Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Predicción de diagnóstico de cáncer de seno mediante biomarcadores sanguíneos usando redes neuronales artificiales

Tesis presentada por

Angel Balam Benítez Mata

para obtener el título de

Bioingeniero

Directora: Dra. Dora Luz Flores Gutiérrez

Ensenada, México, abril de 2019

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO

PREDICCIÓN DE DIAGNÓSTICO DE CÁNCER DE SENO MEDIANTE BIOMARCADORES SANGUÍNEOS USANDO REDES NEURONALES ARTIFICIALES

TESIS
PARA CUBRIR LOS REQUISITOS NECESARIOS PARA OBTENER EL TÍTULO DE
BIOINGENIERO

PRESENTA:
ANGEL BALAM BENÍTEZ MATA

Aprobada por:

Dra. Dora Luz Flores Gutiérrez
Director de Tesis

Dr. David Cervantes Vásquez
Miembro del Comité
(Secretario)

Dra. Eunice Vargas Viveros
Miembro del Comité
(Vocal)

Dr. Franklin David Muñoz Muñoz
Miembro del Comité
(Vocal)

Dra. Michelle A. Digman
Miembro del Comité
(Vocal)

A mis padres y hermanas.

Gracias infinitas por el apoyo y muestras de amor.

Agradecimientos

Estoy agradecido profundamente con mis padres, Apolinar Benítez Betancourt y Siria Carmina Mata Plascencia, quienes siempre me apoyaron en todas las decisiones tomadas durante mi vida académica y me han motivado a salir de mi zona de confort, esto es fruto de su esfuerzo como padres ejemplares y grandes seres humanos. Gracias también a mis hermanas Indira Andrea Benítez Mata, quien soportó mi desorden y me motivaba a su manera durante nuestra vida universitaria juntos en Ensenada, y a Elsa del Rocío Benítez Jaramillo, quien a pesar de la distancia nunca dejó de echarme porras y ánimos. Y por supuesto a mi gato Piñato, por hacerme reír en los momentos de estrés y cansancio.

Quiero agradecer a mis hermanos mayores académicos, Rubén Alfonso Castañeda Martínez, por motivarme a salir de la zona de confort y a ser ambicioso, a Alberto Abaroa Villanueva, por mostrarme paciencia, búsqueda de la excelencia y perfección en todas las actividades a realizar, a Carlos Alberto Castro Estrada, por su equilibrio, ni muy Rubén ni muy Alberto. A los tres les debo mucho de mi aprendizaje y motivación a querer superarme y no conformarme, han sido un ejemplo a seguir.

Un agradecimiento especial a la Dra. Dora Luz Flores Gutiérrez, mi mentora académica, por darme la oportunidad de colaborar en sus proyectos, por mostrar ese deseo de mejora continua y trabajo en equipo, por darnos libertad y al mismo tiempo responsabilidad en cada una de las oportunidades que se han presentado y experiencias a lo largo de estos años, no tengo palabras suficientes para agradecerle.

De igual manera, quiero agradecer a la Dra. María de los Ángeles Cosío León, quien fue mi tutora a lo largo de la licenciatura, por siempre apoyarme en las decisiones académicas y ayudarme a expandir mi panorama sobre la Bioingeniería.

Al Dr. David Cervantes Vásquez, quien me ha apoyado a lo largo de la licenciatura en distintas ocasiones, siempre mostrando su gran calidad humana y energía positiva.

Al Dr. Dante Alberto Magdaleno Moncayo, por mostrarme lo maravilloso y estresante que puede llegar a ser el combinar la biología con las ciencias computacionales.

Al Mtro. Adalberto Avelar García Rojas, jefe del Departamento de Cooperación Internacional e Intercambio Académico, quien fue una persona clave para la decisión de vivir un intercambio estudiantil en Corea del Sur.

Al Ing. Joaquín Heriberto Villavicencio Moreno, por despertar en mi un interés muy bonito, atractivo y especial hacia el hospital, y por siempre apoyar incondicionalmente en las actividades estudiantiles.

Quiero agradecer a la Sociedad Mexicana de Ingeniería Biomédica por tan grandiosas experiencias y amistades, pero sobre todo el despertar el sentimiento de orgullo a mi profesión y a siempre estar en pie de lucha para mejorar a nuestro México.

Un agradecimiento infinito a la Universidad Autónoma de Baja California, por la oportunidad de estudiar la profesión por la que siento pasión, así como a todo el personal académico, administrativo y de intendencia de la Facultad de Ingeniería, Arquitectura y Diseño.

Por último, un agradecimiento al comité revisor de este trabajo, Dr. David Cervantes, Dra. Eunice Vargas, Dr. Franklin Muñoz y Dra. Michelle Digman, por tan valiosas aportaciones y sugerencias realizadas previas a la culminación de este documento.

Resumen

El uso de redes neuronales artificiales (RNAs) ha sido de gran ayuda en una variedad de estudios, llevando a cabo tareas de predicción, clasificación y optimización de datos.

Este trabajo hace uso de RNAs y se compara con otras técnicas de *machine learning* reportadas en la literatura para predecir el diagnóstico de cáncer de seno en una población de mujeres con características específicas, para este trabajo es una población con sobrepeso u obesidad y posible padecimiento de diabetes, aunado a una etapa pre o postmenopáusica. Para validar los modelos se utilizaron los parámetros del área bajo la curva (AUC), desviación estándar (σ), sensibilidad, especificidad e índice de Youden. Los algoritmos con los que se entrenaron los modelos de RNAs son *Scaled Conjugate Gradient* (SCG), *Resilient Backpropagation* (RP), *Conjugate Gradient Backpropagation with Powell-Beale Restarts* (CGB), comparado con los resultados reportados con *Suppor Vector Machine* (SVM), *Logistic Regression* (LR) y *Random Forest* (RF). Los mejores modelos de clasificación obtuvieron un AUC por arriba de 0.95, valores de sensibilidad y especificidad por arriba de 0.90 y 0.95, respectivamente. Los modelos entrenados con el algoritmo CGB demuestran ser superiores a RP y SCG, siendo SCG el algoritmo de entrenamiento con las métricas de evaluación más bajas.

Se sugieren diferentes estrategias para mejorar los resultados usando otros métodos para asignar la importancia relativa de cada predictor, así como los métodos de validación.

Palabras clave: Cáncer de seno, aprendizaje automático, redes neuronales artificiales, modelo de clasificación.

Abstract

The use of artificial neural networks (ANNs) has been of great help in a large number of studies, carrying out tasks of prediction, classification and optimization of data.

This work makes use of ANNs and is compared with other techniques of *machine learning* reported in the literature to predict the diagnosis of breast cancer in a population of women with specific characteristics, for this work the population presents over weight or obesity and possible diabetes, along a stage of pre or postmenopause. To validate the models, the parameters of the area under the curve (AUC), standard deviation (σ), sensitivity, specificity and Youden index were used. The algorithms with which the RNA models were trained are *Scaled Conjugate Gradient* (SCG), *Resilient Backpropagation* (RP), *Conjugate Gradient Backpropagation with Powell-Beale Restarts* (CGB), compared with the results reported with *Support Vector Machine* (SVM), *Logistic Regression* (LR) and *Random Forest* (RF). The best classification models obtained an AUC above 0,95, sensitivity and specificity values above 0,90 and 0,95, respectively. The models trained with the CGB algorithm show to be superior to RP and SCG, with SCG being the training algorithm with the lowest evaluation metrics.

Different strategies are suggested to improve the results using other methods to assign the relative importance of each predictor, as well as the validation methods.

Keywords: Breast cancer, machine learning, artificial neural networks, classification model.

Índice general

Índice de tablas

Índice de figuras

Figura **Página**

Lista de abreviaciones

ACS	*American Cancer Society*
ADN	Ácido desoxirribonucleico
AUC	Área bajo la curva
BC	Cáncer de seno
BRCA1	Gen cáncer de seno 1
BRCA2	Gen cáncer de seno 2
CGB	*Conjugate Gradient Backpropagation with Powell-Beale restarts*
ER	Receptor de estrógeno
FCN	Función de entrenamiento
GEO	Base de datos *Genes Expression Omnibus*
HER1	Receptor 1 de crecimiento epidermal humano
HER2	Receptor 2 de crecimiento epidermal humano
IMC/BMI	índice de masa corporal
LR	*Logistic regression*
miRNAs	Micro ácidos ribonucleicos
ML	*Machine Learning*
NCI	*National Cancer Institute*
OMS	Organización Mundial de la Salud
PMC	Perceptrón multicapa
PR	Receptor de progesterona
RF	*Random forest*
RNA	Red neuronal artificial
ROC	*Receiver Operating Characteristic*
RP	*Resilient Backpropagation*
SCG	*Scaled Conjugate Gradient*
SEEER	*Surveillance, epidemiology and end result program*
SVM	*Support vector machine*
TGCA	*The genome cancer atlas*

Capítulo 1

Introducción

En este capítulo se presenta el panorama multidisciplinario en el que se encuentra un enfoque que ha tomado popularidad en los últimos años al buscar una nueva generación de herramientas que ayuden a atacar los problemas actuales de la salud en la población. Se habla acerca de la problemática a atacar, empezando por una breve introducción a lo que es el cáncer de seno, las actuales estrategias para su detección y diagnóstico, después se introduce cómo el desarrollo tecnológico ha permitido avances en este campo de estudio al recopilar datos crudos y almacenarlos en bases de datos y cómo estos son procesados mediante técnicas de inteligencia artificial. El capítulo finaliza al hacer mención de la importancia y ventajas de usar correctamente la tecnología para tener resultados a favor del sector salud.

1.1. Una nueva generación de soluciones

La necesidad de generar nuevas estrategias en el sector salud a nivel global, ha llevado a la integración de distintas áreas de estudio y creado equipos de trabajo multidisciplinarios, en donde el trabajo y aportación de cada parte es de suma importancia para atacar eficazmente los retos que se afrontan a la hora de innovar y crear soluciones.

La era de la información y el *big data* han permitido tener acceso a grandes cantidades

de información, datos crudos y almacenarlos para su posterior análisis, en combinación con novedosas técnicas en las disciplinas químico biológicas, han dejado al descubierto nuevas oportunidades de exploración en enfermedades que cada día se vuelven más críticas de atender, como lo es el cáncer de seno, que de acuerdo a datos de la Organización Mundial de la Salud (OMS), es una de las enfermedades con mayor incidencia en mujeres.

1.2. Cáncer de seno

La definición de cáncer es dada cuando un grupo de células dentro de un cuerpo empiezan a dividirse descontroladamente, las células se vuelven anormales y en conjunto pueden llegar a formar masas de tejido, mejor conocidas como tumores (NCI, 2018b).

De acuerdo al Instituto Nacional del Cáncer de Estados Unidos (NCI por sus siglas en inglés), un tumor es una masa anormal de tejido que resulta cuando las células se dividen más de lo que deberían o no mueren cuando deberían y se aglomeran en un espacio dentro del cuerpo humano. Los tumores pueden ser benignos (no cancerosos), o malignos (cancerosos). Por lo anterior, las células tumorales, son aquellas células afectadas por sustancias o moléculas iniciadoras de tumores, dichas afectaciones permiten la creación de tumores. Sin embargo, no todas las células tumorales son cancerígenas.

Un tumor benigno, no se extiende o invade tejidos cercanos, pueden ser grandes de tamaño pero una vez que son removidos, usualmente no vuelven a crecer.

A diferencia de un tumor benigno, una vez que un grupo de células forman un tumor maligno, este puede esparcirse o invadir tejidos aledaños, de tal forma que las células cancerígenas pueden viajar a distintas partes del cuerpo a través de la sangre o del sistema linfático y poniendo en riesgo la aparición de nuevos tumores (NCI, 2018a).

La clasificación de un tumor yace en las características de las células que lo conforman, las células cancerígenas, para el caso del tumor maligno, son diferentes en muchos sentidos

a células normales, además de lo mencionado anteriormente, estas suelen ser células menos especializadas, es decir, actividades específicas pueden ser parcial o totalmente no llevadas a cabo (Catherine Sánchez, 2013; Zuber et al., 2012).

Las células cancerígenas tienen la capacidad de ignorar señales a las que normalmente una célula sana tendría respuesta, incluso pueden influenciar a estas últimas, así como a moléculas y vasos sanguíneos que rodean y alimentan al tumor (NCI, 2018b).

Una característica peculiar es la capacidad de evasión al sistema inmunológico, el cual consiste en una red de órganos, tejidos y células especializadas que protegen al cuerpo de infecciones y otras condiciones anormales, estas células cancerígenas se pueden "esconderz evitar ser eliminadas del cuerpo (NCI, 2018b; Catherine Sánchez, 2013).

De acuerdo a la Sociedad Americana del Cáncer, el cáncer de seno (*BC* por sus siglas en inglés) se define como aquellas células de la mama o seno que crecen sin control alguno, estas células usualmente conforman un tumor, el cual se considera maligno si las células tienen la capacidad de invadir tejidos aledaños o propagarse a distintas áreas del cuerpo.

El cáncer es producido por alteraciones en el código genético, en la cadena de doble hélice de ácido desoxirribonucleico (ADN), dichas alteraciones se presentan como rearreglos en el cromosoma, mutaciones y cambios epigenéticos, como la activación de oncogenes o supresión de genes supresores tumorales, provocando que células normales se conviertan en células cancerígenas, algunos genes son heredados por los padres, lo cual incrementa el riesgo de desarrollar cáncer de seno (ACS, 2016).

El estilo de vida de la persona puede aumentar el riesgo de desarrollar cáncer de seno, aunque no está exactamente definido cómo estas acciones cotidianas afectan el desarrollo de las células. Las hormonas también juegan un rol importante en la mayoría de los casos de cáncer de seno, sin embargo, el mecanismo de acción aún no está completamente entendido (Provenzano et al., 2018; ACS, 2016).

1.3. Biomarcadores

Una de las herramientas usadas en los últimos años en el área biológica son los biomarcadores, que de acuerdo al diccionario del Instituto Nacional del Cáncer de Estados Unidos de América, un biomarcador es una molécula biológica encontrada en la sangre, fluidos corporales o tejidos, que es una señal de un proceso normal o anormal, o de una condición o enfermedad. Un biomarcador también puede ser usado para observar qué tan bien responde un cuerpo a un tratamiento, a una enfermedad o condición (NCI, 2018a).

De acuerdo a la Organización Mundial de la Salud (OMS), un biomarcador es cualquier substancia, estructura o proceso que pueda ser medido en el cuerpo o sus productos, e influenciar o predecir la incidencia de aparición o enfermedad (OMS, 2001).

Existen distintos usos para los biomarcadores, a continuación se describen algunos de los más importantes en el área médica. De proyección (*screening* por su nombre en inglés), en donde se detecta un riesgo previo al desarrollo de la enfermedad, es decir, se identifica una probabilidad de incidencia. Para diagnóstico, en donde se identifica la presencia de un estado o proceso patológico y como resultado se obtiene un diagnóstico para el paciente. Biomarcadores de pronóstico, en donde se usan para clasificar estado y etapa de la enfermedad, pero también para asistir en la selección de una modalidad de tratamiento, además, permiten identificar pacientes de alto riesgo de mortalidad o morbilidad. Como monitoreo de enfermedad, en donde la respuesta dinámica de los biomarcadores ayudan a conocer el estadio de la enfermedad y asisten en la generación de un diagnóstico. Biomarcadores para monitoreo de medicamentos, de forma similar al anterior la respuesta dinámica ante la presencia de medicamentos provee información acerca del proceso de respuesta ante una medicación en un paciente. Y como puntos finales sustitutos, en donde la presencia de estos biomarcadores está altamente relacionada con algún resultado de un proceso biológico (Vaughan, 2016; Henry and Hayes, 2012).

Los biomarcadores típicamente permiten diferenciar a un paciente enfermo de una

persona que no presenta la enfermedad en cuestión, y como consecuencia, pueden llegar a afectar la toma de decisiones por parte del personal médico (Henry and Hayes, 2012).

No importa si se miden en unidad o en grupo, los biomarcadores ofrecen asistencia más precisa al diagnosticar pacientes, permitiendo tratamientos específicos para cada paciente y obteniendo mejores resultados (Vaughan, 2016).

En los últimos años se han desarrollado y validado biomarcadores que permiten predecir la respuesta a terapias contra cáncer, así como predecir la probabilidad de desarrollarlo o la etapa de desarrollo del tumor (Núñez, 2019; Nicolini et al., 2018; Catherine Sánchez, 2013).

Entre los biomarcadores clásicos usados para detectar y clasificar el cáncer de seno están los genes *BRCA1*, *BRCA2* (*breast cancer* 1 y 2), el receptor de estrógeno (*ER*), receptor de progesterona (*PR*), receptor 1 de crecimiento epidermal humano (*HER1*) y receptor 2 de crecimiento epidermal humano (*HER2*) (Moo et al., 2018; Schnitt, 2010).

Existe una nueva generación de biomarcadores basados en muestras sanguíneas, los cuales ofrecen una alternativa no invasiva a la hora de mejorar la detección de cáncer, entre estos biomarcadores de nueva generación se encuentran proteínas, anticuerpos contra proteínas asociadas a antígenos tumorales, *miRNAs* (micro ácidos ribonucleicos, secuencias de 20-25 nucleótidos) que regulan la expresión de genes, metilaciones de ácidos nucleicos causando mutaciones que pueden activar oncogenes, metabolitos y lípidos que pueden ayudar a discriminar entre una paciente con BC y una persona sana. Todos los mencionados anteriormente, han mostrado un gran potencial para la detección no solo de BC, también para la detección de estadios pre invasivos y de desarrollo temprano de la enfermedad (Loke and Lee, 2018; Duffy et al., 2018; Uttley et al., 2016).

Estos biomarcadores son conocidos como de detección temprana, los cuales de acuerdo a Yeh (2015) y Levenson (2007) deben cumplir con las siguientes características:

- Detección temprana, proyección de la enfermedad en población asintomática y aplicable

para todo el cuerpo.

- Procedimiento mínimamente invasivo, requiere de poco material médico y normalmente hace muestreo de fluidos biológicos.
- Detección específica del sitio objetivo, es específico para el tejido u órgano.
- Tiempo de procesamiento, requiere poco tiempo para su procesamiento y análisis.
- Observación independiente, medición objetiva de características biológicas.
- Específico, bajo número de falsos positivos.
- Simple y de bajo costo, puede aplicarse a una gran cantidad de población.

La principal ventaja de usar la sangre para muestreo en busca de biomarcadores es que al ser una sustancia que está en contacto íntimo e interacciona con todas las células del cuerpo, integra todos los metabolitos, haciendo al flujo sanguíneo un buen prospecto para análisis en busca de biomarcadores de diagnóstico indirecto. Se sugiere que estos ensayos basados en muestras sanguíneas están dirigidos a pacientes asintomáticos, tomando como principal ventaja la medición de diferentes componentes en el fluído y enfocándose en combinaciones de elementos más que en componentes individuales, ya que las combinaciones pueden significar firmas específicas para la detección de la enfermedad y también su localización (Levenson, 2007).

1.4. Bases de datos

Al tener a disponibilidad una gran cantidad de información y datos crudos sobre biomarcadores, es necesario una herramienta para almacenar dicha información, para ello se hace uso de una base de datos, que es una colección estructurada de registros o datos que

son almacenados en una computadora y que pueden ser consultados mediante un programa para responder a consultas. Consiste en una o más tablas conformadas por filas y columnas, conteniendo valores de datos, cada fila contiene al menos una columna y cada columna contiene al menos un valor perteneciente a un atributo del objeto descrito por la base de datos. De tal forma que las columnas contienen múltiples valores para el mismo atributo, mientras que las filas contienen valores de atributos para una entidad o entrada (Schaeffer et al., 2014).

Estas bases de datos recopilan datos en grandes cantidades, para que después de un procesamiento se obtenga información que pueda ser útil para el desarrollo de nuevas herramientas y la identificación de nuevos objetivos y biomarcadores. Estas bases de datos permiten el análisis e integración de datos, visualización de los mismos, y sobre todo la manipulación de una copia para realizar nuevos descubrimientos con base en lo ya reportado.

El uso de los datos clínicos recopilados en bases de datos han permitido desarrollar herramientas y estrategias que ayudan a hacer frente desde varias perspectivas al cáncer, sin embargo, para poder hacer uso de estos datos, tuvieron que haber pasado por un procesamiento para estar disponibles al público o para los grupos de investigación en cuestión, en dicho proceso, se presentan cuatro principales barreras para el uso efectivo de bases de datos clínicos de acuerdo a Cross et al. (2018); Ward (2004); Rubenfeld et al. (1999).

La primera, refiere a la confidencialidad de los datos, cada dato del paciente debe ser protegido a la hora de ser visto por terceros, por lo que la identidad del paciente debe ser blindada. Regulaciones para lo anterior dependen de cada país y deben cumplirse con los estándares mínimos de privacidad establecidos para hacer uso correcto de dichos datos clínicos.

Segunda, la cantidad de datos es enorme, la frecuencia con la que los parámetros son registrados es alta, cada que se atiende a un paciente se colecta información del mismo ya sea física o electrónicamente, de igual forma cuando pasa por un examen de laboratorio clínico o en imagenología.

Tercera, la organización de los datos es un prerrequisito para la investigación, la estructura de la base de datos es primordial para un buen aprovechamiento de la misma. Es ideal que cada dato este etiquetado al ser guardado, de esta forma, el programa desarrollado puede relacionar tipos de datos sin tener que recorrer toda la base de datos, ahorrando tiempo y recursos.

Y por último, la cuarta barrera está relacionada con la calidad de los datos, la cual en muchas ocasiones es difícil de obtener y de evaluar. Lo anterior debido a dos aspectos importantes acerca de los conjuntos de datos, y tiene que ver con qué tanta cantidad de datos es ingresada de forma ideal o impecable (Hogan and Wagner, 1997). El primero se conoce como qué tan completos (*completeness* en inglés) son los datos, y se refiere a la proporción de observaciones que actualmente están registrados en el sistema y son legibles, es decir, el número de registros anotados completos o incompletos -por incompletos se entienden espacios vacíos, en blanco o ilegibles-, mientras que el segundo es conocido como que tan correctos son los datos (del inglés *correctness*), este se refiere a la proporción de registros anotados que están adecuadamente asignados a su etiqueta en el sistema, y donde el valor registrado para la variable en cuestión es claro y objetivo (Meyfroidt et al., 2009).

Dichas bases de datos registran grandes cantidades de bioinformación (información derivada del análisis físico o biológico, de un sistema biológico) relacionada al cáncer, aportando datos crudos acerca de distintas características de las células cancerígenas, así como características únicas de las mismas que puedan ser de interés para grupos de investigación (Nicolini et al., 2018).

Toda la información antes mencionada ha dado como resultado a una gran recopilación de datos, generando bases de datos como *The Genome Cancer Atlas* (TGCA), *Genes Expression Omnibus* (GEO), *Surveillance, Epidemiology, and End Results Program*(SEER) y *Embase*, las cuales han jugado un papel importante en el descubrimiento de nuevos biomarcadores, así como creación de estrategias para combatir las enfermedades a las cuales están asociadas (Zou

et al., 2015; Yang et al., 2015).

1.5. *Machine Learning*

Para generar soluciones basadas en datos crudos, se requiere un procesamiento, un análisis y por último se obtiene un producto o herramienta para dar solución a una problemática en cualquier sector de interés, en este caso el sector salud. Para ello, uno de los algoritmos más populares a usar es *Machine Learning* (ML), la disciplina en las ciencias computacionales en donde las computadoras son programadas para aprender patrones, de acuerdo a un conjunto de datos proporcionados.

El aprendizaje está definido por una serie de reglas matemáticas y atribuciones estadísticas, que en su conjunto forman reglas de aprendizaje. El principal objetivo de las técnicas de ML es producir un modelo que pueda realizar clasificación, predicción, estimación o cualquier tarea similar (Camacho et al., 2018).

Un clasificador es un sistema que introduce un vector de valores característicos de tipo discreto o continuo, y obtiene como salida un valor discreto. Los componentes de aprendizaje en un clasificador son, de acuerdo a Domingos (2012):

- Representación, un clasificador debe ser representado en un lenguaje formal que la computadora pueda manejar. Elegir una representación permite identificar los tipos de clasificadores que el modelo pudiera aprender.

- Evaluación, una función de evaluación es necesaria para distinguir entre un buen clasificador y uno no tan bueno.

- Optimización, se requiere de un método para buscar de entre todos los clasificadores probados, el que tenga mejor puntaje. Esta opción de optimización es la clave para la

eficiencia del aprendizaje y permite determinar si la función de evaluación tiene más de un clasificador óptimo.

Entre las técnicas de ML más comunes, se encuentran los árboles de decisión (*decision trees*), que son usados para regresión o clasificación, en el que se genera un diagrama de flujo en forma de árbol donde el camino que se decide tomar esta basado en el valor de la instancia en cuestión, resultando en la predicción de un valor o condición. Por otro lado, la técnica de bosque aleatorio (*random forest*), construye múltiples árboles y las predicciones son decididas por votación mayoritaria, es decir, la clasificación se realiza cuando la instancia se valora en cada uno de los árboles y la decisión es acordada con respecto al resultado de la mayoría de los árboles. También es popular hacer uso de máquinas de soporte vectorial (*support vector machines*), donde se crean hiperplanos para cada característica en un espacio dimensional infinito, y ajusta un modelo lineal o no lineal que mejor discrimine entre los valores binarios de una variable de salida. Otra de las técnicas usadas son las redes neuronales artificiales (*artificial neural networks*), donde se recrea la actividad biológica del cerebro humano, creando nodos (neuronas) que asignan pesos a los distintos valores de entrada y producen un valor de salida. Se pueden construir múltiples capas de nodos, cada una con una capa de entrada de datos independiente y una capa de salida con nodos que representa cada uno de los posibles valores de salida, los pesos de cada capa son ajustados hasta encontrar el peso ideal para generar un modelo que su salida se ajuste al valor deseado.

Algunas aplicaciones de las distintas técnicas de ML se pueden observar en seguridad de datos, donde ML puede predecir qué archivos son *malware* o *spam* dentro de la bandeja de entrada en el correo electrónico. Dentro del comercio financiero se usa para la predicción del comportamiento de las distintas monedas. En el cuidado de la salud las técnicas de ML son capaces de detectar patrones en imágenes médicas, identificando anomalías en el tejido. Facebook usa algoritmos que analizan la interacción del usuario con la plataforma y con base en esa información colectada muestra contenido similar que pueda ser de interés para

el usuario. Tesla produce automóviles con capacidades predictivas, gracias a la información colectada mediante sensores y cámaras del vehículo.

La aplicación correcta de las técnicas de ML puede encontrar patrones en bases de datos. En su mayoría, las técnicas de ML usadas para el manejo de enfermedades son redes neuronales artificiales, árboles de decisión, clasificación asociativa, máquinas de soporte vectorial, redes bayesianas y *K-nearest neighbor* por su nombre en inglés; para el diagnóstico, pronóstico y tratamiento de enfermedades como cáncer, hepatitis, enfermedades cardiovasculares, entre otras (El Houby, 2018; Cao et al., 2018; Turki and Wei, 2018; Xiao et al., 2018; Vidyasagar, 2017; Kourou et al., 2015).

Algunos ejemplos de aplicaciones de ML y otros modelos estadísticos en el diagnóstico, pronóstico y respuesta a tratamiento en el cáncer de seno se han trabajado por El Houby (2018); Richter and Khoshgoftaar (2018); Sherafatian (2018); Jafari-Marandi et al. (2018); Liu and Deng (2010).

El uso de las técnicas ya mencionadas, ayuda a abordar importantes desafíos para la salud, con la capacidad de mejorar los procesos y toma de decisiones dentro del sistema de salud, sin embargo, no sustituye al personal médico y profesional que atiende al paciente.

1.5.1. Redes neuronales artificiales

Una de las técnicas de ML con mayor popularidad es la red neuronal artificial (RNA), la cual es una herramienta para modelar datos con la habilidad de captar y representar relaciones complejas entre entradas y salidas en un modelo específico. El objetivo es crear un modelo que tenga la capacidad de relacionar una entrada con su respectiva salida, de tal forma que cuando se introduzca una entrada, el modelo presenta una salida, cuando esta última es desconocida. Una RNA está compuesta por múltiples capas que contienen nodos para relacionarse entre ellas, una capa de entrada, una o más capas ocultas y una capa de

salida para representar el resultado (El Houby, 2018; Hagan et al., 2014).

Los nodos o neuronas, de donde obtiene el nombre esta técnica, se encargan de recibir, procesar y transferir, mediante una función de transferencia, la información desde la capa de entrada hacia la capa de salida para obtener un resultado. Al encontrar relaciones y similitudes entre los datos de entrada del modelo, a cada conexión entre nodo se le es asignado un peso, que le da prioridad a la hora de conectarse con el nodo en la siguiente capa, el valor del peso indica la fuerza de la conexión entre nodos de diferentes capas. Las entradas al modelo son multiplicadas por el peso de cada nodo, al mismo tiempo, dicha entrada es procesada por una función de transferencia. La labor de dicha función es transformar los valores de entrada, limitando su rango de salida a valores que dependen de la función elegida. De esta forma se genera una salida que avanza a la siguiente capa de la red. Este proceso se repite hasta llegar a la capa de salida, obteniendo un resultado. El valor obtenido es comparado con el valor esperado y se calcula un valor para el error, posteriormente los nodos cambian sus pesos hasta que el error sea minimizado (Flores et al., 2017b; Havel et al., 2013; Jiménez, 2012; Agatonovic-Kustrin and Beresford, 2000).

Tratándose de un modelo de una sola neurona con una entrada y una salida, el sistema estaría descrito por la ecuación $a = f(wp + b)$, en donde f es la función de transferencia que se aplica a la entrada p y generando una salida escalar a de la neurona; w corresponde al peso de la conexión escalar de la neurona; y b es el sesgo o error (*bias* por su nombre en inglés) (Hagan et al., 2014). Por lo general, una neurona tiene varias entradas a la vez, por lo que la ecuación del sistema con N entradas se transforma en $a = f(\mathbf{Wp} + b)$, en donde $\mathbf{W}$ representa la matriz de los pesos de cada una de las entradas y $\mathbf{p}$ corresponde a las entradas p_1, p_2, ..., p_N. Por otro lado, la simplicidad de un modelo de una sola neurona no es suficiente para procesos de predicción o clasificación, ya que las variables y relaciones típicas de estos problemas suelen ser complejas. Una red de M neuronas y N entradas está representada por la ecuación $\mathbf{a = f(Wp + b)}$. Donde el vector de salida $\mathbf{a}$, está formado por las salidas

a_i, ..., a_S pertenecientes a cada neurona, mientras que **f** representa al vector que agrupa las funciones de activación, **b** corresponde al vector de b_i errores en donde i señala a la neurona correspondiente, y **W** es la matriz de $w_{S,N}$ pesos de entradas, el índice S indica la neurona y N la salida (Castañeda-Martínez, 2017).

1.6. Justificación

Hay distintos estudios en donde se han usado gran variedad de biomarcadores sanguíneos para la detección de cáncer de seno y en donde se proponen otros tantos. El principal objetivo es encontrar alternativas que mejoren el diagnóstico y detección de cáncer de seno en etapas de temprano desarrollo, cuando las técnicas de imagen son incapaces de proveer suficiente información (Núñez, 2019; Sumbal et al., 2018; Lourenco et al., 2017; Lyng et al., 2016; Yang et al., 2015; Santillán-Benítez et al., 2013; Pruthi et al., 2012; Hanash et al., 2011).

La presencia o ausencia de estos biomarcadores sanguíneos pueden dar un indicativo de presencia de la enfermedad. Lo anterior genera una gran cantidad de datos que se almacenan y que tienen el potencial de proveer información importante para el diagnóstico del cáncer de seno. El uso de técnicas de ML tienen la capacidad de traducir estos datos en información útil y que permita al personal médico tomar una decisión rápida y certera de acuerdo a la información arrojada por la herramienta desarrollada. Como consecuencia se obtiene un ahorro en tiempo y recursos económicos para el diagnóstico de la enfermedad, pero sobre todo el paciente recibirá una respuesta pronta y confiable acerca de su diagnóstico.

De acuerdo a otras investigaciones (Kazarian et al., 2017; Lourenco et al., 2017; Levenson, 2007), es claro que el diagnóstico de BC mediante imágenes médicas es una de las formas de diagnóstico más usadas que mejor resultado han mostrado en los últimos años, principalmente la mamografía, sin embargo, existe una desventaja cuando dicha técnica se aplica en población que tiene alta densidad de tejido, que estima a la población menor de 45-50 años

aproximadamente, ya que la sensibilidad de la mamografía es mucho menor, disminuyendo de un 67 % a un 45 % aproximadamente, y para mujeres en tratamiento hormonal hasta un 25 %. Por lo anterior, los biomarcadores sanguíneos toman importancia ya que si son identificados correctamente pueden detectar mediante un muestreo rutinario de sangre, la presencia o el riesgo de desarrollar cáncer de seno.

Sin embargo, la implementación de estos biomarcadores debe ser validada para que la cantidad de falsos positivos y falsos negativos se reduzcan lo máximo posible. Hay dos maneras de verificar si un grupo de moléculas pueden ser o no usadas como biomarcadores, la primera es llevar a cabo una gran cantidad de estudios a pacientes y verificar mediante biopsia del tejido si dichas moléculas están presentes o no en todos los pacientes comparado con personas sanas. Una desventaja de este método es que requiere de suficiente tiempo para reclutar a las personas que deseen ser parte del estudio. La segunda forma es hacer uso de herramientas computacionales para encontrar patrones de expresión o presencia de las moléculas propuestas como biomarcadores. Los análisis realizados con ML permiten identificar patrones que una persona no sería capaz de reconocer o le tomaría mucho tiempo en deducirlo, además, es necesario un número relativamente pequeño de pacientes para realizar dichos análisis, y la única condición es que todos los datos a usar deben estar completos y sin ningún error.

En distintos estudios se han usado diferentes técnicas de ML para demostrar la capacidad de estas técnicas en la predicción (Wang et al., 2018; Behravan et al., 2018; Tapak et al., 2018) , clasificación (Guo et al., 2019; Xu et al., 2019; Ehteshami Bejnordi et al., 2018; Vandenberghe et al., 2017) y pronóstico (Kourou et al., 2015; Kalinli et al., 2013) de BC mediante biomarcadores o imágenes médicas. Cada uno de ellos obtienen resultados prometedores y en algunos casos aseguran que la sensibilidad del método podría mejorar ya sea por modificaciones de la misma técnica de ML o por el uso en conjunto de dos o más.

Una de las líneas de investigación de gran relevancia es el diagnóstico a pacientes que sufren de sobrepeso u obesidad y están en periodo de pre o postmenopausia, debido a que su

condición física y etapa de adultez arrojan información acerca de una variación en la presencia de moléculas en el torrente sanguíneo. Lo anterior es un producto de una desregulación en pacientes sobre factores que involucran variables como los niveles de adipocitoquinas, el índice de masa corporal (IMC), concentración de glucosa, concentración de insulina, entre otras más (Crisóstomo et al., 2016; Georgiou et al., 2016; Santillán-Benítez et al., 2013; Dalamaga et al., 2013; Grossmann et al., 2010; Chien-An et al., 2010).

Por otra parte, al padecer sobrepeso u obesidad, las pacientes corren el riesgo de desarrollar diabetes. Por separado o en combinación de ambos padecimientos, algunas moléculas presenten concentraciones anormales, que implican un riesgo para desarrollar BC (Dalamaga, 2014, 2013; Vona-davis and Rose, 2012; Rose and Vona-davis, 2012; Cohen and Leroith, 2012; Anderson and Neuhouser, 2012; Xue and Michels, 2007; Lorincz and Sukumar, 2006).

Recientemente se realizó un estudio en donde se proponen las siguientes variables como potencial grupo de biomarcadores para diagnóstico de cáncer de seno (Patrício et al., 2018):

- Índice de masa corporal (IMC o *BMI* por sus siglas en inglés), que es la asociación entre la masa y la talla de un individuo, dada por el cociente entre la masa y la estatura elevada al cuadrado. Si el valor es mayor a 25, se dice la persona padece de sobrepeso y si es mayor a 30 padece de obesidad, $IMC = kg/m^2$.

- Leptina, hormona producida en su mayoría por los adipocitos (células grasas). Está asociada a la regulación del peso corporal.

- Adiponectina, hormona sintetizada por el tejido adiposo, participa en el metabolismo de la glucosa y ácidos grasos. Aumenta la sensibilidad a la insulina, y sus niveles circulantes son inversamente proporcionales al IMC.

- Resistina, proteína secretada por células grasas y tiene presunta implicación en la resistencia a la insulina en personas con obesidad.

- Modelo Homeostático para Resistencia a Insulina (HOMA u HOMA-IR), valora si existe un bloqueo o resistencia a la acción de la insulina.

- Glucosa, indicador de presencia de azúcar presente en la sangre.

- Insulina, hormona secretada por el páncreas, su principal función es la intervención en el aprovechamiento metabólico de nutrientes.

- Proteína Quimiotáctica de Monocitos 1 (MCP1), proteína involucrada en el reclutamiento de células inmunes en sitios de inflamación debido a infección o tejido dañado.

- Edad, puede dar indicativo aproximado del estado de menopausia en el que se encuentra la paciente, pre o postmenopausia.

Dichas moléculas se encuentran presentes en la sangre y pueden ser identificadas mediante un análisis de laboratorio clínico y están en su mayoría relacionadas a obesidad y diabetes. Se añade la edad como un factor clave ya que ésta puede ser indicativo de una etapa pre o post menopáusica.

Lo anterior toma relevancia ya que de acuerdo a datos de la Organización Mundial de la Salud, en el año 2014 había 422 millones de personas con diabetes (OMS, 2018b); en el 2016 más de 1,900 millones de adultos mayores a 17 años tenían sobrepeso u obesidad (OMS, 2018a); en el 2018 el cáncer de seno fue el tipo de cáncer más frecuente en mujeres, causando aproximadamente el 15 % de muertes anuales relacionadas al cáncer, eso es un estimado de 627,000 mujeres, y afectando a un total de casi 2.1 millones (OMS, 2019). Y según datos del Observatorio Global del Cáncer, Figura 1.1, tan sólo en el 2018 se registraron a nivel mundial poco más de 2 millones de nuevos casos de BC en mujeres de todas las edades, representando el 24.2 % del total de incidencia y el más frecuente de los tipos de cáncer en la mujer (International Agency for Research on Cancer, 2019).

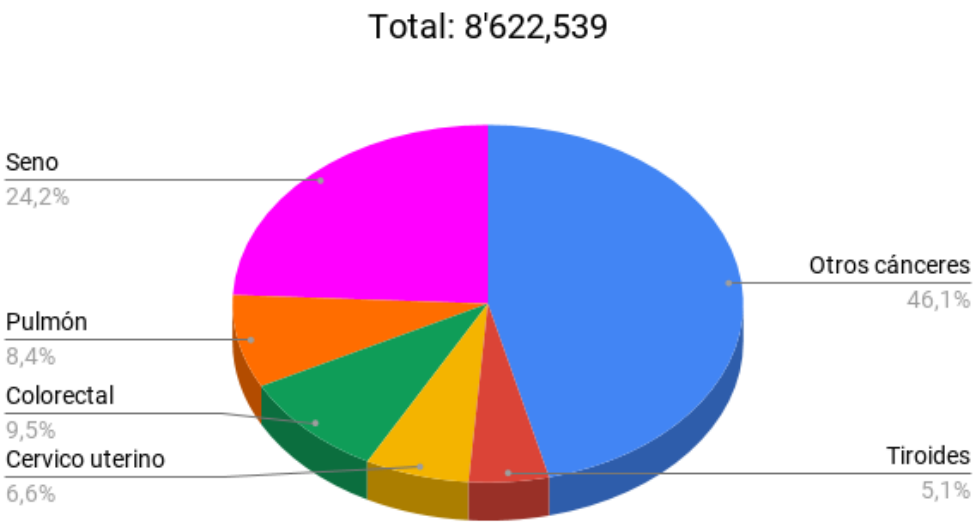

Figura 1.1: Estadística de incidencia de cáncer en mujeres de todas las edades en el 2018. Fuente: (International Agency for Research on Cancer, 2019).

Tomando en cuenta los datos anteriores, la amenaza de desarrollar cáncer de seno por factores como sobrepeso u obesidad y diabetes supone un gran riesgo para la salud de dichas personas, por lo que de gran importancia validar y encontrar nuevas estrategias mediante el uso e implementación de técnicas de ML que permitan un diagnóstico oportuno y eficaz para el cáncer de seno.

1.7. Objetivos

1.7.1. Objetivo general

Usar redes neuronales artificiales como método para validar los biomarcadores edad, índice de masa corporal, leptina, adiponectina, MCP1, resistina, HOMA, glucosa e insulina, en la predicción de cáncer de seno.

1.7.2. Objetivos específicos

- Utilizar las redes neuronales artificiales y construir una configuración con alto valor para el área bajo la curva, en la clasificación de pacientes sanos y con cáncer de seno.
- Mejorar los valores de las métricas de evaluación reportadas por Patrício et al. (2018).
- Comparar con las propuestas reportadas y seleccionar la mejor técnica de *Machine Learning* en la clasificación de cáncer de seno.

1.8. Hipótesis

Los biomarcadores de índice de masa corporal, leptina, adiponectina, MCP1, resistina, HOMA, glucosa e insulina, así como la variable cuantitativa de edad, son validados mediante redes neuronales artificiales para la predicción de diagnóstico de cáncer de seno en pacientes con sobrepeso u obesidad.

Capítulo 2

Metodología

La metodología inicia con un análisis estadístico y preprocesamiento de los datos, con el fin de conocer el comportamiento de los datos y mejorar la capacidad de predicción del modelo. En este capítulo se construyen distintas variantes de modelos, y por último, se hace un análisis de la capacidad de clasificación. El capítulo termina con una comparación de métricas de evaluación.

2.1. Metodología

2.1.1. Conjunto de datos

Basado en los objetivos para este trabajo y en lo reportado anteriormente por Patrício et al. (2018), se hace uso de la propuesta de biomarcadores y del conjunto de datos usado por dicho estudio.

Se descargó el conjunto de datos con nombre *Breast Cancer Coimbra Data Set*, desde el *Machine Learning Repository Center for Machine Learning and Intelligent Systems* de la *University of California in Irvine* (Dua and Karra Taniskidou), en dicho conjunto de datos se recopila información acerca de nueve predictores, todos cuantitativos, y una variable

binaria dependiente, indicando la presencia o ausencia de cáncer. Los predictores son datos antropométricos y parámetros que pueden ser obtenidos durante un análisis de sangre rutinario.

Los atributos cuantitativos son:

- Edad, expresada en años, número entero.
- Índice de masa corporal (IMC), expresado en kilogramos por metro cuadrado, kg/m^2.
- Glucosa, valor presentado en miligramos por decilitro, mg/dL.
- Insulina, dada en microunidades por mililitro, $\mu U/mL$.
- Modelo homeostático para resistencia a la insulina (HOMA), expresado por un número positivo.
- Leptina, expresada en nanogramos por mililitro, ng/mL.
- Adiponectina, dada en microgramos por mililitro, $\mu g/mL$.
- Resistina, expresada en nanogramos por mililitro, ng/mL.
- Proteína quimiotáctica de monocitos 1 (MCP-1), expresada en picogramos por decilitro, pg/dL.

El atributo de clasificación consta de un valor numérico entero, donde el valor de *1* corresponde a paciente sano y el valor numérico *2* a paciente con cáncer de seno. Este último atributo se reemplazó por *0* y *1*, respectivamente, para mayor simplicidad.

El conjunto de datos consta de un archivo tipo Excel (.csv) con 116 filas y 10 columnas. No contiene espacios en blanco o ilegibles.

2.1.2. Análisis estadístico del conjunto de datos

Para el análisis estadístico los datos se dividieron en dos grupos, aquellos datos pertenecientes a pacientes con diagnóstico de cáncer de seno y pacientes sanos. A partir de esta partición se llevaron a cabo pruebas para comprobar que los datos pertenecen a una población con distribución normal mediante el test de *Shapiro Wilk*, ecuación (2.1),

$$W = \frac{1}{ns^2}(\sum_{i=1}^{h} a_j(x_{(i)})) \tag{2.1}$$

si n es par

$$h = \frac{n}{2} \tag{2.2}$$

si n es impar

$$h = \frac{n-1}{2} \tag{2.3}$$

donde W es el estadístico de contraste, n el tamaño de muestra, s^2 corresponde a la varianza, a_j es un coeficiente tabulado, $x_{(i)}$ es el valor en la i-ésima posición tras la ordenación, y se obtiene una probabilidad crítica, valor-p. Como hipótesis nula, H_0, se asume que la muestra proviene de una población con distribución normal, dicha hipótesis se rechaza en caso de que el valor-p sea menor al valor de significancia, α, el cual será de 5 % para este trabajo.

Después de haber realizado las pruebas de normalidad se realizó la prueba U de *Mann-Whitney*, una prueba no paramétrica para analizar la independencia de muestras, para ello se compararon las medianas y los intercuartiles de los dos grupos de paciente, control y con diagnóstico de cáncer de seno, y cada una de las variables. La hipótesis nula, H_0, establece que ambas muestras poblacionales no son independientes una de la otra y pertenecen a la misma población. La hipótesis alternativa, H_A, indica que hay diferencia significativa entre ambas muestras si el valor valor-p es mayor al valor de significancia α, que para este trabajo es del 5 %.

Para llevar a cabo la prueba U de *Mann-Whitney* se hace el cálculo del estadístico U, para cada una de las muestras usando las ecuaciones (2.4) y (2.5).

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{2.4}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{2.5}$$

En donde n_1 y n_2 son los tamaños de cada muestra, mientras que R_1 y R_2 es la suma de los rangos asignados a cada una de las observaciones de cada muestra una vez ordenadas de menor a mayor, y U representa el estadístico a obtener. Se selecciona el estadístico U de menor valor, el cual es usado para obtener nuestro valor z o valor-p ya que el tamaño de la muestra es mayor a 10 y se aproxima a una distribución normal. Para obtener el estadístico z a partir del valor de U se usan las siguientes ecuaciones, (2.6),(2.7),(2.8) ,

$$z = \frac{U - m_U}{\sigma_U} \tag{2.6}$$

$$m_U = \frac{n_1 + n_2}{2} \tag{2.7}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{2.8}$$

Una vez obtenido el valor-p y confirmada la independencia de los datos, se procede a realizar el pre procesamiento de los mismos para la construcción de los modelos de clasificación.

Lo anterior se realizó haciendo uso del software R versión 3.5.1 (2018-07-02) (R Core Team, 2014).

2.1.3. Preprocesamiento de los datos

Los datos en formato excel (.csv) fueron importados al software MATLAB R2016b (MathWorks Inc., Natick, MA, E.U.A.). Se separaron en 10 vectores, nueve de ellos son predictores y uno de clasificación. Para todos los vectores, se realizó una normalización basada en la unidad, en donde en cada vector se identificó el mínimo y máximo local (Han et al., 2011), y los valores se escalaron en el rango [0,1] mediante la ecuación (2.9).

$$X_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{2.9}$$

En donde X_i es el valor escalado en el rango [0,1], x_i corresponde al valor actual a escalar, x_{min} es el mínimo local y x_{max} el máximo local del vector.

De acuerdo al análisis realizado por Patrício et al. (2018), la importancia relativa de los predictores, de mayor a menor importancia se presenta en la Tabla 2.1, al mismo tiempo se les asignó una etiqueta con las que se les identificará a partir de este momento para los experimentos a realizar.

Tabla 2.1: Variables

Biomarcador	Variable
Glucosa (*Glucose*)	V1
Resistina (*Resistin*)	V2
Edad (*Age*)	V3
IMC (*BMI*)	V4
HOMA	V5
Leptina (*Leptin*)	V6
Insulina (*Insulin*)	V7
Adiponectina (*Adiponectin*)	V8
MCP.1	V9

Se hicieron dos grupos de experimentación, el primero consta de las variables V1 a V4 con el propósito de mejorar el desempeño reportado por Patrício et al. (2018) al usar dichos

predictores en algoritmos de *support vector machine*, *random forest* y *logistic regression*; y el segundo grupo consta de las variables V1 a V9, poniendo a prueba todos los predictores disponibles en el conjunto de datos. Ambos grupos hacen uso del mismo vector de clasificación.

Para ambos grupos de experimentación, las entradas al algoritmo se normalizaron a valores de Z en donde cada columna tiene valores de media = 0 y desviación estándar = 1, mediante la función de pre procesamiento *mapstd* descrita por la ecuación (2.10).

$$z_n = \frac{x_n - \bar{X}}{\sigma} \tag{2.10}$$

en donde z_n es el valor estandarizado de la observación n, x_n es el valor original de la observación n, $\bar{X}$ y σ son la media y la desviación estándar de la variable x, respectivamente.

Los datos de *salida* del preprocesamiento son normalizados de una forma centralizada en el rango [-1,1] con la ecuación (2.11).

$$X_n = a + \frac{(x_n - x_{min})(b - a)}{x_{max} - x_{min}} \tag{2.11}$$

donde X_n es el valor escalado de la observación n, x_n es el valor de la observación en cuestión, mientras que a y b son el valor máximo y mínimo, respectivamente, del rango del escalamiento, x_{min} y x_{max} son el valor mínimo y máximo, respectivamente, del conjunto de datos x.

2.1.4. Construcción de modelos de clasificación

2.1.4.1. Clasificación usando redes neuronales artificiales

La topología de RNA usada para los experimentos es la del perceptrón multicapa (PMC), en donde la unidad básica es el perceptrón (Figura 2.1). La función de esta unidad es sumar todas las señales de entrada y multiplicarla por la suma de los pesos previamente inicializados

aleatoriamente. Esta topología es capaz de distinguir patrones complejos en los datos de entrada una vez que se presenta nueva información a la red asociada a su respectiva salida. (Agatonovic-Kustrin and Beresford, 2000)

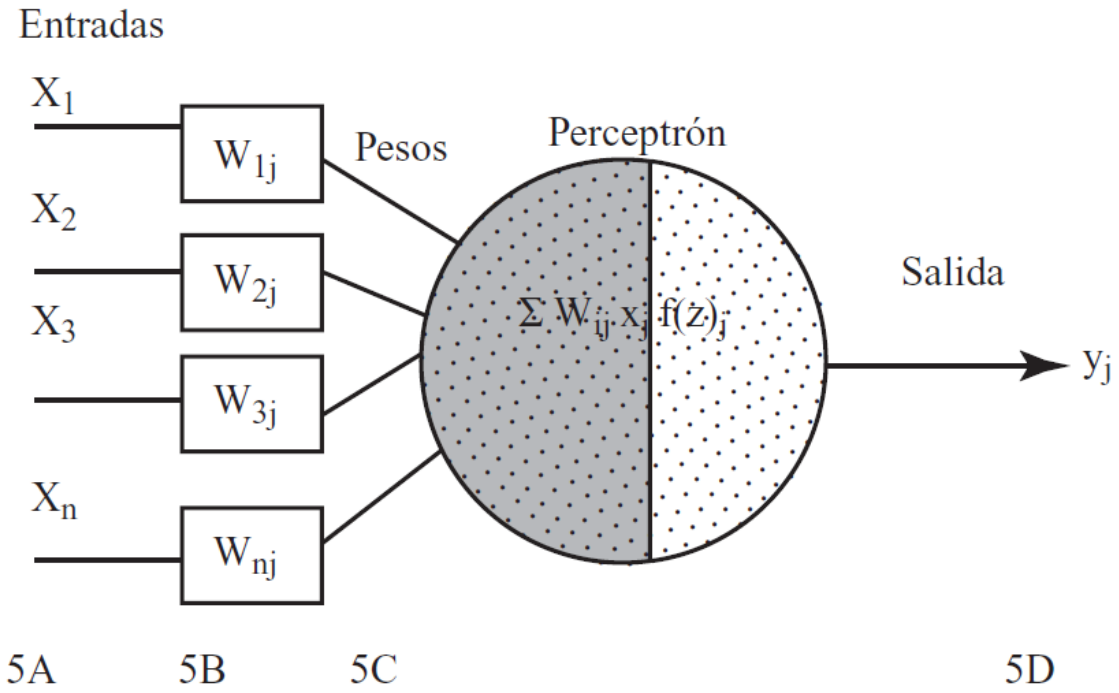

Figura 2.1: El perceptrón, unidad elemental de las redes neuronales artificiales que realiza la suma de los valores numéricos de las entradas para generar la salida. 5A= entradas a la j-ésima neurona, 5B = pesos asignados a cada una de las entradas de la j-ésima neurona, 5C= perceptrón en la j-ésima neurona, 5D= salida generada de la j-ésima neurona.

Se configuraron las redes para contener desde una hasta cinco capas ocultas, con el fin de identificar el mejor desempeño con cada grupo de experimentación con distintas configuraciones. Todas y cada una de las capas utilizaron la tangente hiperbólica como función de activación, descrita por la ecuación (2.12).

$$f(p) = \frac{2}{1 + \exp^{-2n}} - 1 \tag{2.12}$$

en donde p es cada dato de entrada de la red. Esta es una función sigmoide que ayuda a

que las RNA aprendan más rápido en relación con otras funciones y se utiliza cuando la normalización de los datos va de -1 a 1. Para la capa de salida se utilizó una función de transferencia de tipo tangente hiperbólica sigmoidal, la cual está reportada de obtener mejores resultados en problemas de clasificación (Riesel, 2007; Duch and Jankowski, 1999).

Para entrenar a las redes neuronales se utilizaron los vectores extraídos del conjunto de datos *Breast Cancer Coimbra Data Set*, y donde el algoritmo de entrenamiento usado fue el de retropropagación (del inglés *backpropagation*), ya que es uno de los algoritmos más utilizados para el entrenamiento de RNA debido a su efectividad (Castañeda-Martínez, 2017; Flores et al., 2017b,a; Jiménez, 2012). Este algoritmo consiste en propagar el error de las salidas de vuelta a todas las neuronas, de manera que los pesos se actualizan de adelante hacia atrás (Erb, 1993). Se utilizan tres variantes de este algoritmo para comparar su desempeño y seleccionar la mejor configuración: *Scaled Conjugate Gradient* (SCG), *Conjugate Gradient Backpropagation with Powell-Beale Restarts* (CGB) y *Resilient Backpropagation* (RP).

Experimento 1

Este entrenamiento tiene las siguientes características: una configuración PMC, que se describe como 9 - N - 1, en donde 9 y 1 son el número de entradas y salidas, respectivamente, y N es el número de neuronas en la capa oculta del PMC, en este caso se configura un número que va de 4 a 28, con incremento de 2 unidades.

Experimento 2

El entrenamiento 2 tiene las siguientes características: configuración de redes de tipo PMC, 9 - N - N - 1, en donde 9 y 1 son el número de entradas y salidas, respectivamente, y N es el número de neuronas en las capa ocultas del PMC, en este caso la primera capa oculta, asciende desde 16 hasta 24 nodos, con incremento de 4 unidades, mientras que la segunda

capa oculta, asciende desde 16 hasta 24 nodos, con incremento de 2 unidades.

Experimento 3

Para este entrenamiento se usaron las siguientes características: configuración de redes de topología perceptrón multicapa, 9 - N - N - N - 1, en donde 9 y 1 son el número de entradas y salidas, respectivamente, y N es el número de neuronas en las capa ocultas del PMC. En este caso la primera capa oculta, asciende desde 16 hasta 24 nodos, con incremento de 4 unidades, mientras que la segunda y tercera capa oculta, ascienden desde 16 hasta 24 nodos, con incremento de 2 unidades.

Experimento 4

En este entrenamiento se tienen las siguientes características: configuración de redes de topología PMC, 9 - N - N - N - N - 1, en donde 9 y 1 son el número de entradas y salidas, respectivamente, y N es el número de neuronas en las capas ocultas del PMC. En este caso la primera y segunda capa oculta, ascienden desde 16 hasta 24 nodos, con incremento de 4 unidades, mientras que la tercera y cuarta capa oculta, ascienden desde 16 hasta 24 nodos, con incremento de 2 unidades.

Experimento 5

Para este entrenamiento las características son las siguientes: configuración de redes tipo PMC, 9 - N - N - N - N - N - 1, en donde 9 y 1 son el número de entradas y salidas, respectivamente, y N es el número de neuronas en las capas ocultas del PMC. En este caso la primera capa oculta, va desde 20 hasta 24 nodos, con incremento de 4 unidades, mientras que la segunda y tercera capa oculta, ascienden desde 16 hasta 24 nodos, con incremento de 4 unidades y, las capas cuarta y quinta van desde 12 hasta 24 nodos, con incremento de 4

unidades.

La variación del número de nodos e incrementos en las capas ocultas se decidió después de haber analizado los desempeños del entrenamiento *n-1* y seleccionando el número de nodos con mejores valores de evaluación.

Para los experimentos del grupo V1 - V4 se llevaron a cabo los mismos experimentos, la variación es el número de entradas, cuatro entradas y una salida.

Validación

Para validar las RNA se utilizó el método de submuestreo aleatorio (*random sub sampling* por su nombre en inglés), que es una validación cruzada que consta de dividir aleatoriamente los datos en subconjuntos para entrenar y probar las redes construidas, con subconjuntos distintos para cada iteración (Nejatian et al., 2018). Este método se aplicó a todas las RNA para los grupos de experimentación.

A su vez, se utilizó un método conocido como validación *hold out* para el cual se utilizó el criterio de parada *early stopping*, que consiste en detener el entrenamiento de la red cuando el error del subconjunto de validación deja de disminuir y comienza a aumentar (Hagan et al., 2014; Jiménez, 2012). Para esta validación, el conjunto de datos se dividió en tres subconjuntos de la misma proporción usada por Patrício et al. (2018): 68.96 % para el entrenamiento (*train*), 15.52 % para la validación (*validation*) y el 15.52 % restante para probar el desempeño de la red (*test*). Cuando no se utilizó la validación *hold out*, las redes se entrenaron hasta llegar a 50 épocas -así se les conoce a las iteraciones pertenecientes al entrenamiento de una red- usando el 84.48 % de los datos mientras que el 15.51 % restante se usó para validar la red.

Cada configuración de red distinta se corrió 30 veces con y sin validación *hold out*, para cada configuración se identifica el valor promedio máximo de acuerdo al área bajo la curva (AUC) obtenida de la curva ROC. Este proceso es un método adicional de validación cruzada

k-fold, para este caso fue 30-*fold*.

Evaluación

Para evaluar el desempeño de los modelos se usaron los siguientes parámetros usados por Novakovic et al. (2017); Fawcett (2006); Bradley (1997). Uno de ellos es la sensibilidad, en donde se expresa la proporción de casos positivos (con cáncer de seno) correctamente identificados, la ecuación (2.13) muestra como obtener dicho valor. Y la especificidad, que expresa la proporción de casos negativos (pacientes sanos) correctamente clasificados, en la ecuación (2.14) se muestra como obtener dicho valor. Ambos parámetros toman valores de entre 0 y 1, y son obtenidos con ayuda de la matriz de confusión, donde se muestran las diferencias entre las clases verdaderas y predichas para cierto grupo de ejemplos etiquetados, la Tabla 2.2 muestra dicha comparación entre los valores verdaderos y los de predicción.

Tabla 2.2: Matriz de confusión.

		Valor verdadero	
		Positivo	Negativo
Predicción	Positivo	TP	FP
	Negativo	FN	TN

TP = Positivo Verdadero, FP = Falso Positivo, FN = Falso Negativo, TN = Negativo Verdadero

$$Sensibilidad = \frac{TP}{TP + FN} \tag{2.13}$$

$$Especificidad = \frac{TN}{TN + FP} \tag{2.14}$$

El área bajo la curva ROC, es una herramienta gráfica donde se muestra la habilidad de clasificar correctamente los casos positivos y los casos negativos que fueron clasificados incorrectamente. Donde el valor de AUC de 1.0, indica una capacidad perfecta de diferenciar entre

pacientes control y enfermos, y un valor de 0, es la nula capacidad de clasificar correctamente.

Sin embargo, valores de AUC menores a 0.5 se consideran como pruebas no informativas ya que asigna clasificaciones aleatoriamente, 50 % de los positivos son asignados al azar al igual que el 50 % de los negativos, dando como resultado una diagonal principal que va desde [0,0] hasta [1,1].

Por último, el índice de Youden (J), que se muestra en la ecuación (2.15), ya que de acuerdo a Estrada and Luna (2016); Noguera Moreno (2010), es frecuentemente usado en la práctica médica y que se obtiene de la suma de sensibilidad y especificidad menos una unidad.

$$J = Sensibilidad + Especificidad - 1 \tag{2.15}$$

El índice se puede entender como la ganancia promedio de certidumbre neta a la hora de la clasificación en un paciente sano o enfermo. Los valores del índice están en el rango [0,1], donde el valor de 0 corresponde a una prueba incapaz de detectar o descartar el padecimiento, mientras que el valor de 1.0 indica un test capaz de distinguir perfectamente entre pacientes control y enfermos. Por lo anterior, J se define como la distancia vertical máxima entre la diagonal principal y la curva ROC.

Se obtuvo la media para cada uno de los parámetros antes mencionados y para las distintas configuraciones, con el objetivo de obtener un resultado generalizado para cada configuración.

Para evaluar la exactitud del diagnóstico del modelo, se llevó a cabo una prueba de clasificación, introduciendo cada entrada del conjunto de datos a los modelos construidos (Flores et al., 2017b; Babaoglu et al., 2009); las clasificaciones resultantes de la predicción fueron comparadas con el valor real obtenido en las pruebas.

Al tratarse de un problema de clasificación binaria, los resultados de las pruebas se redondean al entero más cercano, cuando un elemento tiene una parte fraccional de exactamente

0.5, se redondea hasta el entero con mayor magnitud (Shanker and Hu, 1996).

Se usó el Neural Network Toolbox TM de MATLAB 8.5 R2016b para construir las RNA-PMC. El entrenamiento se modifica y optimiza de acuerdo con el criterio predeterminado del Toolbox, en donde se asume que a menor valor de *crossentropy*, el modelo construido simula mejor el proceso de diagnóstico y clasificación del paciente.

2.1.5. Selección de modelos de clasificación

Para la selección de las mejores configuraciones de clasificación se evaluó por importancia en el siguiente orden los valores obtenidos de los parámetros, AUC, desviación estándar, los valores mínimo y máximo de AUC, y el índice de Youden. Lo anterior se realizó haciendo uso del software R versión 3.5.1 (2018-07-02)(R Core Team, 2014).

La configuración seleccionada debía cumplir con el valor promedio de AUC más alto posible, el menor valor promedio posible de desviación estándar para la configuración en cuestión, un rango de AUC lo más angosto posible y el valor promedio más alto posible para el índice de Youden.

Lo anterior se realizó para todos los grupos de experimentación sin excepción alguna y se identificaron las mejores configuraciones para cada uno de los algoritmos de entrenamiento con y sin validación *hold out*.

Capítulo 3

Resultados

A continuación se presentan los resultados obtenidos de los experimentos para los distintos modelos construidos. Se muestran las métricas de desempeño y el capítulo termina haciendo una comparación entre los modelos construidos en este trabajo y los reportados por Patrício et al. (2018).

3.1. Análisis estadístico de los datos

En la Tabla 3.1, se muestran los resultados de normalidad de los datos, para todas las variables la hipótesis nula es rechazada y se concluye que dichas muestras no pertenecen a una población con distribución normal. A excepción de las variables de glucosa en pacientes control e IMC en pacientes con cáncer de seno, en donde la hipótesis nula no se rechaza y se asume que las muestras pertenecen a una población con distribución normal.

Para la independencia de los datos realizada por la prueba U de *Mann Whitney*, mostrados en la Tabla 3.2, los resultados obtenidos indican que las variables de edad, IMC, leptina, adiponectina, y proteína quimiotáctica 1, aceptan H_0 al mostrar resultados no significativos entre ambas muestras, pacientes control y con cáncer de seno, por lo que no

Tabla 3.1: Prueba de *Shapiro Wilk* para normalidad de datos

Variable	Valor-p$_{Control}$	Valor-p$_{PacienteBC}$
Edad	0,0048	0,0048
IMC	0.0312	0.1397
Glucosa	0.3944	0
Insulina	0	0
HOMA	0	0
Leptina	0.0001	0
Adiponectina	0	0
Resistina	0	0
MCP.1	0.0209	0

Nivel de significancia adoptado α= 5 % .
Valor-p$_{Control}$ = paciente control,
Valor-p$_{PacienteBC}$ = paciente con cáncer de seno,
HOMA= índice homeostático de resistencia a insulina,
MCP.1= proteína quimiotáctica de monocitos 1.

Tabla 3.2: Prueba de U *Mann Whitney*

Variable	Mediana$_{Control}$	$IQR_{Control}$	Mediana$_{PacienteBC}$	$IQR_{PacienteBC}$	Valor-p
Edad	65	33.2	53	23	0.4789
IMC	28.3	5.4	27	4.6	0.2017
Glucosa	88.2	10.2	105.6	2.6	0
Insulina	6.9	4.9	12.5	12.3	0.0266
HOMA	1.6	1.2	3.6	4.6	0.0029
Leptina	26.6	19.3	26.6	19.2	0.9491
Adiponectina	10.3	7.6	10.1	6.2	0.7665
Resistina	11.6	11.4	17.3	12.6	0.0019
MCP.1	499.7	292.2	563	384	0.5035

Nivel de significancia adoptado α= 5 % .
IQR$_{Control}$ = intercuartil paciente control,
IQR$_{PacienteBC}$ = intercuartil paciente con cáncer de seno,
HOMA= índice homeostático de resistencia a insulina,
MCP.1= proteína quimiotáctica de monocitos 1.

se puede afirmar que exista diferencia significativa entre ambos grupos. Por otra parte, para las variables glucosa, insulina, *HOMA* y resistina, se rechaza la hipótesis nula, y se concluye que ambos grupos pertenecen a poblaciones independientes. Por lo tanto, hay diferencia

significativa entre pacientes control y pacientes con cáncer de seno.

3.2. Preprocesamiento del conjunto de datos

En la Tabla 3.3 se muestra un extracto del conjunto de datos original, en donde se aprecian las nueve variables que lo conforman, así como los valores correspondientes a las primeras cinco filas. Mientras que en la Tabla 3.4 se muestra el mismo extracto con la diferencia de que presenta los datos escalados, producto del pre procesamiento.

Tabla 3.3: Extracto del conjunto de datos original

Age	BMI	Glucose	Insulin	*HOMA*	Leptin	Adiponectin	Resistin	*MCP.1*	Class
48	23.5	70	2.70	0.46	8.80	9.70	7.99	417.11	1
83	20.69	92	3.11	0.70	8.84	5.42	4.06	468.78	1
82	23.12	91	4.49	1.00	17.93	22.43	9.27	554.69	1
68	21.36	77	3.22	0.61	9.88	7.16	12.76	928.22	1
86	21.11	92	3.54	0.80	6.69	4.81	10.57	773.92	1

Age= edad, BMI= índice de masa corporal, Glucose= glucosa, Insulin= insulina, *HOMA*= índice de resistencia a la insulina, Leptin= leptina, Adiponectin= adiponectina, Resistin= resistina, *MCP.1*= proteína quimiotáctica de monocitos 1, Class= clase / clasificación.

Tabla 3.4: Extracto del conjunto de datos después del escalamiento de datos

Age	BMI	Glucose	Insulin	*HOMA*	Leptin	Adiponectin	Resistin	*MCP.1*	Class
0.0709	0.0607	0.3692	0.2539	0	0.0523	0.0049	0.2212	0.2247	0
0.2270	0.0108	0.9077	0.1148	0.0097	0.0527	0.0122	0.1037	0.2559	0
0.2199	0.0769	0.8923	0.2353	0.0221	0.1585	0.0369	0.5710	0.3079	0
0.1206	0.1211	0.6769	0.1483	0.0059	0.0648	0.0142	0.1515	0.5339	0
0.2270	0.0934	0.9538	0.1356	0.0137	0.0278	0.0199	0.0869	0.4406	0

Age= edad, BMI= índice de masa corporal, Glucose= glucosa, Insulin= insulina, *HOMA*= índice de resistencia a la insulina, Leptin= leptina, Adiponectin= adiponectina, Resistin= resistina, *MCP.1*= proteína quimiotáctica de monocitos 1, Class= clase / clasificación.

3.3. Modelos de clasificación generados

Se construyeron 27 diferentes variantes de modelos de clasificación, la Tabla 3.5 muestra todas las variantes de los modelos generados, 15 de ellos para el grupo de variables V1 - V4 y otros 12 más para el grupo de variables V1 - V9. Todos los modelos construidos tuvieron una validación cruzada 30-*fold* y variación en la presencia de validación *hold out*. Para cada algoritmo se obtuvieron las métricas para su respectiva evaluación, generando un archivo tipo .csv, el cual contiene la información y características de cada configuración. El nombre del archivo resultado se genera con el siguiente patrón, *variables_número de capas_función de entrenamiento_HOV o noHOV*.

Tabla 3.5: Modelos de clasificación de RNA construidos

Algoritmos de entrenamiento							
CO	*hold out*			Variables	NO *hold out*		
	SCG	RP	CGB		SCG	RP	CGB
1	✓	✓	✓	V1 - V4	✓	✓	✓
	✓	✓	✓	V1 - V9	✓	✓	✓
2	✓	✓	✓	V1 - V4	✓	✓	✓
	✓	✓	✓	V1 - V9	✓	✓	✓
3	✓	✓	✓	V1 - V4	✓	✓	✓
	✓	✓	✓	V1 - V9	✓	✓	✓
4	✓	✓	✓	V1 - V4	-	-	-
	✓	✓	✓	V1 - V9	-	-	-
5	✓	✓	✓	V1 - V4	✓	✓	✓
	✓	✓	✓	V1 - V9	✓	✓	✓

Modelos de clasificación construidos: se construyeron distintos modelos de clasificación con tres algoritmos de entrenamiento, SCG=*Scaled Conjugate Gradient*, CGB=*Conjugate Gradient Backpropagation with Powell-Beale Restarts* y RP=*Resilient Backpropagation*. Cada algoritmo tuvo variación en la presencia o ausencia de validación *hold out*. Los espacios en blanco (-) corresponden a modelos no realizados. CO = capas ocultas.

3.4. Evaluación de modelos de clasificación generados

Después de haber generado todos los modelos de clasificación, se realizó la evaluación de cada grupo de experimentación con sus variantes de algoritmo de entrenamiento y tipo de validación usada. Se obtuvieron un total de 6,042 distintas configuraciones, cada configuración tuvo 30 iteraciones, correspondientes a la validación cruzada 30-*fold*.

3.4.1. Evaluación por número de capas ocultas

La primera evaluación se realiza al identificar si el número de capas ocultas afecta el promedio del AUC, así como su desviación estándar para cada función de entrenamiento.

En la Tabla 3.6 se muestra el comportamiento de la media del AUC para la función de entrenamiento SCG para los dos grupos de experimentación V1 - V4 y V1 - V9. Para el grupo de experimentación V1 - V4 con validación *hold out* se aprecia como el aumento de capas ocultas provoca un decremento en el valor promedio del AUC y aumento de la σ, al tomar en cuenta a la mejor configuración. La mejor configuración para este caso corresponde a una configuración con una capa oculta y 18 nodos en la misma, con un valor promedio del AUC = 0.8534 y una $\sigma = 0.0263$.

Para el grupo de experimentación V1 - V4 sin validación *hold out*, se observa un aumento en el valor del AUC al aumentar el número de capas ocultas, mientras que la σ aumenta y al aparecer se mantiene en un rango estable. La mejor configuración obtenida para este caso es con cinco capas ocultas con la siguiente cantidad de nodos, 24 - 24 - 24 - 12 - 24, en las capas ocultas C1, C2, C3, C4 y C5 respectivamente, con un valor del AUC = 0.9274 y una $\sigma = 0.0142$.

El grupo de experimentación V1 - V9 con validación *hold out*, muestra el mismo comportamiento de decremento ya descrito anteriormente conforme al aumento en el número de capas ocultas. El valor del AUC y de la σ para la mejor configuración en este caso son de

Tabla 3.6: Evolución del AUC para algoritmo SCG

Algoritmo SCG					
Variables	CO	*hold out*		NO *hold out*	
		AUC_{media}	AUC_{σ}	AUC_{media}	AUC_{σ}
V1 - V4	1	0.8534	0.0263	0.8741	0.0085
	2	0.8333	0.0546	0.9140	0.0143
	3	0.8310	0.0599	0.9244	0.0156
	4	0.8401	0.0479	-	-
	5	0.8426	0.0420	0.9274	0.0142
V1 - V9	1	0.8242	0.0518	0.9449	0.0143
	2	0.8020	0.0774	0.9333	0.0546
	3	0.8120	0.0802	0.9555	0.0126
	4	0.8024	0.0814	-	-
	5	0.7969	0.1016	0.9514	0.0147

Evolución de la media del AUC y σ en varias capas ocultas, con y sin validación *hold out*, función de entrenamiento SCG. Cada valor mostrado representa la media del parámetro a evaluar de la mejor configuración para dicha función de entrenamiento. CO = capas ocultas, AUC= área bajo la curva, σ= desviación estándar. Los espacios en blanco (-) corresponden a modelos no realizados.

0.8242 y 0.0518 respectivamente. La mejor configuración cuenta con 8 nodos en su única capa oculta.

Mientras que para el grupo de experimentación V1 - V9 sin validación *hold out*, los mejores valores de evaluación se encuentran en los modelos de 3 capas ocultas, con 20 nodos en la capa C1, 22 nodos en la capa C2 y la capa C3 con 16 nodos. El valor para AUC = 0.9555 y para la $\sigma = 0.0126$.

En la Tabla 3.7 se muestra el comportamiento de la media del AUC para la función de entrenamiento SCG para los dos grupos de experimentación V1 - V4 y V1 - V9. Para ambos grupos de experimentación con validación *hold out* se observa el mismo patrón de decremento en el valor del AUC y aumento de la σ al incrementarse el número de capas ocultas en las redes. Para el grupo V1 - V4, el mejor desempeño corresponde a una configuración de una sola capa oculta con 20 nodos, una AUC = 0.8625 y una $\sigma = 0.0198$. Mientras que para el

grupo V1 - V9 el número de nodos en la única capa oculta es de 8 y sus respectivos valores para el AUC y la σ son de 0.8255 y 0.0448.

Tabla 3.7: Evolución del AUC para algoritmo CGB

Algoritmo CGB					
Variables	CO	*hold out*		NO *hold out*	
		AUC_{media}	AUC_{σ}	AUC_{media}	AUC_{σ}
V1 - V4	1	0.8625	0.0198	0.8746	0.0111
	2	0.8317	0.0504	0.9270	0.0168
	3	0.8426	0.0467	0.9432	0.0145
	4	0.8472	0.0447	-	-
	5	0.8349	0.0505	0.9606	0.0106
V1 - V9	1	0.8255	0.0448	0.9510	0.0123
	2	0.8150	0.0778	0.9628	0.0140
	3	0.8047	0.0664	0.9623	0.0119
	4	0.8090	0.0714	-	-
	5	0.8010	0.0719	0.9602	0.0145

Evolución de la media del AUC y σ en varias capas ocultas, con y sin validación *hold out*, función de entrenamiento CGB. Cada valor mostrado representa la media del parámetro a evaluar de la mejor configuración para dicha función de entrenamiento. CO= capas ocultas, AUC= área bajo la curva, σ= desviación estándar. Los espacios en blanco (-) corresponden a modelos no realizados.

En cambio para los grupos de experimentación sin validación *hold out* el comportamiento de los parámetros de evaluación es distinto. Para el grupo V1 - V4, a mayor número de capas, mejor resultado obtenido, para este caso el mejor desempeño se obtuvo con una configuración de 5 capas ocultas con el siguiente número de nodos en cada capa, 24 - 20 - 20 - 16 - 24. Un AUC = 0.9606 y una $\sigma = 0.0106$. Y para el grupo V1 - V9 el mejor desempeño se encuentra en una configuración de dos capas ocultas, 20 nodos para ambas capas C1 y C2 respectivamente. Con el mejor valor de AUC = 0.9628 y su respectiva $\sigma = 0.0140$.

En la Tabla 3.8 se muestra el comportamiento de la media del AUC para la función de entrenamiento RP para los dos grupos de experimentación V1 - V4 y V1 - V9. En los grupos de experimentación con validación *hold out*. Para el grupo V1 - V4 el mejor desempeño se

Tabla 3.8: Evolución del AUC para algoritmo RP

Algoritmo RP					
Variables	CO	*hold out*		NO *hold out*	
		AUC_{media}	AUC_{σ}	AUC_{media}	AUC_{σ}
V1 - V4	1	0.8472	0.0191	0.8616	0.0133
	2	0.8653	0.0307	0.9159	0.0107
	3	0.8482	0.0539	0.9404	0.0136
	4	0.8390	0.0562	-	-
	5	0.8270	0.0623	0.9486	0.0126
V1 - V9	1	0.8424	0.0533	0.9252	0.0140
	2	0.8289	0.0906	0.9558	0.0115
	3	0.8353	0.0703	0.9608	0.0133
	4	0.8031	0.0778	-	-
	5	0.7682	0.1076	0.9611	0.0107

Evolución de la media del AUC y σ en varias capas ocultas, con y sin validación *hold out*, función de entrenamiento RP. Cada valor mostrado representa la media del parámetro a evaluar de la mejor configuración para dicha función de entrenamiento. CO= capas ocultas, AUC= área bajo la curva, σ= desviación estándar. Los espacios en blanco (-) corresponden a modelos no realizados.

encuentra en una configuración de doble capa oculta, 24 nodos en C1 y 18 nodos en C2, y sus valores para el AUC es de 0.8653 y la σ de 0.0307. Mientras que para el grupo V1 - V9, una sola capa con 20 nodos obtiene el mejor valor para el AUC = 0.8424 y para la σ = 0.0533.

Para ambos grupos de experimentación sin validación *hold out*, el aumento en el número de capas ocultas mejora el desempeño de los clasificadores. En el grupo V1 - V4, cinco capas ocultas hacen el mejor trabajo, 24 nodos en C1, 24 en C2, 16 en C3, 16 en C4 y 24 en C5, sus valores para AUC y σ son de 0.9486 y 0.0126 respectivamente. Para el grupo V1 - V9, también cinco capas ocultas obtienen el mejor desempeño con un AUC = 0.9611 y una σ = 0.0107.

3.4.2. Evaluación por mejor desempeño

Una vez identificadas las mejores configuraciones para cada función de entrenamiento (FCN) y grupo de experimentación, se compararon entre ellos para evaluar al mejor algoritmo. Se separaron en cuatro grupos, V1 - V4 con validación *hold out*, V1 - V4 sin validación *hold out*, V1 - V9 con validación *hold out* y V1 - V9 sin validación *hold out*. Se compararon al evaluar la media del AUC, la σ asociada a dicha media, el valor mínimo y máximo de AUC alcanzados por el algoritmo y la media del índice de Youden (*J).*

En el grupo V1 - V4 con validación *hold out*, se obtuvieron los siguientes resultados mostrados en la Tabla 3.9, valores de AUC similares, sin embargo, el valor para la σ tiene variaciones importantes de 0.01 unidades afectando considerablemente el rango de la AUC. En la Figura 3.1 se puede observar el comportamiento de la AUC en las iteraciones durante la validación 30-*fold* para los 3 algoritmos de entrenamiento.

Tabla 3.9: Mejores configuraciones para V1 - V4 con validación *hold out*

CO	FCN	AUC_{media}	AUC_{σ}	AUC_{min}	AUC_{max}	J_{media}
1	SCG	0.8534	0.0263	0.7908	0.8894	0.7069
1	CGB	0.8625	0.0198	0.8179	0.9050	0.7250
2	RP	0.8653	0.0307	0.7674	0.9068	0.7307

Comparación de AUC y σ de las mejores configuraciones obtenidas validación *hold out*. CO= capas ocultas, FCN= algoritmo de entrenamiento, AUC= área bajo la curva, σ= desviación estándar.

En la gráfica anteriormente mencionada donde se grafica la mejor configuración para cada algoritmo con validación *hold out* y se observa cómo el algoritmo RP tiene una mayor variación en el comportamiento de la AUC, mientras que el algoritmo CGB se mantiene en un rango más corto. Estas variaciones afectan notablemente la media del AUC, el valor de la σ y la maximización de *J*.

La Tabla 3.10 resume la información estadística del grupo V1 - V4 sin validación *hold out*, en donde se observa una variedad en los valores del AUC, al igual que para el índice de

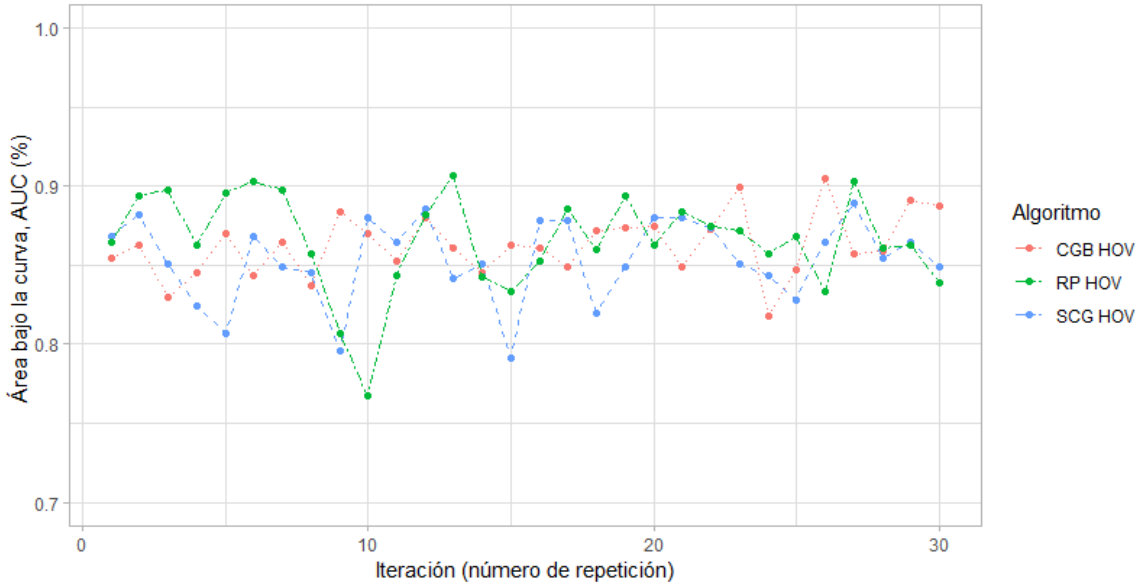

Figura 3.1: Mejores configuraciones para V1 - V4 con validación *hold out*. HOV = validación *hold out*, CGB = *Conjugate Gradient Backpropagation with Powell-Beale restarts*, RP = *Resilient Backpropagation*, SCG = *Scaled Conjugate Gradient*.

Youden, sin embargo, los rangos de AUC son similares y la σ varía por centésimas de unidad.

En la Figura 3.2 se muestra el comportamiento de la AUC durante la validación cruzada 30-*fold* en cada FCN, para todas las funciones se puede observar un patrón escalado de comportamiento en el aumento y disminución del valor del AUC, donde los mínimos y máximos se encuentran casi en las mismas iteraciones, sin embargo, a lo largo de las 30 iteraciones, en contadas ocasiones un algoritmo supera a otro.

Tabla 3.10: Mejores configuraciones para V1 - V4 sin validación *hold out*

CO	FCN	AUC_{media}	AUC_{σ}	AUC_{min}	AUC_{max}	J_{media}
5	SCG	0.9274	0.0142	0.9032	0.9729	0.8548
5	CGB	0.9606	0.0106	0.9399	0.9825	0.9212
5	RP	0.9486	0.0126	0.9266	0.9729	0.8973

Comparación de AUC y σ de las mejores configuraciones obtenidas sin validación *hold out*. CO= capas ocultas, FCN= algoritmo de entrenamiento, AUC= área bajo la curva, σ= desviación estándar.

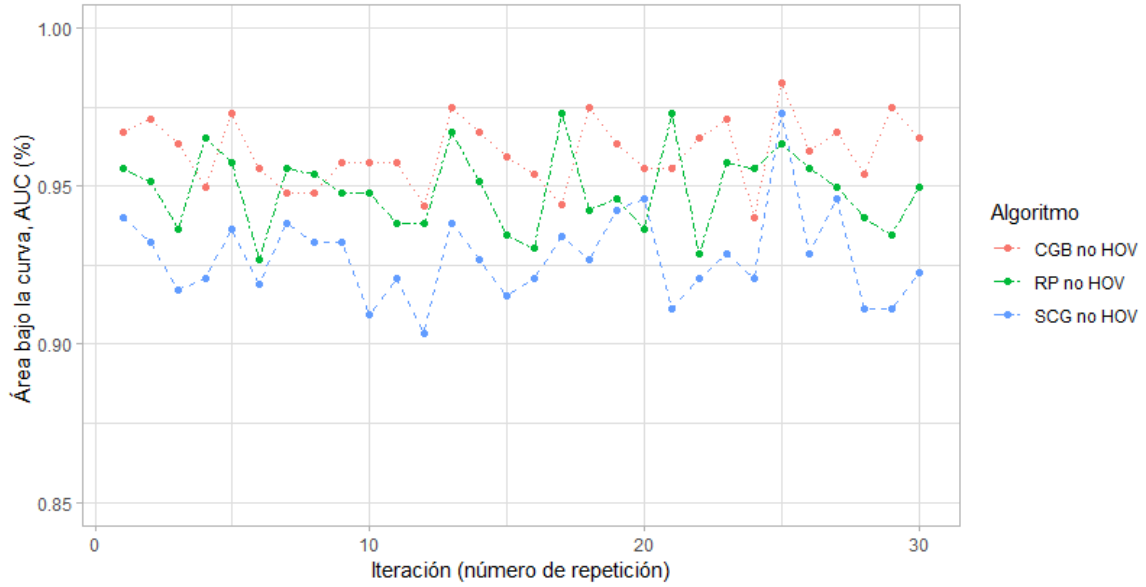

Figura 3.2: Mejores configuraciones para V1 - V4 sin validación *hold out*. no HOV = sin validación *hold out*, CGB = *Conjugate Gradient Backpropagation with Powell-Beale restarts*, RP = *Resilient Backpropagation*, SCG = *Scaled Conjugate Gradient*.

Finalmente, se comparó el mejor algoritmo obtenido con validación *hold out* contra el mejor algoritmo obtenido sin validación *hold out*. En la Figura 3.3 se aprecia dicha comparación durante las iteraciones realizadas por la validación *k-fold*.

Esta última comparación gráfica muestra una superioridad por casi una décima de unidad del algoritmo sin validación *hold out*, a pesar de tratarse de la misma función de entrenamiento en ambos casos.

Para el grupo V1 - V9 con validación *hold out*, el resumen de los parámetros a evaluar se muestran en la Tabla 3.11, en donde se aprecia un desempeño similar para las FCN SCG y CGB, pero ambas menores a RP, aunque con una ligera σ mayor. En la Figura 3.4 se muestra la variación de las AUC obtenidas por las 30 iteraciones de cada algoritmo durante su validación.

La comparación gráfica muestra un desempeño muy similar para las tres mejores configuraciones con validación *hold out*, mostrando nuevamente un patrón similar en los

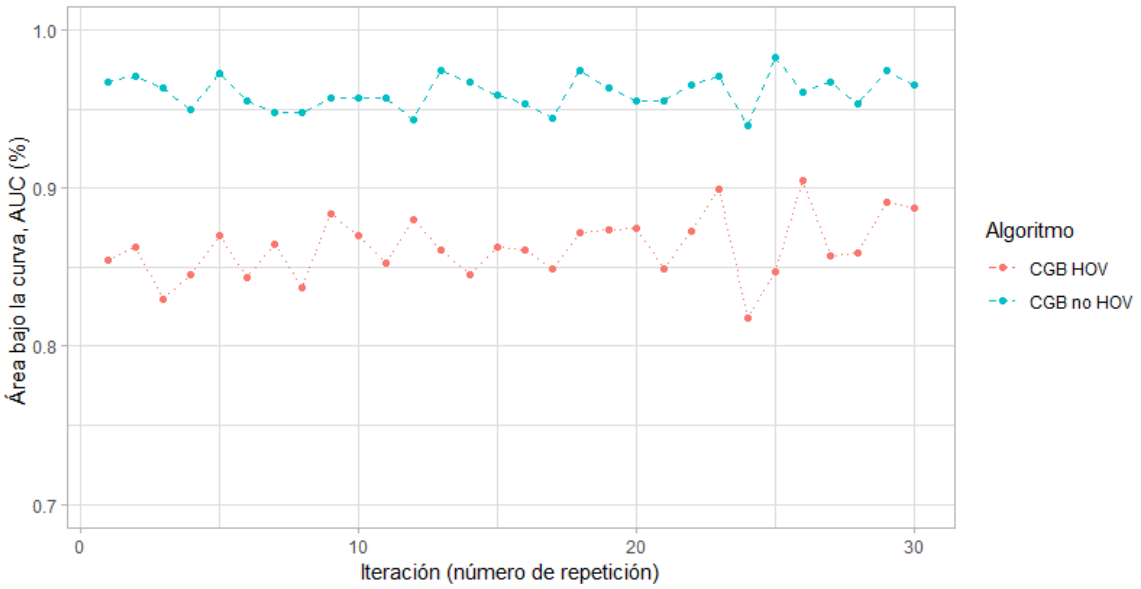

Figura 3.3: Mejor configuración para V1 - V4 con y sin validación *hold out*. HOV = validación *hold out*, no HOV = validación *hold out*, CGB = *Conjugate gradient backpropagation with Powell-Beale restarts*.

Tabla 3.11: Mejores configuraciones para V1 - V9 con validación *hold out*

CO	FCN	AUC_{media}	AUC_{σ}	AUC_{min}	AUC_{max}	J_{media}
1	SCG	0.8242	0.0518	0.6893	0.8894	0.6485
1	CGB	0.8255	0.0448	0.7235	0.9032	0.6511
1	RP	0.8424	0.0533	0.6893	0.9086	0.6849

Comparación de AUC y σ de las mejores configuraciones obtenidas con validación *hold out*. CO= capas ocultas, FCN= algoritmo de entrenamiento, AUC= área bajo la curva, σ= desviación estándar.

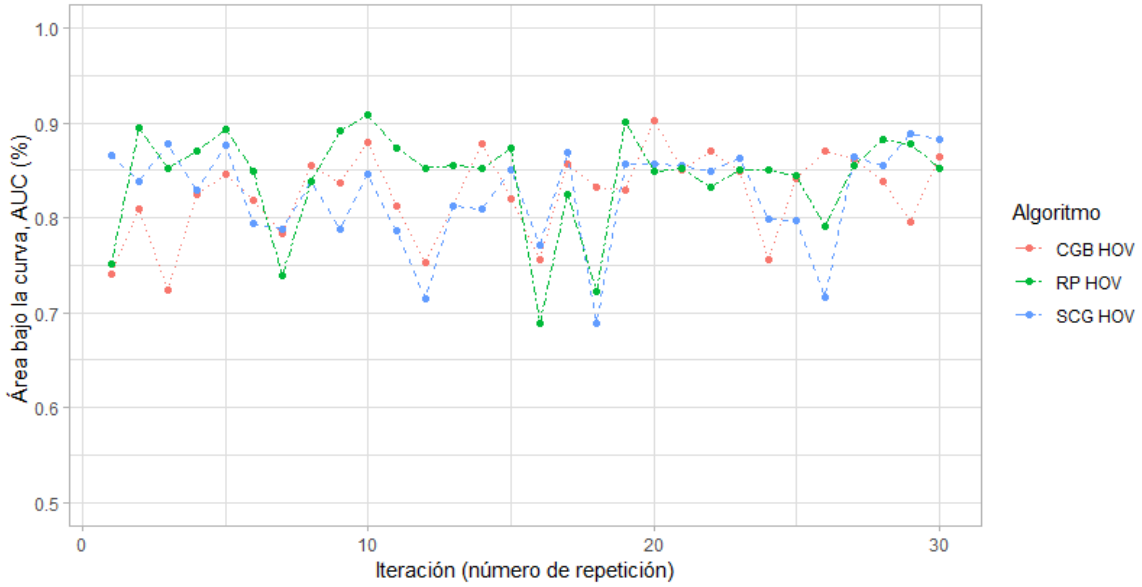

Figura 3.4: Mejores configuraciones para V1 - V9 con validación *hold out*. HOV = validación *hold out*, CGB = *Conjugate Gradient Backpropagation with Powell-Beale restarts*, RP = *Resilient Backpropagation*, SCG = *Scaled Conjugate Gradient*.

máximos y mínimos, y donde para todos los casos la variación del valor para la AUC es evidente en reiteradas ocasiones.

Y por último el grupo V1 - V9 sin validación *hold out* se obtuvieron los resultados mostrados en la Tabla 3.12, y donde se aprecia que el valor del AUC es similar para los tres algoritmos de entrenamiento, con ligeras variaciones en sus desviaciones estándar y obteniendo altos valores para el índice de Youden. La Figura 3.5 muestra el comportamiento del AUC durante la validación *k-fold* para este caso.

Una comparación adicional gráfica entre el mejor algoritmo con validación *hold out* contra el mejor algoritmo sin validación *hold out*, en la Figura 3.3 se aprecia dicha comparación durante las iteraciones realizadas por la validación 30-*fold*.

Claramente se observa como nuevamente la validación sin *hold out* tiene un desempeño notablemente superior en cuanto al valor del AUC, el rango del mismo y la variación de los valores a lo largo de las iteraciones.

Tabla 3.12: Mejores configuraciones para V1 - V9 sin validación *hold out*

CO	FCN	AUC_{media}	AUC_{σ}	AUC_{min}	AUC_{max}	J_{media}
3	SCG	0.9555	0.0126	0.9381	0.9825	0.9110
2	CGB	0.9628	0.0140	0.9302	0.9903	0.9107
5	RP	0.9611	0.0107	0.9417	0.9807	0.9223

Comparación de AUC y σ de las mejores configuraciones obtenidas sin validación *hold out*. CO= capas ocultas, FCN= algoritmo de entrenamiento, AUC= área bajo la curva, σ= desviación estándar.

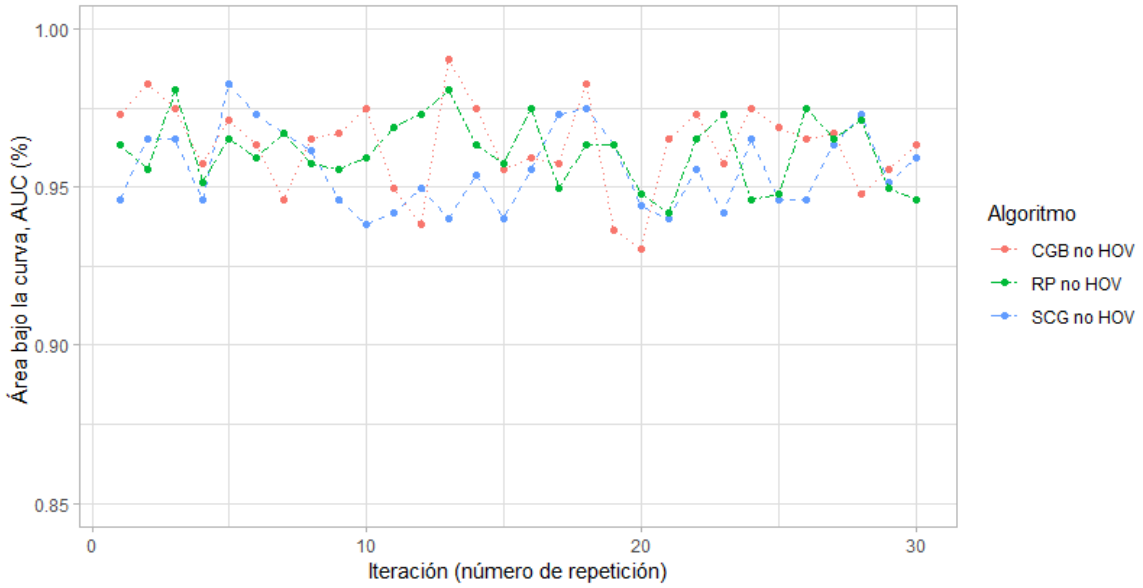

Figura 3.5: Mejores configuraciones para V1 - V9 sin validación *hold out*. no HOV = sin validación *hold out*, CGB = *Conjugate Gradient Backpropagation with Powell-Beale restarts*, RP = *Resilient Backpropagation*, SCG = *Scaled Conjugate Gradient*.

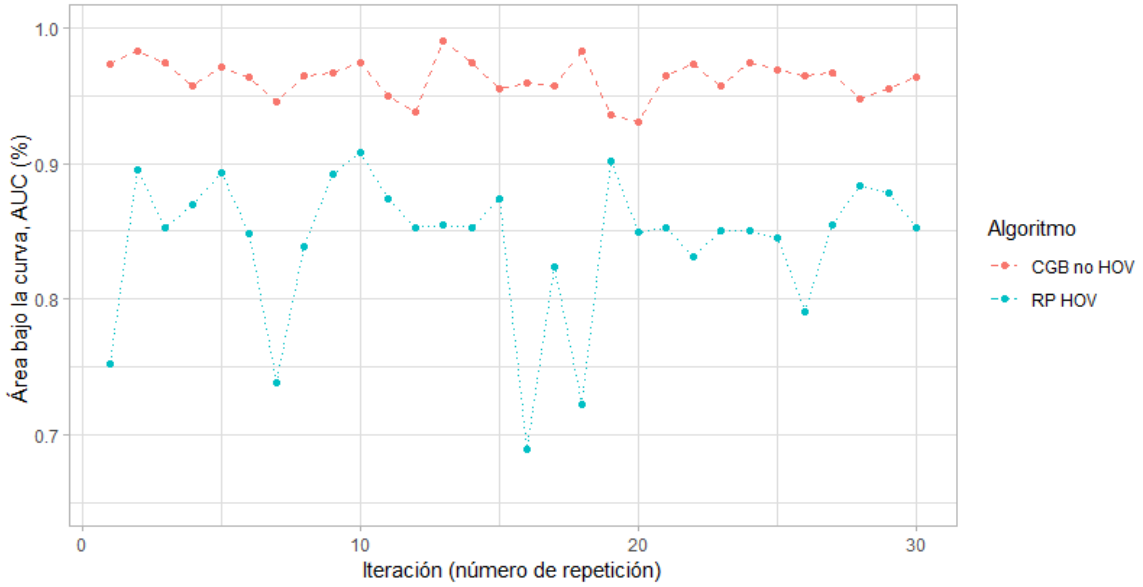

Figura 3.6: Mejores configuración para V1 - V9 con y sin validación *hold out*. HOV = validación *hold out*, no HOV = sin validación *hold out*, CGB = *Conjugate gradient backpropagation with Powell-Beale restarts*, RP = *Resilient backpropagation*.

3.4.3. Evaluación por *Machine Learning*

Como última evaluación y dando seguimiento a los objetivos planteados, se comparan los resultados obtenidos en este trabajo contra los resultados reportados por Patrício et al. (2018).

A continuación se muestra una comparación de algoritmos de *Machine Learning* usados para predecir el diagnóstico de cáncer de seno, en la Tabla 3.13 usando V1 - V4 corresponde a los mejores resultados obtenidos por Patrício et al. (2018), mientras que la Tabla 3.14 hace uso de los 9 predictores contenidos originalmente en el conjunto de datos. Para dicha evaluación se usaron los siguientes parámetros, AUC, sensibilidad, especificidad y el índice de Youden. Cabe mencionar que Patrício et al. (2018) presenta sus resultados en rangos, por lo que se obtuvo la media para el rango reportado y se comparó con lo obtenido en este trabajo.

Los algoritmos usando RNA y validación *hold out* compiten con los reportados en SVM,

Tabla 3.13: Comparación para V1 - V4 con redes neuronales artificiales

	Patrício et al. (2018)			Benítez, HOV			Benítez, NO HOV		
	LR	RF	SVM	SCG	RP	CGB	SCG	RP	CGB
AUC	0.81	0.86	0.89	0.84	0.84	0.86	0.94	0.95	0.96
Sensibilidad	0.75	0.83	0.85	0.84	0.85	0.86	0.92	0.92	0.96
Especificidad	0.84	0.84	0.87	0.84	0.83	0.87	0.95	0.98	0.96
J	0.59	0.67	0.72	0.68	0.67	0.72	0.88	0.90	0.92

HOV= *hold out Validation*, LR= *Logistic Regression*, RF= *Random Forest*, SVM= *Support Vector Machine*, SCG= *Scaled Conjugate Gradient*, RP= *Resilient Backpropagation*, CGB= *Conjugate Gradient Backpropagation with Powell-Beale Restarts*, AUC= área bajo la curva, J= índice de Youden.

LR y RF, pero tienen menor desempeño en el área bajo la curva a pesar de superar o igualar en sensibilidad, especificidad e índice de Youden a los algoritmos ya mencionados. Mientras que los algoritmos de RNA sin *hold out* son notablemente mejores al resto de las técnicas de ML mencionadas anteriormente, superando por casi una decena de unidad a la mayoría del resto de algoritmos en el caso de la mejor configuración de RNA encontrada, de la misma forma los valores para la sensibilidad, especificidad e índice de Youden son mejores por gran diferencia.

Para los algoritmos de ML de Patrício et al. (2018) y las RNA con validación *hold out* los valores son iguales y compiten, pero no superan a los valores obtenidos por el SVM. Solamente en especificidad e índice de Youden, esta variación de RNA resultó ser superior.

Para las RNA sin *hold out* los valores obtenidos en todas las métricas de desempeño son superiores indudablemente al resto de las técnicas de ML en este trabajo, y compitiendo entre los tres mejores algoritmos, aunque con ligeras variaciones en algunos valores. A pesar de que los algoritmos obtienen valores muy similares, algunos corresponden exactamente, el algoritmo CGB muestra ser mejor al hacer uso de solo dos capas ocultas, mientras que el algoritmo SCG y RP hacen uso de tres y cinco capas ocultas, respectivamente.

Se aprecia cómo los algoritmos con RNA obtienen mejores valores en las métricas de

Tabla 3.14: Comparación para V1 - V9 con redes neuronales artificiales

	Patrício et al. (2018)			Benítez, HOV			Benítez, NO HOV		
	LR	RF	SVM	SCG	RP	CGB	SCG	RP	CGB
AUC	0.79	0.81	0.83	0.79	0.80	0.81	0.96	0.96	0.96
Sensibilidad	0.73	0.82	0.78	0.71	0.73	0.76	0.95	0.96	0.95
Especificidad	0.83	0.74	0.81	0.87	0.87	0.87	0.97	0.96	0.97
Índice de Youden	0.56	0.55	0.59	0.58	0.60	0.63	0.92	0.92	0.92

HOV= *hold out Validation*, LR= *Logistic Regression*, RF= *Random Forest*, SVM= *Support Vector Machine*, SCG= *Scaled Conjugate Gradient*, RP= *Resilient Backpropagation*, CGB= *Conjugate Gradient Backpropagation with Powell-Beale Restarts*, AUC= área bajo la curva, J= índice de Youden.

evaluación.

Capítulo 4

Discusiones

Durante la evaluación de los modelos que tuvieron valores para la AUC similares y en los que dicho valor variaba por centésimas o milésimas de unidad, se le dio prioridad a aquellos modelos con configuraciones usando menos capas ocultas, esto debido a que los recursos computacionales usados para el entrenamiento de las redes neuronales usando el menor número de capas ocultas posible es menor comparado con un mayor número de capas ocultas. El mismo razonamiento se aplicó para evaluar los demás parámetros.

El aumento de capas ocultas no fue benéfico para todos los casos de experimentación, principalmente en aquellos con validación *hold out*, en donde los mejores valores de desempeño se obtuvieron por modelos de una capa oculta y para el caso del grupo V1 - V4, únicamente el algoritmo RP mejoró al usar dos capas ocultas, sin embargo, la mejora fue por pocas centésimas de unidad y su desviación estándar aumentó considerablemente por lo que se omitió su selección como mejor algoritmo de entrenamiento en este grupo de experimentación. En cambio el algoritmo CGB obtuvo la σ más baja y valores para la AUC y J muy similar a RP, pero con menor número de capas ocultas. En el grupo V1 - V9 el algoritmo RP superó por casi dos decenas de unidad a los demás algoritmos y manteniendo un valor para la σ similar al resto de los algoritmos, pero su valor para J fue por varias decenas de unidad el

mejor.

En el caso de los grupos de experimentación sin validación *hold out*, el aumento en el número de capas ocultas resultó con mejores valores en los parámetros de evaluación, para el grupo V1 - V4 el algoritmo CGB sobresalió notablemente al ser mejor que el resto por varios pares de decena de unidad para el AUC y manteniendo una baja σ, lo anterior con una configuración de cinco capas ocultas. Para el grupo V1 - V9 el algoritmo CGB resultó el mejor con el AUC más alto, aunque con la σ más alta, con una configuración de dos capas ocultas, comparado con el algoritmo RP que obtuvo un AUC ligeramente por debajo de CGB, y una menor desviación estándar, pero con cinco capas ocultas, lo que requiere mucho más poder computacional.

De los modelos construidos, es claramente visible que el mejor desempeño lo obtienen aquellos que no llevan a cabo la validación *hold out*. El objetivo de la validación es evitar una selección sesgada a la hora de separar en subconjuntos y de esta forma asegurar la robustez del modelo. Una doble validación (*k-fold* y *hold out*) asegura la robustez del modelo frente a una selección sesgada más fina de los datos de los subconjuntos de prueba y validación. Sin embargo, puede que esta doble validación evite que los modelos tengan disponibles datos cruciales para generalizar el problema enfrentado. Es decir, a la hora de dividir en tres subconjuntos (entrenamiento, validación y prueba) puede que datos conteniendo características clave para el aprendizaje sean seleccionadas para la muestra de validación o prueba, y como consecuencia el grupo de entrenamiento no sea lo suficiente informativo para entrenar al modelo.

Lo mencionado anteriormente toma aún mayor relevancia ya que Patrício et al. (2018) reporta haber usado una validación cruzada tipo Monte Carlo y el uso de una técnica en la que generan datos artificiales basados en el conjunto de datos original, la cual genera aleatoriamente nuevas particiones de entrenamiento y de prueba en cada nueva configuración. De acuerdo a la metodología implementada en este trabajo, la validación 30-*fold* con *random sub-sampling* se acerca bastante a la definición de la validación cruzada tipo Monte Carlo.

Se pueden inferir dos cosas derivado de lo ya mencionado, la primera es que una doble validación para este caso no es viable puesto que limita de información crucial a la hora de entrenar a los modelos, y segunda, una variación en los porcentajes usados para conformar las particiones de entrenamiento, prueba, y en su debido caso de validación, tienen un alto efecto en los resultados del modelo de clasificación.

Ahora, al comparar los resultados de acuerdo al algoritmo de *Machine Learning* usado, las RNA usadas en este trabajo, sin doble validación, son notablemente superiores a lo reportado por Patrício et al. (2018). En todos los sentidos, el AUC mejora considerablemente, al igual que los valores de sensibilidad y especificidad. El índice de Youden también mejoró. Lo anterior aplica para ambos grupos de experimentación, usando cuatro y nueve predictores.

Capítulo 5

Conclusiones y trabajo futuro

Se identificaron dos configuraciones de RNA notablemente superiores a los algoritmos presentados por Patrício et al. (2018) al usar cuatro y nueve predictores del conjunto de datos *Breast Cancer Coimbra Data Set*. Cada configuración varía en el número de capas ocultas y el algoritmo de entrenamiento usado.

Los algoritmos CGB sin validación *hold out* sobrepasan al resto de los algoritmos RP y SCG, con y sin validación *hold out*, para ambos grupos de experimentación. A pesar de que estos últimos tuvieron desempeños similares o ligeramente superiores como en el caso en concreto del grupo V1 - V9 sin validación *hold out*, pero el poder computacional requerido fue mayor (fueron necesarias más capas ocultas).

El mejor modelo para cuatro predictores corresponde a una RNA con algoritmo de entrenamiento CGB de cinco capas ocultas, sin validación *hold out*, obteniendo un AUC = 0.96, sensibilidad = 0.96, especificidad = 0.96 e índice de Youden = 0.92.

El mejor modelo para nueve predictores corresponde a una RNA con algoritmo de entrenamiento CGB de dos capas ocultas, con validación *hold out*, valores de AUC = 0.96, sensibilidad = 0.95, especificidad = 0.97 e índice de Youden = 0.92.

Como trabajo a futuro se propone realizar un análisis distinto para obtener la importancia relativa de cada variable, en donde cada una recibe un peso de importancia o aportación, para realizar la predicción en cuestión. Anteriormente, Patrício et al. (2018) asignó la importancia de cada predictor mediante el coeficiente de Gini, que se basa en medir la desigualdad en cuanto a la aportación a un resultado por parte de un conjunto de variables, es decir, mide el grado de desigualdad de la distribución de la aportación a un total por parte de varias variables independientes y una variable dependiente. Los métodos sugeridos para obtener este peso de importancia son descritos por Maozhun and Ji (2017), de Oña and Garrido (2014), Olden et al. (2004) y Olden and Jackson (2002). En dichos trabajos se usan algoritmos mejorados o nuevas propuestas de algoritmo que resultan con altos valores de desempeño específicamente para redes neuronales artificiales, por lo que se esperaría obtener resultados diferentes.

Debido a que se trata de un problema de clasificación binaria y la naturaleza de los algoritmos usados son de reconocimiento de patrones para clasificación, los resultados de la clasificación se redondean al entero más cercano de acuerdo a un valor de corte, dicho valor puede ser modificado a juicio, y obtener distintos valores en las métricas de evaluación con el fin de mejorar el umbral de decisión entre una paciente sano y una paciente enfermo. Encontrar un valor de corte óptimo mejora la capacidad de clasificación del algoritmo.

Modificar a juicio los porcentajes para las particiones de entrenamiento, prueba y validación cuando sea necesario, puede modificar considerablemente los resultados obtenidos en los parámetros a evaluar, puesto que a mayor disponibilidad de datos en el entrenamiento, más características clave pueden ser aprendidas por el modelo y generando mejores resultados.

Bibliografía

ACS. What Is Breast Cancer?, 2016. URL https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

S. Agatonovic-Kustrin and R. Beresford. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22(5):717–727, 2000. ISSN 07317085. doi: 10.1016/S0731-7085(99)00272-1.

Garnet L Anderson and Marian L Neuhouser. Obesity and the Risk for Premenopausal and Postmenopausal Breast Cancer. *Cancer Prevention Research*, (17):515–522, 2012. doi: 10.1158/1940-6207.CAPR-12-0091.

Ismail Babaoglu, Omer Kaan Baykan, Nazif Aygul, Kurtulus Ozdemir, and Mehmet Bayrak. Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization. *Expert Systems with Applications*, 36(2 PART 1):2562–2566, 2009. ISSN 09574174. doi: 10.1016/j.eswa.2007.11.013.

Hamid Behravan, Jaana M. Hartikainen, Maria Tengström, Katri Pylkäs, Robert Winqvist, Veli–Matti –M Kosma, and Arto Mannermaa. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Scientific Reports*, 8(1):1–13, 2018. ISSN 20452322. doi: 10.1038/s41598-018-31573-5.

Andrew P. Bradley. THE USE OF THE AREA UNDER THE ROC CURVE IN THE EVALUATION OF MACHINE LEARNING ALGORITHMS. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 14316730.

D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins. Next-Generation Machine Learning for Biological Networks. 2018.

Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics and Bioinformatics*, 16(1):17–32, 2018. ISSN 22103244. doi: 10.1016/j.gpb.2017.07.003. URL https://doi.org/10.1016/j.gpb.2017.07.003.

Rubén A. Castañeda-Martínez. Modelo Farmacodinámico in silico de Almeja Pismo con Digoxina Usando Técnicas de Inteligencia Artificial, 2017.

N Catherine Sánchez. Conociendo y comprendiendo la célula cancerosa: Fisiopatología del cáncer. 2013.

Sun Chien-An, Wu Mei-Hsuan, Chu Chi-Hong, Chou Yu-Ching, Hsu Giu-Cheng, Yang Tsan, Chou Wan-Yun, Yu Cheng-Ping, and Yu Jyh-Cherng. Adipocytokine resistin and breast cancer risk. *Breast Cancer Res Treat*, pages 869–876, 2010. doi: 10.1007/s10549-010-0792-4.

Dara Hope Cohen and Derek Leroith. Obesity , type 2 diabetes , and cancer : the insulin and IGF connection. *Endocrine-Related Cancer*, pages 27–45, 2012. doi: 10.1530/ERC-11-0374.

Joana Crisóstomo, Paulo Matafome, Daniela Santos-Silva, Ana L. Gomes, Manuel Gomes, Miguel Patrício, Liliana Letra, Ana B. Sarmento-Ribeiro, Lelita Santos, and Raquel Seiça. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. *Endocrine*, 53(2): 433–442, 2016. ISSN 15590100. doi: 10.1007/s12020-016-0893-x.

Simon S. Cross, Ian R. Palmer, and Timothy J. Stephenson. How to design and use a research database. *Diagnostic Histopathology*, 24(4):149–153, 2018. ISSN 18767621. doi: 10.1016/j.mpdhp.2017.09.005. URL https://doi.org/10.1016/j.mpdhp.2017.09.005.

Maria Dalamaga. Obesity, insulin resistance, adipocytokines and breast cancer: New biomarkers and attractive therapeutic targets. *World Journal of Experimental Medicine*, 3(3):34–43, 2013. doi: 10.5493/wjem.v3.i3.34.

Maria Dalamaga. Resistin as a biomarker linking obesity and inflammation to cancer : potential clinical perspectives. *Biomarkers Med*, 8:107–118, 2014.

Maria Dalamaga, George Sotiropoulos, Konstantinos Karmaniolas, Nicolaos Pelekanos, Evangelia Papadavid, and Antigoni Lekka. Serum resistin : A biomarker of breast cancer in postmenopausal women ? Association with clinicopathological characteristics , tumor markers , in fl ammatory and metabolic parameters. *Clinical Biochemistry*, 46(7-8):584–590, 2013. ISSN 0009-9120. doi: 10.1016/j.clinbiochem.2013.01.001. URL http://dx.doi.org/10.1016/j.clinbiochem.2013.01.001.

Juan de Oña and Concepción Garrido. Extracting the contribution of independent variables in neural network models: A new approach to handle instability. *Neural Computing and Applications*, 25(3-4):859–869, 2014. ISSN 09410643. doi: 10.1007/s00521-014-1573-5.

P. Domingos. A Few Useful Things to Know about Machine Learning . 2012.

D. Dua and E. Karra Taniskidou. UCI Machine Learning Repository, year = 2017, url = http://archive.ics.uci.edu/ml.

Włodzisław Duch and Norbert Jankowski. Survey of neural transfer functions. *Neural Computing Surveys*, 2:163–212, 1999. URL ftp://ftp.icsi.berkeley.edu/pub/ai/jagota/vol2{_}6.pdf.

Michael J. Duffy, Enda W. McDermott, and John Crown. Blood-based biomarkers in breast cancer: From proteins to circulating tumor cells to circulating tumor DNA. *Tumor Biology*, 40(5):1–11, 2018. ISSN 14230380. doi: 10.1177/1010428318776169.

Babak Ehteshami Bejnordi, Maeve Mullooly, Ruth M. Pfeiffer, Shaoqi Fan, Pamela M. Vacek, Donald L. Weaver, Sally Herschorn, Louise A. Brinton, Bram van Ginneken, Nico Karssemeijer, Andrew H. Beck, Gretchen L. Gierach, Jeroen A.W.M. van der Laak, and Mark E. Sherman. Using deep convolutional neural networks to identify and classify tumor-associated stroma in

diagnostic breast biopsies. *Modern Pathology*, 31(10):1502–1512, 2018. ISSN 15300285. doi: 10.1038/s41379-018-0073-z.

Enas M.F. El Houby. A survey on applying machine learning techniques for management of diseases. *Journal of Applied Biomedicine*, 16(3):165–174, 2018. ISSN 12140287. doi: 10.1016/j.jab.2018.01.002. URL http://dx.doi.org/10.1016/j.jab.2018.01.002.

Randall J. Erb. Introduction to Backpropagation Neural. *Pharmaceutical Research*, 10(2):165–170, 1993.

Jorge Estrada and Juan Luna. El índice de Youden y su aplicación a la determinación del punto de corte en un test cuantitativo, 2016. URL http://masteres.ugr.es/moea/pages/curso201516/tfm1516/estrada{_}alvarez{_}tfm/!

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, (27):861–874, 2006. ISSN 11769106. doi: 10.1016/j.patrec.2005.10.010.

D. Flores, C. Gómez, D. Cervantes, A. Abaroa, C. Castro, and R. Castañeda-Martínez. Prediciendo la Actividad Cardíaca de la Almeja Tivela stultorum con Digoxina Utilizando Redes Neuronales Artificiales. *Revista Mexicana de Ingeniería Biomédica*, 38(1):208–216, 2017a. doi: dx.doi.org/10.17488/RMIB.38.1.15Prediciendo.

Dora Luz Flores, Claudia Gómez, David Cervantes, Alberto Abaroa, Carlos Castro, and Rubén A. Castañeda-Martínez. Predicting the physiological response of Tivela stultorum hearts with digoxin from cardiac parameters using artificial neural networks. *BioSystems*, 151:1–7, 2017b. ISSN 18728324. doi: 10.1016/j.biosystems.2016.11.002.

Georgia P Georgiou, Xeni Provatopoulou, Eleni Kalogera, Gerasimos Siasos, Evangelos Menenakos, George C Zografos, and Antonia Gounaris. Serum resistin is inversely related to breast cancer risk in premenopausal women. *The Breast*, 29:163–169, 2016. ISSN 0960-9776. doi: 10.1016/j.breast.2016.07.025. URL http://dx.doi.org/10.1016/j.breast.2016.07.025.

Michael E. Grossmann, Amitabha Ray, Katai J. Nkhata, Dmitry A. Malakhov, Olga P. Rogozina, Soner Dogan, and Margot P. Cleary. Obesity and breast cancer: Status of leptin and adiponectin in pathological processes. *Cancer and Metastasis Reviews*, 29(4):641–653, 2010. ISSN 01677659. doi: 10.1007/s10555-010-9252-1.

Yang Guo, Xuequn Shang, and Zhanhuai Li. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. *Neurocomputing*, 324:20–30, 2019. ISSN 18728286. doi: 10.1016/j.neucom.2018.03.072. URL https://doi.org/10.1016/j.neucom.2018.03.072.

Martin Hagan, Howard Demuth, Mark Hudson, and Orlando De Jesus. *Neural Network Design*. 2014.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data Transformation by Normalization*. 2011. ISBN 978-0-12-381479-1. doi: 10.1016/B978-0-12-381479-1.00001-0. URL http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

Samir M. Hanash, Christina S. Baik, and Olli Kallioniemi. Emerging molecular biomarkers-blood-based strategies to detect and monitor cancer. *Nature Reviews Clinical Oncology*, 8(3):142–150, 2011. ISSN 17594774. doi: 10.1038/nrclinonc.2010.220. URL http://dx.doi.org/10.1038/nrclinonc.2010.220.

Josef Havel, Filippo Amato, Petr Vaňhara, Aleš Hampl, Alberto López, and Eladia María Peña-Méndez. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2): 47–58, 2013. ISSN 1214021X. doi: 10.2478/v10136-012-0031-x.

N Lynn Henry and Daniel F Hayes. Cancer biomarkers 5. *Molecular Oncology*, 6(2):140–146, 2012. ISSN 1574-7891. doi: 10.1016/j.molonc.2012.01.010. URL http://dx.doi.org/10.1016/j.molonc.2012.01.010.

W. R. Hogan and M. M. Wagner. Accuracy of data in computer-based patient records . 1997.

International Agency for Research on Cancer. World cancer statistics, 2019. URL http://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf.

Ruholla Jafari-Marandi, Samaneh Davarzani, Maryam Soltanpour Gharibdousti, and Brian K. Smith. An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. *Applied Soft Computing Journal*, 72:108–120, 2018. ISSN 15684946. doi: 10.1016/j.asoc.2018.07.060. URL https://doi.org/10.1016/j.asoc.2018.07.060.

Fernando Mateo Jiménez. *Redes neuronales y preprocesado de variables para modelos y sensores en bioingeniería*. PhD thesis, Universidad Politécnica de Valencia, 2012.

Adem Kalinli, Fatih Sarikoc, Hulya Akgun, and Figen Ozturk. Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples. *Computer Methods and Programs in Biomedicine*, 110(3):298–307, 2013. ISSN 01692607. doi: 10.1016/j.cmpb.2012.12.005. URL http://dx.doi.org/10.1016/j.cmpb.2012.12.005.

Anna Kazarian, Oleg Blyuss, Gergana Metodieva, Aleksandra Gentry-Maharaj, Andy Ryan, Elena M. Kiseleva, Olga M. Prytomanova, Ian J. Jacobs, Martin Widschwendter, Usha Menon, and John F. Timms. Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples. *British Journal of Cancer*, 116(4):501–508, 2017. ISSN 15321827. doi: 10.1038/bjc.2016.433. URL http://dx.doi.org/10.1038/bjc.2016.433.

Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015. ISSN 20010370. doi: 10.1016/j.csbj.2014.11.005. URL http://dx.doi.org/10.1016/j.csbj.2014.11.005.

Victor V. Levenson. Biomarkers for early detection of breast cancer: What, when, and where? *Biochimica et Biophysica Acta - General Subjects*, 1770(6):847–856, 2007. ISSN 03044165. doi: 10.1016/j.bbagen.2007.01.017.

Lijuan Liu and Mingrong Deng. An evolutionary artificial neural network approach for breast cancer diagnosis. *3rd International Conference on Knowledge Discovery and Data Mining, WKDD 2010*, 25:593–596, 2010. ISSN 09333657. doi: 10.1109/WKDD.2010.148.

S. Y. Loke and A. S. G. Lee. The future of blood-based biomarkers for the early detection of breast cancer. 2018.

A M Lorincz and S Sukumar. Molecular links between obesity and breast cancer. *Endocrine-Related Cancer*, pages 279–292, 2006. doi: 10.1677/erc.1.00729.

Ana P. Lourenco, Kasey L. Benson, Meredith C. Henderson, Michael Silver, Elias Letsios, Quynh Tran, Kelly J. Gordon, Sherri Borman, Christa Corn, Rao Mulpuri, Wendy Smith, Josie Alpers, Carrie Costantini, Nitin Rohatgi, Rebecca Yang, Ali Haythem, Shah Biren, Michael Morris, Fred Kass, and David E. Reese. A Noninvasive Blood-based Combinatorial Proteomic Biomarker Assay to Detect Breast Cancer in Women Under the Age of 50 Years. *Clinical Breast Cancer*, 17(7): 516–525.e6, 2017. ISSN 19380666. doi: 10.1016/j.clbc.2017.05.004. URL https://doi.org/10.1016/j.clbc.2017.05.004.

Maria B. Lyng, Annette R. Kodahl, Harald Binder, and Henrik J. Ditzel. Prospective validation of a blood-based 9-miRNA profile for early detection of breast cancer in a cohort of women examined by clinical mammography. *Molecular Oncology*, 10(10):1234, 2016. ISSN 18780261. doi: 10.1016/j.molonc.2016.10.004. URL http://dx.doi.org/10.1016/j.molonc.2016.10.004.

Sun Maozhun and Liu Ji. Improved Garson algorithm based on neural network model. *Proceedings of the 29th Chinese Control and Decision Conference, CCDC 2017*, pages 4307–4312, 2017. doi: 10.1109/CCDC.2017.7979255.

G. Meyfroidt, F. Güiza, J. Ramon, and M. Bruynooghe. Machine Learning Techniques to Examine Large Patient Databases. 2009.

T. A. Moo, R. Sanford, C. Dang, and M. Morrow. Overview of Breast Cancer Therapy. 2018.

NCI. NCI Dictionary of Cancer Terms, 2018a. URL http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

NCI. What Is Cancer ? Differences between Cancer Cells and Normal Cells What Is Cancer ? - National Cancer Institute, 2018b. URL https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

Samad Nejatian, Hamid Parvin, and Eshagh Faraji. Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. *Neurocomputing*, 276:55–66, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2017.06.082. URL https://doi.org/10.1016/j.neucom.2017.06.082.

A. Nicolini, P. Ferrari, and M. J. Duffy. Prognostic and predictive biomarkers in breast cancer: Past, present and future . 2018.

TL Noguera Moreno. Metodología ROC en la Evaluación de Medidas Antropométricas como Marcadores de la Hipertensión Arterial. Aplicación a Población Gallega Adulta, 2010. URL http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto{_}418.pdf.

Jasmina Dj. Novakovic, Alempije Veljovic, Sinisa S. Ilic, Zeljko Papic, and Milica Tomovic. Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, pages 39–46, 2017. ISSN 17450926. doi: 10.1080/17450918.2017.1406394.

Cristina Núñez. Clinica Chimica Acta Blood-based protein biomarkers in breast cancer. *Clinica Chimica Acta*, 490(December 2018):113–127, 2019. ISSN 0009-8981. doi: 10.1016/j.cca.2018.12.028. URL https://doi.org/10.1016/j.cca.2018.12.028.

Julian D Olden and Donald A Jackson. Illuminating the "black box": Understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154:135–150, 2002. ISSN 03043800. doi: 10.1016/S0304-3800(02)00064-9. URL https://doi.org/10.1016/S0304-3800(02)00064-9.

Julian D. Olden, Michael K. Joy, and Russell G. Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3-4):389–397, 2004. ISSN 03043800. doi: 10.1016/j.ecolmodel.2004.03.013.

OMS. Biomarkers In Risk Assessment: Validity And Validation, 2001. URL http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

OMS. Obesidad y sobrepeso, 2018a. URL https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight.

OMS. Diabetes, 2018b. URL https://www.who.int/es/news-room/fact-sheets/detail/diabetes.

OMS. Cancer: Breast cancer, 2019. URL https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.

Miguel Patrício, José Pereira, Joana Crisóstomo, Paulo Matafome, Manuel Gomes, Raquel Seiça, and Francisco Caramelo. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, 18(1):1–8, 2018. ISSN 14712407. doi: 10.1186/s12885-017-3877-1.

E. Provenzano, G. A. Ulaner, and S. F Chin. Molecular Classification of Breast Cancer. 2018.

Sandhya Pruthi, Li Yang, Nicole P. Sandhu, James N. Ingle, Cheryl L. Beseler, Vera J. Suman, Ercole L. Cavalieri, and Eleanor G. Rogan. Evaluation of serum estrogen-DNA adducts as potential biomarkers for breast cancer risk. *Journal of Steroid Biochemistry and Molecular Biology*, 132(1-2):73–79, 2012. ISSN 09600760. doi: 10.1016/j.jsbmb.2012.02.002. URL http://dx.doi.org/10.1016/j.jsbmb.2012.02.002.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

A. N. Richter and T. M. Khoshgoftaar. A Review of Statistical and Machine Learning Methods for Modeling Cancer Risk Using Structured Clinical Data. 2018.

David Riesel. *Selecting an activation function*. 2007. ISBN 9780849371943. doi: 10.1016/S0140-6736(95)92880-4.

David P Rose and Linda Vona-davis. The cellular and molecular mechanisms by which insulin influences breast cancer risk and progression. *Endocrine-Related Cancer*, pages 225–241, 2012. doi: 10.1530/ERC-12-0203.

G. D. Rubenfeld, D. C. Angus, M.R. Pinsky, J.R. Curtis, A.F. Jr. Connors, G.R. Bernard, and TheMembersoftheOutcomes ResearchWorkshop. Outcomes Research in Critical Care. Results of the American Thoracic Society Critical Care Assembly Workshop on Outcomes Research. 1999.

Jonnathan G. Santillán-Benítez, Hugo Mendieta-Zerón, Leobardo M. Gómez-Oliván, Juan J. Torres-Juárez, Juan M. González-Bañales, Lorena V. Hernández-Peña, and Angel Ordóñez-Quiroz. The Tetrad BMI, Leptin, Leptin/Adiponectin (L/A) Ratio and CA 15-3 are Reliable Biomarkers of Breast Cancer. *Journal of Clinical Laboratory Analysis*, 27(1):12–20, 2013. ISSN 08878013. doi: 10.1002/jcla.21555.

Mary L. Schaeffer, Taner Z. Sen, and Carolyn J. Lawrence. Databases. *Genetics, Genomics and Breeding of Maize*, pages 215–235, 2014. ISSN 18787584. doi: 10.1201/b17274.

S. J. Schnitt. Classification and prognosis of invasive breast cancer: From morphology to molecular taxonomy. 2010.

M. Shanker and M. Hu. Cutoff values for two-group classification using neural networks, 1996. ISSN 00198528.

Masih Sherafatian. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene*, 677(July):111–118, 2018. ISSN 18790038. doi: 10.1016/j.gene.2018.07.057. URL https://doi.org/10.1016/j.gene.2018.07.057.

Sumbal Sumbal, Aneeqa Javed, Bakht Afroze, Hafiza Fizzah Zulfiqar, Faqeeha Javed, Sobia Noreen, and Bushra Ijaz. Circulating tumor DNA in blood: Future genomic biomarkers for cancer detection. *Experimental Hematology*, 65:17–28, 2018. ISSN 18732399. doi: 10.1016/j.exphem.2018.06.003. URL https://doi.org/10.1016/j.exphem.2018.06.003.

Leili Tapak, Nasrin Shirmohammadi-Khorram, Payam Amini, Behnaz Alafchi, Omid Hamidi, and Jalal Poorolajal. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*, (September):1–7, 2018. ISSN 22133984. doi: 10.1016/j.cegh.2018.10.003. URL https://doi.org/10.1016/j.cegh.2018.10.003.

Turki Turki and Zhi Wei. Boosting support vector machines for cancer discrimination tasks. *Computers in Biology and Medicine*, 101(July):236–249, 2018. ISSN 18790534. doi: 10.1016/j.compbiomed.2018.08.006. URL https://doi.org/10.1016/j.compbiomed.2018.08.006.

Lesley Uttley, Becky L. Whiteman, Helen Buckley Woods, Susan Harnan, Sian Taylor Philips, and Ian A. Cree. Building the Evidence Base of Blood-Based Biomarkers for Early Detection of Cancer: A Rapid Systematic Mapping Review. *EBioMedicine*, 10:164–173, 2016. ISSN 23523964. doi: 10.1016/j.ebiom.2016.07.004. URL http://dx.doi.org/10.1016/j.ebiom.2016.07.004.

Michel E. Vandenberghe, Marietta L.J. Scott, Paul W. Scorer, Magnus Söderberg, Denis Balcerzak, and Craig Barker. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, 7(March):1–11, 2017. ISSN 20452322. doi: 10.1038/srep45938. URL http://dx.doi.org/10.1038/srep45938.

Louella Vaughan. Biomarkers in acute medicine Key points. *Medicine*, 45(3):150–156, 2016. ISSN 1357-3039. doi: 10.1016/j.mpmed.2016.12.005. URL http://dx.doi.org/10.1016/j.mpmed.2016.12.005.

Mathukumalli Vidyasagar. Annual Reviews in Control Machine learning methods in computational cancer biology. *Annual Reviews in Control*, 43:107–127, 2017. ISSN 13675788. doi: 10.1016/j.arcontrol.2017.03.007.

Linda Vona-davis and David P Rose. Type 2 Diabetes and Obesity Metabolic Interactions : Common Factors for Breast Cancer Risk and Novel Approaches to Prevention and Therapy. pages 116–130, 2012.

Haifeng Wang, Bichen Zheng, Sang Won Yoon, and Hoo Sang Ko. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, 267 (2):687–699, 2018. ISSN 03772217. doi: 10.1016/j.ejor.2017.12.001. URL https://doi.org/10.1016/j.ejor.2017.12.001.

N. S. Ward. The Accuracy of Clinical Information Systems. 2004.

Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153:1–9, 2018. ISSN 18727565. doi: 10.1016/j.cmpb.2017.09.005.

Yuan Xu, Yuxin Wang, Jie Yuan, Qian Cheng, Xueding Wang, and Paul L. Carson. Medical breast ultrasound image segmentation by machine learning. *Ultrasonics*, 91(July 2018):1–9, 2019. ISSN 0041624X. doi: 10.1016/j.ultras.2018.07.006. URL https://doi.org/10.1016/j.ultras.2018.07.006.

Fei Xue and Karin B Michels. Diabetes, metabolic syndrome, and breast cancer: a review of the current evidence 1– 4. *Am J Clin Nutr*, 86, 2007.

Yadong Yang, Xunong Dong, Bingbing Xie, Nan Ding, Juan Chen, Yongjun Li, Qian Zhang, Hongzhu Qu, and Xiangdong Fang. Databases and web tools for cancer genomics study. *Genomics, Proteomics and Bioinformatics*, 13(1):46–50, 2015. ISSN 22103244. doi: 10.1016/j.gpb.2015.01.005. URL http://dx.doi.org/10.1016/j.gpb.2015.01.005.

Chen-Hsiung Yeh. Circulating Cell-Free DNA: The Blood Biopsy in Cancer Management. *MOJ Cell Science & Report*, 2(2):21–24, 2015. ISSN 23746912. doi: 10.15406/mojcsr.2015.02.00021. URL http://medcraveonline.com/MOJCSR/MOJCSR-02-00021.php.

Dong Zou, Lina Ma, Jun Yu, and Zhang Zhang. Biological databases for human research. *Genomics, Proteomics and Bioinformatics*, 13(1):55–63, 2015. ISSN 22103244. doi: 10.1016/j.gpb.2015.01.006. URL http://dx.doi.org/10.1016/j.gpb.2015.01.006.

Johannes Zuber, Dominique Bonnet, Harald Herrmann, Jan Jacob Schuringa, Jean Soulier, Gerrit Jan Schuurhuis, Alexander Roesch, Christine Chomienne, Sabine Cerny-Reiterer, Michel Arock, Connie Eaves, Junia V. Melo, Stefan Wöhrer, Ruggero De Maria, Peter Valent, Brian Huntly, Giorgio Stassi, Fumihiko Ishikawa, Tsvee Lapidot, Hans E. Johnsen, Mhairi Copland, and Michael Andreeff. Cancer stem cell definitions and terminology: the devil is in the details. *Nature Reviews Cancer*, 12(11):767–775, 2012. ISSN 1474-175X. doi: 10.1038/nrc3368. URL http://dx.doi.org/10.1038/nrc3368.

Apéndice A

Código fuente

A.1. Código fuente en Matlab para redes neuronales artificiales

A.1.1. Script principal BC_class.m

```
% Este script resuelve un problema de clasificaci'on por reconocimiento de
% patrones con redes neuronales artificiales
% Script base generado por Neural Pattern Recognition app
% Red neuronal para predicci'on de diagn'ostico de c'ancer de seno usando los
% biomarcadores
% edad (AGE), 'indice de masa corporal (BMI), glucosa (Glucose), insulina
% (Insulin), 'indice de resistencia a insulina (HOMA), leptina (Leptin),
% adiponectina (Adiponectin), resistina (Resistin), prote'ina
% quimiot'actica 1 (MCP1)
clear
clc
% En esta parte se imporan los datos desde un archivo xlsx, el cual
% contiene los valores numricos de cada biomarcador y su clasificacin
% Age,BMI,Glucose,Insulin,HOMA,Leptin,Adiponectin,Resistin,MCP1,Classification
% importacin  de los datos y normalizacin  con funci'on auxiliar scale.m
% cada biomarcador es una entrada a la red
data =(xlsread('dataR2.xlsx', 'BC', 'A2:J117'));
Age = scale(data(:,1));
BMI = scale(data(:,2));
Glucose = scale(data(:,3));
Insulin = scale(data (:,4));
HOMA = scale(data (:,5));
Leptin = scale(data (:,6));
Adiponectin = scale(data (:,7));
Resistin = scale(data (:,8));
MCP1 = scale(data (:,9));
Classification = scale(data(:,10));

% Creaci'on de grupos de experimentaci'on, para habilitar un grupo de
% experimentaci'on: remover s'imbolo porcentual
%V1-V9
% data = [Glucose,Resistin,Age,BMI,HOMA,Leptin,Insulin,Adiponectin,MCP1];
%V1-V4
data = [Glucose,Resistin,Age,BMI];

%asignaci'on de datos de entrada y datos objetivo
```

```
input = data';
target = Classification';

%validaci'on hold-out, 1= habilitado, 0= deshabilitado
% HOV = 1;
HOV = 0;

% Para habilitar la funci'on de entrenamiento, descomentar la l'inea
% trainFcn = 'trainscg';
trainFcn = 'trainrp';
% trainFcn = 'traincgb';

%Para guardar los estad'isticos de las redes generadas
values = zeros;

% ID de la red (iteraci'in)
nn = 1;

% k representa el n'umero de repeticiones o iteraciones (validaci'on k-fold)
% Para habilitar una capa oculta, quitar s'imbolo porcentual

% capa oculta 1
for i=16:4:24

%     capa oculta 2
%     for ii=16:4:24

%       capa oculta 3
%       for iii=16:2:24

%         capa oculta 4
%         for iiii=16:2:24

%            capa oculta 5
%            for iiiii=12:4:24

                for k = 1:30

            % Funci'on para entrenar y generar los datos estad'isticos de la red
            % para cambiar el n'umero de capas ocultas usar el formato de vector [L1,L2,L3]
            % ejemplo-> RedValid([i,ii],inputs,targets,trainFcn);
            [tr, performance, net ] = ...
                RedValid_classif([i],input,target,trainFcn,HOV);

%           valores de predicci'on-> pred=net(input)
%           valores realues-> real=targets
            pred = net(input)';
            real = target';

%           redondeando al entero m'as cercano, valor de corte 0.5
            prediction = round(pred) ;

%           Calculando errores producidos por la red en la predicci'on
            MSE = immse(real, prediction);
            sse = sum((real - prediction).^2);
            MAE= mae(prediction - real);
            RMSE= sqrt(MSE);

%           Obtenci'on de la AUC, sensibilidad, especificidad e 'indice de
%           Youden
            [X,Y,T,AUC] = perfcurve(real,prediction,'1');
            [c,cm,ind,per] = confusion(real',prediction');
            TP=cm(1,1);
            FN=cm(1,2);
            TN=cm(2,2);
```

```
            FP=cm(2,1);
            sensitivity=TP/(TP+FN);
            specificity=TN/(TN+FP);
            j_index=sensitivity+specificity-1;

%           Matriz con informaci'on de la red a mostrar en el archivo csv
            values(nn,2) = tr.num_epochs;
            values(nn,3) = tr.best_epoch;
            values(nn,4) = sse;
            values(nn,5) = MSE;
            values(nn,6) = RMSE;
            values(nn,7) = MAE;
            values(nn,8) = performance;
            values(nn,9) = AUC;
            values(nn,10) = sensitivity;
            values(nn,11) = specificity;
            values(nn,12) = i;
%             values(nn,13) = ii;
%             values(nn,14) = iii;
%             values(nn,15) = iiii;
%             values(nn,16) = iiiii;
            values(nn,17) = TP;
            values(nn,18) = FP;
            values(nn,19) = TN;
            values(nn,20) = FN;
            values(nn,21) = j_index;
            nn = nn + 1;
                end
%           end
%         end
%       end
%     end

% Se genera csv con la informaci'on de las distintas configuraciones creadas
  csvwrite('nombre_archivo.csv',values,1,0);
end
```

BC_class.m

A.1.2. Función principal RedValid_classif.m

```
function[tr, performance, net] = RedValid_classif(NEU,input,target,FCN,HOV)

x = input;
t = target;
% Crea red de reconocimiento de patrones NEU= capas ocultas a usar, FCN=
% funci'on de entrenamiento
net = patternnet(NEU,FCN);

% Funci'on de pre/postprocesamiento para las entradas y salidas
net.layers{length(NEU)+1}.transferFcn = 'tansig';
net.input.processFcns = {'mapstd'};
net.output.processFcns = {'mapminmax'};

% Divisi'on de datos para el entrenamiento, validaci'on y prueba
net.divideFcn = 'dividerand';  % divide aleatoriamente los datos
net.divideMode = 'sample';  % divide cada muestra

if HOV == 1
% radio de divisi'on para datos de entrenamiento, validaci'on y prueba
    net.divideParam.trainRatio = 80/116;
    net.divideParam.valRatio = 18/116;
```

```
22      net.divideParam.testRatio = 18/116;
23  else
24      net.divideParam.trainRatio = 98/116;
25      net.divideParam.valRatio = 0/116;
26      net.divideParam.testRatio = 18/116;
27  end
28
29  % Funcin  de desempeo
30  net.performFcn = 'crossentropy';
31  net.performParam.regularization = 0.5;
32
33  % Choose Plot Functions
34  % For a list of all plot functions type: help nnplot
35  net.plotFcns = {'plotperform','plottrainstate','ploterrhist', ...
36      'plotconfusion', 'plotroc'};
37
38  % 'epocas de entrenamiento, early-stop, 1= habilitado, 0= deshabilitado
39  if HOV == 1
40      net.trainParam.epochs = 100;
41  else
42      net.trainParam.epochs = 50;
43  end
44
45  % Entrenamiento de la red, x=entradas, t=objetivos
46  [net,tr] = train(net,x,t);
47
48  % Prueba de la red, x=entradas, y=salida predicci'on
49  y = net(x);
50
51  % (t) menos predicci'on (y), c'alculo del error=e
52  e = gsubtract(t,y);
53
54  % Evaluaci'on del desempeo  de la red
55  performance = perform(net,t,y);
56
57  % convirtiendo vectores a  dices  para obtener errores porcentuales
58  tind = vec2ind(t);
59  yind = vec2ind(y);
60  percentErrors = sum(tind ~= yind)/numel(tind);
61
62  % Recalculando desempeo  de entrenamiento, validaci'on y prueba
63  trainTargets = t .* tr.trainMask{1};
64  valTargets = t .* tr.valMask{1};
65  testTargets = t .* tr.testMask{1};
66  trainPerformance = perform(net,trainTargets,y);
67  valPerformance = perform(net,valTargets,y);
68  testPerformance = perform(net,testTargets,y);
```

RedValid_classif.m

A.1.3. Función auxiliar scale.m

```
1  function [scaled]= scale(data)
2  %obteniendo mins y maxs para normalizar con base en la unidad
3  minimum = min(data(:));
4  maximum = max(data(:));
5  scaled = ((data - minimum) / (maximum - minimum));
```

scale.m

A.2. Código fuente en R para análisis estadístico

A.2.1. Script normality_test.r

```
library(doBy)
BCdata=read.csv("dataR2.csv")
table(BCdata$Classification)

#primero se hace un test de normalidad para los datos con shapiro test
x=BCdata$Age
Agenormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test(
    subset(x, BCdata$Classification=="1"))$p.value)

x=BCdata$BMI
BMInormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test(
    subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$Glucose
Glucosenormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test
    (subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$Insulin
Insulinnormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test
    (subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$HOMA
HOMAnormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test(
    subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$Leptin
Leptinnormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test(
    subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$Adiponectin
Adiponectinnormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.
    test(subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$Resistin
Resistinnormality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.
    test(subset(x, BCdata$Classification=="2"))$p.value)

x=BCdata$MCP.1
MCP1normality=c(shapiro.test(subset(x, BCdata$Classification=="1"))$p.value, shapiro.test(
    subset(x, BCdata$Classification=="2"))$p.value)

normalitytests=rbind(Agenormality, BMInormality, Glucosenormality, Insulinnormality,
    HOMAnormality, Leptinnormality, Adiponectinnormality, Resistinnormality, MCP1normality)
colnames(normalitytests)=c("Control", "Patient")
normalitytests=round(normalitytests, digits=4)
rownames(normalitytests)=c("Age", "BMI", "Glucose", "Insulin", "HOMA", "Leptin", "
    Adiponectin", "Resistin", "MCP.1")
write.csv(normalitytests,"shapiro_test_BC.csv")

#test de mann u whitney wilcox
Agedata=summaryBy(Age ~ Classification, data=BCdata, FUN=c(median,IQR))
Agedata=cbind(Agedata[1, 2:3], Agedata[2, 2:3], wilcox.test(Age ~ Classification, data=
    BCdata)[3])
colnames(Agedata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-
    value")

BMIdata=summaryBy(BMI ~ Classification, data=BCdata, FUN=c(mean,sd))
BMIdata=cbind(BMIdata[1, 2:3], BMIdata[2, 2:3], wilcox.test(BMI ~ Classification, data=
    BCdata)[3])
```

```
colnames(BMIdata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Glucosedata=summaryBy(Glucose ~ Classification, data=BCdata, FUN=c(mean,sd))
Glucosedata=cbind(Glucosedata[1, 2:3], Glucosedata[2, 2:3], wilcox.test(Glucose ~ Classification, data=BCdata)[3])
colnames(Glucosedata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Insulindata=summaryBy(Insulin ~ Classification, data=BCdata, FUN=c(mean,sd))
Insulindata=cbind(Insulindata[1, 2:3], Insulindata[2, 2:3], wilcox.test(Insulin ~ Classification, data=BCdata)[3])
colnames(Insulindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

HOMAdata=summaryBy(HOMA ~ Classification, data=BCdata, FUN=c(mean,sd))
HOMAdata=cbind(HOMAdata[1, 2:3], HOMAdata[2, 2:3], wilcox.test(HOMA ~ Classification, data=BCdata)[3])
colnames(HOMAdata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Leptindata=summaryBy(Leptin ~ Classification, data=BCdata, FUN=c(mean,sd))
Leptindata=cbind(Leptindata[1, 2:3], Leptindata[2, 2:3], wilcox.test(Leptin ~ Classification, data=BCdata)[3])
colnames(Leptindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Adiponectindata=summaryBy(Adiponectin ~ Classification, data=BCdata, FUN=c(mean,sd))
Adiponectindata=cbind(Adiponectindata[1, 2:3], Adiponectindata[2, 2:3], wilcox.test(Adiponectin ~ Classification, data=BCdata)[3])
colnames(Adiponectindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Resistindata=summaryBy(Resistin ~ Classification, data=BCdata, FUN=c(mean,sd))
Resistindata=cbind(Resistindata[1, 2:3], Resistindata[2, 2:3], wilcox.test(Resistin ~ Classification, data=BCdata)[3])
colnames(Resistindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

MCP.1data=summaryBy(MCP.1 ~ Classification, data=BCdata, FUN=c(mean,sd))
MCP.1data=cbind(MCP.1data[1, 2:3], MCP.1data[2, 2:3], wilcox.test(MCP.1 ~ Classification, data=BCdata)[3])
colnames(MCP.1data)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

summarytable=rbind(Agedata, BMIdata, Glucosedata, Insulindata, HOMAdata, Leptindata, Adiponectindata, Resistindata, MCP.1data)
rownames(summarytable)=c("Age", "BMI", "Glucose", "Insulin", "HOMA", "Leptin", "Adiponectin", "Resistin", "MCP.1")
summarytable[, 1:4]=round(summarytable[, 1:4], digits=1)
summarytable[, 5]=round(summarytable[, 5], digits=4)
write.csv(summarytable,"wilcox_test_BC.csv")
```

normality_test.R

A.2.2. Script stats_data.r

```
# Script para generar estad'isticos
rm(list = ls())
cat("\014")
library(ggplot2)

```

```
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
# Importacion del csv de la variante del modelo a evaluar
data=read.csv('V1-V4_1-layer_trainscg_noHOV-2.csv')
# Extracci'on de valores de AUC para obtener media,
# desviaci'on est'andar, m'inimos, m'aximos
AUC=data$AUC
AUC_datos=split(AUC, ceiling(seq_along(AUC)/30))
avg_AUC=sapply(AUC_datos, mean, na.rm = TRUE)
stdev_AUC=sapply(AUC_datos, sd, na.rm = TRUE)
median_AUC=sapply(AUC_datos, median, na.rm = TRUE)
min_AUC=sapply(AUC_datos, min, na.rm = TRUE)
max_AUC=sapply(AUC_datos, max, na.rm = TRUE)
j_index=data$j_index
j_index_datos=split(j_index, ceiling(seq_along(AUC)/30))
avg_j_index=sapply(j_index_datos, mean, na.rm = TRUE)
Num_config=c(1:length(AUC_datos))
Num_nodos_L1=c(seq(20,60,by=2))
write.csv(cbind(Num_config,Num_nodos_L1,avg_AUC,avg_j_index,stdev_AUC,median_AUC,min_AUC,max
    _AUC),
            "V1-V4_1-layer_trainscg_noHOV-2_stats.csv")

data=read.csv('V1-V4_1-layer_trainscg_noHOV-2_stats.csv')
# genera imagen jpeg para el promedio de AUC y se desviaci'on est'andar
png(filename="V1-V4_1-layer_trainscg_noHOV-2_avg_AUC_stdev.png")
ggplot(data, aes(x=Num_nodos_L1, y=avg_AUC)) +
  geom_errorbar(aes(ymin=avg_AUC-stdev_AUC, ymax=avg_AUC+stdev_AUC), width=.2) +
  geom_point()+
  ggtitle(" Valor promedio para AUC y su desviacin  estndar")+
  xlab('Nmero de nodos en capa oculta') +
  ylab('Valor AUC')
dev.off()

#grafica los mejores HOV para V1-V4
rm(list = ls())
cat("\014")
library(ggplot2)
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
data1=read.csv('V1-V4_1-layer_traincgb_HOV_best.csv')
data2=read.csv('V1-V4_1-layer_trainscg_HOV_best.csv')
data3=read.csv('V1-V4_2-layer_trainrp_HOV_best.csv')
# genera imagen para el promedio de AUC y se desviaci'on est'andar
#filename V1-V4_best_HOV
ggplot(data=data1, aes( x=seq_along(1:30), y=AUC) ) +
  geom_point( aes(color="CGB HOV") )+
  geom_line(data=data1,aes(y=AUC, color="CGB HOV"),linetype="dotted")+
  geom_point( data=data2, aes( x=seq_along(1:30), y=AUC, color="SCG HOV"))+
  geom_line(data=data2,aes(y=AUC, color="SCG HOV"), linetype="dashed")+
  geom_point( data=data3, aes( x=seq_along(1:30), y=AUC, color="RP HOV"))+
  geom_line(data=data3,aes(y=AUC, color="RP HOV"),linetype="twodash")+
  ylim(0.7,1.0) +
  xlab('Iteracin') +
  ylab('AUC')+
  theme_light()+
 labs(color = "Algoritmo")

#grafica los mejores no HOV para V1-V4
rm(list = ls())
cat("\014")
library(ggplot2)
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
data1=read.csv('V1-V4_5-layer_traincgb_noHOV_best.csv')
data2=read.csv('V1-V4_5-layer_trainscg_noHOV_best.csv')
data3=read.csv('V1-V4_5-layer_trainrp_noHOV_best.csv')
# genera imagen para el promedio de AUC y se desviaci'on est'andar
#filename V1-V4_best_no_HOV
```

```
ggplot(data=data1, aes( x=seq_along(1:30), y=AUC) ) +
  geom_point( aes(color="CGB no HOV") )+
  geom_line(data=data1,aes(y=AUC, color="CGB no HOV"),linetype="dotted")+
  geom_point( data=data2, aes( x=seq_along(1:30), y=AUC, color="SCG no HOV"))+
  geom_line(data=data2,aes(y=AUC, color="SCG no HOV"), linetype="dashed")+
  geom_point( data=data3, aes( x=seq_along(1:30), y=AUC, color="RP no HOV"))+
  geom_line(data=data3,aes(y=AUC, color="RP no HOV"),linetype="twodash")+
  ylim(0.85,1.0) +
  xlab('Iteracin') +
  ylab('AUC')+
  theme_light()+
  labs(color = "Algoritmo")

#grafica los mejores no HOV para V1-V9
rm(list = ls())
cat("\014")
library(ggplot2)
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
data1=read.csv('V1-V9_2-layer_traincgb_noHOV_best.csv')
data2=read.csv('V1-V9_3-layer_trainscg_noHOV_best.csv')
data3=read.csv('V1-V9_5-layer_trainrp_noHOV_best.csv')
# genera imagen para el promedio de AUC y se desviaci'on est'andar
#filename V1-V9_best_no_HOV
ggplot(data=data1, aes( x=seq_along(1:30), y=AUC) ) +
  geom_point( aes(color="CGB no HOV") )+
  geom_line(data=data1,aes(y=AUC, color="CGB no HOV"),linetype="dotted")+
  geom_point( data=data2, aes( x=seq_along(1:30), y=AUC, color="SCG no HOV"))+
  geom_line(data=data2,aes(y=AUC, color="SCG no HOV"), linetype="dashed")+
  geom_point( data=data3, aes( x=seq_along(1:30), y=AUC, color="RP no HOV"))+
  geom_line(data=data3,aes(y=AUC, color="RP no HOV"),linetype="twodash")+
  ylim(0.85,1.0) +
  xlab('Iteracin') +
  ylab('AUC')+
  theme_light()+
  labs(color = "Algoritmo")

#grafica los mejores HOV para V1-V9
rm(list = ls())
cat("\014")
library(ggplot2)
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
data1=read.csv('V1-V9_1-layer_traincgb_HOV_best.csv')
data2=read.csv('V1-V9_1-layer_trainscg_HOV_best.csv')
data3=read.csv('V1-V9_1-layer_trainrp_HOV_best.csv')
# genera imagen para el promedio de AUC y se desviaci'on est'andar
#filename V1-V9_best_HOV
ggplot(data=data1, aes( x=seq_along(1:30), y=AUC) ) +
  geom_point( aes(color="CGB HOV") )+
  geom_line(data=data1,aes(y=AUC, color="CGB HOV"),linetype="dotted")+
  geom_point( data=data2, aes( x=seq_along(1:30), y=AUC, color="SCG HOV"))+
  geom_line(data=data2,aes(y=AUC, color="SCG HOV"), linetype="dashed")+
  geom_point( data=data3, aes( x=seq_along(1:30), y=AUC, color="RP HOV"))+
  geom_line(data=data3,aes(y=AUC, color="RP HOV"),linetype="twodash")+
  ylim(0.5,1.0) +
  xlab('Iteracin') +
  ylab('AUC')+
  theme_light()+
  labs(color = "Algoritmo")

#Mejor performance con y sin HOV de V1-V4
rm(list = ls())
cat("\014")
library(ggplot2)
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
data1=read.csv('V1-V4_1-layer_traincgb_HOV_best.csv')
```

```
data2=read.csv('V1-V4_5-layer_traincgb_noHOV_best.csv')
# genera imagen para el promedio de AUC y se desviaci'on est'andar
#filename V1-V4_HOV_vs_noHOV
ggplot(data=data1, aes( x=seq_along(1:30), y=AUC) ) +
  geom_point( aes(color="CGB HOV") )+
  geom_line(data=data1,aes(y=AUC, color="CGB HOV"),linetype="dotted")+
  geom_point( data=data2, aes( x=seq_along(1:30), y=AUC, color="CGB no HOV"))+
  geom_line(data=data2,aes(y=AUC, color="CGB no HOV"), linetype="dashed")+
  ylim(0.7,1.0) +
  xlab('Iteracin') +
  ylab('AUC')+
  theme_light()+
  labs(color = "Algoritmo")

#Mejor performance con y sin HOV de V1-V9
rm(list = ls())
cat("\014")
library(ggplot2)
setwd("C:/Users/Balam Benitez/Dropbox/Tesis Balam Benitez/matlab")
data1=read.csv('V1-V9_1-layer_trainrp_HOV_best.csv')
data2=read.csv('V1-V9_2-layer_traincgb_noHOV_best.csv')
# genera imagen para el promedio de AUC y se desviaci'on est'andar
#filename V1-V9_HOV_vs_noHOV
ggplot(data=data1, aes( x=seq_along(1:30), y=AUC) ) +
  geom_point( aes(color="RP HOV") )+
  geom_line(data=data1,aes(y=AUC, color="RP HOV"),linetype="dotted")+
  geom_point( data=data2, aes( x=seq_along(1:30), y=AUC, color="CGB no HOV"))+
  geom_line(data=data2,aes(y=AUC, color="CGB no HOV"), linetype="dashed")+
  ylim(0.65,1.0) +
  xlab('Iteracin') +
  ylab('AUC')+
  theme_light()+
  labs(color = "Algoritmo")
```

stats_data.R