método del filtrado dado que se adapta mejor a nuestras necesidades de estudiar dos variables.

El método del filtrado permite asignar un puntaje a cada característica de acuerdo con un criterio univariable, y seleccionar así un subconjunto de características con los mejores puntajes según dicho criterio. Dado que nuestra base de datos contiene variables categóricas, resulta fundamental utilizar correlación de chi-cuadrado en lugar de correlación de Pearson, ya que esta última se aplica únicamente a variables numéricas continuas.

3.4.1 Selección de características correlacionadas a depresión y enfermedades cardiovasculares y creación de nuevo dataframe.

Una vez definido el método de selección de características a utilizar, se procedió a realizar la programación con este fin, primero se definió una variable auxiliar para la variable objetivo, que alterno su valor para la característica depresión y tipos de eventos cardiovasculares dentro del procedimiento del método del filtrado.

En primer lugar, se utilizó un objeto llamado selector, con la función *SelectKBest* y los argumentos *score_func=Chi2*, para realizar la correlación con chi cuadrado y un k=50 para seleccionar las 50 características más correlacionadas a la variable objetivo, posteriormente se aplica el método *fit_transform* generando un subconjunto de las características correlacionadas en la variable *caract_selected*. En segundo lugar, se procede a reconocer los nombres de las variables correlacionadas, para esto se genera una máscara booleana con el método *get_support* del selector, lo que concluye en la creación de una lista con las columnas correspondientes a las características seleccionadas, que fueron almacenadas en la variable *caracteristicas_selected_names*. Por último, se realizó un nuevo dataframe que almacena el subconjunto de características encontrados para depresión en el dataframe "depresion_features" y enfermedades cardiovasculares contenido en el dataframe

“Evtype_features”, y se unieron en un nuevo dataframe llamado “union_features” mediante el método de *pandas.merge*.

3.5 Análisis exploratorio de variables depresión y enfermedades cardiovasculares

En esta sección, se llevó a cabo la codificación para generar gráficos y reconocer los niveles de depresión y los tipos de eventos cardiovasculares en el conjunto de datos union_features. Se crearon histogramas para representar la frecuencia de cada nivel de depresión y los tipos de eventos cardiacos comunes en el dataframe. Además, se realizó un recuento preciso de los casos relacionados con eventos cardiacos específicos, que brindan información importante sobre la distribución de la depresión y los eventos en el conjunto de datos.

3.5.1 Histograma de niveles de depresión y frecuencia

Se utilizó la biblioteca seaborn para trazar el histograma de los niveles de depresión contenidos en la variable DEPSCR05. Se etiquetó el eje X como "nivel de depresión" y se trazó una línea vertical en x=8 para indicar el umbral que define el riesgo de depresión clínica para aquellos pacientes con un valor mayor. En el eje Y, se etiquetó la frecuencia, que representa el número de pacientes con el nivel de depresión mostrado, dando por resultado el histograma de la Fig.3.1. Además, mediante el uso de la función "len", se calculó la longitud del subconjunto "union_features" para valores mayores a 8, resultando en un total de 868 pacientes con riesgo de depresión clínica.

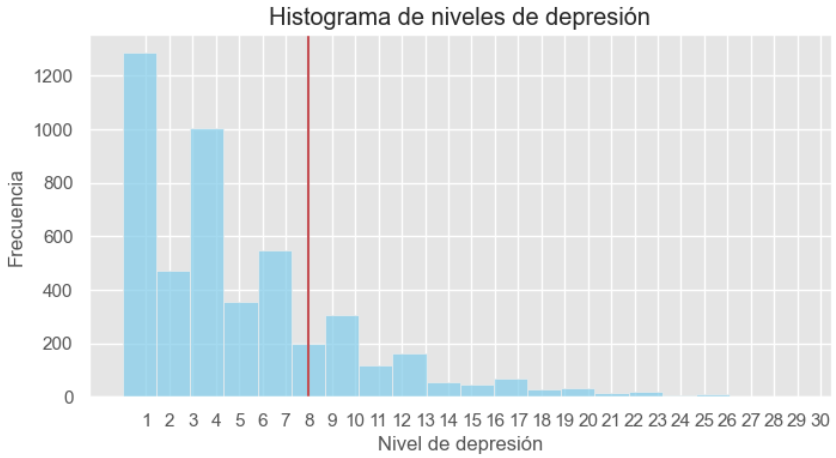

Fig 3.1 Histograma de niveles de depresión y su frecuencia absoluta. Fuente: Elaboración propia.

3.5.2 Histograma de tipos de eventos cardiacos y frecuencia

Utilizando nuevamente la biblioteca seaborn se trazó el histograma que representa las frecuencias de los diferentes tipos de eventos cardiovasculares del dataframe "union_features". Se etiqueta el eje X como "tipos de eventos cardiovasculares" y el eje Y como "frecuencia de los eventos cardiovasculares", obteniendo el gráfico de la Fig. 3.2. Luego, se imprimieron las cantidades exactas de casos para cada uno de los eventos mediante el uso de la función "len". Resultando que existen 3 eventos cardiovasculares fatales en el archivo events.csv, estos son: IAM, accidente cerebrovascular e insuficiencia cardíaca congestiva, información detallada en TABLA 3.1.

Fig. 3.2 Histograma de tipos de eventos y frecuencia absoluta. Fuente: Elaboración propia.

Tabla 3.1 Tipos de eventos cardiovasculares y su frecuencia.

Categoría	Evento	Frecuencia absoluta
1	Infarto agudo al miocardio	234
3	Accidente cerebrovascular	245
4	Insuficiencia cardíaca congestiva	154
9	Otras muertes (no CHD)	3277
11	Otras muertes (CHD)	826

3.6 Creación de nuevas variables apuntando a realizar una clasificación binaria.

Se crearon nuevas columnas en el dataframe "union_features" con el objetivo de tener variables binarias que representara a los pacientes con depresión clínica, así como a aquellos pacientes con muerte por infarto agudo al miocardio, accidente cerebrovascular o falla cardíaca congestiva, a diferencia de las antiguas variables categóricas.

En primer lugar, se creó una nueva variable binaria llamada "DEPSCR05_RISK" con valor 1 si el nivel de depresión en la variable previa "DEPSCR05" era mayor a 8, indicando riesgo de depresión clínica, y tomando el valor 0 en caso contrario. En segundo lugar, se crearon nuevas variables binarias para representar los infartos agudos al miocardio como "X_MI", los accidentes cerebrovasculares como "X_STROKE" y la falla cardíaca congestiva como "X_CHF". Este proceso se resume en la TABLA A.4 (ver anexo A).

3.7 Diseño y entrenamiento algoritmo de Random Forest.

3.7.1 Selección de variable objetivo

Como primer paso, se realizó la preparación de los datos para clasificar cada una de las diferentes variables objetivo mediante el algoritmo de aprendizaje automático RF. Para cada variable objetivo, como por ejemplo la variable "Depresión", se llevó a cabo un procedimiento específico. En este caso, se eliminó la columna 'DEPSCR05_RISK' del conjunto de características (X) y se asignó dicha columna como etiqueta (y). Este proceso se repitió para las demás variables objetivo (IAM, ACV, ICC). De esta manera, se logró la separación adecuada de las características y etiquetas para cada variable, permitiendo posteriormente realizar la clasificación binaria.

3.7.2 División en entrenamiento y prueba.

Se realizó la división del conjunto de datos en conjuntos de entrenamiento y prueba utilizando la función "train_test_split". Las características se asignaron a "X" y las etiquetas se asignaron a "y", estableciendo que el 20% de los datos se utilizará para prueba y el 80 en entrenamiento. Posteriormente, se utilizó el remuestreo SMOTE para abordar el desbalance entre las clases. Se aplicó la función "*fit_resample*" de SMOTE tanto al conjunto de entrenamiento como al conjunto de prueba original, generando conjuntos de datos re-muestreados para el entrenamiento y la prueba, es así como

este procedimiento de remuestreo permitió obtener una distribución de clases más equilibrada y representativa del conjunto de datos completo.

3.7.3 Evaluación de los mejores parámetros con GridSearch.

Se utilizó la función *GridSearchCV* para realizar una búsqueda de hiperparámetros a utilizar en el ajuste del modelo RandomForestClassifier. Se definen diferentes valores de hiperparámetros, como el número de estimadores, la profundidad máxima del árbol y las características máximas a considerar en cada división. De este modo, se ajusta el modelo utilizando los datos de entrenamiento previamente re-muestreados. Posteriormente, se imprime un diccionario con los mejores parámetros encontrados, que son los siguientes: 'max_depth': None, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200. Además, se muestra la mejor puntuación obtenida durante la clasificación binaria. Esto permite identificar los valores óptimos de los hiperparámetros para el modelo *RandomForestClassifier* y obtener una configuración que mejore el rendimiento en la clasificación realizada con el modelo.

3.7.4 Entrenamiento del algoritmo y evaluación en conjunto de entrenamiento y prueba

Se desarrolló el entrenamiento del modelo Random Forest, con el objetivo principal de evitar el sobreajuste y mejorar la precisión en la clasificación, se aplicaron las siguientes técnicas para prevenir el sobreajuste: i) regularización de los datos mediante los parámetros *min_samples_split* y *min_samples_leaf*, ii) sobre muestreo de datos utilizando SMOTE para mejorar la capacidad de generalización, iii) la reducción de complejidad disminuyendo la profundidad de los árboles y la selección de características relevantes en el dataframe. Los hiperparámetros del modelo fueron mejorados mediante una combinación de la búsqueda con *GridSearchCV* y una posterior sintonización manual que definió los hiperparametros que mejoraban el rendimiento en el conjunto de entrenamiento y prueba. Luego, los valores se fijaron en: n_estimators=180, max_features='log2', max_depth=6, random_state=42, min_samples_split=3, min_samples_leaf=3.

3.7.5 Curva de aprendizaje del modelo.

Posterior al entrenamiento del algoritmo, se generó la curva de aprendizaje del clasificador Random Forest utilizando cross-validation con 5 pliegues, con el objetivo de evaluar el clasificador utilizado. Calculando la precisión para el conjunto de entrenamiento y prueba en diversos tamaños del

conjunto de entrenamiento. Mediante este método, se crearon los gráficos que se observan en la Fig.4.2, los cuales permitieron conocer el comportamiento del algoritmo para todas las variables objetivo, evaluando la capacidad de generalización del modelo y permitiendo detectar posibles sobreajustes, De los resultados, se puede concluir que no existe un sobreajuste significativo, puesto que las curvas convergen a medida que se aumenta el tamaño del conjunto de entrenamiento.

3.8 Diseño y entrenamiento algoritmo Red Neuronal Artificial

3.8.1 Selección de variable objetivo

Como primer paso, se realizó la preparación de los datos para clasificar cada una de las diferentes variables objetivos, mediante el algoritmo de aprendizaje automático RNA. Para cada variable objetivo, como por ejemplo la "Depresión", se llevó a cabo un procedimiento específico. En este caso, se eliminó la columna 'DEPSCR05_RISK' del conjunto de características (X) y se asignó dicha columna como etiqueta (y). Este proceso se repitió para las demás variables objetivo (IAM, ACV, ICC). De esta manera, se logró la separación adecuada de las características y etiquetas para cada variable, permitiendo posteriormente realizar la clasificación binaria.

3.8.2 División en entrenamiento y prueba.

En esta sección del código, se realizó la división del conjunto de datos en conjuntos de entrenamiento y prueba utilizando la función "*train_test_split*". Las características se asignaron a "X" y las etiquetas se asignaron a "y", estableciendo que el 30% de los datos se utilizará para prueba.

Posteriormente, se utilizó el remuestreo SMOTE para abordar el desbalance entre las clases. Se aplicó la función "*fit_resample*" de SMOTE tanto al conjunto de entrenamiento como al conjunto de prueba original. Como resultado se generaron conjuntos de datos remuestreados para el entrenamiento y la prueba, permitiendo una distribución de clases más equilibrada y representativa del conjunto de datos completo.

3.8.3 Normalización.

Se utilizó la función *StandardScaler* de scikit-learn, creando así un objeto scaler que se utilizó para realizar la transformación a los conjuntos de datos remuestreados previamente por medio del método *fit_transform*. De esta forma, las características se presentan en una escala similar, evitando

de este modo que las características que tengan valores grandes dominen a características con valores pequeños en el modelo.

3.8.4 Diseño de tipo de Red Neuronal Artificial

Una vez separado el conjunto de entrenamiento y prueba ya se puede proceder con el diseño de una red neuronal secuencial utilizando el método *sequential.* A continuación, se definieron las capas de la red. La primera capa de entrada tiene 73 neuronas, que es equivalente al número de características del conjunto de datos, y utiliza la función de activación ReLU. Adicionalmente, se añadieron en esta capa una regularización de tipo L1 y L2, que permiten eliminar características menos relevantes y conducir a una distribución más suave de los pesos en la red, respectivamente. Después de esta capa, se aplicó una regularización de tipo dropout con una tasa del 0.3, lo que implica que durante el entrenamiento se desactivan aleatoriamente el 30% de las neuronas. Esta técnica ayuda a prevenir el sobreajuste y mejora la capacidad de generalización del modelo. A continuación, se agregaron dos capas intermedias, cada una con 50 neuronas y función de activación ReLU. Estas capas proporcionan al modelo una mayor capacidad de abstracción y mejoran la precisión tanto en el entrenamiento como en la prueba. La última capa, es una capa de salida con una sola neurona y con función de activación sigmoidea, utilizada para la clasificación binaria.

3.8.5 Entrenamiento del algoritmo y evaluación en conjunto de entrenamiento y prueba.

Posterior al diseño de la arquitectura de la red neuronal, se realizó el entrenamiento de esta, efectuando primeramente un ajuste de la tasa de aprendizaje del optimizador Adam, utilizando un learning_rate=0.00001. Este paso permite disminuir la velocidad del aprendizaje para evitar sobreajuste. Luego se compilo el modelo utilizando una función de perdida "*binary_crossentropy*" y se especifica la medición de precisión durante el entrenamiento. A continuación, se definen dos objetos, el primero "*reduceLROnPlateau*" que monitorea la perdida en el conjunto de validación y reduce la tasa de aprendizaje si no hay mejoras después de 20 épocas de entrenamiento. El segundo es "*earlyStopping*", que monitorea la perdida en el conjunto de validación y detiene el entrenamiento si no hay mejoras después de 20 épocas, evitando así el sobreajuste.

3.9 Variables predictoras de variables objetivos en algoritmos Random Forest y Red Neuronal Artificial

A continuación, se desarrolló el entrenamiento de cada uno de los algoritmos, una codificación que permitió conocer el nombre de las características que mayor peso aportaron a la clasificación binaria para predecir riesgo de depresión clínica, fatalidad por infarto agudo al miocardio, accidente cerebrovascular e insuficiencia cardíaca congestiva, para ambos algoritmos. Se organizaron estás variables predictoras de acuerdo con su peso en orden decreciente y se realizaron gráficos con la biblioteca matplotlib que permitieron visualizar las variables predictoras para cada variable objetivo según Random Forest y Red neuronal Profunda, las que se pueden observar en la Fig. 4.2 y Fig. 4.3 respectivamente.

3.10 Intersección de variables predictoras entre algoritmo Random Forest y Red Neuronal Artificial

Una vez graficados los valores de variables predictoras para las variables objetivo, visualizados gráficamente en las Fig. 4.2 y Fig.4.3 del estudio, se procedió para conocer si ambos algoritmos confirmaban las mismas variables como las de mayor aporte a la clasificación. Para esto, se realizó una codificación que tomo en primer lugar, las 15 características de mayor peso para el algoritmo Random Forest y luego las 15 características de mayor peso del algoritmo Red Neuronal Artificial. Una vez obtenidas las listas de características de mayor peso para ambos algoritmos, se realizó una intersección utilizando la biblioteca Numpy y el método *intersect1d*. Dando por resultado una lista de características coincidentes entre ambos algoritmos. Las características coincidentes fueron estudiadas posteriormente, mediante el uso de la biblioteca seaborn, realizando un gráfico de barras y gráfico de cajas para éstas, con el fin de conocer los valores promedios y su distribución cuartilica en la base de datos CHS, tal como se muestra en la Fig.B.5, Fig.B.6, Fig.B.7 y Fig.B.8 (Ver ANEXO B).

Capítulo 4. Resultados

4.1 Introducción

Este capítulo, comprende los resultados obtenidos en el desarrollo del modelo de aprendizaje automático. En primer lugar, se analizan los resultados para el rendimiento de los modelos Random Forest y Redes Neuronales Artificiales, evaluando su curva de aprendizaje para las distintas variables objetivos. Posteriormente se analizan las métricas exactitud, precisión, recall y valor F1 que se desprenden de la matriz de confusión, y, por último, la curva ROC-AUC para cada una de las variables objetivo: depresión, IAM, ACV y ICC (ver TABLA 4.1).

Luego de haber analizado el desempeño de los algoritmos, se procede a estudiar los resultados obtenidos analizando las variables predictoras para cada algoritmo, así como también, para ambos algoritmos en conjunto, utilizando en esta tarea los gráficos obtenidos. Por último, se realiza una discusión y conclusión sobre los resultados obtenidos en la realización de la memoria de título, destacando el buen desempeño de los algoritmos RF y RNA según la curva ROC-AUC y métricas de evaluación para la tarea de clasificación llevada a cabo, además se hace expresa la concordancia de los estudios científicos de los últimos 4 años y algunas de las variables predictoras resultantes obtenidas por el modelo de aprendizaje automático.

4.2 Resultados Random Forest

4.2.1 Curva de aprendizaje para variables objetivo.

La curva de aprendizaje de la Fig. 4.1, representa la evolución en el rendimiento del modelo Random Forest al aumentar el tamaño del conjunto de entrenamiento. De este gráfico se observa que la línea verde representa la exactitud en el conjunto de prueba, mientras que la línea roja denota la exactitud en el conjunto de entrenamiento. El tamaño varía en cada una de las curvas desde 0 hasta aproximadamente 5000, dada la dimensión de nuestro dataframe por la cantidad de pacientes de la base de datos CHS.

La Fig.4.1 (a), muestra la curva de aprendizaje del modelo para la variable objetivo depresión. Se observa un crecimiento poco significativo de la exactitud del conjunto de prueba hasta el tamaño 3000, luego del cual existe un aumento lineal de la exactitud hasta llegar a su punto de inflexión en tamaño 5000, al igual que el conjunto de entrenamiento. No se observa una convergencia entre ambas

curvas, pero si una estabilización. De este gráfico podemos concluir que no existe un sobreajuste, dado que la exactitud del conjunto de prueba se mantiene estable en relación con el conjunto de entrenamiento, por lo tanto, el modelo está realizando una buena generalización. La Fig. 4.1 (b), denota la curva de aprendizaje para la variable IAM, donde se observa que a partir del tamaño 3500 existe un aumento en el rendimiento del modelo, resultando en el tamaño 5000 una convergencia entre ambas curvas, lo que indica una óptima generalización del modelo, sin presentar un sobreajuste. La Fig.4.1 (c), muestra la curva de aprendizaje para la variable ACV, se observa un crecimiento del rendimiento en el conjunto de prueba a partir del tamaño 4000, hasta llegar a un punto de convergencia con el rendimiento del conjunto de entrenamiento en el tamaño 5000, lo que indica que el modelo generaliza de forma óptima para el tamaño de nuestro dataframe, sin presentar sobreajuste. Mientras que Fig. 4.1 (d), representa la curva de aprendizaje para la variable ICC, la cual comienza a crecer en rendimiento al comienzo del tamaño 3000, llegando a converger la curva de entrenamiento y prueba para el tamaño 5000 del dataframe. Por lo que se desprende que el algoritmo RF presenta optima generalización sin presencia de sobreajuste.

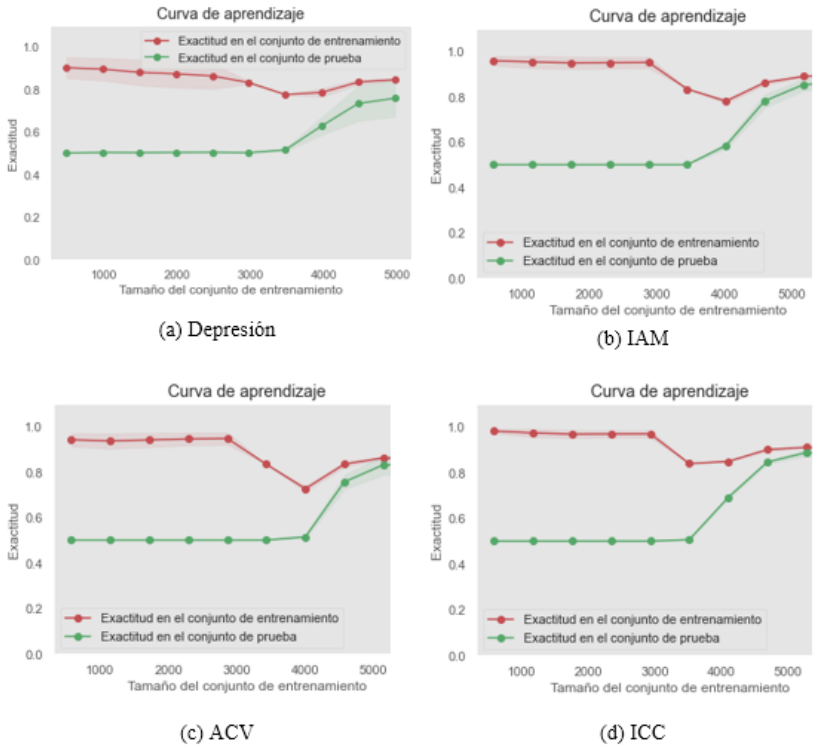

Fig 4.1 Curva de aprendizaje en modelo Random Forest para variables objetivo (a): depresión, (b): Infarto agudo al miocardio, (c): accidente cerebrovascular y (d) insuficiencia cardíaca congestiva. Fuente: Elaboración propia.

4.2.2 Evaluación algoritmo RF con matriz de confusión, métricas y curva ROC-AUC.

Dados los resultados contenidos en la TABLA 4.1, se puede concluir que el algoritmo Random Forest muestra un rendimiento prometedor en términos de precisión, siendo mayor estricto a 0.74 en su clasificación para verdaderos positivos (por ejemplo, verdaderos pacientes con riesgo de depresión) en relación con falsos positivos (pacientes sin depresión, clasificados como depresivos) para todas las variables objetivos. En cuanto a la exactitud, se obtuvo valores mayores de 0.722 para todas las variables objetivo, lo que también indica que el modelo tiene un buen rendimiento general para la clasificación. Por otra parte, los valores F1 Score y Recall, tuvieron como límite inferior 0.709 y 0.679, respectivamente valores que representan una buena discriminación para el modelo.

La métrica más general sobre el funcionamiento del modelo como el actual, en el que existe desbalance de clases es la curva ROC-AUC, estos valores se encuentran en la Fig. B.3 para cada variable objetivo, y se obtuvo como límite inferior el valor 0.8 en AUC, lo que indica que es un buen modelo, capaz de separar de forma eficiente las clases a las que pertenece cada variable objetivo, según los intervalos mostrados en la TABLA 2.2.

Tabla 4.1 Evaluación del rendimiento algoritmo Random Forest para variables objetivos.

Variable objetivo	Accuracy	Precision	Recall	F1 Score	AUC
Depresión	0.722	0.743	0.678	0.709	0.80
Infarto agudo al miocardio	0.768	0.789	0.732	0.76	0.86
Accidente cerebro vascular	0.749	0.755	0.738	0.746	0.83
Insuficiencia Cardíaca Congestiva	0.784	0.808	0.744	0.775	0.88

4.2.3 Variables predictoras de Depresión, IAM, ACV, ICC.

El resultado obtenido para las variables predictoras de depresión se condensa en la Fig. 4.2. En esta figura se observa la importancia de la variable que depresión en la predicción de la muerte por

infarto agudo al miocardio, accidente cerebrovascular o insuficiencia cardíaca congestiva. Por contraparte, se puede extraer si las enfermedades cardiovasculares son o no predictoras de riesgo de depresión clínica.

La depresión clínica mostrada en la Fig. 4.2 (a), no muestra en sí misma una relación con las enfermedades cardiovasculares con que hemos trabajado (IAM, ACV, ICC), sin embargo, como variables predictoras se encuentra; MAJABN en la quinta posición de importancia, MAJABN representa anormalidades mayores en el ECG; MIBLMOD en la treceava posición, donde MIBLMOD es el estado inicial de infarto agudo al miocardio en el comienzo del estudio CHS.

La variable IAM mostrada en la Fig. 4.2 (b), posee en su posición treceava de mayor peso predictivo a la variable riesgo de depresión clínica DEPSCR05_RISK, es decir, la depresión clínica se muestra como un predictor de muerte por IAM, según el algoritmo Random Forest; La variable ACV mostrada en la Fig. 4.2 (c) tiene al riesgo de depresión clínica en su posición número 20 de mayor peso predictivo para fatalidad por accidente cerebro vascular; En cuanto a la variable ICC, mostrada en la Fig. 4.2 (d), se observa que el riesgo de depresión clínica es el 5° predictor de mayor peso para fatalidad por insuficiencia cardíaca congestiva.

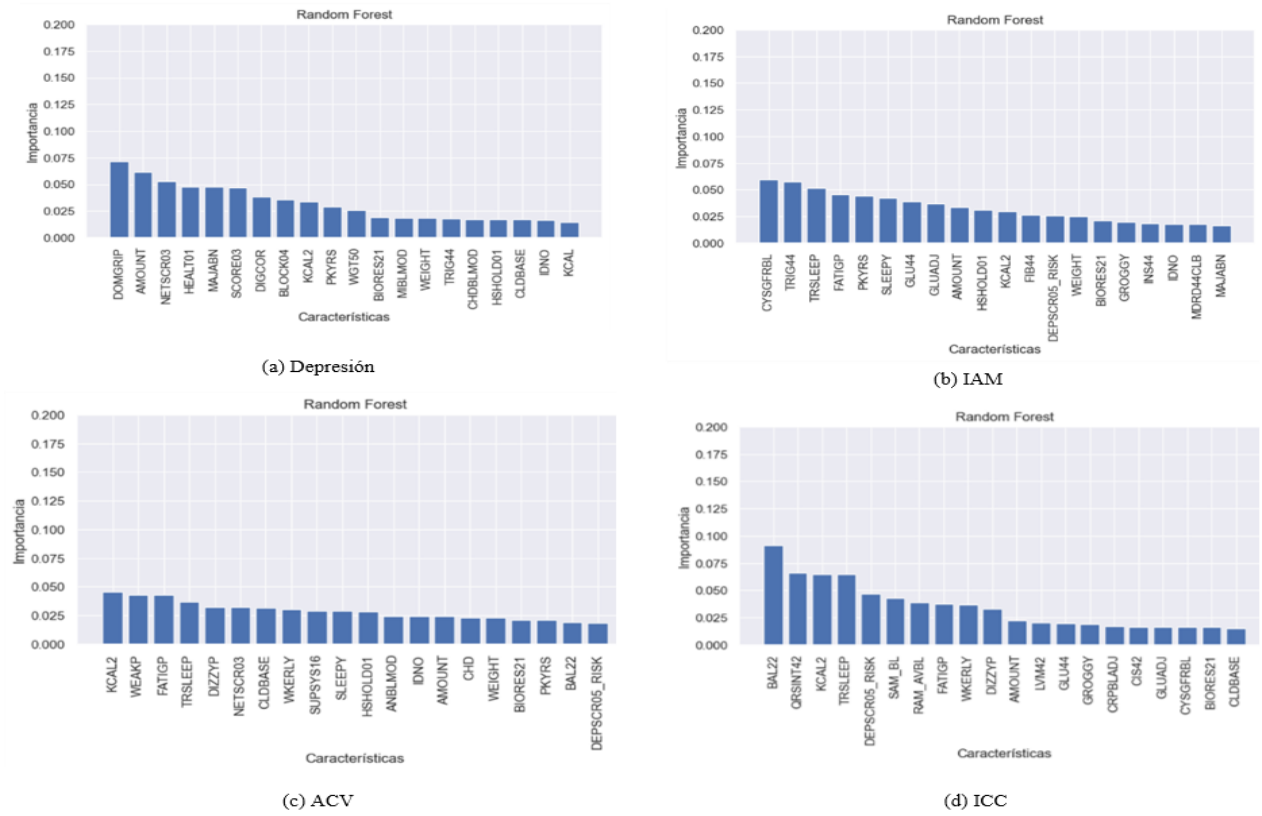

Fig 4.2 Variables predictoras según el modelo Random Forest para variables objetivos (a): depresión, (b): Infarto agudo al miocardio, (c): accidente cerebrovascular y (d): insuficiencia cardíaca congestiva. Fuente: Elaboración propia.

4.3 Resultados Red Neuronal Artificial

4.3.1 Evaluación algoritmo RNA con métricas y curva ROC-AUC.

Dado los resultados contenidos en la TABLA 4.2, se puede concluir que el algoritmo de Red Neuronal Artificial muestra un rendimiento bueno en términos de precisión, con valores de Precisión que son mayores estrictos a 0.73, es decir, para verdaderos positivos en relación con falsos positivos en cada variable objetivo. Además, el modelo alcanza un nivel de exactitud mayor a 0.66 para todas las variables, lo que indica un rendimiento general moderado en la clasificación.

En cuanto a las métricas de F1 Score, se observa que los valores mínimos son de 0.60 y 0.60, respectivamente. Estos valores representan una capacidad de discriminación moderada para el modelo en términos de equilibrar la precisión y la capacidad de recuperación de los casos positivos.

La métrica más general sobre el funcionamiento del modelo es la curva ROC-AUC, la cual se encuentra representada en la Fig. B.4 para cada variable objetivo. Los valores obtenidos tienen un límite inferior de 0.79, lo cual indica que el modelo es considerado bueno para clasificar cada variable objetivo en su clase correspondiente, se considera bueno según los intervalos establecidos en la TABLA 2.2.

Tabla 4.2 Evaluación del rendimiento algoritmo Red Neuronal Artificial para variables objetivos.

Variable objetivo	Accuracy	Precision	Recall	F1 Score	AUC
Depresión	0.68	0.73	0.65	0.69	0.79
Infarto agudo al miocardio	0.70	0.80	0.61	0.69	0.84
Accidente cerebrovascular	0.66	0.76	0.60	0.67	0.82
Insuficiencia Cardíaca Congestiva	0.69	0.83	0.49	0.60	0.81

4.3.2 Variables predictoras de Depresión, IAM, ACV, ICC.

El resultado obtenido para las variables predictoras condensa en la Fig. 4.3, de la cual se puede extraer la información sobre cuál es la posición de la depresión como variable predictora de muerte por infarto agudo al miocardio, accidente cerebrovascular o insuficiencia cardíaca congestiva. Por contraparte, se puede extraer también si las enfermedades cardiovasculares son o no predictoras de riesgo de depresión clínica, según el algoritmo de Red Neuronal Artificial.

La depresión clínica mostrada en la Fig. 4.3 (a), muestra en sí misma una relación con las enfermedades cardiovasculares con que hemos trabajado, puesto que la décima variable de mayor peso corresponde a insuficiencia cardiaca congestiva (X_CHF), es decir, esta variable es un predictor de depresión, además que, se encuentra en la treceava y quinceava posición de mayor peso, el infarto agudo al miocardio (X_MI) y la variable accidente cerebrovascular (X_STROKE), respectivamente. Por otra parte, la variable IAM mostrada en la Fig. 4.3 (b), tiene en su posición cuarentava de mayor peso predictivo a la variable riesgo de depresión clínica DEPSCR05_RISK, es decir, la depresión clínica no se muestra como un predictor relevante de muerte por IAM, según el algoritmo Red Neuronal Artificial. En cuanto a la variable ACV mostrada en la Fig.4.3 (c), tiene al riesgo de

depresión clínica en su posición número 20 de mayor peso predictivo para fatalidad por accidente cerebro vascular. En tanto que, a la variable ICC, mostrada en la Fig.4.4 (d), se observa que el riesgo de depresión clínica es el 36° predictor de mayor peso para fatalidad por insuficiencia cardíaca congestiva, por lo que no se verifica como un predictor claro según RNA.

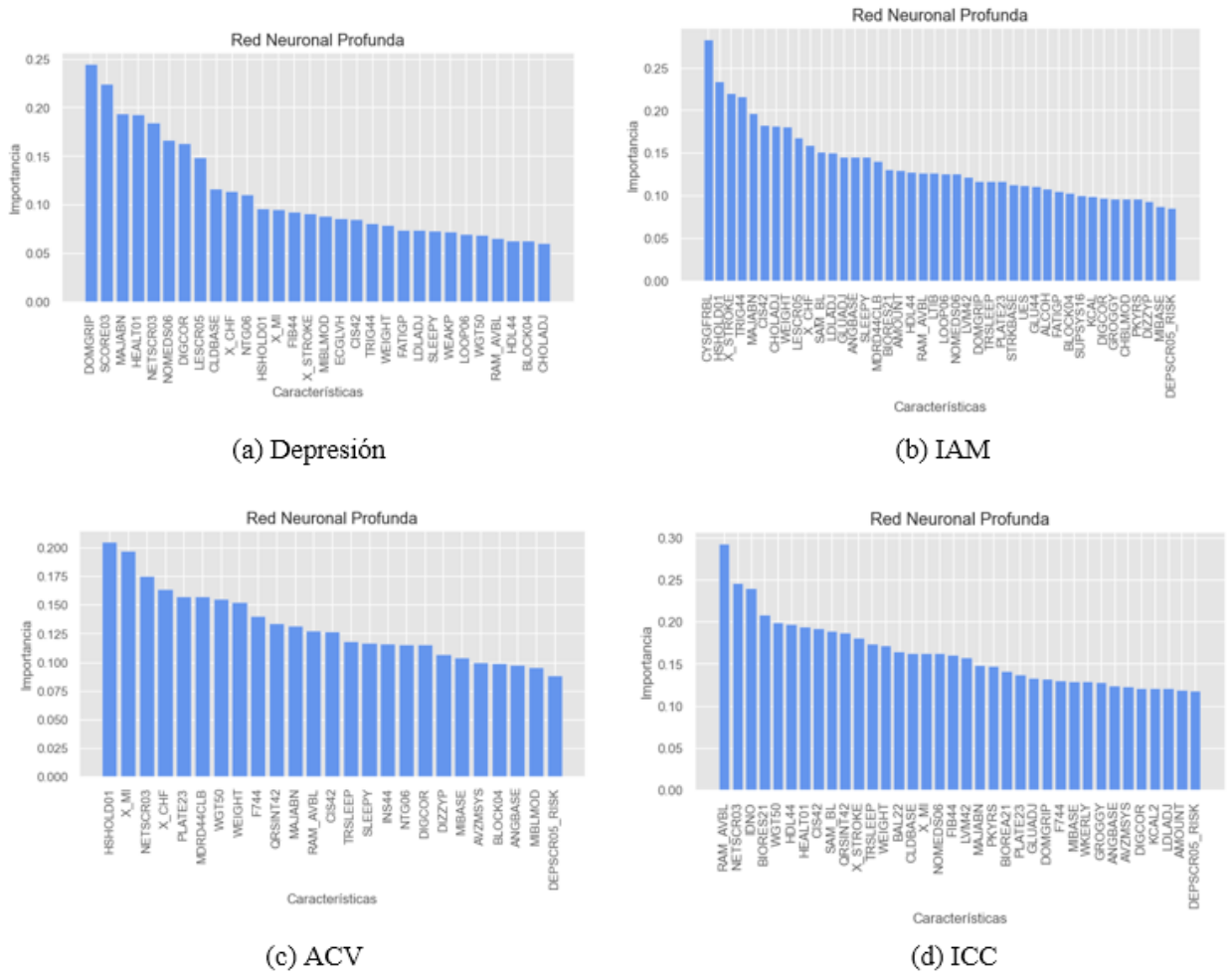

Fig 4.3 Variables predictoras según el modelo Red Neuronal Artificial para variables objetivos (a): depresión, (b): Infarto agudo al miocardio, (c): accidente cerebrovascular y (d): insuficiencia cardíaca congestiva. Fuente: Elaboración propia.

4.4 Resultados en común según Random Forest y Redes Neuronales artificiales.

Las variables predictoras resultantes mediante el algoritmo Random Forest y Red Neuronal Artificial, fueron interceptados tomando en cuenta las primeras 15 variables de mayor peso encontradas por ambos algoritmos, lo que dio por resultado vectores de características más

importantes para cada variable objetivo, para visualizar este resultado se crearon las siguientes tablas: TABLA 4.3, TABLA 4.4, TABLA 4.5 y TABLA 4.6, que contienen las variables predictoras para: depresión, infarto agudo al miocardio, accidente cerebrovascular e insuficiencia cardíaca congestiva respectivamente.

4.4.1 Variables predictoras de Depresión

Los resultados obtenidos mediante ambos algoritmos de aprendizaje automático para la variable objetivo depresión, se muestran en la **TABLA 4.3,** y Fig 4.4. se puede apreciar el gráfico de barras correspondiente a cada variable predictora de depresión junto con un gráfico de caja que permite realizar comparaciones entre las distribuciones de datos en población con riesgo de depresión clínica y sin riesgo de depresión clínica.

En cuanto a la variable predictora DIGCOR, es una medida de la función cognitiva, según se observa en la Fig.4.4, tanto depresivos como no depresivos no se diferencian notablemente en su distribución según su gráfico de caja puesto que sus rangos intercuartílicos que indican el 50% de los datos están separados pero no significativamente, al igual que se visualiza en la mediana, sin embargo, el gráfico de barras indica que aquellos pacientes sin depresión tuvieron una mayor puntuación promedio de función cognitiva, en comparación con aquellos pacientes que tienen depresión, que obtuvieron un puntaje promedio menor. Según la literatura, un estudio del año 2018 se encargó de comprobar la función cognitiva mediante 4 escalas de puntaje distintas en una población de 657 adultos mayores, encontrando en sus resultados que, para las distintas escalas, el puntaje de función cognitiva promedio fue menor en pacientes con depresión o con antecedentes de depresión [21].

La variable HEALT01 se refiere a la calidad de vida auto percibida, clasificada en 5 categorías, donde 1 representa una calidad de vida excelente y 5 una calidad de vida pobre. El gráfico de caja muestra una alta separabilidad entre depresivos y no depresivos, dado que los rangos intercuartílicos de ambos grupos están completamente separados. Por otro lado, el grafico de barras revela una puntuación promedio más alta en la categoría de depresivos, lo cual se relaciona con una peor calidad de vida auto percibida. Estos hallazgos son consistentes con estudios actuales en esta área. Una revisión sistémica del año 2020 identificó 1231 estudios relevantes en esta área, de los cuales 15 cumplían los criterios de selección e inclusión. De estos, 4 demostraron que existe una asociación estadística entre el riesgo de depresión y el deterioro de la calidad de vida [22].

La variable DOMGRIP representa la fuerza de agarre promedio. Según Fig. 4.4, se muestra un gráfico de caja con un rango intercuartílico más disperso y una mediana más alta en pacientes sin depresión, en comparación con aquellos con riesgo de depresión clínica. Por su parte, el grafico de barras indica que la fuerza de agarre promedio fue menor en pacientes con depresión. Según un estudio del año 2018, se aplicó a 34.129 pacientes mayores de 50 años pertenecientes a 6 países de ingresos bajos y medianos, con el objetivo de determinar la importancia de la fuerza de agarre como un marcador simple y económico del riesgo de salud y mortalidad. Los resultados mostraron que para todos los países la fuerza de agarre débil estaba asociada con una probabilidad mayor de depresión [23].

Tabla 4.3 Variables predictoras de depresión según RF y RNA

Variable predictora	Descripción
DIGCOR	Puntuación de función cognitiva
HEALT01	Calidad de vida percibida
DOMGRIP	Fuerza de agarre promedio
NETSCR03	Puntuación de redes sociales
NOMEDS06	Número de medicamentos
SCORE03	Puntuación de apoyo social
CLDBASE	Claudicación
MIBLMOD	I.A.M inicial
MAJABN	Anormalidades mayores ECG

4.4.2 Variables predictoras de Infarto agudo al miocardio

Los resultados obtenidos mediante ambos algoritmos de aprendizaje automático para la variable objetivo IAM, se muestran en la **TABLA 4.4** y en la Fig. 4.5, se puede apreciar el gráfico de barras correspondiente a cada variable predictora de IAM junto con un gráfico de caja que permite realizar comparaciones entre las distribuciones de datos en población con riesgo de depresión clínica y sin riesgo de depresión clínica.

La variable CYSGFRBL representa la función renal según el examen sanguíneo de filtrado glomerular, donde un valor mayor a 90 representa una función renal normal y un valor menor puede llegar a indicar insuficiencia renal. Según se observa en la Fig.4.5, los gráficos de caja de ambos grupos presentan baja separabilidad, teniendo una mediana mínimamente menor en pacientes con IAM, además el rango intercuartílico se distribuye hacia valores menores de CYSGFRBL en los pacientes con IAM. Por otra parte, el valor máximo para pacientes sin IAM es mayor que pacientes con IAM, lo que indica que los pacientes sin IAM podrían tener una función renal mejor. Un artículo científico publicado el año 2017 aplicado en población cubana, buscaba estudiar el valor pronóstico de la función renal a corto plazo para pacientes con IAM, se analizaron 284 pacientes con IAM en la UCI del hospital universitario Clínico-Quirúrgico "Dr. Miguel Henríquez", los resultados fueron que cualquier incremento en creatinina o disminución del filtrado glomerular significaba un aumento de la probabilidad de muerte a corto plazo para pacientes con IAM [24].

Las variables TRIG44 y GLU44, hacen referencia a triglicéridos y glucosa respectivamente. En la figura B.6, se observa en el gráfico de cajas de la variable TRIG44 que un 50% de pacientes con IAM tienen valores más altos que el 50% de pacientes sin IAM, esto se afirma observando el valor mayor del Q1 y Q3 en pacientes con IAM, además los valores mínimos y máximos sitúan al grupo con IAM en valores mayores de triglicéridos, por otra parte, dada la diferencia en las medianas de ambos grupos se afirma una buena separabilidad. La variable GLU44, presenta mediana superior en pacientes con IAM y también un rango intercuartílico ubicado en niveles mayores de glucosa, como también un valor máximo mayor, esto indica su distribución hacia valores superiores de glucosa. Para ambas variables el gráfico de barras permite observar un valor mayor promedio de triglicéridos y glucosa en pacientes con IAM. Según afirma un estudio multicéntrico del año 2021 que estudio la importancia del indicador triglicéridos-glucemia en la mortalidad intrahospitalaria por IAM en 1123 pacientes que un valor alto de las variables triglicéridos y glucemia constituye un factor de riesgo independiente de mortalidad intrahospitalaria, y se plantea que podrían resultar una herramienta de poco costo, simple y de uso sencillo para predecir efectos adversos en pacientes con IAM [25].

Tabla 4.4 Variables predictoras de IAM según modelo RF y RNA.

Variable predictora	Descripción
CYSGFRBL	Filtración glomerular estimada
GLU44	Glucosa
TRIG44	Triglicéridos
WEIGHT	Peso en [Lb]
MAJABN	Anormalidades ECG
SLEEPY	Somnolencia en el día

4.4.3 Variables predictoras de Accidente cerebrovascular

La variable WEIGHT hace referencia al peso en libras de un paciente. En la Fig. 4.6, se observa según el gráfico de barras que el valor promedio de peso en libras es mayor en pacientes sin ACV, por su parte, el gráfico de cajas muestra una mediana mayor en pacientes sin ACV lo que confirma un valor promedio de peso en este grupo, así como también, una separabilidad entre ambos grupos. Además, el gráfico de cajas denota un rango intercuartílico mayor en pacientes sin ACV, a diferencia de pacientes con ACV, que presentan el 50% de los pesos en un rango más acotado hacia inferior. Cabe destacar, que el valor promedio de peso en libras en estados unidos es de 194.7 para adultos mayores, por lo que ambos grupos se ubican bajo el promedio, sin embargo, no se puede identificar un bajo o alto peso en la muestra dada la ausencia de la altura de los pacientes, para así evaluar una métrica más representativa como el índice de masa corporal. La literatura por su parte, asocia la obesidad como un factor de riesgo para ACV, sin embargo, existe una controversia puesto que estudios sugieren que las personas con un exceso de peso pueden incluso mejorar luego de un ACV, uno de estos estudios fue una revisión sistemática de 25 artículos, en donde se estudió esta asociación entre exceso de peso y pronostico por ACV, se encontró que la tasa de mortalidad era menor en individuos que tienen sobrepeso, pero en muchos otros se encontró que la mortalidad era mayor en individuos excepcionalmente obesos y con bajo peso [26].

La variable STDSYS16, representa la presión arterial sistólica. La figura Fig. 4.6 muestra un valor promedio mayor en pacientes con ACV a diferencia de los pacientes sin ACV. Por su parte el gráfico de caja muestra una separabilidad entre ambos grupos, teniendo una mediana superior en el

grupo con ACV y un rango intercuartílico más acotado y con valores mayores de presión arterial sistólica en el grupo de pacientes con ACV. Un estudio realizado en Chile, en el año 2021, específicamente en el servicio de salud metropolitano sur, contemplo 135 pacientes con ACV e identificó como el factor de riesgo modificable de mayor relevancia a la hipertensión arterial, constituyendo la principal causa para enfermedades cerebrovasculares [27].

La variable NETSCR03, denota la puntuación de redes sociales de los pacientes. La observación de la Fig.4.6 demuestra según el gráfico de barras, una puntuación de redes sociales menor en pacientes con ACV, y mediante la observación del gráfico de cajas, se muestra una separabilidad entre ambos grupos, teniendo el grupo con ACV una mediana menor y un rango intercuartílico con puntuaciones menores de redes sociales en comparación con el grupo que no tuvo ACV.

La variable SLEEPY representa la somnolencia durante el día, esta variable tuvo un valor mayor en pacientes con ACV a diferencia de pacientes sin ACV, según se verifica en el gráfico de barras de la figura B.7. Un artículo del año 2023 comienza destacando la prevalencia de trastornos del sueño en pacientes con ACV en un 50%, destacando que, solo se ofrecen pruebas formales de sueño a un 6% de los supervivientes, de los cuales solo un 2% completa las pruebas y así da cuenta del objetivo de la revisión en el artículo, que fue evaluar el papel de los trastornos del sueño, los trastornos respiratorios del sueño y trastornos del ciclo sueño-vigilia en pacientes con ACV. Se concluye con los datos anatómicos, fisiológicos y clínicos una coherencia entre los trastornos del sueño como riesgo de ACV, y su vínculo con un mal pronóstico por ACV, también concluyendo, se realiza un énfasis en la necesidad de que investigadores y médicos tomen en cuenta con vigor este factor de riesgo que suele pasarse por alto [28].

La variable NTG06 representa la ingesta de medicamentos nitratos, y según lo observado en el gráfico de barras, aquellos pacientes que tuvieron un accidente cerebrovascular tuvieron una ingesta menor de esta familia de medicamentos. Un artículo científico realizado en china en el año 2022 estudio mediante una revisión narrativa el metabolismo del nitrato y el ACV isquémico, destaca la evidencia existente para el uso de nitratos, destacando su importancia en estados fisiológicos y fisiopatológicos por su potencial en el sistema vascular, y realza la importancia de obtener distintos tipos de nitratos por diversas vías, además, añade que la suplementación es conveniente, económica y

efectiva para reducir el riesgo de enfermedad cerebrovascular, añadiendo que, su transformación permite inhibir inflamación y proteger el endotelio [29].

Tabla 4.5 Variables predictoras de ACV según RF y RNA.

Variable predictora	Descripción
WEIGHT	Peso en [Lb]
NETSCR03	Redes de apoyo
STDSYS16	Presión arterial sistólica
MIBLMOD	I.A.M inicial.
MAJABN	Anormalidades mayores ECG
NTG06	Toma de medicamentos Nitratos
SLEEPY	Somnolencia en el día

4.4.4 Variables predictoras de Falla cardíaca congestiva

La variable CIS42, representa la puntuación de lesión cardíaca obtenida en un ECG de 12 derivaciones. El grafico de caja de la Fig. 4.7 muestra una buena separabilidad entre ambos grupos, dado que sus rangos intercuartílicos se encuentran claramente delimitados y separados, al igual que sucede con la mediana, se afirma con este gráfico que, el grupo con ICC presentó una puntuación de lesión cardíaca mayor en el electrocardiograma, situación confirmada con el gráfico de barras de la figura.

Las variables QRSINT42, RAM_AVBL y SAM_BL codifican la información obtenida en un electrocardiograma de 12 derivaciones, correspondiendo respectivamente a; Intervalo QRS en milisegundos, Amplitud de onda R y amplitud de onda S. Según se verifica en el gráfico de cajas de la Fig. 4.7, la distribución del intervalo QRS en ambos grupos tiene una alta separabilidad, dada la diferencia en su mediana y en su rango intercuartílico, por lo que se concluye observando el gráfico de barras, que los pacientes con insuficiencia cardíaca congestiva tienen un intervalo QRS más dilatado en el tiempo. La indagación científica, muestra en una tesis doctoral realizada el año 2018, la cual buscó estudiar los efectos electrofisiológicos del aumento de la presión intraventricular en un

modelo experimental de insuficiencia cardíaca en cerdos, y halló que, en el modelo animal, realizado con corazones aislados de cerdos y cartografía óptica, se mostró que la presión intraventricular alta y/o el remodelado electrofisiológico en la ICC reducen la velocidad de conducción significativamente en las fibras de Purkinje, lo que se correlaciona con un ensanchamiento del complejo QRS, y también el aumento del estrés parietal por la poscarga ventricular izquierda produce ensanchamiento del complejo QRS, sugiriendo como conclusión que, la duración del complejo QRS es una herramienta útil en la detección de pacientes con riesgo arrítmico elevado [30]. En cuanto a las variables correspondientes a amplitud de onda R y S, encontramos en el gráfico de cajas que ambas tienen una alta separabilidad dada la mediana de los grupos con y sin ICC, además de la ausencia de superposición total de sus rangos intercuartílicos, lo que indica buena separabilidad en ambos grupos y permite concluir observando la distribución y el grafico de barras correspondiente que para pacientes que sufrieron ICC, la amplitud de las ondas R y S en el electrocardiograma fueron mayores. Destacando que una alta amplitud de la onda R puede indicar una hipertrofia ventricular derecha severa, pausa en el nódulo sinusal o bloqueos fascicular [31].

Tabla 4.6 Variables predictoras de ICC con RF y RNA.

Variable predictora	Descripción
CIS42	Puntuación de lesión cardíaca
GLUADJ	Glucosa
LVM42	Acortamiento del ventrículo izquierdo
QRSINT42	Intervalo QRS [ms]
RAM_AVBL	Amplitud onda R
SAM_BL	Amplitud onda S
BAL22	Prueba de pérdida del equilibrio
TRSLEEP	Problemas para dormir

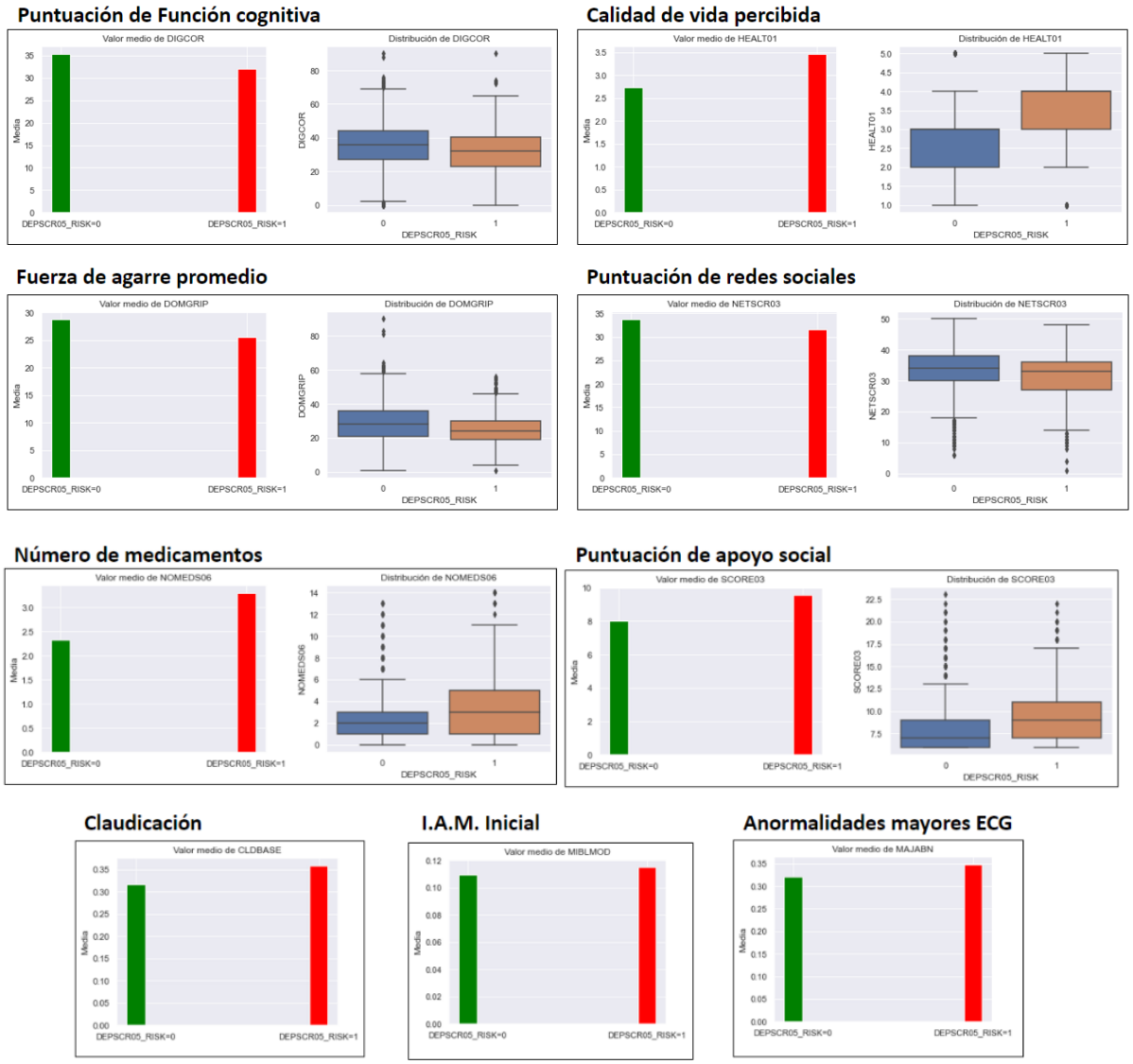

Fig. 4.4 Gráfico de barras y cajas para variables predictoras de depresión según RF y RNA

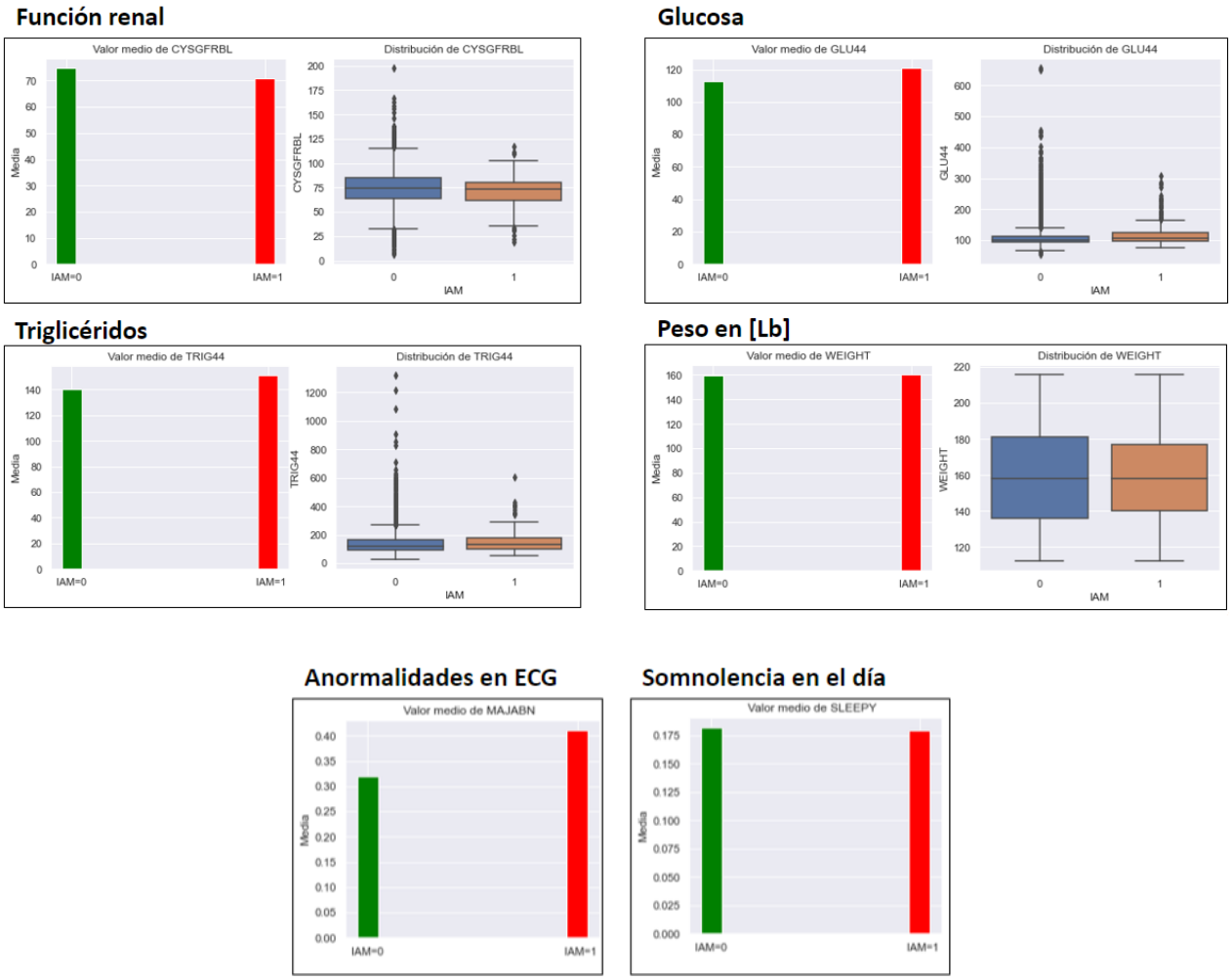

Fig. 4.5 Gráfico de barras y cajas para variables predictoras de IAM según RF y RNA

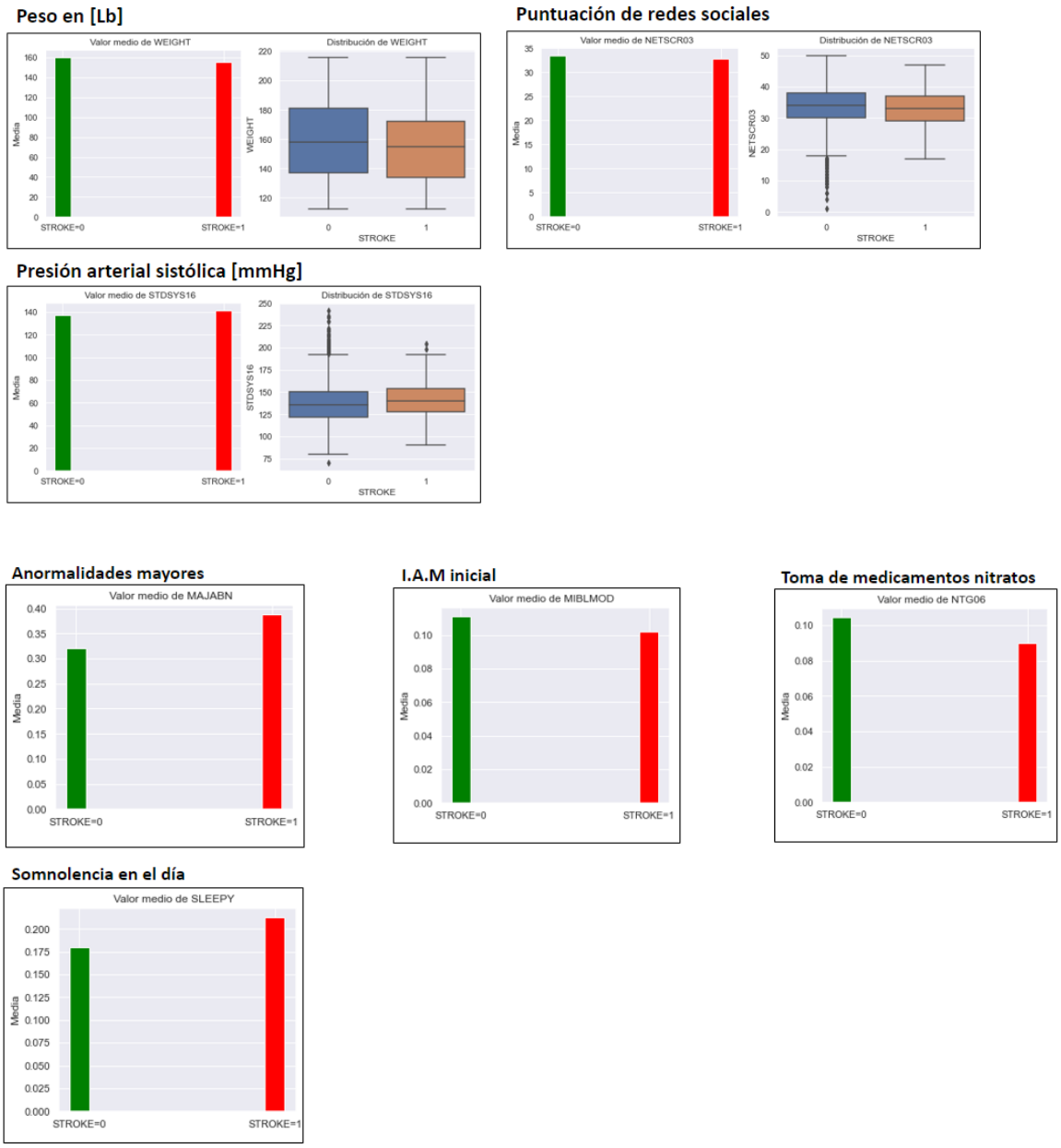

Fig. 4.6 Gráfico de barras y cajas para variables predictoras de ACV según RF y RNA

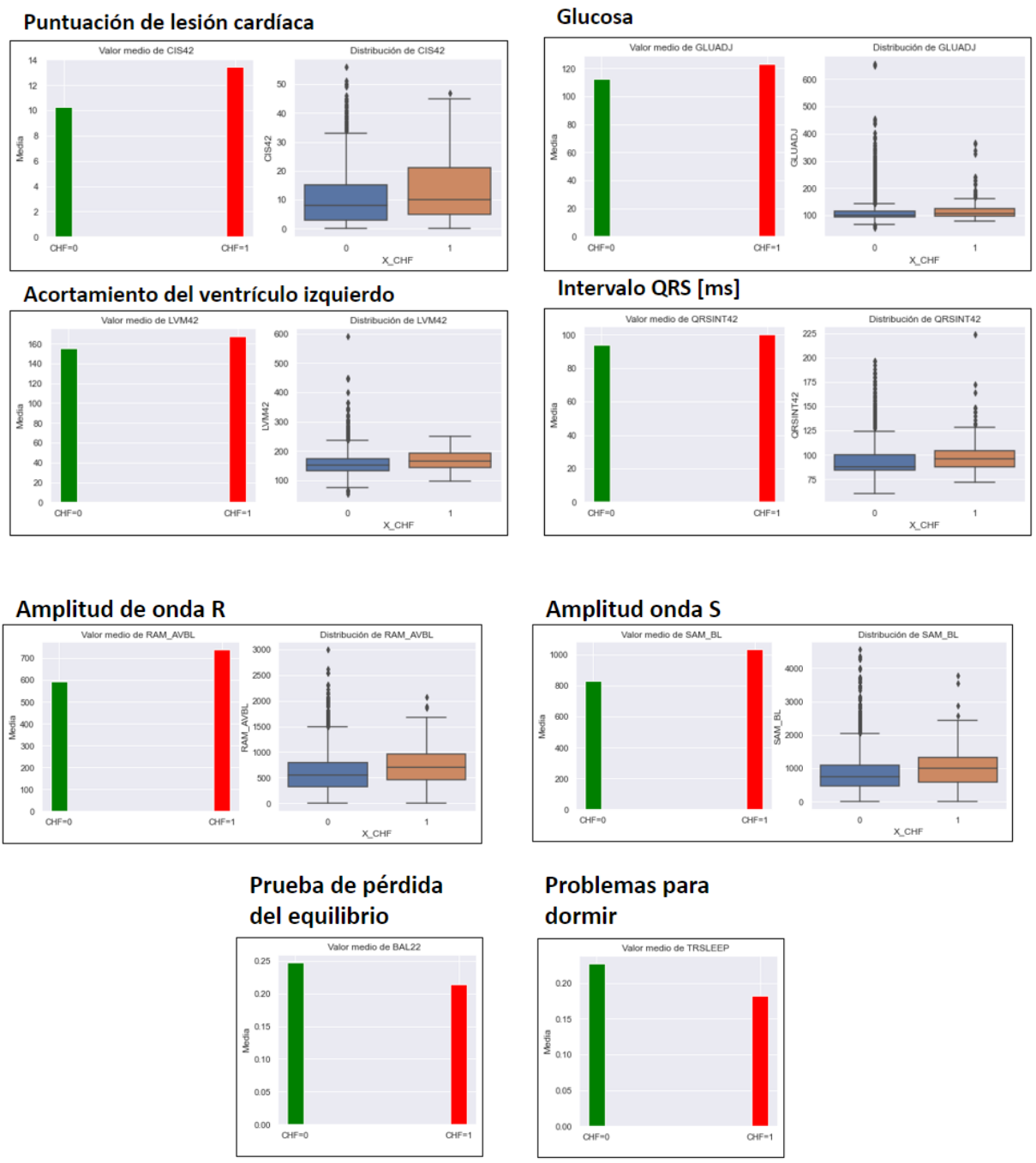

Fig. 4.7 Gráfico de barras y cajas para variables predictoras de ICC según RF y RNA

Capítulo 5. Discusión, Conclusión y Trabajo futuro.

5.1 Discusión

En esta memoria de título, se desarrolló un modelo de aprendizaje automático utilizando el algoritmo Random Forest (RF) y Redes Neuronales Artificiales (RNA) para la predicción de riesgo de depresión clínica, infarto agudo al miocardio (IAM), accidente cerebrovascular (ACV) e insuficiencia cardíaca congestiva (ICC).

El rendimiento del modelo Random Forest fue bueno, obteniendo una buena generalización y un desempeño en las métricas evaluadas F1 Score mayor a 0.71 y una curva ROC-AUC con valores mayores a 0.8, demostrando un desempeño sólido. Además, su curva de aprendizaje mostró ausencia de sobreajuste y buena generalización del modelo, es decir, clasificar correctamente las clases de cada variable objetivo sin presentar sobreajuste a los datos de entrenamiento. Por otro lado, el modelo de Red Neuronal Artificial mostró un rendimiento mayor a 0.6 para la métrica de F1 Score, que representa un equilibrio razonable entre precisión y recall. Además, la curva ROC-AUC para la RNA indicó que era bueno para la clasificación de las variables objetivo, teniendo un valor mínimo de 0.79 en la variable depresión y valores superiores en las demás variables, por lo que el modelo es aceptable para discriminar entre clases en cada variable objetivo.

El algoritmo Random Forest reconoció el riesgo de depresión clínica como el 5° predictor de ICC, 13° predictor de IAM y 20° predictor de ACV. Por contraparte, las anormalidades mayores de un ECG fueron el 5° predictor de depresión y un infarto agudo al miocardio previo al estudio CHS, fue el predictor 13° de depresión. En cuanto al algoritmo de Red Neuronal Artificial, este reconoció como el 3° predictor de depresión a las anormalidades mayores de ECG, 9° predictor fue haber padecido claudicación, 10° insuficiencia cardíaca congestiva, 13° IAM, 15° haber padecido accidente cerebrovascular. Por contraparte, la depresión fue el 25° predictor de ACV, 42° predictor de IAM, 37° predictor de ICC.

5.2 Conclusión

El trabajo realizado logró preprocesar la base de datos de tal manera que no existieran valores faltantes, además, se logró reducir la dimensionalidad del problema y mejorar la eficiencia computacional realizando una selección de variables óptima mediante el método del filtro. El diseño y desarrollo de los algoritmos de aprendizaje logró una buena capacidad de predicción para cada variable objetivo sin presentar un sobreajuste.

Dada la identificación de variables predictoras de orden sociodemográficas clínicas y psicológicas encontradas y visualizadas en gráficos, se concluye según RF y RNA, que el padecer infarto agudo al miocardio, anormalidades mayores en un examen de ECG y claudicación, son predictores del riesgo de depresión clínica, estableciendo que, las ECV predicen que un paciente tenga depresión clínica. Por otra parte, la depresión no es un predictor directo ECV, sin embargo, sus variables predictoras están relacionadas, siendo las anormalidades mayores en un ECG un predictor para depresión y ACV, además, una baja puntuación de redes interpersonales es predictor tanto de depresión como de ACV.

Es relevante para el campo de la medicina incluir algoritmos de aprendizaje automático, especialmente en la extracción de grandes bases de datos, que entregan información fidedigna de las cuales los seres humanos no alcanzamos a abstraer la información que, si nos permite abstraer la computación, entregando información relevante para la salud, permitiendo comprender mejor la etiología de enfermedades que afectan a gran parte de la población.

5.3 Trabajo futuro

Los alcances del presente estudio permitieron diseñar, desarrollar y evaluar un algoritmo de aprendizaje automático que predice ECV y riesgo de depresión clínica, de donde se obtienen las principales variables predictoras de estas enfermedades. Estas variables predictoras han demostrado su relevancia siendo objeto de estudio en los últimos años, por lo que el trabajo futuro por seguir será reconocer las variaciones que tuvieron estas variables predictoras a lo largo de los 24 años del estudio CHS y analizar su relación con aquellos eventos cardiovasculares que no fueron fatales y con el riesgo de depresión clínica. Posterior a esto, se trabajará con profesionales del área médica para realizar un análisis de los resultados y toma de decisiones basadas en la información obtenida.

Bibliografía

[1] Organización de Naciones Unidas. (2020, diciembre 10). “Las diez principales causas de muerte en el mundo." Disponible en: Las diez principales causas de muerte en el mundo, una lista que varía entre países ricos y pobres | Noticias ONU [Última revisión: 30 de agosto de 2023]

[2] Clínicas de Chile. (2020, diciembre 10) "Estudio global revela drástico aumento de las muertes por enfermedades cardiovasculares." Disponible en: Estudio global revela drástico aumento de las muertes por enfermedades cardiovasculares - Clínicas de Chile (clinicasdechile.cl) [Última revisión: 30 de agosto de 2023]

[3] Instituto Nacional de Estadística. (2020). “Estadísticas vitales año 2020." Disponible en: Estadísticas Vitales (ine.gob.cl) [Última revisión: 30 de agosto de 2023]

[4] M. T. Corea Del Cid, "Depression and its impact in public health," REV MÉD HONDUR, vol. 89, Supl. No. 1, pp. S1-68, 2021. DOI: 10.5377/rmh.v89isupl.1.12047

[5] Ciper Chile. (2021, marzo 18). “Cómo se vive la depresión y por qué nos demoramos tanto en reconocerla." Disponible en: Cómo se vive la depresión y por qué nos demoramos tanto en reconocerla - CIPER Chile. [Última revisión: 30 de agosto de 2023]

[6] Y. Zhang, Y. Chen, and L. Ma, "Depression and cardiovascular disease in elderly: Current understanding," Journal of Clinical Neuroscience, vol. 47, pp. 1-5, 2018. https://doi.org/10.1016/j.jocn.2017.09.022

[7] A. L. Silverman, A. A. Herzog, and D. I. Silverman, "Hearts and Minds: Stress, Anxiety, and Depression: Unsung Risk Factors for Cardiovascular Disease," Cardiology in Review, vol. 27, no. 4, pp. 202-207, 2019. https://doi.org/10.1097/crd.0000000000000228

[8] BESH. "Machine Learning y Big Data en el campo de la salud y medicina personalizada." Disponible en: Machine Learning y Big Data en el campo de la salud y la medicina personalizada - (bes-h.com) [Última revisión: 30 de agosto de 2023]

[9] A. López and C. Macaya, "Capítulo 9: Fármacos cardiovasculares," en R. Freire y A. Moreno, Eds., Libro de la salud cardiovascular del hospital clínico San Carlos y la Fundación BBVA.

Disponible en: Libro de la salud cardiovascular del Hospital Clínico San Carlos y la Fundación BBVA (fbbva.es) [Última revisión: 30 de agosto de 2023]

[10] Middlesex Health. (2022, marzo 5) . "Peligros sobre la hipertensión: Efectos de la hipertensión sobre tu cuerpo," Disponible en: https://middlesexhealth.org/learning-center/espanol/articulos/peligros-sobre-la-hipertensi-n-efectos-de-la-hipertensi-n-sobre-tu-cuerpo [Última revisión: 30 de agosto de 2023]

[11] Middlesex Health.(2022, mayo 18). "Angina de pecho, perspectiva general." Disponible en: Angina de pecho // Middlesex Health [Última revisión: 30 de agosto de 2023]

[12] Middlesex Health (2020, mayo 29) "Síndrome coronario agudo, perspectiva general." Disponible en: Síndrome coronario agudo // Middlesex Health [Última revisión: 30 de agosto de 2023]

[13] Middlesex Health. (2023, mayo 12) "Insuficiencia cardíaca, perspectiva general." Disponible en: Insuficiencia cardíaca // Middlesex Health [Última revisión: 30 de agosto de 2023]

[14] RPubs. (2017, mayo 10) "Evaluación de modelos de clasificación." Disponible en: RPubs - Matriz de Confusión - Evaluación de modelos de predicción [Última revisión: 30 de agosto de 2023]

[15] Spyder, "Spyder visión general." Disponible en: Home — Spyder IDE (spyder-ide.org) [Última revisión: 30 de agosto de 2023]

[16] IBM, "Análisis exploratorio de los datos." Disponible en: ¿Qué es el análisis exploratorio de datos? | IBM [Última revisión: 30 de agosto de 2023]

[17] R. Figueroa, "Selección de características" RPyAA, Universidad de Concepción, 2021.

[18] Historia de la empresa. "Qué es el procesamiento de los datos." Disponible en: ¿Qué es el preprocesamiento de datos y quién lo utiliza? - Historiadelaempresa.com. [Última revisión: 30 de agosto de 2023]

[19] Toward Data Science.(2019, junio 12) "Comprender el bosque aleatorio." Disponible en: Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science [Última revisión: 30 de agosto de 2023]

[20] iWannaBeDataDriven. (2020, septiembre 17) “Regresión logística – Machine Learning." Disponible en: Regresión Logística I — Machine Learning | by Brayan Buitrago | iWannaBeDataDriven | Medium. [Última revisión: 30 de agosto de 2023]

[21] J. Apaza, M. Valer, y F. M. Runzer-Colmenares, "Depresión y disminución de la función cognitiva en adultos mayores de un hospital peruano, 2010-2015," Acta Médica Peruana, vol. 35, pp. 191-192, 2018. Disponible en: http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1728-59172018000300010&nrm=iso. [Última revisión: 19 de junio de 2023]

[22] M. Gálvez Olivares, C. Aravena Monsalvez, H. Aranda Pincheira, C. Ávalos Fredes y F. López-Alegría, "Salud mental y calidad de vida en adultos mayores: revisión sistémica," Revista chilena de neuro-psiquiatría, vol. 58, pp. 384-399, 2020. Disponible en: Salud mental y calidad de vida en adultos mayores: revisión sistémica (scielo.cl) [Última revisión: 19 de junio de 2023]

[23] G. Ashdown-Franks et al., "Handgrip strength and depression among 34,129 adults aged 50 years and older in six low- and middle-income countries" Journal of Affective Disorders, vol. 243, pp. 448-454, 2019. https://doi.org/10.1016/j.jad.2018.09.036. [Última revisión: 19 de junio de 2023]

[24] H. B. Gutiérrez y F. D. Martos Benítez, "Valor pronóstico de la función renal a corto plazo en pacientes con infarto agudo de miocardio," Revista Colombiana de Cardiología, vol. 25, pp. 26-32, 2018. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-56332018000100026&nrm=iso. [Última revisión: 19 de junio de 2023]

[25] G. Martínez-García, et al., "Impacto del índice triglicéridos-glucemia en la mortalidad intrahospitalaria por infarto agudo de miocardio. Resultados del registro multicéntrico RECUIMA," Gaceta médica de México, vol. 158, no. 2, pp. 86-92, 2022. Epub 16 de mayo de 2022. https://doi.org/10.24875/gmm.21000628. [Última revisión: 19 de junio de 2023]

[26] InBody (2023) "La conexión entre accidente cerebrovascular y obesidad," Disponible en: https://www.inbodymexico.com/2019/05/01/la-conexion-entre-accidente-cerebrovascular-y-obesidad/ [Última revisión: 10 de julio de 2023].

[27] J. Sepúlveda-Contreras, "Caracterización de pacientes con accidente cerebrovascular ingresados en un hospital de baja complejidad en Chile," Universidad y Salud, vol. 23, pp. 8-12, 2021. Disponible en: http://orcid.org/0000-0002-7060-2475. [Última revisión: 10 de julio de 2023].

[28] A. Ferre et al., "Los ictus y su relación con el sueño y los trastornos del sueño," Neurología, vol. 28, no. 2, pp. 103-118, 2013. Disponible en: https://doi.org/10.1016/j.nrl.2010.09.016. [Última revisión: 10 de julio de 2023].

[29] Y. Wang et al., "Nitrate Metabolism and Ischemic Cerebrovascular Disease: A Narrative Review [Review]," Frontiers in Neurology, vol. 13, 2022. Disponible en: https://doi.org/10.3389/fneur.2022.735181. [Última revisión: 10 de julio de 2023].

[30] Javier, M. P. (2018). Efectos electrofisiológicos del aumento de la presión intraventricular en un modelo experimental de insuficiencia cardiaca. https://digitum.um.es/digitum/handle/10201/57371 [Última revisión: 10 de julio de 2023]

[31] Spiegato. "En cardiología, ¿qué son las ondas R?" Disponible en: En cardiología, ¿qué son las ondas R? - Spiegato [Última revisión: 10 de julio de 2023].

[32] Springer Link. "Data Science and Predictive Analytics: Biomedical and Health Applications using R." Disponible en: Data Science and Predictive Analytics: Biomedical and Health Applications using R | SpringerLink [Última revisión: 10 de julio de 2023].

[33] Barcelona Geeks. "Prueba Chi cuadrado de Pearson." Disponible en: Python – Prueba Chi-Cuadrado de Pearson – Barcelona Geeks/ [Última revisión: 10 de julio de 2023].

[34] The Data Schools. "K-nearest neighbors KNN en python." Disponible en: K-nearest neighbors (KNN) en Python (thedataschools.com) [Última revisión: 10 de julio de 2023].

[35] Ciencia de Datos. "Redes neuronales en python." Disponible en: Redes neuronales con Python (cienciadedatos.net) [Última revisión: 10 de julio de 2023].10 de julio de 2023]

ANEXO A. Tablas

TABLA A.1 Comparación de algoritmos.

	Sensibilidad	Especificidad	Precisión
SVM	84.8%	79.13%	82.17%
R. L	88.41%	79.13%	84.15%
A. D	80.48%	71.94%	76.56%

Tabla A.1: A. K. De la Hoz Manotas, U. J. Martínez-Palacio and F. E. Mendoza-Palechor (2013).

TABLA A.2 Descripciones de variables y estadísticos descriptivos.

Variable	Estadísticas descriptivas
Sociodemográfico(N=5201)	
Edad al inicio	Media:72.8 (Rango, 65-100); D.E:5.6; Mediana:71.0.
Sexo	Femenino: 2962(57%), Masculino: 2239(43%)
Raza	Blanco: 4926 (94.7%), No blanco: 275(5.3%)
Nivel educacional (N=5187)	Media:13.9 (rango, 0-21); D.E: 4.7; mediana: 12.0.
Estado civil	Casado: 3596 (69.1%); Divorciado o separado: 196(3.8%); Viudo: 1197(23%); Soltero: 212(4.1%)
Acontecimientos estresantes de la vida	Media: 1,1(rango, 0-7); D.E: 1.1; mediana: 1,0.
Síntomas depresivos:	
Bajo (puntuación 0-7)	4156(80%)
Alto (puntuación ≥8)	1036(20%)

Enfermedad clínica prevalente	5201(100%)
Infarto de miocardio	504(9.7%)
Angina de pecho	814(15.7%)
Insuficiencia cardíaca congestiva	217(4.2%)
Claudicación intermitente	135(2.6%)
Derrame cerebral	190(3.7%)
Ataque isquémico transitorio	130(2.5%)
Diabetes:	
Normal (Referencia)	5201(100%)
Alteración de la glucosa	1434(27.6%)
Diabético	1173(22.6%)
Hipertensión:	5185(100%)
Normotenso(referencia)	2273(43%)
Hipertenso limítrofe	774(14.9%)
Hipertenso	2138(41.2%)
Indicadores de enfermedad subclínica(N=5196)	
Cuestionario de Rose para Claudicación	100(1.9%)
Cuestionario de Rose para Angina	320(6.2%)
Volumen espiratorio forzado en 1s (N=5111)	Media: 2.1(Rango, 0.3-4.5); SD: 0.7; Mediana: 2.0
Relación tobillo-brazo	5087(100%)

0->0.9	4455(87.6%)
1-≤0.9	632(12.4%)
Fracción de eyección del ventrículo izquierdo	5152(100%)
Normal: 0	4962(96.3%)
Anormal: 1	190(3.7%)
Electrocardiograma mayor anormal	1461/5028(29.1%)
Estenosis Carotidea	2540/5171(49.1%)
<u>Factores de riesgo biológicos o conductuales</u>	
Presión arterial sistólica mmHg (N=5191)	Media: 135.8 (rango, 77-230); DE: 21.5; Mediana: 134.
Nivel de colesterol total, mmol/L(mg/dl) (N=5173)	Media: 5.55 (214.6) (rango, 62-433); SD: 1.02 (39.3); Mediana: 5.51(213)
Lipoproteínas de baja densidad mg/dl (N=5101)	Media: 133.1 (rango, 28-340); DE: 35.7; Mediana: 131.0
Nivel de glucosa en ayunas mmol/L (mg/dl) (N=5165)	Media: 6.1 (110.2) {(rango, 2.9-36.5) (53-657)}; SD: 1.9(34.6); Mediana: 5.6(101)
Nivel de triglicéridos, mmol/L (mg/dl) (N=5173)	Media: 1.61 (142.70) {(rango0.42-14.94) (37-1323)}; SD: 0.89(78.5); Mediana: 1.39(123)
Tabaquismo:	5198(100%)
Nunca fumo	2397(46.1%)
Exfumador	2200(42.3%)

Actual fumador	601(11.6%)
índice de masa corporal (N=5185)	Media: 26.4 (rango,14.7-53.2); DE: 4.5; Mediana: 25.9
Consumo de alcohol:	5183(100%)
Bebidas a la semana:	
0	2489(48%)
1-7	2026(39.1%)
>7	668(12.9%)

Tabla A.2: Schulz, R., S. R. Beach, D. G. Ives, L. M. Martire, A. A. Ariyo and W. J. Kop (2000).

TABLA A.3 Exploración de datos.

Carpeta	Subcarpeta	Tipo de archivo	Registros	Variables	Contenido
Ancillary_studies					
	Capolla_t	.CSV	5888	41	Thyroid
	Crouse_	Word			Endothelial
	Daniels_	.CSV	899	18	SNPs
	Defilippi_	.CSV	5888	6	PRoBNP
	Kanstenbaum_	.CSV	2341	7	Mineral metabolism and vit D

	Kurosawa	.CSV	980	3	EPCR
	Mozaffarian_	.CSV	5888	228	Fatty acid
	Mukamal_	.CSV	5888	11	Glucose harmonization
	Mukamal_Yr5	.CSV	588	6	Year 5 analytes
	Shores_	.CSV	1299	7	Testosterone
BASELINE					
	Base1final	.CSV			
	Base2final	.CSV			
	Basebothfinal	.CSV	5888	322	
EVENTS					
	CRITERIA	WORD			Criterios de asignación para eventos de CHS.
	EVENTS	WORD			Informa asignación numérica a variables e incluye consideraciones.
	Events	.Doc			Asignación numérica para cada evento
	Events	.CSV	18729	32	

FORMS		JPG			Escaneo de formularios analógicos.
ICD9					
	CHS HOSPITALIZATIONS_11	WORD			Codificación de hospitalizaciones y modificación de estos datos.
	Drhosp11	.CSV	28472	25	
INTRO					
	DATABLB	WORD			Nombres de las carpetas de base de datos y descripción breve.
	OVERVIEW	WORD			Explicación del diseño y razón fundamental del estudio.
	RYCYRTBL	WORD			Registros por año de exámenes realizados.
LDD					
	LDD_LADS_2011	WORD			Tabla con nombres de variables para todos los

					formularios de CHS.
	YR18_ALLSTAR	WORD			Contiene los nombres asignados a todas las variables, por categorías.
MOOPS		WORD			Varios Word con explicación de cuestionarios realizados y sus escalas.
MRRECS					
	Danielsnp	.CSV	899	18	
	Mr2recs	.CSV	5888	125	Información cerebral.
	Mrrecs	.CSV	588	134	Información cerebral.
NOFORM					
	Medbame	WORD			Nombre de medicamentos y uso
	NOFORM	WORD			Criterios utilizados en formularios de

					toma de medicamentos
ULTRA					
	ultrabl	.CSV	5201	59	
	Ultrayr11	.CSV	5888	92	
	Ultry11	.CSV	5888	52	
VARIABLE INFO					
	AnalysisTips	WORD			Consejos para análisis de variables no informadas y auto informadas.
	CALVARS	WORD			Cálculo de variables.
	VALUELBL	WORD			Establece etiquetas de valor para variables que cambian con los años

TABLA A.3 Medición de desempeño con matriz de confusión

Medida	Fórmula	Significado
Exactitud (Accuracy)	(VP+VN) / (VP + VN + FN + FP)	Porcentaje de los datos clasificados correctamente
Tasa de error (Misclassification Rate)	(FP + FN) / (VP + VN + FN + FP)	Porcentaje de los datos clasificados incorrectamente
Sensibilidad (Recall)	VP / (VP + FN)	Cuando la clase es positiva, ¿qué porcentaje logra clasificar?
Especificidad (True Negative Rate)	VN / (VN + FP)	Cuando clase es negativa, ¿qué porcentaje logra clasificar?
Precisión	VP / (VP+FP)	Cuando predice positivos, ¿qué porcentaje clasifica correctamente?
Valor de predicción negativo	VN/ (VN + FN)	Cuando predice negativo, ¿qué porcentaje clasifica correctamente?

Tabla A.4 Creación de variables binarias desde variables categóricas.

Variable	Valor	Nueva variable binaria	Nuevos valores binarios
Depresión	Depresion ≥8	Depresión clínica (DEPSCR05_RISK)	0: No tiene depresión clínica 1: Tiene depresión clínica.
Tipo de evento cardiovascular	Tipo de evento= 1	Infarto agudo al miocardio (X_MI)	0: no tuvo IAM 1: Tuvo IAM
	Tipo de evento=3	Accidente cerebrovascular (X_STROKE)	0: No tuvo ACV 1: Tuvo ACV
	Tipo de evento=4	Insuficiencia cardíaca congestiva (X_CHF)	0: No tuvo ICC 1: Tuvo ICC

ANEXO B. Gráficos

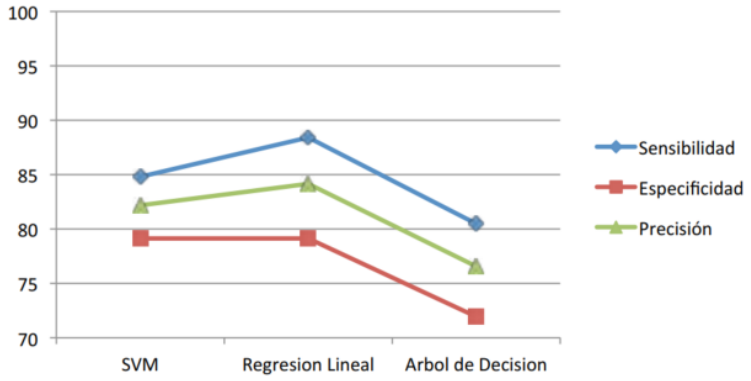

Fig. B.1 Gráfica comparativa de los algoritmos SVM, Regresión Lineal y árbol de Decisión

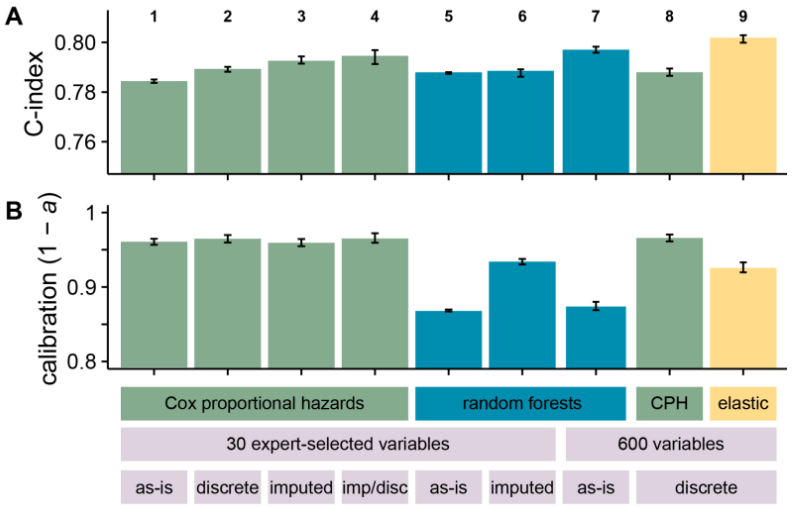

Fig. B.2 Desempeño general de discriminación y calibración para los diferentes modelos y conjuntos de datos utilizados.

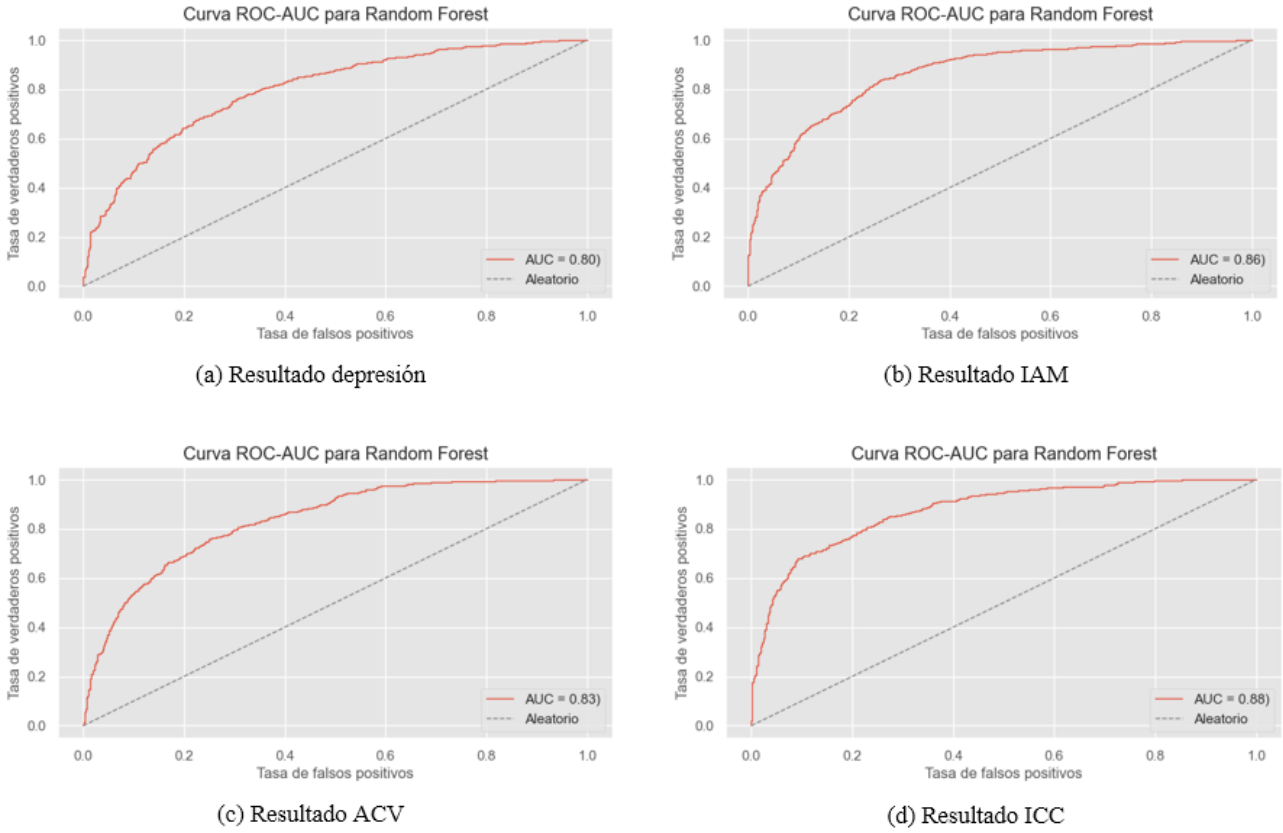

Fig. B.3 Resultados de curvas ROC-AUC en modelo Random Forest, para variables objetivos (a): depresión, (b): Infarto agudo al miocardio, (c): accidente cerebrovascular y (d): insuficiencia cardíaca congestiva.

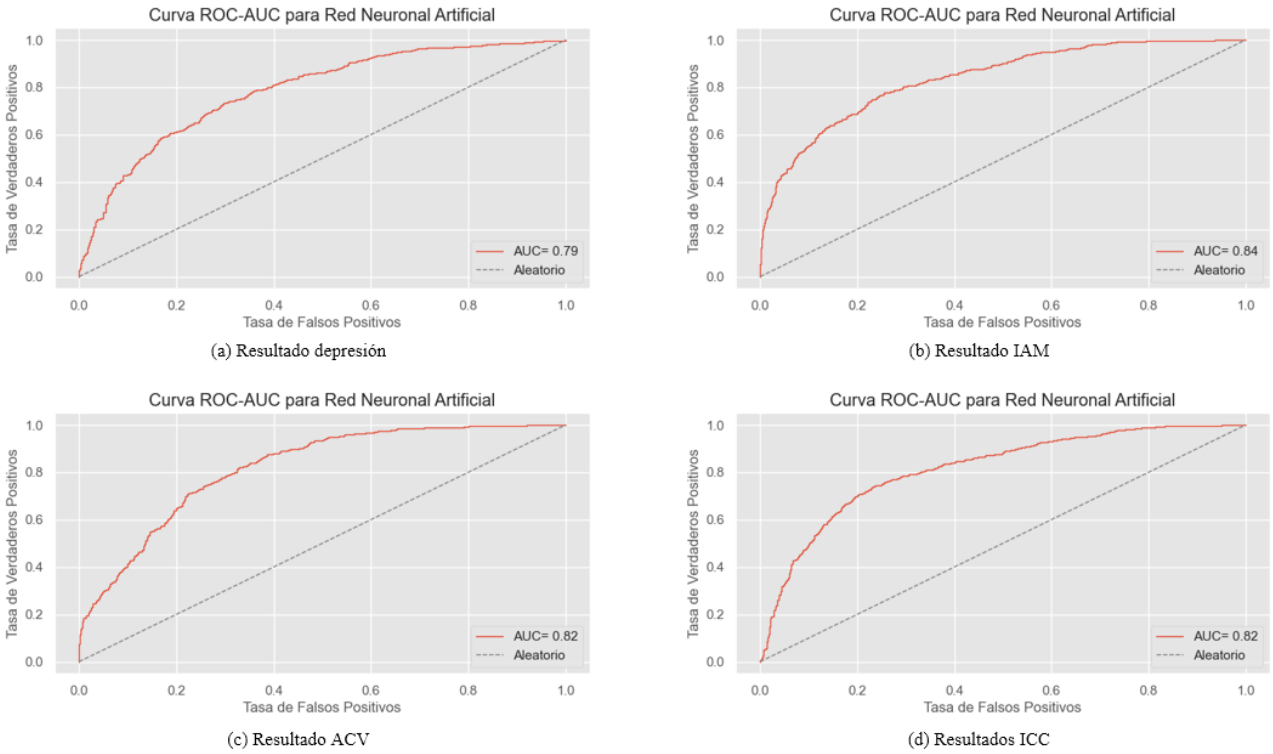

Fig. B.4 Variables predictoras según el modelo Red Neuronal Artificial para variables objetivos (a): depresión, (b): Infarto agudo al miocardio, (c): accidente cerebrovascular y (d): insuficiencia cardíaca congestiva.

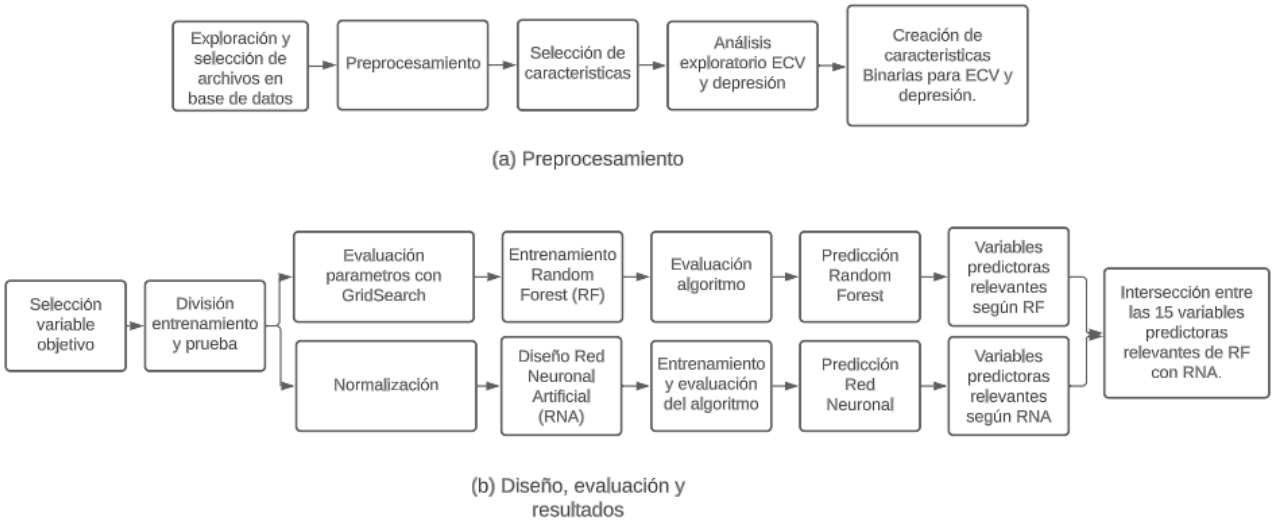

Fig. B.5 Esquemas de metodología para desarrollo de algoritmos en: (a) preprocesamiento base de datos, y en (b) Diseño, evaluación y resultados algoritmos.

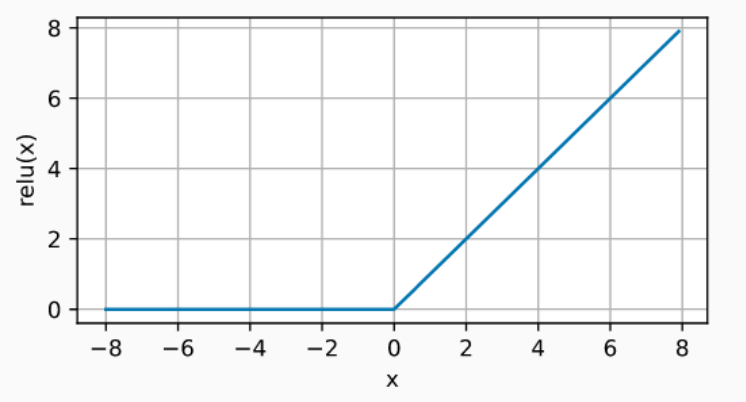

Fig B.6 Representación de función de activación ReLu.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fig. B.7 Matriz de confusión. Fuente: RPubs by RStudio.

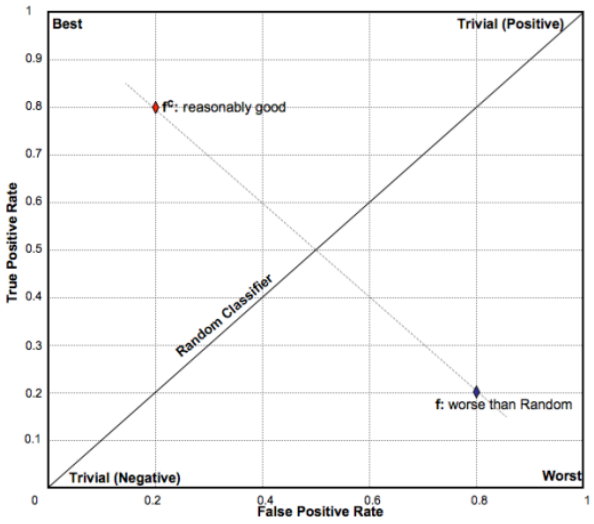

Fig B.8. Representación de Curva ROC.

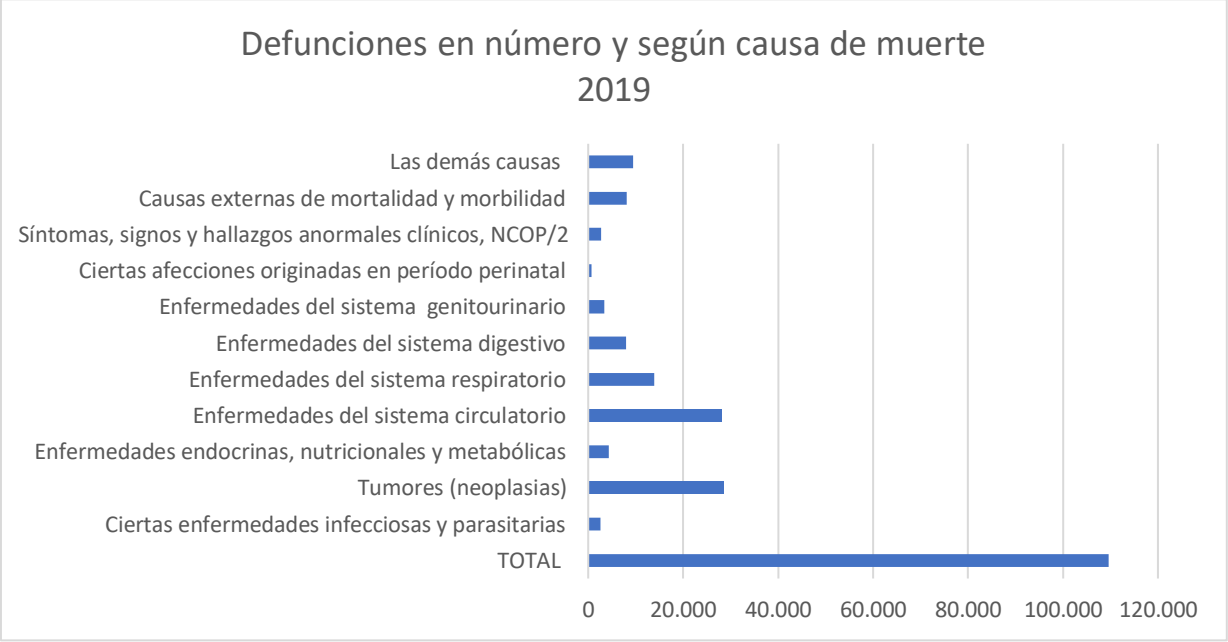

Fig B.9 Principales causas de muerte en chile para el año 2019 son atribuidas a enfermedades del sistema circulatorio y tumores. Fuente: Información obtenida del INE, Elaboración propia.

ANEXO D. RESUMEN DE MEMORIA DE TÍTULO

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA
RESUMEN DE MEMORIA DE TÍTULO

Departamento : Departamento de Ingeniería eléctrica
Carrera : Ingeniería Civil Biomédica
Nombre del memorista : Rodrigo Ignacio Navarro Araneda
Título de la memoria : Algoritmo de aprendizaje automático para el estudio de la asociación entre enfermedad cardiovascular y depresión.
Fecha de la presentación oral : 1 de septiembre de 2023.

Profesor(es) Guía : Rosa Liliana Figueroa Iturrieta.
Profesor(es) Revisor(es) : Pablo Aqueveque Navarro y Sergio Sobarzo Guzmán.
Concepto :
Calificación :

Resumen (máximo 200 palabras)

Desde hace 20 años las enfermedades cardiovasculares son la principal causa de muerte a nivel global y en Chile son las responsables de un 25.6% de defunciones. Por su parte, la depresión es la principal causa de baja laboral, afectado a más de 450 millones de personas a nivel global, y en Chile se estima que un 15.8% de la población la padece. La base de datos CHS posee datos de 20 años, más de 300 variables y un aproximado de 5000 pacientes adultos mayores con enfermedades cardiovasculares, siendo al año 2000 el estudio longitudinal más extenso realizado nunca. Se utilizaron los algoritmos de aprendizaje automático Random Forest y Red Neuronal Artificial para realizar un estudio de las variables que predicen la fatalidad por enfermedades cardiovasculares y aquellas que predicen la depresión. El ensamble de ambos algoritmos permitió encontrar variables predictoras de depresión cuyo origen es cardiovascular, sin encontrar a la depresión como un predictor directo de las enfermedades cardiovasculares.