Enfoque de la teoría de juegos en detección de cáncer de mama, asistido por un algoritmo clasificador

Gancino Chacha, Katherine Michelle

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Trabajo de titulación, previo a la obtención del título de Ingeniera en Electrónica y Telecomunicaciones

Ing. Bernal Oñate, Carlos Paúl MSc.

16 de marzo del 2021

Curiginal

Document Information

Analyzed document	Gancino_Katherine_tesis.docx (D98683887)
Submitted	3/17/2021 7:59:00 PM
Submitted by	Bernal Oñate Carlos Paúl
Submitter email	cpbernal@espe.edu.ec
Similarity	1%
Analysis address	cpbernal.espe@analysis.urkund.com

Sources included in the report

W	URL: https://repositorio.unbosque.edu.co/bitstream/handle/20.500.12495/3297/Jord%C3%A1n ... Fetched: 1/7/2021 11:16:49 PM	1
W	URL: https://la.mathworks.com/discovery/machine-learning.html) Fetched: 3/17/2021 8:06:00 PM	3
W	URL: https://la.mathworks.com/discovery/neural-network.html). Fetched: 3/17/2021 8:06:00 PM	1

Firmado electrónicamente por:
CARLOS PAUL
BERNAL ONATE

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"Enfoque de la teoría de juegos en detección de cáncer de mama, asistido por un algoritmo clasificador"** fue realizado por la señorita **Gancino Chacha, Katherine Michelle** el cual ha sido revisado y analizado en su totalidad por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 17 de marzo del 2021

Firma:

Firmado electrónicamente por:
CARLOS PAUL
BERNAL ONATE

Ing. Carlos Paúl Bernal Oñate MSc.

C. C. 1709775637

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

Responsabilidad de Autoría

Yo, **Gancino Chacha, Katherine Michelle**, con cédula de ciudadanía n°1723114516, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Enfoque de la teoría de juegos en detección de cáncer de mama, asistido por un algoritmo clasificador"** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 17 de marzo del 2021.

Firma:

Gancino Chacha, Katherine Michelle

C.C.: 1723114516

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

Autorización de Publicación

Yo **Gancino Chacha, Katherine Michelle,** con cédula de ciudadanía n°1723114516, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Enfoque de la teoría de juegos en detección de cáncer de mama, asistido por un algoritmo clasificador"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi/nuestra responsabilidad.

Sangolquí, 17 de marzo del 2021.

Firma:

.....

Gancino Chacha, Katherine Michelle

C.C.: 1723114516

Dedicatoria

Esta investigación dedico a las personas más importantes de mi vida, mi familia.

A mi padre, Willian, quien es mi fortaleza y el piloto de mi vida, el mejor modelo de valor y esfuerzo.

A mi madre, Nati, mi refugio de amor y paz, que con su ternura y calidez me da valentía y motivación.

A mi hermanita, Dani, quien es mi gran soporte de vida y manantial de aprendizaje.

A mi hermanita, Abi, quien es amparo y calor en mi vida y me protege y cuida siempre.

A mi dulce y pequeña Molly, mi fiel compañía.

Y a mis amigos, con quienes compartí muchos momentos maravillosos y han sido incondicionales, tienen un gran espacio en mi corazón.

Con amor para ustedes,

Katherine Gancino

Agradecimientos

Agradezco primero a Dios, por darme salud y fuerza cada mañana de este recorrido para cumplir mi objetivo, a pesar de las adversidades y por mantener a las personas que amo conmigo.

Gracias infinitas mis padres, porque de su mano caminé todos los días hasta este momento y me han dado los mejores recursos profesionales, personales y de vida para alcanzar este primer logro. Gracias porque me educaron siempre con amor, paciencia y sabiduría.

A mis hermanitas, gracias porque que juntas enfrentamos muchos momentos y situaciones difíciles, y que solo nos han hecho más valientes, son mis guerreras; así también gracias por compartir felicidad, risas, consejos y conocimientos.

Gracias a toda mi familia, especialmente a mi preciosa tía Raquelita que es un angelito que me apoya y cuida siempre.

Gracias amigos queridos, tuve el honor de conocerlos en esta etapa universitaria y han sido parte fundamental de este proceso, gracias por las experiencias compartidas. Quiero agradecer especialmente a Anthony, Carolina, Stefany y Andrés por su valiosa amistad y por ser un apoyo importante en lo académico y deportivo.

Un agradecimiento a todos los docentes de la Universidad de las Fuerzas Armas ESPE por la formación profesional y deportiva, especialmente a mi tutor académico Ing. Paúl Bernal por bridarme la oportunidad de culminar mis estudios de ingeniería con su apoyo.

Finalmente, un agradecimiento muy especial al Ing. Rodolfo Gordillo, por todo el conocimiento brindado y el aporte significativo en esta investigación.

Contenidos

Índice de Tablas

Índice de Figuras

Resumen

La ciencia del último siglo ha centrado su atención sobre la teoría de juegos algorítmica en investigaciones sobre inteligencia artificial, ensayos médicos, economía, telecomunicaciones, etc., en general es un área de inminente crecimiento. Los algoritmos basados en la teoría de juegos son estudios actuales que tienen como objetivo analizar la dinámica de toma de decisiones en entornos donde existen partes que compiten entre sí y que requieren construcción de estrategias. Una parte importante de estos estudios se dirigen a aplicaciones médicas como el diseño de sistemas computacionales de apoyo al profesional para la detección de patologías. En la presente investigación se propone el desarrollo de un algoritmo clasificador modelado en base a la teoría de juegos para detectar la presencia de cáncer de mama, utilizando el software Matlab; se emplea una base de datos que contiene las características celulares de una muestra de tejido obtenida de pacientes con sospecha, y diagnóstico confirmado de cáncer de mama. Con el objetivo de validar la confiabilidad del algoritmo propuesto se desarrollaron modelos con *Machine Learning,* Máquina de Vectores de Soporte y Redes Neuronales Artificiales. A partir de las técnicas de optimización: reducción de dimensionalidad y selección de características, se definieron modelos eficientes y mayor capacidad predictiva. La comparación de los resultados obtenidos entre los parámetros de rendimiento del algoritmo de teoría de juegos y los de Machine Learning, demuestran que es apto para aplicaciones de clasificación.

Palabras clave:

- **TEORÍA DE JUEGOS**
- **MACHINE LEARNING**
- **CÁNCER DE MAMA**

Abstract

Last century's science has focused its attention on the algorithm games theory in research about artificial intelligence, medical trials, economics, telecommunications, etc., generally it's an area in constant growth. The algorithms based in the game theory are current studies that aim to analyze the dynamic oh decision taking in environments where exist parts that compete between them and that require to build strategies. An important part in this studies is directed to medical applications such as the design of computing support systems to the professional in order to detect pathologies. The current investigation aims to develop an classifier algorithm model based in the game theory to detect the presence of breast cancer, using the software Matlab; it's used in data base that contains cellular characteristics of tissue sample obtained from patients with suspect and confirmed diagnosis of breast cancer. The purpose is to validate the reliability of the suggested algorithm Machine Learning, Support Vector Machine and Artificial Neural Network model were developed. From the optimization techniques/ dimensional reduction and characteristics selection, efficient model were defined and a higher predictive capacity. The comparison of the obtained results between the performance parameters of the algorithm from the game theory and the Machine Learning, demonstrate that is suitable for the applications of classification.

Keywords:

- **GAME THEORY**
- **MACHINE LEARNING**
- **BREAST CANCER**

Capítulo I

Introducción

Antecedentes

El cáncer ocasionó la muerte de 8,8 millones de personas en 2015 según la OMS, es decir, una de cada seis defunciones en el mundo se debe a esta enfermedad. El cáncer de mama se presenta con más frecuencia en mujeres que en hombres, solo en América más de 462.000 mujeres son diagnosticadas con cáncer de mama, y casi 100.000 fallecen cada año a causa de esta enfermedad. En América Latina y el Caribe, el cáncer de mama es el más común en las mujeres y también considerado el segundo tipo de cáncer con mayor tasa de mortalidad (Organización Mundial de la Salud, 2018).

El diagnóstico precoz ha sido clave en la estadística tangible de una mayor supervivencia para las mujeres con cáncer de mama. Sin embargo, muchos países de América Latina continúan teniendo un acceso limitado a estas intervenciones, es por esta razón que la mayoría de los casos se diagnostican en fases avanzadas (WHO, 2008).

La primera descripción de un estudio de detección de masas anómalas en las mamas fue determinada por la mamografía, este estudio realizado en 1913 corresponde a Albert Salomón. Mediante exposiciones bajas de rayos X, utilizó tres mil piezas precedentes de mastectomías.

A partir de este estudio surgen nuevas técnicas como la ecografía en 1951, Wild J. y Neal D. describen la primera ecografía de la glándula mamaria, sin embargo, apenas en 1970 adquiere relevancia debido a la preocupación por la radiación de estudios anteriores (Monte, 1995).

Si bien el método de referencia de las imágenes de mama es la mamografía, en el intento de evitar las radiaciones ionizantes, con efectos deletéreos comprobados sobre el organismo, se han desarrollado otros métodos, con fuentes de energía alternas.

Elise Fear y colaboradores en su artículo "*Microwaves for breast cancer*", mencionan que las propiedades eléctricas de los tejidos están relacionadas con su estado fisiológico y por lo tanto los cambios en dichas propiedades indican la presencia de enfermedades. De aquí el estudio de detección de cáncer con imágenes de microondas, este define una vista de la estructura interna del seno a través de la exposición de las mamas a un campo electromagnético utilizando un transmisor para iluminarlo con microondas, que viajan a través del seno y pueden ser detectadas en receptores situados en el lado opuesto del mismo y los reflejos pueden ser registrados en la antena transmisora. Con un tumor presente, las ondas que viajan a través del seno encuentran un cambio en las propiedades eléctricas, causando que la onda incidente se disperse (Fear, Meaney, & Stuchly, 2003).

Existen otros planteamientos diferentes para la detección de cáncer de mama, están los métodos implementados a través de la inteligencia artificial y aún más de algoritmos de reconocimiento de patrones altamente desarrollados y de programas avanzados que manejan complejos conjuntos de datos (especialmente imágenes) y que funcionan a muy altas velocidades, dando lugar a la aparición de sistemas informáticos capacitados para realizar tareas complejas en bioinformática, imágenes médicas, robótica médica y capaces de predecir eficazmente los resultados futuros de un tipo de cáncer (Goldenberg, Nir, & Salcudean, 2019).

Un ejemplo alterno a un diagnóstico por equipo médico es el estudio desarrollado por Samik Real, a partir de la implementación de una red neuronal convolucional para la clasificación de lesiones de mamas en mamografías. Aquí interviene también procesamiento de

imágenes y técnicas de segmentación aplicadas en las mismas para así entrenar a la red neuronal y clasificar los casos en presencia o ausencia de cáncer de mama. (Real, 2019)

En general una variedad de estas técnicas, las redes neuronales artificiales o RNA (*Artificial Neural Networks, ANN*), las Máquinas Vectoriales de Apoyo (SVM, *Support Vector Machine*) y los Árboles de Decisión (*Decision Trees*) se han aplicado ampliamente en la investigación del cáncer para el desarrollo de modelos de predicción, sin embargo es evidente que existen rasgos que se pueden mejorar en esta investigación y obtener un nivel apropiado de validación de forma que se puedan considerar en la práctica clínica diaria (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015).

Por otra parte, en el pasado el desarrollo de aplicaciones basado en teoría de juegos se restringía a temas de economía en el análisis de mercado y comportamiento social sin embargo las técnicas de teoría de juegos han recibido una atención creciente en los últimos años, ya que se pueden adoptar para modelar y comprender escenarios competitivos y cooperativos entre los responsables de la toma de decisiones, (Trestian, Ormond, & Muntean, 2012) esta situación es ubicua en la práctica de la ingeniería, por ejemplo en el campo de las telecomunicaciones en un entorno inalámbrico tan heterogéneo un desafío importante es habilitar mecanismos de selección de red para mantener a los usuarios móviles siempre mejor conectados en cualquier lugar y en cualquier momento.

Según científicos de la Universidad Johns Hopkins, quienes comenzaron a visualizar los procesos tumorales desde la perspectiva de la teoría de juegos, la aplicación de dicha teoría en biología suele referirse al estudio de la capacidad de reproducción, cambios y etapas que siguen las células. Uno de los expertos, Kenneth J. Pienta, indicó que su estrategia es determinar cómo las células cancerosas cooperan con el objetivo de reunir más de ellas y energía de otras células (Cashin-Garbutt, 2014).

Para Archetti M. y Pienta K. la cooperación celular en beneficio del tumor o de reunir energía para incrementarse, está ligada a métodos y conceptos de la teoría de los juegos evolutivos, que se ha utilizado con éxito en otros problemas similares en torno a otras ramas de la biología, pero se ha infrautilizado en la investigación del cáncer. La teoría de juegos puede proporcionar información sobre la estabilidad de la cooperación entre las células de un tumor y sobre el diseño de terapias potencialmente a prueba de evolución que interrumpen esta cooperación (Archetti & Pienta, 2018).

Justificación

En Ecuador y en el mundo, el cáncer de mama constituye la tercera causa de muerte en mujeres. En 2018, según el Ministerio de Salud Pública, se atendieron 28.058 nuevos casos. Además, se evidenció que el acceso a un diagnóstico temprano incrementa la probabilidad de supervivencia del paciente de manera drástica. (Ministerio de Salud Pública , 2020)

Por el número importante de vidas que ha tomado esta enfermedad, la evolución de los distintos métodos de detección de cáncer es una constante para médicos, biólogos y demás profesionales adyacentes al tema. En este sentido, el siguiente paso de la investigación es mejorar la precisión de la predicción del cáncer de mama con la implementación de un algoritmo desarrollado en base a la teoría de juegos, y así también estimar un diagnóstico oportuno de recurrencia y supervivencia.

El foco de la investigación se centra en cubrir un tema poco abordado y novedoso de gran valor científico, ya que el desarrollo de algoritmos de aprendizaje y clasificación se restringe a metodologías tradicionales cuál sea el fin de la aplicación; y por otra parte, un problema común que se ha observado en varios trabajos de investigación es la falta de validación o comprobación externa del rendimiento predictivo de sus modelos, métodos y algoritmos.

La teoría de juegos busca superar estos aspectos describiendo un proceso biológico a través de un algoritmo de detección de cáncer de mama, utilizando una base de datos confiable que contiene características tomadas de casos reales. El resultado de los parámetros que describen el nivel de eficiencia del algoritmo clasificador será validado en contraste con los de los modelos de *Machine Learning.* Por lo expuesto anteriormente, este sistema de detección de cáncer de mama constituye un método fiable con parámetros de altas exigencias, validación y principalmente con potencial soporte al diagnóstico médico.

Alcance del Proyecto

La primera fase es de carácter exploratorio en la cual se realiza un estudio y análisis del estado del arte sobre aplicaciones y funcionamiento teórico de los algoritmos con teoría de juegos, así como los conceptos principales, metodologías y parámetros involucrados en un modelo de juego.

A partir de la base de datos *Wisconsin Prognostic Breast Cancer (WPBC)* obtenida del repositorio de dominio público del Departamento de Ciencias Clínicas de la Universidad de Wisconsin (University of Wisconsin, 1995) se desarrollan los modelos de clasificación con *Machine Learning*: Máquina de Vectores de Soporte y Redes Neuronales Artificiales en el software Matlab. Se aplican técnicas de reducción de dimensionalidad PCA y selección de características para abstraer los patrones idóneos de clasificación y optimizar los modelos de entrenamiento.

El algoritmo de teoría de juegos se implementa en Matlab entorno a los parámetros que comprende un modelo de juego y posteriormente se aplica la estrategia Minimax que establece la clase o diagnóstico del paciente a través de las recompensas obtenidas por los jugadores. Mediante métricas se evalúan y comparan los resultados obtenidos entre los distintos modelos y su rendimiento como clasificador para definir cuán rentable es el algoritmo propuesto.

Objetivos

General

Desarrollar un algoritmo clasificador con teoría de juegos aplicado a la detección de cáncer de mama.

Específicos

- Investigar y analizar aplicaciones existentes de algoritmos basados en la teoría de juegos.
- Desarrollar modelos de clasificación convencionales con Machine Learning: Máquina de Vectores de Soporte y Redes Neuronales Artificiales.
- Aplicar técnicas de reducción de dimensionalidad y selección de características para optimizar los modelos de Machine Learning.
- Implementar el algoritmo clasificador con teoría de juegos considerando los parámetros y métodos que intervienen en el juego.
- Analizar y comparar los resultados obtenidos del desempeño del algoritmo de teoría de juegos frente a los modelos de Machine Learning e identificar las prestaciones del mismo.

Capítulo II

Fundamentación Teórica

Teoría de Juegos

Frente a los posibles escenarios de incertidumbre y riesgos en torno a la Guerra Fría, los Estados Unidos de América funda la *Rand Corporation*, una organización que surgió para realizar procesos de análisis, estudio y toma de decisiones ante diversas situaciones críticas que pudieran presentarse. El principal promotor en esta línea de investigadores implicados en aplicaciones para el Departamento de Defensa de los Estados Unidos fue John Von Neumann (Herrero & Pinedo del Campo, 2005). A partir de aquí se forman grupos de investigación respecto al comportamiento de los "jugadores" o de quienes participan en un conflicto y las estrategias que formulan para ganar.

Con la obra de Von Neumann y Oskar Morgenstern*, Theory of Games and Economic Behaviour* (Teoría de juegos y comportamiento económico) de 1944, la teoría de juegos logró establecerse como una disciplina científica independiente cuya influencia trascendería el dominio estrictamente matemático en poco tiempo, para originar un amplio abanico de campos de conocimiento, desde el estudio de patologías, biología evolutiva, ciencia política, psicología, estrategias militares, hasta actualmente en las telecomunicaciones.

La teoría de juegos tiene dos interpretaciones, la primera una interpretación normativa que busca predecir acerca del comportamiento estratégico óptimo de cada uno de los jugadores involucrados. Y la segunda, la interpretación descriptiva, que trata de explicar cada movimiento de cualquier jugador en base a la racionalidad de estos y determinar cuál es el resultado estable para ellos (Cremades, 2016).

De forma general la Teoría de Juegos es un concepto matemático el cual estudia la formulación de la estrategia más óptima que permitirá a un individuo o entidad (jugador) tener éxito en abordar un desafío complejo en un ecosistema competitivo (Al-Raweshidy, 2010).

Elementos que Intervienen en la Teoría de Juegos

Agente o jugador. Representa a la persona o entidad teniendo sus propios objetivos o preferencias.

Acciones o Estrategias. Son cada una de las alternativas o "formas de jugar" que el jugador puede adoptar cuando debe decidir, buscando maximizar su utilidad, y teniendo en cuenta el estado de las posibles acciones de la contraparte.

Recompensa. También llamada pago, y se refiere a la utilidad que cada jugador espera obtener por cada acción permitida en el juego.

Información. Hace referencia a la información que posee cada jugador para decidir sus estrategias y puede ser completa, simétrica o perfecta de acuerdo con las acciones, recompensas y secuencia de movimientos que se manifiesten en el juego.

Formas de representación

A continuación, se utiliza un mismo juego para deducir las dos formas de representación.

Normal o Matricial. Se suele utilizar este método de representación, aunque no de manera exclusiva, cuando hay únicamente dos jugadores. De tal manera que se colocan las posibles estrategias del jugador 1 en filas y en columnas las del jugador 2, como se muestra en la Tabla 1**.** Los pagos o ganancias se colocan dentro de los casilleros de intersección siendo el pago del jugador 1 el que se ubica más cerca de él (Pa) y continuación el pago del jugador 2

seguido de una coma (Px). Las posibles estrategias del jugador 1 son A y B, y por lo tanto X e Y son estrategias del jugador 2.

Tabla 1

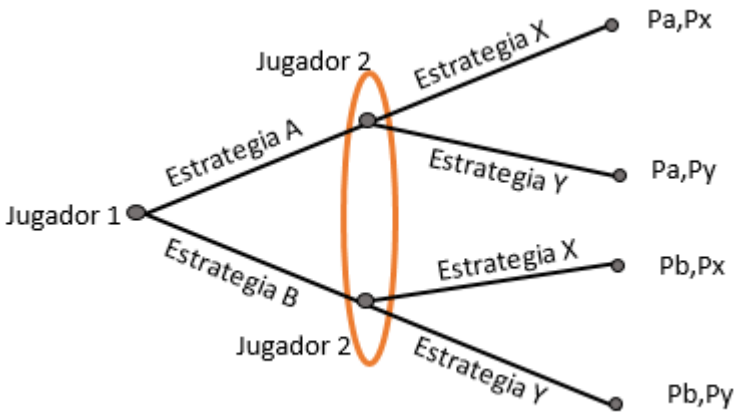
Representación matricial de un juego

		Jugador 2	
		Estrategia X	Estrategia Y
Jugador 1	Estrategia A	Pa,Px	Pa,Py
	Estratega B	Pb,Px	Pb,Py

Extensiva o de árbol. Tomando en consideración las mismas cuatro posibilidades de la explicación anterior, se forman ramas a partir de cada jugador como se observa en la Figura 1 y en estas se colocan respectivamente las estrategias así como los pagos al final de cada ramificación, para evitar que se interprete que primero elige el jugador 1 y a continuación el jugador 2 (puesto que leyendo el gráfico de izquierda a derecha pareciera que es así), se representa la elipse que lleva el nombre de “Conjunto de Información” la cual une los dos puntos los cuales el jugador 2 va a elegir, de esta forma se quiere indicar que el jugador 2 hará su elección única desconociendo la elección del jugador 1.

Figura 1

Representación extensiva o de árbol de un juego



Tipos de juegos

Para definir un conjunto de juegos y modelos existentes se considera fundamentalmente los métodos que se aplicarán para resolverlos, cantidad de jugadores, el cómputo de recompensas y pérdidas que obtienen los jugadores, entre otros parámetros, a continuación se muestran los tipos de juegos en la Figura 2**.**

Figura 2

Tipos de juegos respecto de la teoría de juegos

Nota. Adaptado de *Economics Discussion*, de Nitisha, 2011, Types of Games (https://www.economicsdiscussion.net/game-theory).

Juegos cooperativos y no cooperativos. Los juegos no cooperativos se refieren a los juegos en los que los jugadores eligen su estrategia de forma independiente para mejorar su propio rendimiento (utilidad) o reducir sus pérdidas(costes). Para resolver estos juegos existen varios conceptos como el Equilibrio de Nash. El mejor ejemplo de un juego no cooperativo es "El Dilema del prisionero" (Saad, Han, Debbah, Hjorungnes, & Basar, 2009).

Mientras que en los juegos cooperativos los jugadores pueden formar compromisos vinculantes a través de negociaciones y acuerdos, también referidos como coaliciones.

Juegos forma normal y extensiva. Se refiere básicamente a la representación de un juego, los juegos de forma normal se describen en una matriz donde se demuestra las estrategias adoptadas por los jugadores y los posibles resultados, también permite identificar las estrategias.

Los juegos de forma extensiva se representan a través de un árbol de decisiones en el que los jugadores están representados en diferentes nodos. Una particularidad de esta descripción es que ayudan en la representación de eventos que pueden ocurrir por casualidad.

Juegos simultáneos y juegos secuenciales. La información sobre el juego, es decir, las reglas y los pagos son de conocimiento común para todos los jugadores y las decisiones que estos tomen obligadamente son simultáneas, se refiere a que los jugadores no tienen conocimiento sobre el movimiento que hará el otro. Los juegos simultáneos se describen a través de la forma normal.

En los juegos secuenciales se realizan alternadamente una serie de decisiones, y el resultado de cada una de ellas altera las posibilidades del otro jugador. Es decir, los jugadores son conscientes de los movimientos de los jugadores que ya han adoptado una estrategia. Los

juegos secuenciales se describen en árboles de decisión, con nodos sucesivos en cada punto de decisión.

Juegos de suma cero y no cero. En los juegos de suma cero el beneficio de uno solo supone la pérdida del otro y en consecuencia, no es posible la colaboración entre las partes (Cremades, 2016). Un jugador gana precisamente lo que el otro pierde.

En los juegos de suma no cero, la ganancia de un jugador no necesariamente implica la pérdida del otro y el desenlace de estos juegos tiene resultados mayores o menores para cada jugador de acuerdo a las estrategias aplicadas. El ejemplo más clásico es El Dilema del Prisionero.

Juegos simétricos y asimétricos. En un juego simétrico, las estrategias adoptadas por los jugadores deben ser las mismas. El pago adquirido depende de las estrategias que utilicen los otros jugadores y no de quien las juegue.

En los juegos asimétricos, por el contrario, las estrategias adoptadas por todos los jugadores son diferentes. Si una estrategia en particular beneficia a un jugador, no quiere decir que sea favorable para otro.

El Dilema del Prisionero

Ideado por los investigadores de RAND Corporation en 1950, Merril Flood y Melvin, y formalizado por Albert W. Tucker de la Universidad de Princeton, es el problema más popularizado por los expertos para explicar la Teoría de Juegos (Cremades, 2016). En este juego, dos prisioneros fueron arrestados y acusados de un crimen; la policía no tiene suficiente evidencia para condenar a ninguno de ellos, a menos que uno de los sospechosos confiese. La policía mantiene a los criminales en celdas separadas, por lo tanto, cada uno debe optar por una de dos alternativas, la colaboración o la defección, sin contar con información alguna

acerca de la elección que hará su par. Eventualmente, a cada sospechoso se le dan tres posibles resultados (Al-Raweshidy, 2010):

1. Si uno confiesa y el otro no, el confesor será liberado y el otro permanecer tras las rejas durante diez años (es decir, -10);
2. Si ninguno de los dos lo admite, ambos serán encarcelados por un corto período de tiempo (es decir, -2,-2); y
3. Si ambos confiesan, ambos serán encarcelados por un período de tiempo intermedio (es decir, seis años en prisión, -6).

Las posibles acciones y las correspondientes sentencias de los criminales se dan en la Tabla 2**.**

Tabla 2

Representación del juego "El Dilema del Prisionero"

		Segundo criminal	
		Coopera	No coopera
Primer criminal	Coopera	-2,-2	-10,0
	No coopera	0,-10	-6,-6

Para resolver este juego, se busca la estrategia dominante de cada jugador, es decir la mejor opción para cada jugador independientemente de lo que el otro decida. Desde el punto de vista de cada jugador, si su oponente coopera (es decir, no confiesa), entonces tendría ventaja confesando (es decir, culpando a su compañero), funciona en ambos sentidos. Al final, ambos prisioneros concluyen que la mejor decisión es no confesar, y ambos son enviados a prisión por un corto periodo.

Estrategias dominantes y dominadas

Una estrategia dominante es la estrategia óptima para un jugador, independientemente de lo que hagan los demás. Si un jugador sigue una estrategia dominante, obtendrá la mejor recompensa posible independientemente de lo que hagan los demás jugadores.

Así como una estrategia dominante es una estrategia que es mejor que cualquier otra estrategia que un jugador pueda elegir de su conjunto de acciones, una estrategia dominada es una estrategia que es peor que otra estrategia disponible para el jugador (Kubilay & Anderson, 2010).

Equilibrio de Nash

El Equilibrio de Nash describe un resultado del juego en el que ningún jugador tiene un incentivo para cambiar su estrategia dadas las estrategias de los otros jugadores (Doufene & Krob, 2015), también descrita como la opción más óptima elegida por aquel jugador.

Este concepto pertenece a la teoría de juegos debido a que el matemático creador, John Nash, logró demostrar en 1951 que en todo juego en donde los participantes tienen la opción de escoger entre un número finito de estrategias existirá al menos un equilibrio.

Estrategia Minimax

En un juego de suma cero entre dos personas, si un jugador I intenta tomar acción (es) para reducir la recompensa del otro jugador (jugador II), el jugador II tomará la (s) acción (es) que le darán el pago mínimo máximo. Como cada jugador no puede mejorar su posición respecto a la utilidad, esta estrategia es el equilibrio (minimax). Este criterio está basado en el teorema minimax que fue demostrado por John von Neumann en 1928. Él llegó a la conclusión de que el mínimo de la ganancia máxima es igual al máximo de la ganancia mínima (Kubilay & Anderson, 2010).

Cáncer de Mama

El cáncer de mama representa uno de los tipos de cáncer líderes en número de casos diagnosticados y se ubica en segundo lugar después del cáncer de pulmón, como la mayor causa de muerte en mujeres (American Cancer Society, 2016).

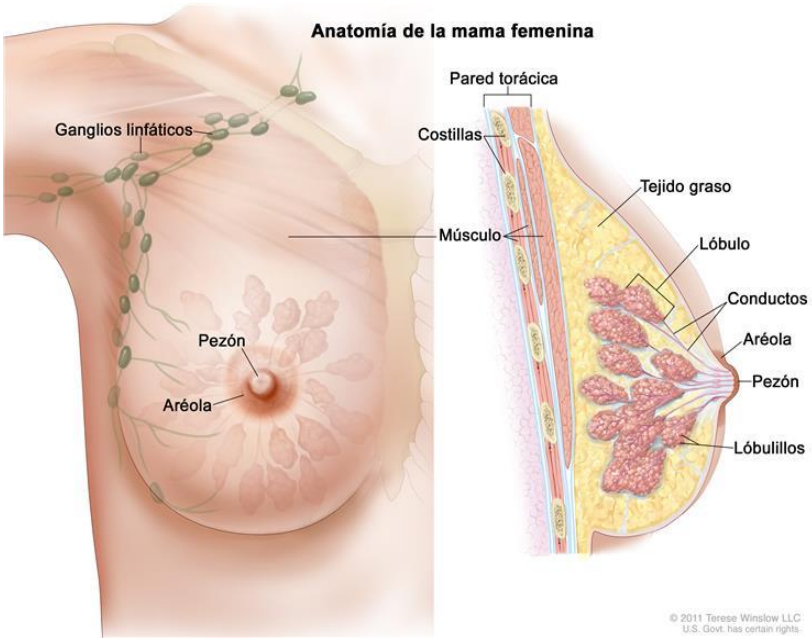
El cáncer de mama es una enfermedad caracterizada por el crecimiento excesivo, desorganizado e invasivo de células anormales propias de la mama, es decir, en el proceso normal de división celular existen ciertas células anómalas que carecen de control en su división, dicho desbordamiento de células (pueden ser o no malignas) forman un tumor que a menudo se puede ver en una radiografía o sentir como un bulto; a continuación dichas células se pueden extender hasta diferentes partes del cuerpo a través de los vasos sanguíneos y los canales linfáticos, provocando metástasis y la muerte, no solamente en mujeres, también los hombres lo padecen pero en inferior porcentaje.

Algunas de las causas del cáncer son debido a factores externos como el tabaco, organismos infecciosos, una dieta no saludable y factores internos como: mutaciones genéticas heredadas, hormonas y condiciones inmunes (American Cancer Society, 2016).

Existen diferentes tipos de cáncer de mama, sin embargo, están determinados por las células específicas de la mama que se ven afectadas o sección de origen, como se aprecia en la Figura 3. La mayoría de los casos de cáncer de mama están clasificados como: in situ, o invasivos (American Cancer Society , 2019) (Emory Univeristy, Whiship Cancer Institute, 2020).

Figura 3

Anatomía de la mama femenina



Nota. Adaptado de *Breast Cancer Sreening Patient Version,* de National Cancer Institute*, 2020, (https://www.cancer.gov/types/breast/patient/breast-screening-pdq).*

Diagnóstico

La perseverante dedicación del científico-humanos al tratamiento de esta enfermedad, ha llevado a desarrollar constantemente equipos de carácter multidisciplinar con tecnología potencial en todo el mundo, con el objetivo de diagnosticar oportunamente más casos de cáncer.

En ciertos casos se presentan signos físicos que pueden ser posibles síntomas de cáncer de mama, sin embargo y debido a la contribución tecnológica en este campo de la medicina oncológica, se puede discriminar estos síntomas; además, la disminución de la mortalidad se debe en parte a los métodos de detección precoz y técnicas avanzadas de diagnóstico (Emory Univeristy, Whiship Cancer Institute, 2020).

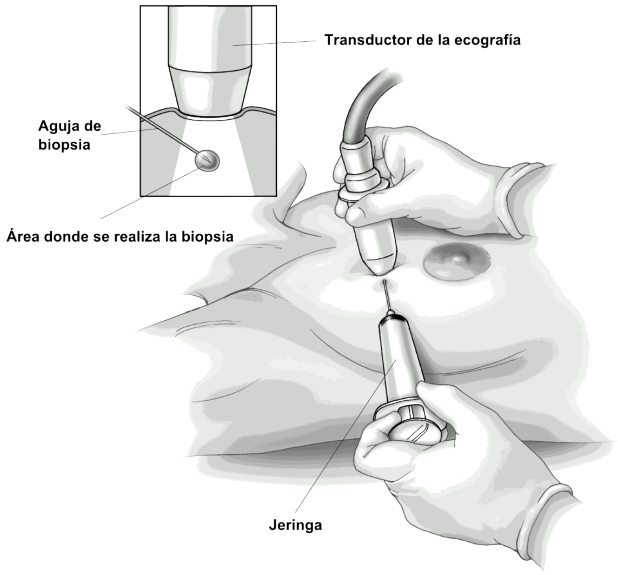
Entre las pruebas y procedimientos para detección de cáncer de mama más comunes por costo y facilidad, sin embargo, no concluyentes son los exámenes físicos que consiste en la verificación de signos generales de salud, hábitos de salud, genética familiar respecto del cáncer, así como la palpación cuidadosa de las mamas y debajo de las axilas con el objetivo de buscar bultos o protuberancias inusuales.

A continuación, se descarta o afirma mediante una mamografía (radiografía de las mamas) o ecografía para detectar presencia de masas sólidas o líquidas inusuales. Un procedimiento menos común es la resonancia magnética por su alto costo, sin embargo, la imagen muestra resultados son más contundentes y detallados.

Finalmente, está la biopsia por aspiración con aguja fina (FNA, Fine-needle Aspirate) o gruesa (CNB, Core-needle Biopsy), en este procedimiento se extrae un pequeño fragmento de tejido mamario o líquido de la región que causa sospecha guiado por un ultrasonido, y se estudia las irregularidades en las células y se conoce si son cancerosas. La Figura 4 ilustra este procedimiento.

Figura 4

Aspiración con aguja fina usando ultrasonido



Nota. Adaptado de *Fine Needle Aspiration (FNA) Biopsy of the Breast*, de American Cancer Society, 2016.

Procesamiento Digital de Imágenes

De forma general el procesamiento digital de imágenes es una herramienta versátil en el desarrollo de aplicaciones donde involucran manipulación de datos. El objetivo del PID es, a través de un conjunto de técnicas, obtener información fiable de las imágenes.

En la medicina, es un logro poseer tecnología que sea capaz de, por sí sola, emitir un diagnóstico médico a través de imágenes. Técnicas como el ajuste de intensidad y contraste, compactación de la imagen, filtros para suavizar la misma y algoritmos para la extracción de propiedades, facilitan el reconocimiento de patrones y anomalías en la anatomía humana.

Para el presente estudio, previo a la manipulación de los datos de entrada y extracción de características para *Machine Learning*, se realizó un procesamiento digital de las imágenes

capturadas de las muestras de tejido mamario sospechoso extraídas en aspiraciones con aguja fina en un portaobjetos.

Extracción de Características para Machine Learning

El sistema de diagnóstico por visión artificial extrae diez características diferentes de los límites de los núcleos celulares. Todas las características están modeladas numéricamente de tal manera que valores más grandes indicarán típicamente mayor probabilidad de malignidad (Street, Wolberg, & Mangasarian, 1993). Por lo tanto se calculan el valor medio, el peor (media de los tres valores más grandes), y el error estándar SE (*Standard error of the mean*) de cada característica resultando en un total de treinta características (Wolberg W. , Street, Heisey, & Mangasarian, 1995).

Media. Es la proporción entre la suma de todas las observaciones y número total de las mismas. En la ecuación (*1*, donde *n* es el número de las x_i observaciones.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

Error Estándar. Cuantifica las oscilaciones de la media de la muestra en torno a la media poblacional. En la ecuación (*2*, donde σ es la desviación estándar de la población, y N número de observaciones de la muestra .

$$SE = \frac{\sigma}{\sqrt{N}} \qquad (2)$$

Machine Learning

El reconocer un rostro, comprender el habla, discriminar objetos, especies y clasificarlos son actividades que conocemos bien, pero nuestra capacidad de reflexionar sobre cómo las realizamos es pobre. Hoy en día los algoritmos de Machine Learning han hecho posible

entrenar sistemas informáticos robustos para que sean más precisos y capaces de codificar muchas tareas y procesos a nivel industrial.

Machine Learning es una aplicación de inteligencia artificial que se centra en desarrollo de algoritmos de predicción partiendo de la experiencia, es decir desde los datos, sin importar la fuente de la que provienen ya que en la actualidad cubre muchas áreas de la investigación (por lo tanto el mismo algoritmo puede ser aplicado a cualquier cantidad de conjuntos de datos diferentes a pesar de que pueden tener aplicaciones prácticas completamente diferentes); estos datos podrían presentarse en forma de conjuntos digitalizados y etiquetados por el ser humano, u otro tipo de información obtenida mediante la interacción con el medio ambiente. A continuación cumplen un proceso definido donde se agrupan o dividen a través de una valoración de características similares, esto con el objetivo de generar patrones que clasifiquen datos nuevos sin esfuerzo adicional de la máquina o software (Kajaree & Rabi, 2017) (Mehryar, Afshin, & Ameet, 2018).

Sin embargo, en el aprendizaje automático es necesario una noción de complejidad de una muestra para evaluar el tamaño global de datos requeridos para que el algoritmo aprenda una familia de conceptos. Evidentemente, cuanto más grande sea la muestra o base de datos, más fácil será la tarea del algoritmo. En términos más generales, las garantías de aprendizaje teórico de un algoritmo dependen de la complejidad de las clases de conceptos consideradas y del tamaño de la muestra. (Mehryar, Afshin, & Ameet, 2018)

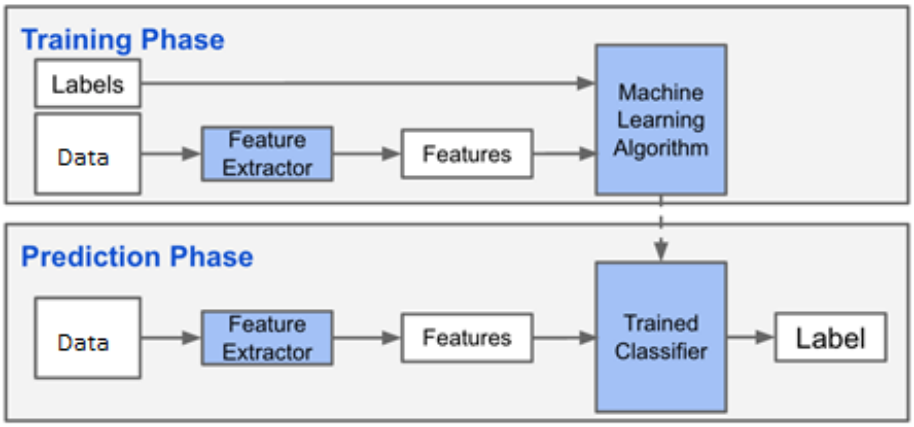
Las técnicas basadas en el aprendizaje por máquina se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por computadora, la ingeniería de naves espaciales, la ciencia, el entretenimiento y la biología computacional hasta las aplicaciones biomédicas y médicas. (El Naqa & Murphy, 2015)

Técnicas de Aprendizaje de Machine Learning

Los problemas o aplicaciones en general pueden ser clasificados en: aprendizaje supervisado, el cual involucra la construcción de un modelo predictivo que toma un conjunto conocido de datos de entrada y respuestas conocidas para estos (salidas), extrae características de dichos datos y entrena un modelo para generar predicciones razonables como respuesta a datos nuevos; este proceso se resume la Figura 5. El aprendizaje supervisado emplea técnicas de clasificación y regresión para desarrollar modelos predictivos, dependiendo del tipo de variable, es decir, continua o discreta. (Mathworks, 2019).

Figura 5

Fases de Machine Learning



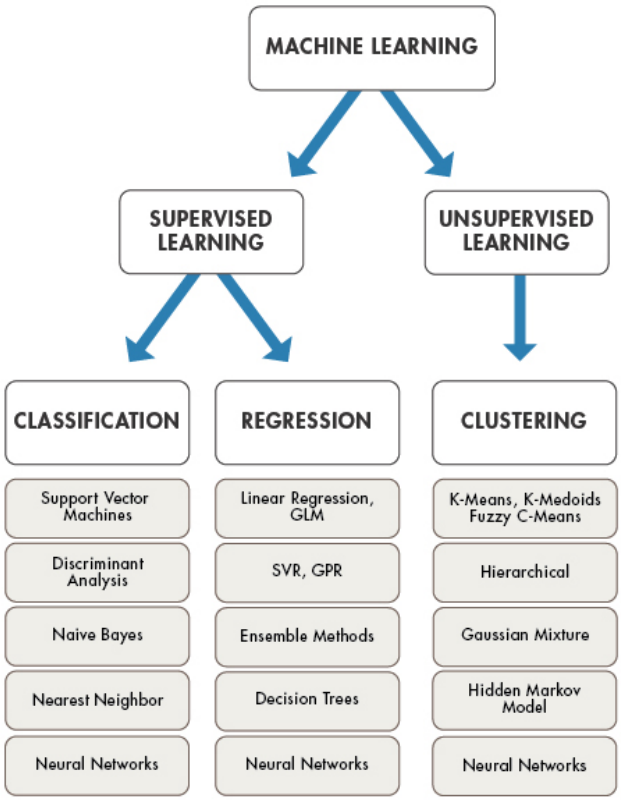
Nota. Adaptado de *A Practical Introduction to Deep Learning with Caffe and Python,* de Adil Moujahid, 2016, (http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/).

Y el aprendizaje no supervisado, involucra la búsqueda una estructura o patrón intrínseco dentro de un conjunto de datos. Se utiliza para inferir información a partir de conjuntos de datos que constan entradas no etiquetadas utilizando algoritmos de *clustering* (Mathworks, 2019).

Para implementar un algoritmo se considera varios parámetros: el tamaño y el tipo de los datos disponibles, velocidad de entrenamiento, uso de memoria, flexibilidad, interpretabilidad y principalmente la aplicación u objetivo de estudio, etc.; por lo tanto, los tipos de algoritmos que se puede aplicar son los que muestra Figura 6**.**

Figura 6

Técnicas de aprendizaje de Machine Learning



Nota. Tomado de *Machine Learning, de* Mathworks, 2019-2021.(https://la.mathworks.com/discovery/machine-learning.html)

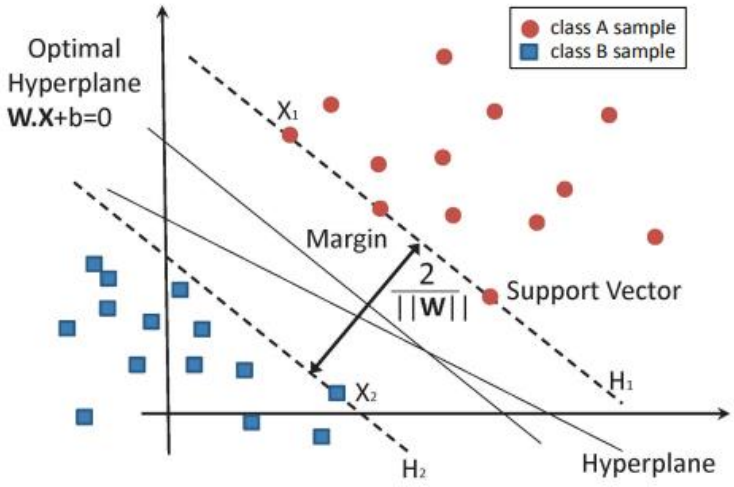
En el presente estudio, de los datos adquiridos se conoce tanto la entrada como la salida, también se sabe que únicamente existen dos posibilidades: presencia o ausencia de cáncer (clases discretas) se empleará un algoritmo de clasificación de aprendizaje supervisado.

Support Vector Machine

Las Máquinas de Vectores de Soporte (en adelante, SVM) son aplicadas tanto para problemas de clasificación como de regresión. Una particularidad es que puede resolver problemas no lineales y de alta dimensión de manera efectiva. Se desempeña como clasificador discriminatorio definido por un hiperplano de separación de los datos de entrenamiento etiquetados; el principal objetivo de SVM es la selección del hiperplano óptimo, para lo cual se aplica el concepto de margen máximo el cual se define por los vectores de cada clase más cercanos a este, estos se denominan vectores soporte (Figura 7).

Figura 7

Clasificación mediante máquinas de vectores de soporte



Nota. Adaptado de "*Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers Classification data by support vector machine (SVM)"* (p.7), por E. García, 2016, Materials.

El conjunto de datos entrenado es separado por el hiperplano $\boldsymbol{w}^T\boldsymbol{x_i} + b = 0, \in \mathbb{R}^d$, donde $\boldsymbol{w}$ es el vector de pesos, $\boldsymbol{x_i}$ vector de muestras y *b* es el sesgo. Si los datos no son

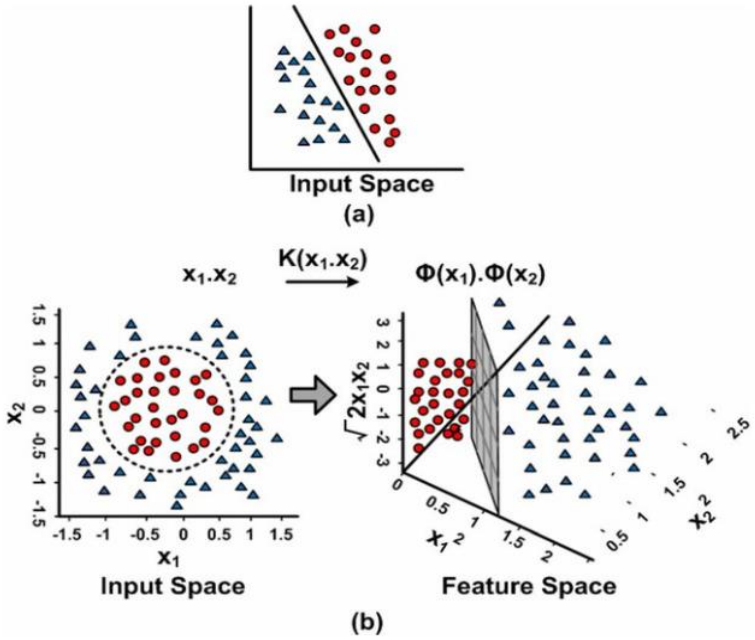
lineales, el SVM puede mapear los puntos de entrenamiento, usando la función ϕ, y a continuación aplicar la función Kernel, ecuación (*3*,

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = [\phi(\boldsymbol{x_i}) \cdot \phi(\boldsymbol{x_j})], \in \mathbb{R}^D \quad (3)$$

donde se llevan los datos a un espacio de alta dimensión y es posible la separación lineal, esto se demuestra en la Figura 8 (García, Fernández, Garcia, & Bernando, 2016).

Figura 8

Clasificador no lineal SVM



Nota. Adaptado de "*A 1.83 J/Classification, 8-Channel, Patient-Specific Epileptic Seizure Classification SoC Using a Non-Linear Support Vector Machine"* (p.51), por Muhammad B., 2015, IEEE Transactions on Biomedical Circuits and Systems, Vol. 10.

En el presente trabajo se utilizó SVM por la estructura de los datos, el desempeño del algoritmo ante una clasificación binaria, es decir únicamente existe la posibilidad de presencia o ausencia de células cancerígenas.

Redes Neuronales

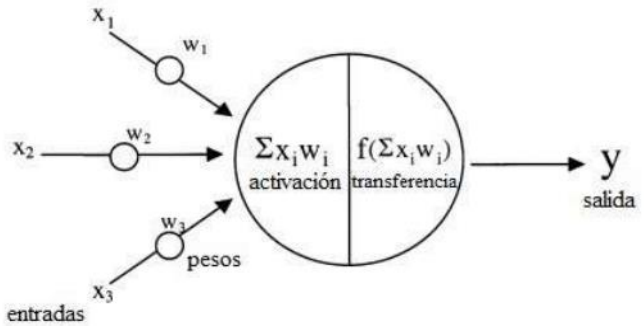
Similar a una neurona biológica, la neurona artificial tiene conexiones de entrada a través de las que reciben estímulos externos: los valores de entrada, y con estos valores la neurona realiza un cálculo interno y genera las salidas.

La neurona artificial es el componente de construcción de las Redes Neuronales Artificiales (en adelante, RNA) diseñado para simular la función de la neurona biológica. Los estímulos externos que llegan, llamados entradas, multiplicadas por los pesos de conexión (que indican la intensidad con que cada entrada afecta a la neurona) se suman y luego pasan por una función de transferencia para producir la salida de esa neurona (.

Figura **9**). La función de activación es la que distorsiona el proceso lineal interno y es igual a la suma ponderada de las entradas de la neurona y la función de transferencia, la cual puede ser sigmoidea, Tanh (tangente hiperbólica), ReLu (Agatonovic-Kustrin & Beresford, 1999).

Figura 9

Modelo de una neurona artificial



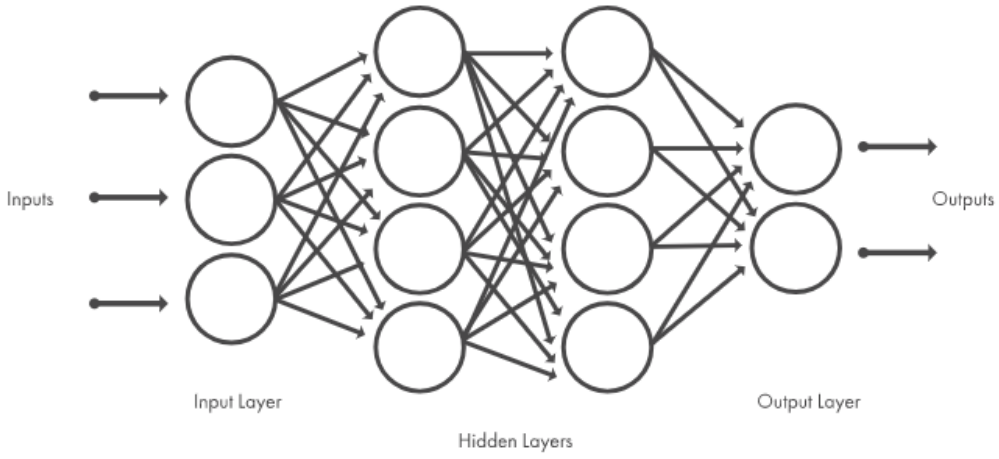
Nota. Adaptado de "*Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research"* (p.719), por S. Agatonovic-Kustrin, R. Beresford, 1999, Journal of Pharmaceutical and Biomedical Analysis 22 (2000).

Aunque una sola neurona puede realizar ciertas funciones simples de procesamiento de información, la potencia de los cálculos neuronales proviene de la conexión de las neuronas en una red, llamadas hidden units.

Arquitectura de la red neuronal. Una red neuronal combina diversas capas o *hidden layers* de procesamiento que operan en paralelo y por lo tanto el número de capas es proporcional a su robustez. La estructura de la red involucra una capa de entrada, una o varias capas ocultas y la capa de salida, interconectadas entre sí mediante nodos, o neuronas. Estos son los parámetros que definen la complejidad de la red, así como el rendimiento y tasa de aprendizaje de la misma. En la Figura 10 se representa una red neuronal (Mathworks, 1994-2021).

Figura 10

Arquitectura típica de una red neuronal



Nota. Adaptado de *Redes Neuronales*, de Mathworks, 1994, Cómo funcionan las redes neuronales (https://la.mathworks.com/discovery/neural-network.html).

Se pueden identificar dos tipos de arquitectura de una red neuronal debido a la aplicación que se busca, pueden identificarse en función de la presencia o ausencia de

conexión de retroalimentación en una red.

La arquitectura *feedforward* no tiene una conexión de retroalimentación desde la salida hacia la entrada y, por lo tanto, no mantiene un registro de los sus valores de salida anteriores. Por otra parte, la arquitectura de *feedback* tiene conexiones desde la salida hacia las neuronas de entrada.

Métricas desempeño para modelos de regresión RNA. El entrenamiento de una red neuronal consiste principalmente en minimizar una función de costo o error en la salida de la red. Esta función de costo es el MSE (Mean Square Error) o promedio de las distancias cuadradas entre el valor real y_j y el valor pronosticado $\widehat{y_j}$ de las N observaciones, expresado en la ecuación (*4*. Es un indicador de ajuste que permite estimar la calidad del modelo a través del error cuadrado promedio de las predicciones. Los valores más bajos de MSE indican un mejor ajuste.

$$MSE = \frac{1}{N}\sum_{j=1}^{n}\left(y_j - \widehat{y_j}\right)^2 \tag{4}$$

Otra métrica es el coeficiente de correlación R que determinar el ajuste del modelo, es decir, la correlación existente entre la salida del modelo y los datos reales. El rango de valores de R está comprendido entre 0 y 1 , 1 indica que el modelo se ajusta bien a los datos reales (Demuth & Beale, 2004). Su valor solo puede mejorar a medida que se agregan predictores al modelo de regresión, se calcula con la ecuación (*5*.

$$R = 1 - \frac{\sum_{j=1}^{n}\left(y_j - \widehat{y_j}\right)^2}{\sum_{j=1}^{n}\left(y_j - \overline{y_j}\right)^2} \tag{5}$$

Criterio de detección temprana. El criterio de detección temprana (ES, early stopping) permite mejorar la generalización de un modelo y reducir el sobreajuste (Echeverría, 2020).

Para aplicar este método, es necesario reservar un conjunto de datos para validación y por este motivo se dividen los datos en tres grupos:

- **Training Set.** Es el subconjunto de datos empleados en el ajuste de los parámetros de la red neuronal. Son la muestra más significativa y generalmente se seleccionan aleatoriamente.
- **Validation set**. Es el subconjunto de datos empleados después de cada iteración o época en el entrenamiento con el fin de comprobar si se genera el sobreaprendizaje.
- **Test set.** Se emplean al final del entrenamiento para validar el desempeño de la red.

El método permitirá detener el entrenamiento cuando el MSE comienza a incrementarse indicando sobreajuste y conservará los parámetros del menor valor obtenido para el subconjunto de datos de validación (Demuth & Beale, 2004).

Algoritmos de Aprendizaje. Los algoritmos de aprendizaje en una red neuronal (neural network, NN) permiten calcular los parámetros de la red; de forma general se los puede clasificar en dos grupos superiores: de primer orden y de segundo orden. Para el presente proyecto se aplica el algoritmo Levenberg Marquardt que pertenece al segundo grupo; se considera de segundo orden ya que determina su mejor desempeño en la segunda derivada del indicador de error (Echeverría, 2020), escogiendo la curva que permita llegar al punto mínimo más velozmente.

Métodos de Validación

Una alternativa para medir la capacidad de un clasificador es la matriz de confusión, permite discriminar en cada caso concreto los distintos tipos de error que pueden resultar de la aplicación de un algoritmo. La matriz de confusión (Tabla 3) muestra el número total de observaciones en cada celda, las filas corresponden a la clase verdadera y las columnas

corresponden a la clase predicha. Los resultados de la diagonal principal y fuera de la diagonal corresponden a observaciones clasificadas correcta e incorrectamente, respectivamente.

Tabla 3

Matriz de confusión

Matriz de confusión		**Estimado por el modelo**	
		Negativo	Positivo
Dato real	Negativo	VN	FN
	Positivo	FP	VP

Los elementos de la tabla anterior se definen a continuación.

Verdaderos Positivos (VP). Representa la cantidad de casos diagnosticados como positivos que fueron clasificados correctamente como positivos.

Verdadero Negativo (VN). Representa la cantidad de casos diagnosticados como negativos que fueron clasificados correctamente como negativos.

Falso Negativo (FN). Representa la cantidad de casos diagnosticados como positivos que fueron clasificados incorrectamente como negativos.

Falso Positivo (FP). Representa la cantidad de casos diagnosticados como negativos que fueron clasificados incorrectamente como positivos.

Parámetros de Rendimiento

El análisis de la matriz de confusión muestra la eficiencia del sistema de clasificación a través de los siguientes parámetros de rendimiento:

Exactitud. Es la capacidad que posee el sistema de determinar la cantidad de predicciones correctas tanto positivas como negativas, expresada en la ecuación (*6*:

$$\boldsymbol{Accuracy}(\%) = \frac{VP + VN}{VP + FP + VN + FN} \times 100 \quad (6)$$

Precisión. Es la capacidad que posee el sistema de determinar el porcentaje de casos positivos detectados, expresada en la ecuación (*7*:

$$\boldsymbol{Precision}(\%) = \frac{VP}{VP + FP} \times 100 \quad (7)$$

Sensibilidad/ Exhaustividad. Es la capacidad que posee el sistema de determinar el número de casos positivos de cáncer detectados acertadamente sobre el número de casos positivos reales (Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008), expresada en la ecuación (*8*:

$$\boldsymbol{Recall}(\%) = \frac{VP}{VP + FN} \times 100 \quad (8)$$

Especificidad. Es la capacidad que posee el sistema de determinar el número de casos negativos de cáncer detectados acertadamente sobre el número de casos negativos reales (Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008), expresada por la ecuación (*9*:

$$\boldsymbol{Specifity}(\%) = \frac{VN}{VN + FP} \times 100 \quad (9)$$

BER. Balanced error rate, o tasa de error equilibrada no es más que el promedio de los errores en cada clase y es expresada por la ecuación (*10*:

$$\boldsymbol{BER} = 1 - \frac{Recall + Specifity}{2 \times 100} \quad (10)$$

Análisis de Componentes Principales

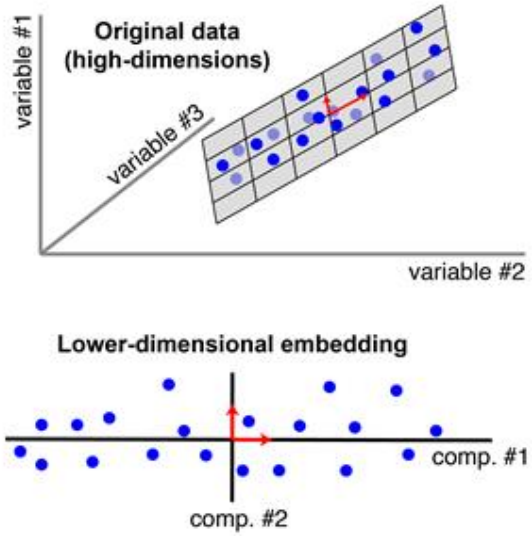
Principal Component Analysis (en adelante, PCA) es una técnica estadística no paramétrica que se utiliza principalmente para la reducción de dimensionalidad (referida a una gran cantidad de datos) en el aprendizaje automático, ya que el principal problema asociado con la alta dimensionalidad en el entrenamiento de un modelo es el sobreajuste u *overfitting* que reduce la capacidad de generalizar. Otra función que se atribuye a este método es la capacidad de filtrar conjuntos de datos ruidosos, como la compresión de imágenes (Goonewardana, 2019).

El método genera un nuevo conjunto de variables, denominadas componentes principales, donde cada una es una combinación lineal de las variables originales. Dichas componentes son ortogonales entre sí, por lo que no hay información redundante (The MathWorks, Inc., 1994). PCA intenta encontrar componentes que capturen la máxima varianza dentro de los datos y la contribución de cada variable a una cierta componente se basa en la magnitud de su varianza. Un ejemplo básico para datos tridimensionales se ilustra en la Figura 11.

Una buena práctica es normalizar los datos previamente a realizar un PCA, ya que los datos sin escala con diferentes unidades de medida pueden distorsionar la comparación relativa de la varianza entre características.

Figura 11

Reducción a dos dimensiones con PCA



Nota. Adaptado de "*Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers Classification data by support vector machine (SVM)"* (p.7), por E. García, 2016, Materials.

Capítulo III

Metodología e Implementación

Descripción general

Los modelos de clasificación de ML y Teoría de Juegos se desarrollan partiendo de un procesamiento de imágenes, las cuales son tomadas de una muestra de tejido mamario a 198 pacientes, entre ellos con sospecha y diagnóstico confirmado de cáncer de mama, se extraen las características idóneas para desarrollar, por un lado, un modelo predictivo de clasificación con Machine Learning en la herramienta *Classification Learner*, y por otro el modelo con Teoría de Juegos, y a continuación comparar los parámetros de rendimiento entre ambos métodos. Las dos técnicas se han implementado en el software MATLAB® R2020b. La eficiencia de los algoritmos radica en la dimensionalidad de los datos; por lo cual se aplican técnicas de reducción y selección de las características relevantes en el estudio.

A continuación, en la Figura 12 se presenta el diagrama de bloques del proceso para la clasificación supervisada con Machine Learning.

Figura 12

Diagrama de bloques del proceso con Machine Learning

El modelo en el cual se basa el algoritmo de la Teoría de Juegos es el mencionado Dilema del Prisionero. En este proceso primero se definen los componentes o parámetros del juego, alrededor de los cuales se establecen las estrategias de decisión y la metodología de estrategia Minimax anteriormente definida.

En la Figura 13**¡Error! No se encuentra el origen de la referencia.** se muestra el diagrama de bloques del proceso para la clasificación con Teoría de Juegos.

Figura 13

Diagrama de bloques del proceso con Teoría de Juegos

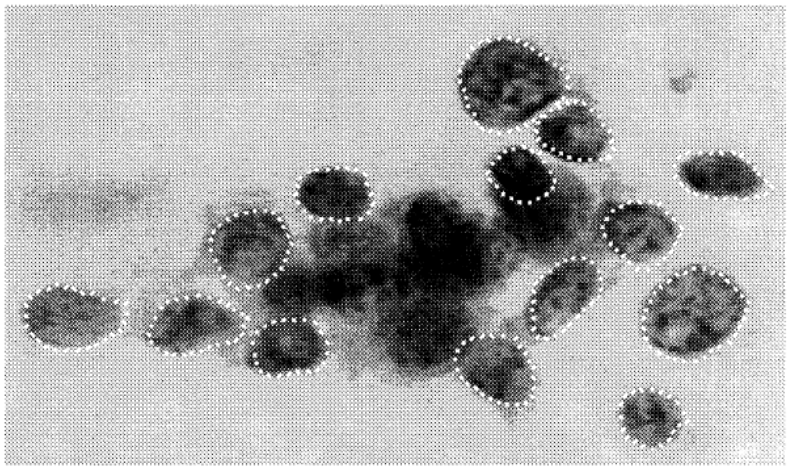
Bases de Datos

La base de datos empleada, *Wisconsin Prognostic Breast Cancer (WPBC)*, fue obtenida del repositorio de dominio público del Departamento de Ciencias Clínicas de la Universidad de Wisconsin (University of Wisconsin, 1995), en el cual participaron un conjunto de 569 que proporcionaron los datos para el estudio diagnóstico (212 muestras malignas y 357 benignas). Otro conjunto de 198 pacientes proporcionó los datos para el estudio pronóstico.

Los datos corresponden a un procesamiento de imágenes obtenidas a partir de un sistema informático que evalúa y diagnostica en base a las características citológicas derivadas directamente de una exploración digital de portaobjetos que contienen una pequeña muestra de aspirado con aguja fina (*FNA*) del tejido mamario (Figura 14) (Wolberg, Street, & Mangasarian, 1994).

Figura 14

Microfotografía de tejido mamario: límites de los núcleos celulares



Nota. Adaptado de *"Nuclear feature extraction for breast tumor diagnosis"* (p.862), por W. Street, W. Wolberg y 0. L. Mangasarian, 2015, Departments of Computer Sciences, Surgery, and Human Oncology, University of Wisconsin, Madison.

Además, el diámetro máximo del tumor y el número de ganglios linfáticos axilares afectados se consideraron en el estudio. Se administró quimioterapia en pacientes con ganglios positivos y en ambos casos (paciente con diagnóstico positivo y negativo) se realizó un seguimiento a intervalos de tres meses durante dos años para conocer si presenta recidiva, en consecuencia, la recurrencia antes de 24 meses indica diagnóstico positivo y la no recurrencia más allá de 24 meses, negativo.

Extracción de características

El área de tejido seleccionada contiene células de apariencia anormal, de manera que utilizando la técnica de ranura o *spline* en las imágenes se extrajeron diez características nucleares para cada célula. Con el procesamiento de estas imágenes se consigue un modelo de tal manera que los valores más altos son típicamente asociados con la malignidad, dichos

valores son el valor medio, el peor (media de los tres valores más grandes) y el error estándar calculado para cada imagen y característica, el resultado fue un total de 30 características. (Wolberg, Street, & Mangasarian, 1994). La descripción de las diez características se encuentra en el Anexo 1. Adicionalmente se considera el tiempo de recurrencia en meses, el diámetro del tumor y el número de ganglios positivos. En la Tabla 4 se aprecian las características involucradas en el estudio.

Tabla 4

Características para el desarrollo de los modelos predictivos

ID	Característica
Time	Tiempo
Radio	Radio
Perimeter	Perímetro
Area	Área
Compactness	Compacidad
Smoothness	Suavidad
Concavity	Concavidad
ConPoint	Puntos cóncavos
Simmetry	Simetría
FractalDimens	Dimensión fractal
Texture	Textura
TumorSize	Diámetro del tumor
LymphNode	Número de ganglios

Modelos con *Machine Learning*

Con la base de datos generada se entrenan los modelos con SVM y RNA; se importan las variables de entrada y de salida, es decir las 33 características o predictores y sus

respectivas etiquetas. Los algoritmos aprenden de los datos etiquetados y generan patrones para predecir que etiqueta o clase asignan a los datos de los nuevos pacientes.

Entrenamiento y clasificación con SVM

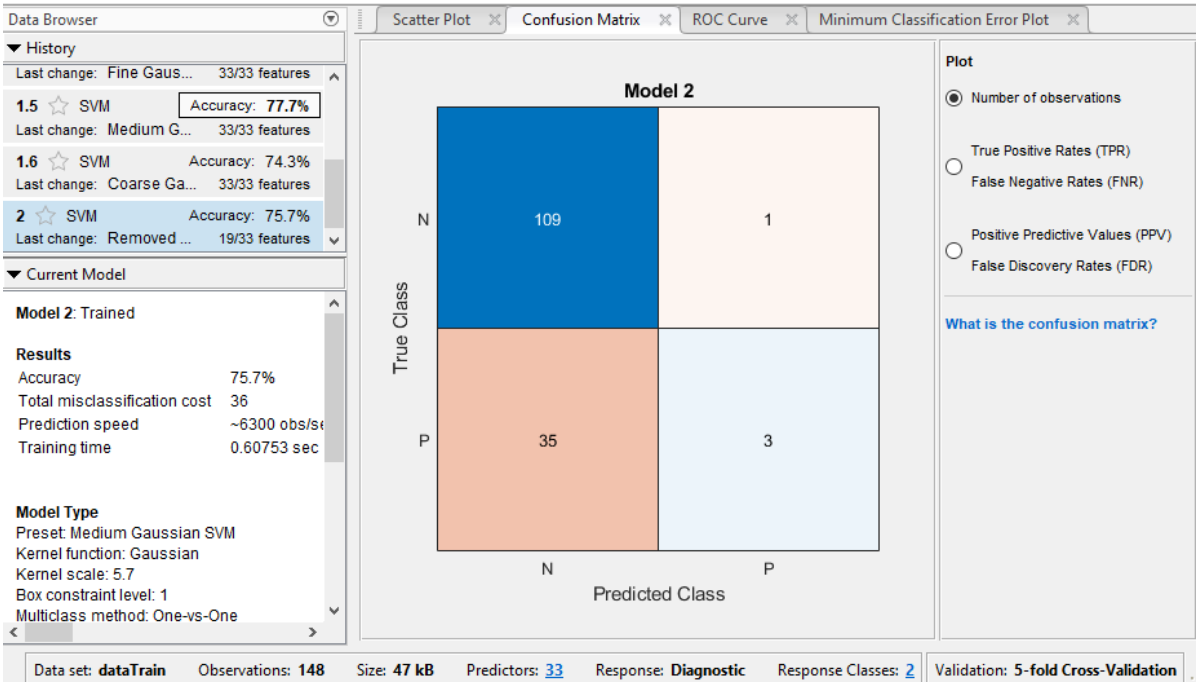
A través de la herramienta de Matlab, *Classification Learner*, se entrena un subconjunto de datos, 148 muestras que corresponde al 75% de los datos totales o *training data* previamente etiquetados con N pertenece a la clase negativo o ausencia de cáncer de mama y P o clase positivo indica presencia de cáncer.

Con el modelo generado a través de esta herramienta se evalúa la capacidad de predicción del clasificador con el subconjunto de datos complementarios, 50 muestras que corresponde al 25% de la base de datos o *data test.* Se aplica la técnica de validación cruzada como protección contra el *overfitting* mediante la partición de datos en *folds* o capas y así estimar la precisión en cada una.

La herramienta muestra los parámetros de evaluación como el porcentaje de precisión de predicción de cada modelo, tasa de error de clasificación, velocidad de precisión, tiempo de entrenamiento, y principalmente la gráfica de la matriz de confusión, como se muestra en la Figura 15, con estos valores se calculan los parámetros de rendimiento que son trascendentales en la fase de validación del modelo.

Figura 15

Parámetros de validación de los modelos de clasificación en Matlab



Se realizan experimentos con los diferentes algoritmos de SVM, dependiendo del kernel usado: lineal, cuadrático, cúbico o gaussiano, con el fin de observar cual es el hiperplano más idóneo donde se adaptaron los datos disponibles.

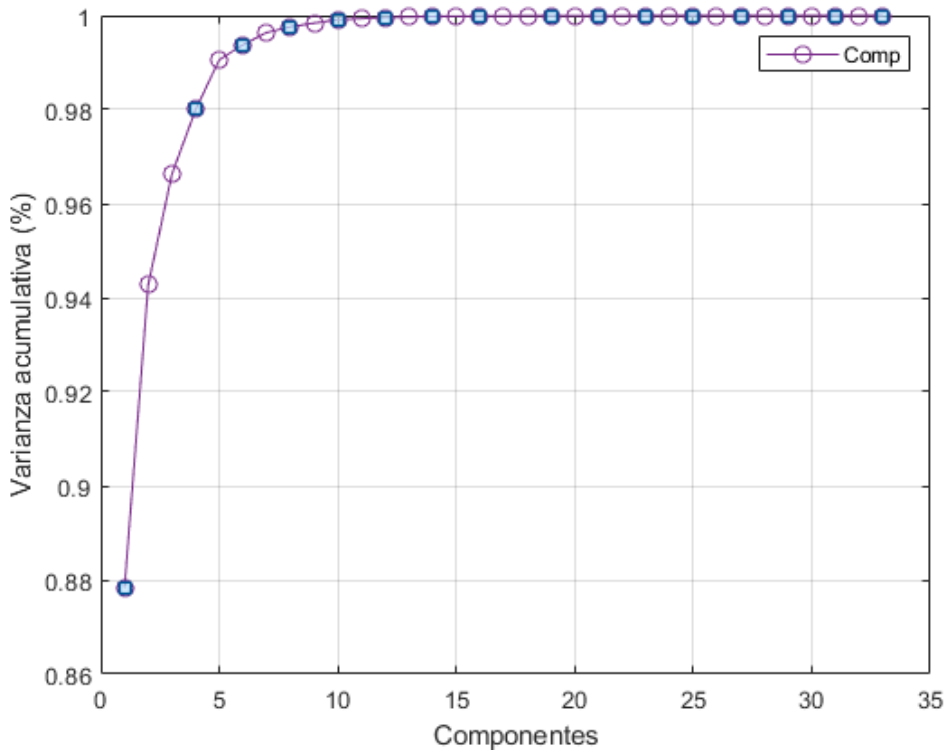
Reducción de dimensionalidad con PCA

Con el fin de optimizar recursos y obtener un modelo eficiente y de mayor precisión se implementa la técnica PCA en Matlab, se descompone la matriz de datos **D** con las características de entrenamiento en valores singulares a través del comando *svd* y se obtiene las matrices **U**, **S** y **V**; a continuación, se extrae de la matriz de covarianza la diagonal principal que contiene la varianza proyectada en las 33 dimensiones correspondientes, este proceso da como resultado los vectores propios y magnitudes de las nuevas componentes.

Para establecer un mínimo de componentes adecuadas se analiza la varianza acumulada en función de los componentes principales, Figura 16.

Figura 16

Grafica Varianza acumulada vs Componentes PCA



A través del comando *pca* de Matlab que contiene los parámetros mostrados en la Tabla 5, se selecciona aquellos predictores de mayor *score* o que más contribuyen a la clasificación.

Tabla 5

Parámetros de la función PCA de Matlab

Parámetro	Descripción
pcs	Coeficientes de las componentes principales. La contribución de cada predictor a los n componentes principales.

Parámetro	Descripción
scrs	Puntaje de las componentes principales. Las representaciones de las m observaciones en los predictores en el espacio de componentes principales.
pctExp	Porcentaje de varianza en los datos proyectados por cada componente principal.

Entrenamiento con RNA

De modo similar dentro del método de aprendizaje automático están las redes neuronales y a través de la herramienta de Matlab, *Neural Net Fitting*, se entrenan una cantidad de neuronas artificiales paulatinamente según el comportamiento de la red. El proceso consiste en: distribución de datos, selección del modelo de entrenamiento, definición de la estructura de la red, algoritmo de entrenamiento, criterio de parada y por último los indicadores de desempeño.

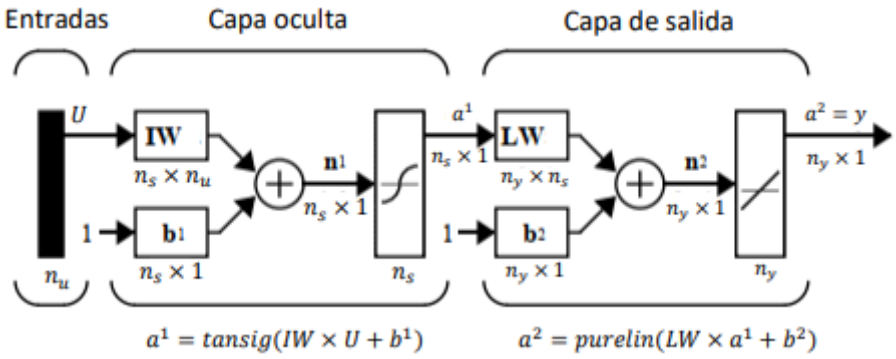
Distribución de datos. De acuerdo al criterio de detección temprana, **l**a herramienta de Matlab recomienda distribuir los subconjuntos de datos de la siguiente manera: training set que contiene los valores de las características de las células, validation set y test set, con 70% 15% y 15% respectivamente, las muestras serán seleccionadas aleatoriamente para cada subconjunto de datos; en este diseño se adoptó la recomendación del software.

Estructura de la Red Neuronal. Se implementa una red *forward feed* multicapa: la capa de salida y una capa oculta. La red,

Figura 17, se entrenará con el algoritmo de retro-propagación de Levenberg-Marquard, se utiliza este algoritmo debido a que requiere menos tiempo, aunque más memoria por esta razón es el más idóneo para los datos disponibles. El entrenamiento se detiene automáticamente (criterio de parada) cuando la generalización deja de mejorar, un indicador de esto es el aumento en el error cuadrático medio de las muestras de validación.

Figura 17

Estructura de red neuronal



Nota. Adaptado de *Aproximador General de Funciones* (p.5-7), por Demuth & Beale,2004, The MathWorks, Inc

Donde n_u es el número de entradas, n_s número de neuronas de la capa oculta, n_y el número de salidas, **U** son entradas de la red, y es la salida, **IW** es la matriz de pesos en la capa oculta, **LW** es la matriz de pesos en la capa de salida, $a^{1,2}$ las salidas de las capas (Echeverría, 2020).

Los pesos de la red están asociados a los valores de las 33 características que servirán para definir con qué intensidad cada variable de entrada afecta a la neurona. El número de capas ocultas y unidades ocultas es determinante en este modelo, ya que la precisión de la red depende de esto y por lo tanto se realizan varios experimentos y continuamente se observa el cambio en el rendimiento en función de la complejidad de la red (Gallegos & Aguirre, 2019).

Los indicadores para evaluar el sobreajuste de datos y la capacidad de generalizar de la red son: el MSE que como ya se mencionó, una vez comienza a ascender su valor, el subconjunto de validación detiene el entrenamiento, y R que es la métrica de correlación entre la salida del modelo y los datos reales (targets).

Técnicas con Teoría de Juegos

El algoritmo se desarrolla en tres fases: primero se establecen los parámetros del juego, a continuación, la dinámica en sí del juego con la combinación de estrategias y asignación de recompensas o atributos de las clases, y por último la aplicación de la metodología Minimax para determinar las clases.

Parámetros del algoritmo

Para el desarrollo de este algoritmo lo principal es determinar los parámetros del juego: los jugadores, son los 198 pacientes que participaron en la toma de las muestras para el extracción de las 33 características definidas anteriormente, estas a su vez son las estrategias que el jugador utiliza y compara con las de su oponente para maximizar su recompensa; el atributo o valor de cada una de las características tiene un peso o utilidad que mide la importancia de dicha estrategia y que le da al jugador la ventaja para ganar. Y la información, en este juego particularmente es imperfecta ya que los movimientos se dan simultáneamente y los jugadores desconocen el próximo movimiento de su contrincante.

De acuerdo a los elementos y la información disponible se estima un modelo de juego de suma cero, es decir la ganancia de un jugador supone la pérdida del otro, en otras palabras, sólo existe diagnóstico positivo o negativo como conclusión del juego.

Desarrollo y metodología del algoritmo

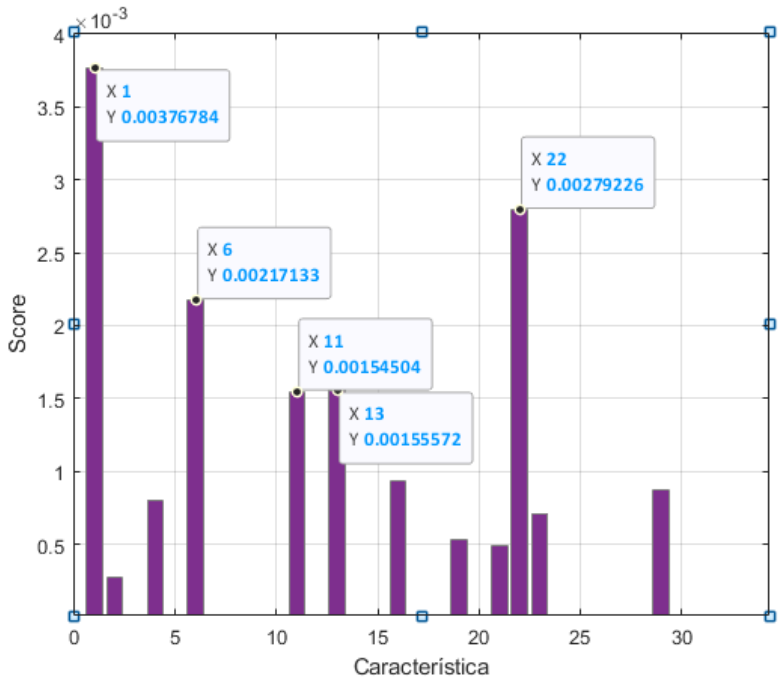
La estructura del algoritmo es alusiva al famoso ejemplo del juego “El Dilema del Prisionero” donde se comparan todas las posibles estrategias que poseen los jugadores con el objetivo de buscar cada uno la estrategia dominante, es decir la mejor opción para cada jugador independientemente de lo que el otro decida.

Para establecer un referente de los umbrales de decisión, se utilizó la función *fitctree* de Matlab que toma los 33 predictores, crea un árbol de decisión (Figura 19) y muestra los valores umbrales que determinan la clase correspondiente: P positivo (1 en el juego) o N negativo (0 en el juego), que a su vez estas son las recompensas que reciben los jugadores.

En este sentido y para el diseño de este algoritmo, al ser un modelo clasificador que busca ser eficiente en cuanto a recursos, se cuantifica la importancia de cada predictor mediante el comando *predictorImportance* (Figura 18), este suma los cambios en el riesgo de clasificación debido a las divisiones que se dan para formar las ramas de decisión, y al final divide dicha suma para el número de nodos.

Figura 18

Puntuación de importancia de los predictores



Por lo tanto en el juego participan solo las cinco características de mayor puntuación (detalladas en la Tabla 6) que en el algoritmo son las estrategias que aplican los jugadores. Se

realizan posteriores análisis aplicando más predictores para evaluar el comportamiento del algoritmo.

Tabla 6

Estrategias del algoritmo

Estrategia	Característica	Umbral de decisión
E1	Time	24
E2	MeanSmoothness	0.111
E3	MeanDimenFractal	0.0565
E4	ErrorTexture	1.44
E5	WorstRadio	24.3

El algoritmo descrito en el diagrama de flujo de la Figura 20, muestra el proceso en el cual los valores de las estrategias E1, E2, E3, E4, E5 son comparadas respecto al umbral y de acuerdo al resultado se define el pago, es decir, 1 cuando es un caso positivo de cáncer de mama y 0 caso negativo.

A continuación, se suman los pagos obtenidos por cada jugador y se aplica el método de la Estrategia Minimax, en el que los jugadores de mayor puntaje, en un acuerdo de mayor a 3 estrategias positivas, tienen como resultado un diagnóstico positivo lo cual no es necesariamente una “ganancia” para el jugador en la realidad sino más bien en el juego; esto significa que el otro jugador buscará la máxima ganancia mínima. Finalmente se comparan los resultados obtenidos en contraste con los modelos de clasificación de ML antes implementados.

Figura 19

Valores umbrales de decisión

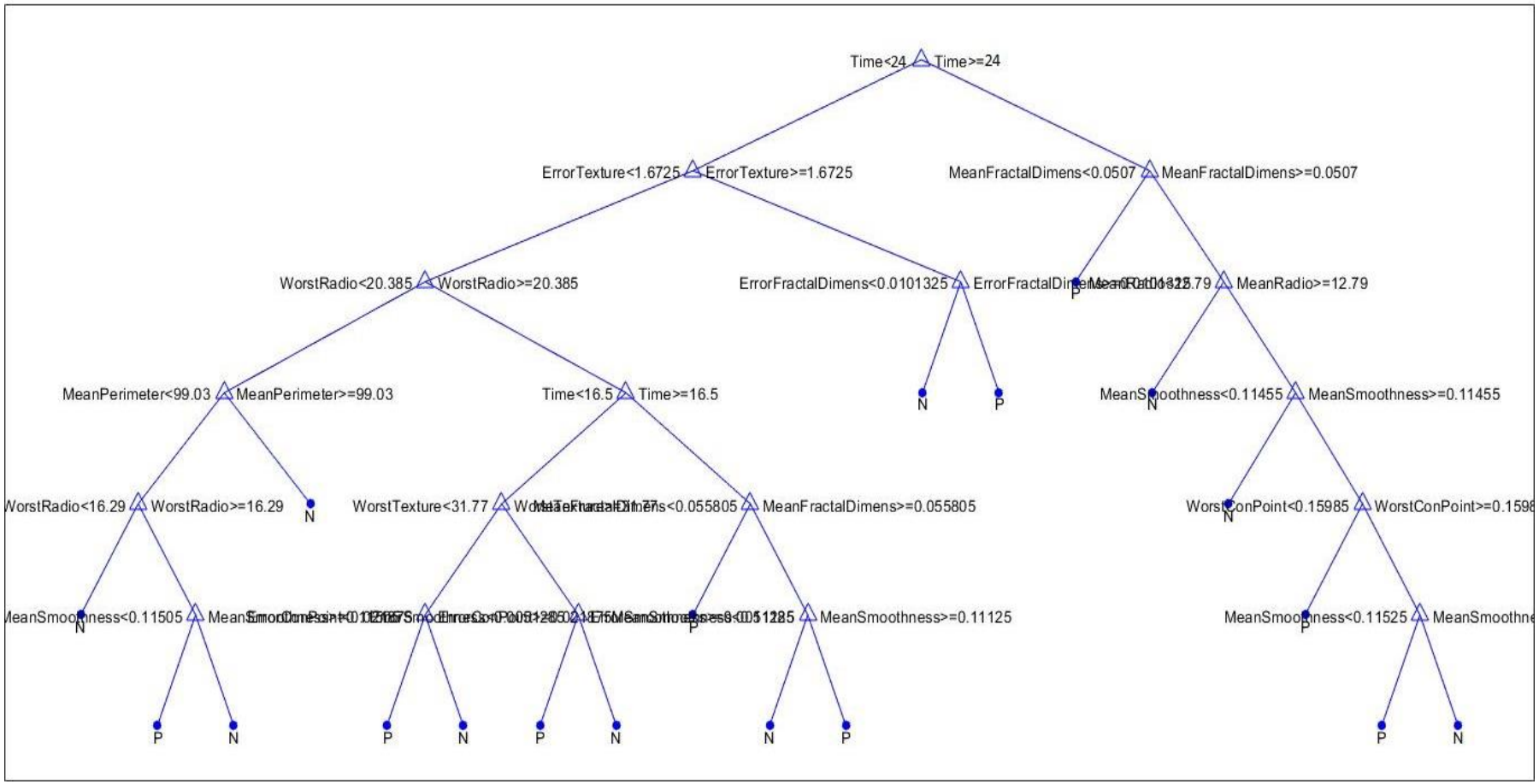
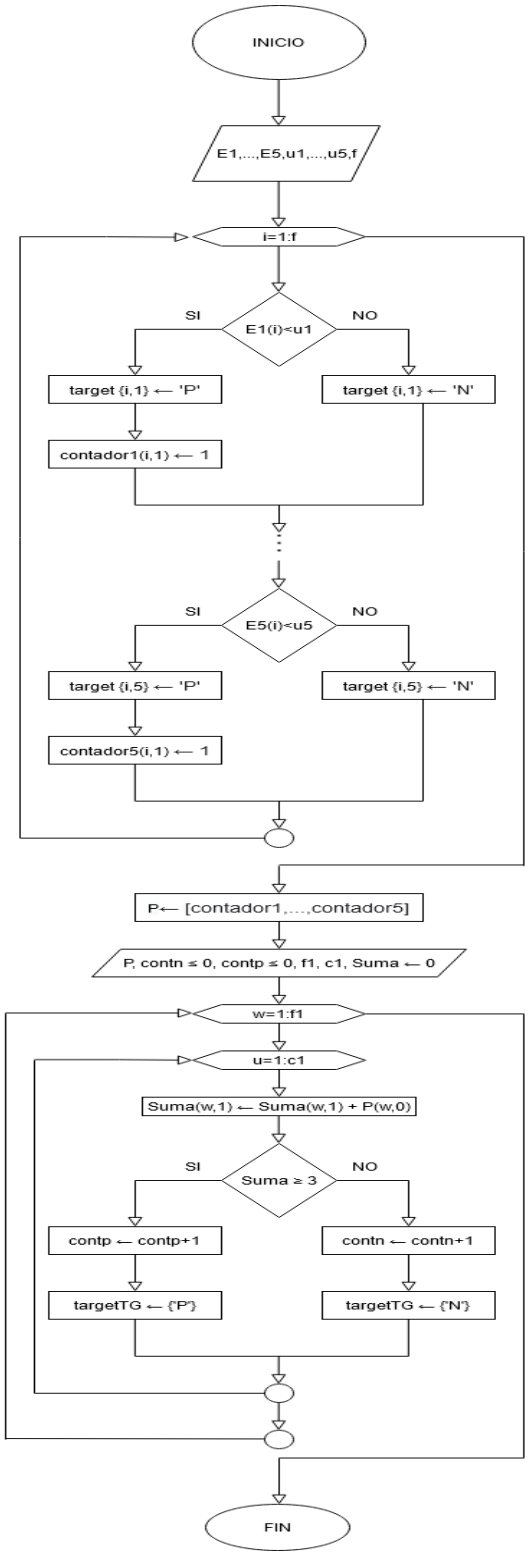


Figura 20

Diagrama de flujo del algoritmo



Capítulo IV

Análisis de Resultados

Resultados con Machine Learning

Se evaluaron los resultados por separado con los modelos de *Machine Learning*: SVM y RNA, los experimentos se realizan con un total de 198 pacientes y 33 características celulares. Los parámetros bajo los cuales se analizan los resultados óptimos de clasificación y predicción son: exactitud, precisión, sensibilidad, especificidad, BER para SVM, y las métricas de evaluación de RNA son: el MSE y el coeficiente de correlación R. Los resultados a continuación son el fundamento de evaluación para el algoritmo basado en Teoría de Juegos.

Resultados con SVM

Primer experimento. Mediante la aplicación *Classification Learner* de Matlab se entrenaron los modelos SVM con distintos kernel: lineal, cuadrático, y cúbico o gaussiano. Como se observa en la Tabla 7 el hiperplano más idóneo, para los 33 predictores, es el gaussiano con 77.70% de exactitud, 15,79% de precisión que significan 32 falsos positivos (FP), 85.71% de sensibilidad que es 1 falso negativo (FN) y 77.30% de especificidad en su predicción que corresponde 109 casos negativos clasificados correctamente (VN), en consecuencia el BER es de 0.1849, sin embargo se han seleccionado aleatoriamente 15 y luego 5 características para evaluar nuevamente las métricas.

Tabla 7

Parámetros de evaluación de los algoritmos SVM en función del número de características

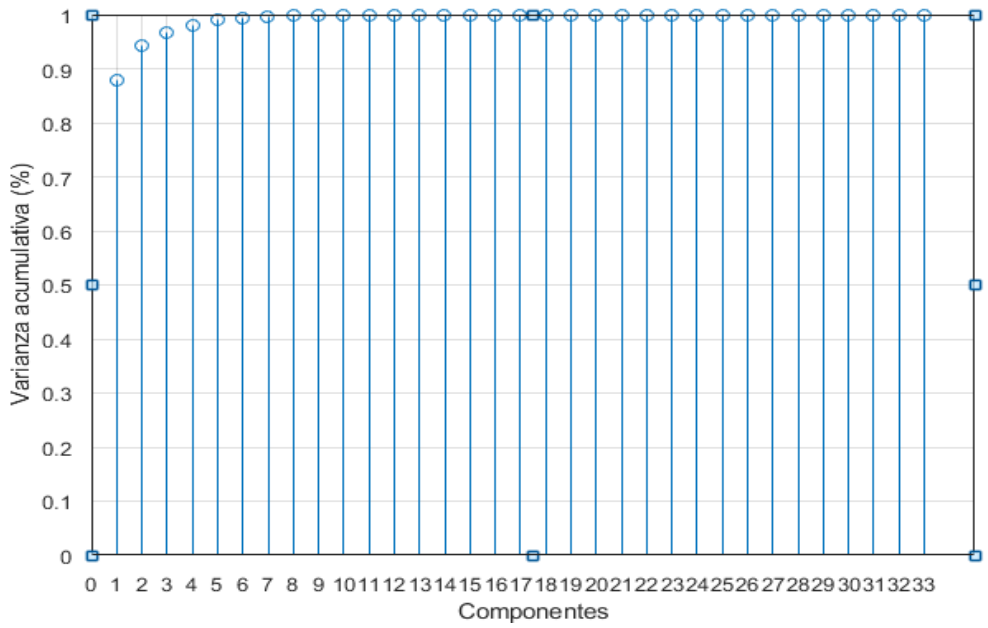
No. Predictor	Clasificador	Exactitud [%]	Precisión [%]	Sensibilidad [%]	Especificidad [%]	BER
33	Linear SVM	71.62	21.05	40.00	76.56	0.4172
	Quadratic SVM	71.62	42.11	44.44	80.36	0.3760
	Gaussian SVM	77.70	15.79	**85.71**	77.30	**0.1849**
15	Linear SVM	72.97	28.95	45.83	78.23	0.3797
	Quadratic SVM	74.32	**47.37**	50.00	**82.14**	0.3393
	Gaussian SVM	***79.05***	23.68	81.82	78.83	0.1967
5	Linear SVM	77.03	42.11	57.14	81.67	0.3060
	Quadratic SVM	75.00	44.74	51.52	81.74	0.3337
	Gaussian SVM	79.05	36.84	66.67	81.10	0.2612

Según los resultados obtenidos en este primer experimento, el modelo de clasificador óptimo es el SVM Gaussiano con 15 características, se aprecia un mejor rendimiento de clasificación debido a que la exactitud incrementa a 79.05%, la precisión es 23.68% esto significa 24 falsos positivos que es un número importante de casos erróneamente clasificados. La sensibilidad de 66.67% que son únicamente 7 falsos negativos y la especificidad de 81.10% indica que 103 casos negativos fueron clasificados correctamente, en consecuencia, el BER es de 0.1967 por la cantidad notable de falsos positivos. Se destaca el efecto de ciertas características que tienen más capacidad de generalidad y si bien se tomaron aleatoriamente, es necesario un análisis de valor de las mismas.

Segundo experimento. Del experimento anterior se comprueba ciertas características que definen mejor las etiquetas de clasificación. Para optimizar aún más el modelo SVM, en este experimento se utiliza la técnica PCA y se observa en la Figura 21 que 88.9% de la varianza acumulada se concentra en las primeras dos componentes principales. La sexta componente con el 100% de la varianza abarca prácticamente todos los casos, hasta los más particulares.

Figura 21

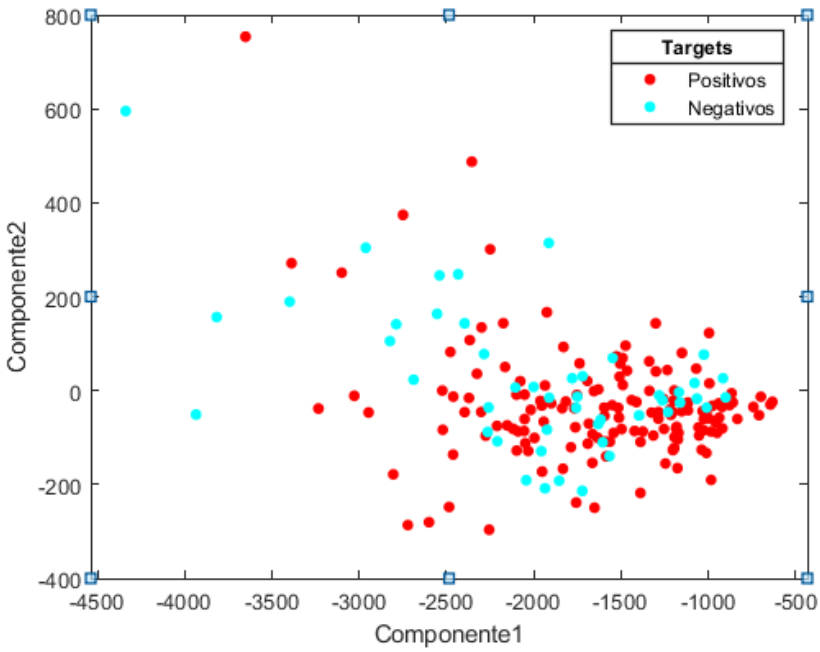
Varianza acumulada de las componentes de PCA



Efectivamente al proyectar los datos en las dos componentes principales, en la Figura ***22*** se observa como la mayor cantidad de casos se ubican en estas dos nuevas dimensiones, identificados con colores por los targets: 151 casos reales positivos y 47 negativos. Sin embargo, existen casos puntuales que precisan información adicional para ser etiquetados.

Figura 22

Proyección de los datos sobre dos componentes principales de PCA



Utilizando el comando *pca* Matlab se obtuvieron los puntajes de las componentes que contribuyen a los predictores significativamente, con los cuales se realiza un nuevo experimento. Las características relevantes respecto al *score* son las detallados a continuación en la **Tabla *8***.

Tabla 8

Puntaje de continuación de los predictores con PCA

Característica	Score	Porcentaje[%]
Time	4.6299647	92.60
ErrorTexture	1.7694958	35.39
MeanSmoothness	1.4218595	28.44
MeanDimenFractal	1.2753267	25.51
WorstRadio	1.1917568	23.84

Todos los predictores tienen un puntaje, en un rango de 1 a 5, no distante entre sí excepto el tiempo, por esto vale puntualizar que el resultado del diagnóstico depende en un

92.60% de esta característica, puesto que, a partir de las observaciones de los patrones en la distribución de los datos se identifica que la mayor parte de casos diagnosticados se basan en este atributo. Los resultados de las métricas obtenidas de la selección de estas características según el *score* aplicando PCA se muestran a continuación en la **Tabla *9***.

Tabla 9

Parámetros de evaluación de los algoritmos SVM utilizando PCA (2 componentes)

No. Predictor	Clasificador	Exactitud [%]	Precisión [%]	Sensibilidad [%]	Especificidad [%]	BER
33	Linear SVM	74.32	2.63	50.00	74.66	0.3767
	Quadratic SVM	74.32	2.63	50.00	74.66	0.3767
	Gaussian SVM	74.32	5.26	50.00	75.00	0.3750
15	Linear SVM	75.00	47.37	51.43	82.30	0.3314
	Quadratic SVM	75.68	36.84	53.85	80.33	0.3291
	Gaussian SVM	75.00	36.84	51.85	80.17	0.3399
5	Linear SVM	**81.08**	**57.89**	**64.71**	**85.96**	**0.2466**
	Quadratic SVM	79.05	47.37	62.07	83.19	0.2737
	Gaussian SVM	78.38	50.00	59.38	83.62	0.2850

Con relación a los resultados obtenidos en este experimento, se observa que al seleccionar las dos componentes principales y dentro de estas únicamente las 5 características de mayor contribución; el modelo óptimo es SVM lineal con un porcentaje de exactitud de predicción 81.08%, muestra un mejor rendimiento respecto al experimento anterior. El porcentaje de precisión es de 57.89% que corresponde a 16 falsos positivos, la sensibilidad de

64.71% con 12 falsos negativos y el BER es 0.2466 que indiscutiblemente muestra que persisten una cantidad considerable de casos negativos que son clasificados erróneamente.

Tercer experimento. En este experimento se busca obtener un mejor desempeño del modelo SVM con PCA y tres componentes o tres dimensiones para los 5 predictores mejor puntuados, a continuación las métricas se muestran en la Tabla 10.

Tabla 10

Parámetros de evaluación de los algoritmos SVM utilizando PCA (3 componentes)

No. Predictor	Clasificador	Exactitud [%]	Precisión [%]	Sensibilidad [%]	Especificidad [%]	BER
5	Linear SVM	78.38	**47.37**	60.00	83.05	0.2847
	Quadratic SVM	75.68	42.11	53.33	81.36	0.3266
	Gaussian SVM	**78.38**	42.11	**61.54**	**81.97**	**0.2825**

Como se observa al seleccionar tres componentes principales, el porcentaje de exactitud para un kernel lineal es 78.38%, considerablemente menor que en el experimento anterior, y en consecuencia el BER es mayor con 0.2825. Corroborando lo ya mencionado, donde las dos componentes principales contienen la mayor cantidad de casos influyentes dentro del estudio de clasificación.

Resultados con RNA

En la aplicación *Neural Net Fitting* de Matlab se entrenaron las neuronas artificiales con el subconjunto correspondiente de 70% de los datos, es decir 138 muestras tomadas aleatoriamente, las siguientes 30 muestras fueron utilizadas para evaluar el error cuadrático medio MSE interpretando gráficamente en que iteración o época aparece un sobreajuste. Las

30 muestras restantes fueron utilizadas para evaluar la capacidad predicción en la salida de la NN, a través del coeficiente de correlación R. De acuerdo a la teoría de redes neuronales, es trascendental el número de *hidden units* y *hidden layers*, dicho esto se realizaron varios experimentos incrementando el número de *hidden units* para determinar la versión más óptima de la NN. Los resultados de este experimento se muestran a continuación en la Tabla 11.

Tabla 11

Métricas de rendimiento de RNA

Hidden units	Iteración	MSE Validation	Coeficiente R
5	1	0.1616	0.3123
10	2	0.1384	0.3803
12	**3**	**0.1191**	**0.4277**
14	3	0.1737	0.2189
16	2	0.1904	0.2626
20	4	0.2947	0.3865

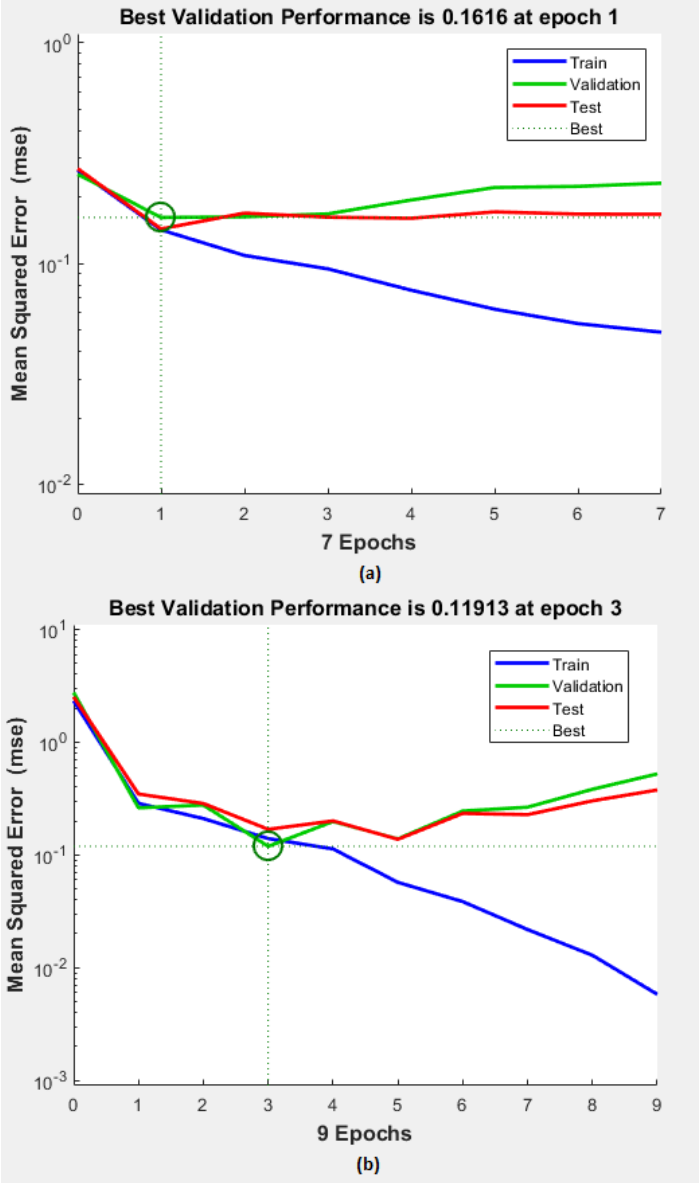
Los resultados muestran que inicialmente al entrenar la RNA con 5 neuronas, el MSE fue de 0.1616 y R de 0.3123 en la primera iteración, si bien el objetivo es mejorar el desempeño de la red, se agregaron nuevas neuronas y partir de aquí se comprueba como el a MSE se reduce a 0.1191 y se considera el valor más óptimo con 12 neuronas y un coeficiente de correlación igual a 0.4277, siendo este el valor más cercano a 1 de los experimentos realizados.

En el gráfico superior (a) de la Figura 23 se observa el desempeño de la red con 5 neuronas donde la curva en verde representa el performance del MSE que se detiene en la primera iteración cuando se presenta el sobreajuste, y por el criterio de parada o el mejor valor de MSE de entrenamiento (curva azul) se detiene en la iteración 7; por otro lado en la gráfico inferior (b) el desempeño de la red con 12 neuronas, considerado el mejor debido al MSE que

se detiene en la tercera iteración en un valor más cercano a cero (0.119 exactamente), y este sentido el performance muestra menor sobreajuste y mejor estabilidad cuando el MSE de entrenamiento (curva azul) se detiene en la iteración 9.

Figura 23

Desempeño del modelo RNA a través del MSE

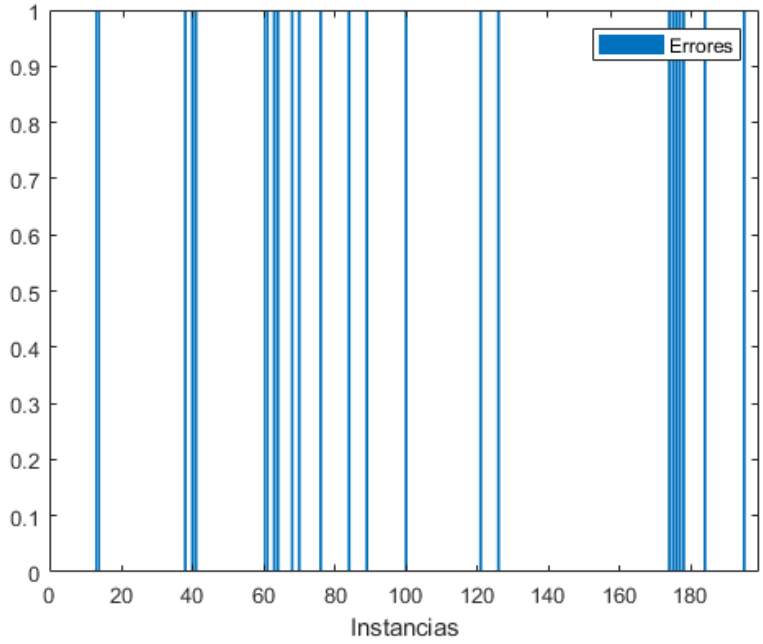


Después de este análisis se extrajo la red neuronal como una función en Matlab para estimar la capacidad de predicción y el porcentaje de diagnósticos errados. Con el comando

bar(targetsestimados-targetsreales) se graficaron y cuantificaron los falsos positivos y negativos, 25 en total, que representa un 12.63% de error en la predicción y 87.37% de precisión del modelo, esto se puede observar en la Figura 24 en el eje horizontal se encuentras las instancias o predictores.

Figura 24

Errores de predicción con RNA

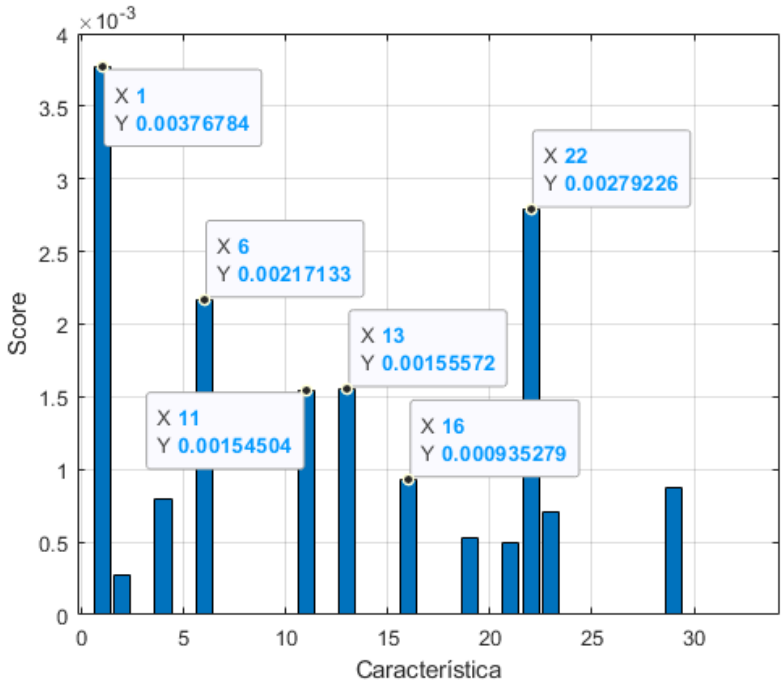


Resultados con Teoría de Juegos

El algoritmo se implementó a partir de las estrategias seleccionadas para los jugadores, el criterio de selección de las mismas está basado en el análisis de importancia de predicción realizado en Matlab con el comando *predictorImportance*. En el gráfico de barras de la Figura 25 se observan etiquetadas las cinco características de mayor puntaje.

Figura 25

Score de importancia de predicción de las características



La gráfica anterior se resumen en la Tabla 12 con: la característica que corresponde a cada estrategia, el porcentaje de importancia de cada una y el umbral de decisión obtenido del árbol de decisión implementado también en Matlab de la Figura 19.

Tabla 12

Estrategias del algoritmo y porcentaje de importancia

Estrategia	Característica	Score	Porcentaje de Importancia[%]	Umbral de decisión
E1	Time	0.00376784	94.19	24
E2	WorstRadio	0.00279226	69.80	24.3
E3	MeanSmoothness	0.00217133	54.28	0.111
E4	ErrorTexture	0.00155572	38.89	1.44
E5	MeanDimenFractal	0.00154504	38.62	0.0565
E6	ErrorSmoothness	0.00093528	23.38	0.0490

Primer experimento

A medida que los jugadores van incluyendo sus estrategias, con la convicción de ir obteniendo la mayor recompensa, se van valorando las métricas de desempeño; y en este primer experimento se observa en la Tabla 13 los valores obtenidos de la matriz de confusión del proceso de clasificación de cada una de las estrategias.

Tabla 13

Matriz de confusión por estrategias

Estrat.	VN	FN	VP	FP	Exact. [%]	Precisión [%]	Sensibili. [%]	Especific. [%]	BER
E1	113	19	28	38	**71.21**	**42.42**	**59.57**	74.83	**0.3280**
E2	125	33	14	26	70.20	35.00	29.79	**82.78**	0.4372
E3	115	34	13	36	64.65	26.53	27.66	76.16	0.4809
E4	107	39	8	44	58.08	15.38	17.02	70.86	0.5606
E5	118	34	13	33	66.16	28.26	27.66	78.15	0.4710

Los resultados indican un gran porcentaje de falsos positivos en todas las estrategias, esto se produce debido a que los valores de los predictores están muy dispersos respecto al diagnóstico, y en efecto la característica con más aptitud de clasificación es el tiempo o estrategia E1 con 71.21% de capacidad de clasificación, y el menor BER de 0.3280.

La combinación de estas 5 estrategias define la recompensa final, que su vez es la que clasifica al jugador o paciente y da el diagnóstico. Es decir, se suman las recompensas obtenidas (1 o 0, de acuerdo al umbral de decisión) de manera que si obtienen un resultado mayor a 3 se trata de un caso positivo de cáncer, este criterio está basado en la estrategia Minimax. Las métricas de este experimento se muestran en la

Tabla **14**.

Tabla 14

Parámetros de rendimiento para un algoritmo de 5 estrategias

Estrategias	Exactitud [%]	Precisión [%]	Sensibilidad [%]	Especificidad [%]	BER
5	73.23	43.18	40.43	83.44	0.3807

Los resultados obtenidos fueron un total de 145 casos correctamente clasificados de 198, en consecuencia 28 casos positivos fueron diagnosticados como negativos (falso negativo) y 25 casos negativos fueron diagnosticados como positivos (falso positivo). En definitiva, esto representa un 73.23% de exactitud de clasificación y un BER de 0.3807, que en comparación con los modelos ML se considera apto ya que se puede identificar puntualmente cada caso mal clasificado.

Segundo experimento

Así como en los modelos con ML, cada predictor tiene un valor de importancia dentro del algoritmo propuesto. En la teoría de juegos estos son los pesos que se asignan a cada estrategia con el objetivo de mejorar la ganancia de cada jugador. Dichos pesos corresponden al porcentaje de importancia acumulado dividido entre las estrategias aplicadas (Tabla 15) estos se multiplican por las recompensas obtenidas y se evalúa nuevamente el algoritmo aplicando el método Minimax.

Tabla 15

Peso de las estrategias

Estrategia	Característica	Porcentaje de Importancia[%]	Peso
E1	Time	94.19	31.84
E2	WorstRadio	69.80	23.60

Estrategia	Característica	Porcentaje de Importancia[%]	Peso
E3	MeanSmoothness	54.28	18.35
E4	ErrorTexture	38.89	13.15
E5	MeanDimenFractal	38.62	13.06
			$\sum 100$

A continuación en la Tabla 16 se observan las métricas obtenidas de este experimento. Los resultados obtenidos muestran que la capacidad predictiva del algoritmo incrementa a 76.26% debido a que las estrategias E1 y E2 tienen un 31.48% y 23.60% respectivamente del peso total. La precisión es de 50%, sensibilidad 38.30%, especificidad 88.80% y el BER se reduce a 0.3681.

Tabla 16

Parámetros de rendimiento con asignación de pesos

Estrategias	Exactitud [%]	Precisión [%]	Sensibilidad [%]	Especificidad [%]	BER
5	76.26	50.00	38.30	88.08	0.3681

Tercer experimento

En este experimento se incluye una estrategia adicional: ErrorSmoothnes, que es la siguiente característica de mayor aporte con 23.38% de importancia (Tabla 12), con el objetivo que comprobar si existe un cambio favorable en la capacidad predictiva del algoritmo y cómo influye esta nueva estrategia. Los resultados se muestran en la Tabla 17.

Tabla 17

Parámetros de rendimiento para un algoritmo de 6 estrategias

Estrategias	Exactitud [%]	Precisión [%]	Sensibilidad [%]	Especificidad [%]	BER
6	76.26	50.00	38.30	88.08	0.3681

En contraste con el experimento anterior, esta última estrategia incluida no aporta utilidad de los jugadores, pero mantiene el desempeño del algoritmo en 76.26% en su capacidad de predicción. Este resultado trasmite que las estrategias anteriores definen propiamente la clasificación.

Análisis comparativo de desempeño entre los modelos

Dentro de los experimentos antes realizados y los resultados obtenidos en este capítulo, se puede apreciar que tanto ML y teoría de juegos tiene un desempeño admisible para clasificar los datos propuestos. Sin embargo, es importante considerar ciertos parámetros como: la complejidad del modelo, manejo y control de los datos, parámetros individuales de los modelos, técnicas de optimización, etc.

Tabla 18

Comparación de desempeño clasificadores ML y algoritmo de teoría de juegos

Clasificador	Exactitud [%]	BER
Gaussian SVM	79.05	0.1967
QSVM-PCA	81.08	0.2466
RNA	87.37	0.1263
Teoría de Juegos	76.26	0.3681

Como se evidencia en la **Tabla 1**Tabla 18, y tras el análisis individual de los experimentos, el modelo de ML: RNA con 87.37% de exactitud de predicción, es el modelo de mayor desempeño tomando en consideración la arquitectura del mismo. La capacidad predictiva del algoritmo de teoría de juegos no está lejana a un modelo ML sin el acondicionamiento de las metodologías de redimensionamiento y optimización (modelo SVM

con kernel gaussiano con 79.05% de exactitud), ya que con un 76.26% de exactitud se considera apto como clasificador respecto a los modelos estudiados.

Respecto al BER, claramente el algoritmo de teoría de juegos aun con una tasa de 0.3681 se considera aceptable al ser un prototipo de clasificador para una base de datos compleja, es decir, los datos extremadamente dispersos no se adecuaron al algoritmo ya que los umbrales de decisión son clave en este proceso. Otras metodologías de clusterización como hiperplanos, folds y árboles de decisión, mejoran el desempeño del algoritmo sin embargo el juego es quien decide cual estrategia es penalizada o cual obtiene más utilidad.

Capítulo V

Conclusiones y Recomendaciones

En este primer planteamiento se investigó una solución que sea parsimoniosa e interpretable, a pesar de no haber alcanzado porcentajes comparables con los métodos de ML.

Se logró una nueva estrategia para resolver un problema de clasificación optimizando recursos y con un enfoque real respecto al comportamiento de las partes en una competencia.

La principal ventaja del algoritmo planteado, es el control absoluto de su desempeño ya que se puede mejorar la etapa de clasificación estableciendo los mejores umbrales de decisión.

Los resultados obtenidos de los modelos ML implementados superan en cuanto a la capacidad de predicción (SVM y RNA con 79.05% y 87.37% de capacidad predictiva respectivamente) al algoritmo de teoría de juegos, sin embargo, se considera un clasificador apto frente a los antes mencionados con 76.26% de exactitud en su capacidad de predicción.

El algoritmo de teoría de juegos aun con una tasa de 0.3840 se consideró aceptable al ser un prototipo de clasificador para una base de datos compleja, es decir, los datos extremadamente dispersos no se adecuaron al algoritmo ya que los umbrales de decisión son clave en este proceso.

Las técnicas de reducción de dimensionalidad PCA y selección de características para abstraer los patrones idóneos de clasificación y optimizar los modelos de entrenamiento mejoraron el desempeño del modelo SVM de a 79.05% de exactitud de predicción 81.08%.

Se destacó que el resultado del diagnóstico depende en un gran porcentaje del predictor tiempo de recurrencia, puesto que a partir de las observaciones de los patrones en la distribución de los datos y el entrenamiento (con esta característica) de los modelos

planteados, se identificó que la mayor parte de casos diagnosticados se basan en este atributo, cuyo umbral de decisión es 24 meses de recurrencia de la enfermedad.

Basado en el estudio se recomienda realizar una investigación de metodologías que definan mejor los umbrales de decisión o a su vez se pueden incluir otros métodos como la clusterizacion o detección de patrones, que logren métricas de desempeño mejores.

Capítulo VI

Trabajos Futuros

Como trabajos futuros se propone mejorar el desempeño del algoritmo, investigando propuestas de optimización u otra forma de estructuras el modelo de juego; procurando que se adapte a las mismas reglas de la teoría de juegos.

Diseñar una interfaz gráfica capaz de resolver procesos de clasificación, con la opción de ingresar bases de datos de distintas dimensiones para cualquier aplicación.

Discriminar médicamente con el apoyo de un especialista que tan factible es el prototipo planteado.

Bibliografía

The MathWorks, Inc. (1994). *Análisis de componentes principales (PCA)*. Obtenido de https://la.mathworks.com/help/stats/principal-component-analysis-pca.html

Agatonovic-Kustrin, S., & Beresford, R. (1999). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 717-719.

Al-Raweshidy, D. O. (2010). Theory pf Games: An Introduction. *Brunel University-West London, UK*, 1-2.

American Cancer Society . (2019, September 20). *Types of Breast Cancer*. Obtenido de https://www.cancer.org/content/dam/CRC/PDF/Public/8580.00.pdf

American Cancer Society. (2016). Cancer Facts & Figures 2016. Obtenido de http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf

Archetti, M., & Pienta, K. (2018). Cooperation among cancer cells: applying game theory to cancer. *Nature Reviews Cancer*.

Cashin-Garbutt, A. (2014, Octubre 29). *Teoría del juego y el ecosistema del cáncer: una entrevista con profesor Pienta, Johns Hopkins*. Obtenido de https://www.news-medical.net/news/20141029/6/Spanish.aspx

Cremades, Á. (2016). Teoría de Juegos y Análisis estratégico: Una revisión metodológica en torno a la toma de desiciones y el conflicto internacional. *Instituto Españo de Estudios Estratégicos*.

Demuth, H., & Beale, M. (2004). Neural Network Toolbox User's Guide. *Natick, MA: The MathWorks, Inc.*

División de Prevención y Control del Cáncer, Centros para el Control y la Prevención de Enfermedades. (2018, Septiembre 11). *Breast Cancer*. Obtenido de https://www.cdc.gov/spanish/cancer/breast/basic_info/what-is-breast-cancer.htm

Doufene, A., & Krob, D. (2015). Pareto Optimality and Nash Equilibrium for Building Stable Systems. *Anual IEEE System Conference (SysCon) Proceedings*, 2-3.

Echeverría, G. (2020). Modelos híbridos basados en datos para la predicción de carga eléctrica a corto y mediano plazo.

El Naqa, I., & Murphy, M. (2015). What Is Machine Learning? *Machine Learning in Radiation Oncology*, 3-11. Obtenido de https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1

Emory Univeristy, Whiship Cancer Institute. (2020). *Cancer Quest*. Obtenido de Breast Cancer: https://www.cancerquest.org/es/para-los-pacientes/cancer-por-tipo/cancer-de-mama#toc-deteccin-y-Vq8Y7F7T

Fear, E., Meaney, P., & Stuchly, M. (2003). Microwaves for breast cancer detecction? *IEEE Potentials*, 12-18.

Friedl, M., & Brodley, C. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environmet* , 399-409.

Gallegos, S., & Aguirre, H. (2019, junio). *Desarrollo y análisis de sistemas de estimación y detección de objetos de radar mediante algoritmos de Machine Learnign y Deep Learning".* Obtenido de Repositorio ESPE: https://repositorio.espe.edu.ec/bitstream/21000/20499/1/T-ESPE-039346.pdf

García, E., Fernández, Z., Garcia, P., & Bernando, A. (2016). Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers. *Materials* , 6-8.

Goldenberg, L., Nir, G., & Salcudean, S. (2019). A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, 391–403.

Goonewardana, H. (2019, Feb 8). *Apprentice Journal*. Obtenido de PCA: Application in Machine Learning: https://medium.com/apprentice-journal/pca-application-in-machine-learning

Herrero, M., & Pinedo del Campo, J. (2005). Pensamiento Estratégico, Teoría de Juegos y Comportamiento Humano. *Indivisa. Boletin de Estudios e Investigación.*, 37-67.

Kajaree, D., & Rabi, N. (2017). A Survey on Machine Learning: Concept, Algorithms and Applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 1-2.

Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and predicttion. *Computational and Structural Biotechnology Journal*, 8-17.

Kubilay, I., & Anderson, P. (2010). *Applied Game Theory and Strategic Behavior.* NW: Taylor and Francis Group, LLC.

Mathworks. (1994-2021). *Redes Neuronales*. Obtenido de https://la.mathworks.com/discovery/neural-network.html

Mathworks. (2019). *Machine Learning with MATLAB*. Obtenido de https://la.mathworks.com/help/stats/machine-learning-in-matlab.html?s_tid=srchtitle

Mehryar, M., Afshin, R., & Ameet, T. (2018). *Foundation of Machine Learning.* Cambridge, Massachusetts: The MIT Press.

Ministerio de Salud Pública . (2020). *Cifras de Ecuador – Cáncer de Mama*. Obtenido de https://www.salud.gob.ec/cifras-de-ecuador-cancer-de-mama

Monte, R. (1995). Historia del Diagnostico por la imagen de la mama. *Medicina Balear*, 155-158.

National Cancer Institute. (2019, Octubre 11). *PDQ Breast Cancer Screening*. Obtenido de https://www.cancer.gov/types/breast/patient/breast-screening-pdq

Organización Mundial de la Salud. (2018, Septiembre 12). *Cáncer: Datos y Cifras*. Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/cancer

Parikh, R., Mathai, A., Parikh, S., Sekhar, C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 45.

Real, S. (2019, Abril 22). Clasificación de lesiones en mamografías mediante red neuronal. *Universidad San Francisco de Quito*. Quito.

Roffo, G. (2018, Agosto). *Feature Selection Library*. Obtenido de Mtlab Toolbox .

Saad, W., Han, Z., Debbah, M., Hjorungnes, A., & Basar, T. (2009). Coalitional Game Theory for Communication Networks. *IEEE SIGNAL PROCESSING MAGAZINE*, 77-78.

Street, N., Wolberg, W., & Mangasarian, O. (1993). Nuclear feature extraction for breast tumor diagnosis. *Departments of Computer Sciences, Surgery, and Human Oncology. University of Wisconsin, Madison*, 861-863.

Trestian, R., Ormond, O., & Muntean, G.-M. (2012). Game Theory — Based Network Selection:Solutions and Challenges. *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, 1-2. Obtenido de https://ieeexplore.ieee.org/abstract/document/6144681

Universidad Rey Juan Carlos. (2015). Introducción a la Teoría de los Juegos: Equilibrio de Nash. Madrid, España. Obtenido de https://online.urjc.es/es/quienes-somos/106-espanol/mooc/cursos-mooc/421-teoria-de-juegos

University of Wisconsin. (1995, Diciembre). *Wisconsin Prognostic Breast Cancer (WPBC).* Obtenido de https://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WPBC/

WHO. (2008). *Global Burden of Disease 2004*.

Wolberg, W., Street, N., & Mangasarian, O. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 163-171.

Wolberg, W., Street, N., Heisey, D., & Mangasarian, O. (1995). Computerized Breast Cancer Diagnosis and Prognosis From Fine-Needle Aspirates. *Archives of Surgery, University of Wisconsin*.

Anexos

Anexo A: Características celulares