Comparación de modelos de aprendizaje automático para la predicción de células cancerígenas a partir del complejo MHC I

Mateo Navas Luquez

Trabajo Dirigido

Tutor
Alvaro David Orjuela Cañón (PhD)

UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ D.C
2020

AGRADECIMIENTOS

En primer lugar, deseo expresar mi agradecimiento a Dios, al director de esta tesis, al Dr. Alvaro David Orjuela Cañón, por la dedicación y apoyo que ha brindado a este trabajo, por el respeto a mis sugerencias e ideas, la dirección, el rigor que ha facilitado a las mismas y la confianza ofrecida. Un trabajo de investigación es siempre fruto de ideas, proyectos y esfuerzos previos que corresponden a otras personas, que nos impulsan, sin las cuales no tendríamos la fuerza y energía que nos anima a crecer como personas y como profesionales. Por eso le doy gracias a mi madre, mis padrinos Julia y Tomas por su apoyo, económico y emocional, fundamental para el logro de mis objetivos, a mi familia, abuelos, tíos, primos, que siempre me han prestado un gran apoyo moral y humano, necesarios en los momentos difíciles, a mi abuela quien me apoya con sus oraciones. A todos, muchas gracias.

Contenido

TABLA DE FIGURAS

1. INTRODUCCIÓN

En la actualidad el cáncer es la segunda causa de muerte en el mundo, en 2015 la OMS registró que una de cada seis muertes a nivel global estuvo provocada por esta enfermedad, mientras que en 2018 provocó más de 9.6 millones de muertes a nivel global [1]. Según la sociedad americana de oncología clínica uno de los factores más importantes para la supervivencia de la población con esta condición es: la detección y tratamiento en una fase temprana de enfermedad [2], y acorde a las estadísticas mundiales solo el 30 % de los países de bajos ingresos cuenta con un sistema de salud que soporte tratamientos oncológicos, mientras que en los países de mediano y alto ingreso asciende esta cifra hasta el 65 y 90 % respectivamente [1].

Según el DANE en Colombia el 8% de las muertes de origen natural son ocasionadas por cáncer de mama, colon y estómago, mientras que otros tipos de cáncer, localizados en áreas diferentes, abordan solo el 2 % de las muertes naturales en el país [3]. De acuerdo con el reporte de defunciones del último trimestre del año 2019, se presentaron aproximadamente de 3000 casos relacionados con cáncer [3].

El ministerio de salud y protección social ha promovido en la última década planes de prevención y tratamiento para esta enfermedad, en consecuencia, al incremento en las tasas de cáncer poblacional en las últimas dos décadas, convertido así en un problema de salud pública con un crecimiento alarmante [4]. Según los datos del ministerio de salud, entre el 2007 y el 2013 las enfermedades relacionadas con cáncer alcanzaron un pico histórico en el número de defunciones por año, llegando a los 30.000 habitantes, con un promedio de 96 personas fallecidas al día [4] [5]. Para el sistema de salud colombiano el costo de tratamientos para el cáncer asciende a más de 415 mil millones cada año, mientras que la inversión en políticas de prevención y detección ronda los 9.000 millones [3]. A nivel operacional es mucho más rentable solventar tratamientos oncológicos para tratar las primeras etapas de la enfermedad. En 2016 el instituto colombiano de cancerología realizó una comparación del gasto operativo que conlleva tomar un tratamiento oncológico y su proceso de seguimiento en un lapso de 5 años. Los resultados arrojaron que a la fecha un tratamiento de estas características con un cáncer de primera etapa alcanza un estimado promedio de 30.1 millones de pesos, donde el mayor rubro se va en radioterapia, en cambio en tratamiento de la misma duración, pero con un cáncer con metástasis puede ascender a 350 millones de pesos, donde las cirugías y la quimioterapia son los gastos más grandes [6].

Basado en la necesidad de detectar el cáncer en la etapa más temprana posible, no solo para crear tratamientos más efectivos, sino que también para alivianar el peso económico que acarrean los procedimientos más agresivos. Debido a esto el uso de aplicaciones bioinformáticas se han vuelto herramientas importantes para el diagnóstico de este tipo de patologías. Estas herramientas basan su análisis en un biomarcador o en un conjunto de biomarcadores para predecir un posible diagnóstico. La técnica que se usa para cada problema en bioinformática depende en sí de la aplicación. Existen procedimientos que utilizan la espectrometría de masas, la secuenciación de proteínas o ADN o imágenes radiológicas para desarrollar una predicción del diagnóstico [7] [8].

Para poder desarrollar aplicativos bioinformáticos que ayuden a la detección de células cancerígenas es necesario conocer cómo es el comportamiento del sistema, proteínas y de

más componentes biológicos que se pretenden analizar en un entorno computacional. Por lo tanto, para tener un precedente de la actividad cancerígena a nivel celular es necesario conocer en el metabolismo celular, y entender los procesos inmunitarios que intervienen en el perfil de esta patología.

1.1 Marco Teórico

1.1.1 El metabolismo celular

La célula es una de las unidades funcionales más pequeñas en el cuerpo humano, dichas estructuras poseen una numerosa cantidad de atributos que les permiten adaptarse al ecosistema, por medio de la agrupación de diversas unidades celulares y modificaciones de sus estructuras nativas [9].

Cuando se explora la complejidad del cuerpo humano se encuentra que las unidades microescala, como la célula, pueden formar los modelos y sistemas complejos que existen en la naturaleza. Una única unidad funcional como lo es la célula no está capacitada para realizar todas las funciones que soporten la vida, pero al agrupar cantidades significativas de estas se obtienen tejidos especializados, los cuales pueden empezar a realizar funciones metabólicas más complejas. En una escala un poco más alta en esta jerarquía se encuentran los órganos, los cuales son estructuras compuestas por diferentes tejidos, permitiéndole a estas unidades especializarse en una enorme cantidad de procesos. Una vez que se interconectan diferentes órganos se obtienen los diferentes sistemas biológicos, los cuales se pueden interpretar como una vasta red de funciones vitales. Al juntar todos los niveles de estructuras anteriormente mencionados se forman seres vivos, organismos estructuralmente complejos como lo es el cuerpo humano [9].

Para el correcto funcionamiento de diversos sistemas biológicos, estos necesitan reemplazar un cierto número de células en intervalos finitos de tiempo, con el fin de preservar la integridad de sus tejidos y mantener un nivel de homeóstasis biológica (Cooper, 2014). Para producir las nuevas unidades celulares que el cuerpo necesita, el organismo recurre a 3 mecanismos biológicos, conocidos como: división, diferenciación y muerte celular. [10].

La división o ciclo celular es la etapa en donde una célula madre o progenitora disminuye su tasa de actividad extracelular y concentra toda su energía para promover de manera controlada las fases: G1, S, G2 y M del ciclo celular como se observa en la Figura 1. La fase G1 es una etapa de crecimiento, en esta se sobreexpresa las proteínas y el citoplasma celular y la célula duplica su volumen. La fase S procede con la replicación del material genético contenido en el núcleo y la formación de dos cromátidas idénticas, las cuales contienen todo el material genético de las células en formación. En la fase G2 se termina de ensamblar las proteínas y organelos necesarios para la división de núcleos y citoplasma. La fase M y última etapa marca la división celular. En este proceso la célula madre separa todo su material genético al formar dos núcleos independientes y agruparlos para formar a dos células hijas con una distribución de; organelos, citoplasma y material genético simétrico o asimétrico. Dicha característica depende totalmente del tipo de señalización que recibió la célula progenitora antes de entrar al ciclo celular [9] [11].

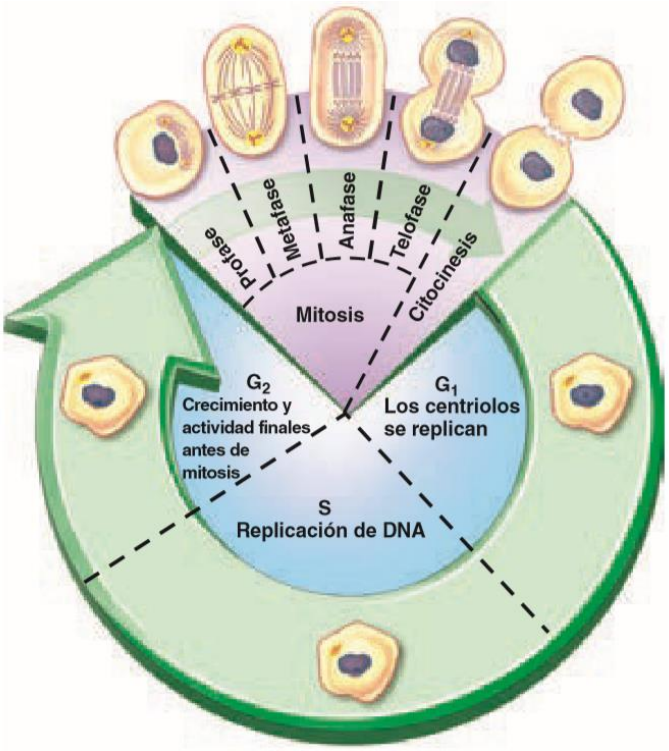

Figura 1. Fases del ciclo celular [11].

Pese a que el cuerpo puede recuperar un volumen significativo de células por medio del ciclo celular, no indica que dichos organismos estén preparados para cumplir una función específica en un determinado tejido. Por lo tanto, otro proceso vital en el reemplazo celular consiste en la especialización o diferenciación de las células y sus líneas celulares [10].

La expresión génica es un proceso indispensable para la diferenciación celular. Este mecanismo se asocia al nivel de “potencia celular” que tiene cada célula progenitora, y en sí mismo es una jerarquía la capacidad de modificarse estructuralmente, como se aprecia en la Figura 2. Los organismos con la mayor potencia celular adquieren el nombre de totipotentes debido a que pueden diferenciarse en cualquier célula del cuerpo, más sin embargo estas células solo están presentes en las primeras semanas de la gestación humana. El segundo grado es la pluripotencia, la cual le permite a un organismo como las células madres diferenciarse en cualquiera de las capas germinativas: endodermo, mesodermo y ectodermo. El penúltimo nivel es la multipotencia, la cual permite a las células diferenciarse únicamente dentro de una línea celular. Por último, quedan las células unipotentes que pierden la capacidad de diferenciarse a una especie diferente [11].

Para lograr que la célula se especialice en una estructura se debe reducir su potencia génica, y para ello se requieren de proteínas de soporte (histonas) encargadas de acetilación y desacetilación del ADN. En una terminología más simple son los marcadores que indican las porciones del ADN que se sobreexpresan, mientras que otras se empaquetan y se inhiben en su actividad transcripcional, lo cual provoca modificaciones en la estructura nativa de las células y de sus membranas [10].

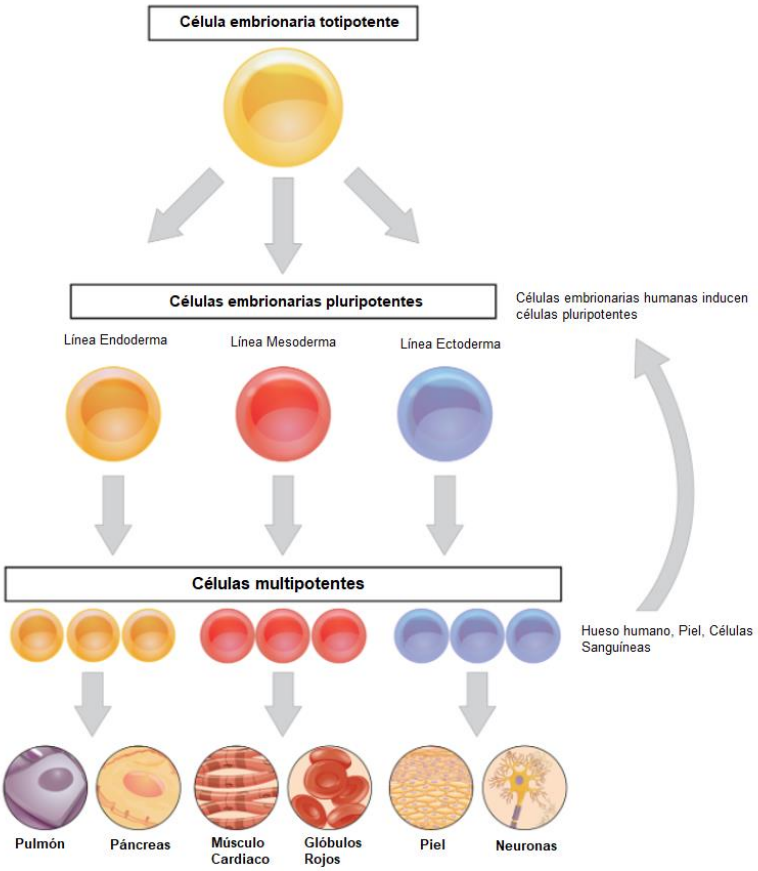

Figura 2. Modelo general de la diferenciación celular.

El último mecanismo que interviene en el reemplazo celular se denomina muerte celular. Esto ocurre debido a que una unidad celular entra en un estado de senescencia, ya sea porque posee una mutación considerable en sus proteínas o debido a que no es capaz de cumplir con sus actividades metabólicas [10]. Para inducir a una célula a la muerte se puede realizar de diversas maneras, lo cual ocasiona diversos modos de muerte como se aprecia en la Figura 3. La muerte programada o apoptosis es una secuencia en donde la célula mitiga la creación de ATP, reduce su volumen celular y empieza a descomponerse en pequeñas micelas que son fagocitadas por los macrófagos. Este modo de muerte permite que las estructuras laterales de la célula continúen con su función. El tipo de señalización que desencadena toda esta respuesta es un factor pre-apoptótico que desnaturaliza las mitocondrias en el citoplasma. El segundo modo de muerte celular es la autofagia, en este existe una cantidad de organelos y citoplasma dañado al interior de la célula, lo cual activa la respuesta de los lisosomas y las vacuolas para degradar los elementos dañinos, en caso de contener un número de daños significativos se puede inducir a apoptosis o necrosis dependiendo de qué tan comprometida esté la membrana celular [10]. El último método es la necrosis, la cual consiste en una desnaturalización de la membrana celular, el citoplasma y los organelos en su interior, por lo general ocurre como respuesta a una condición patológica o lesiones externas, en donde no hay mecanismos de adaptación funcionales que le permitan a la célula inducir otro modo de muerte [12].

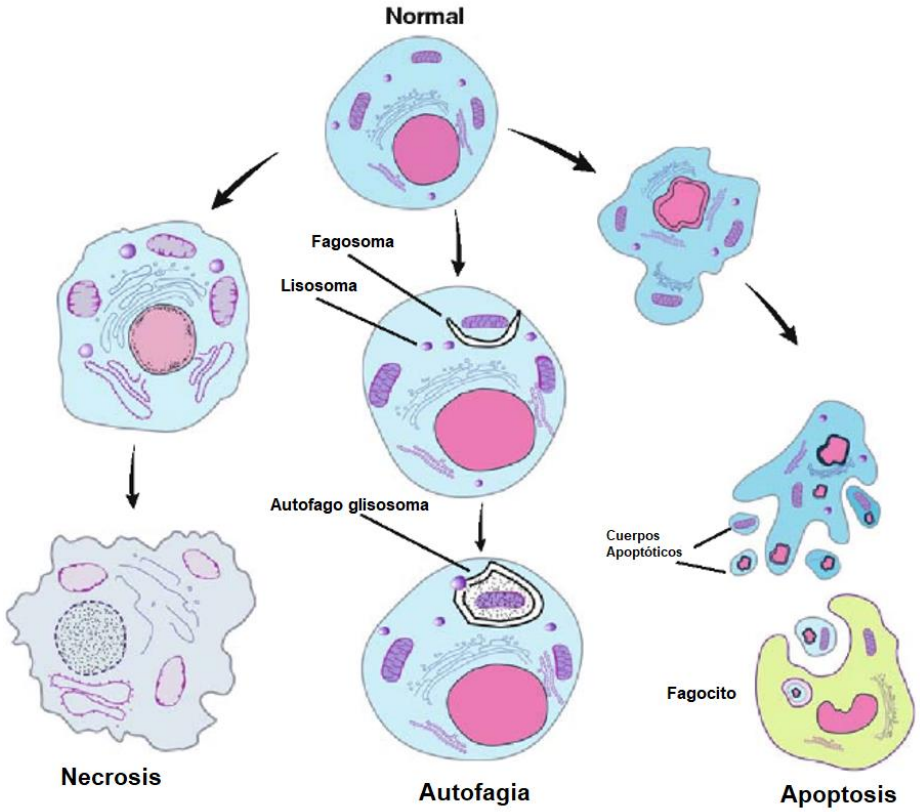

Figura 3. Modelos de muerte celular [12].

1.1.2 El cáncer a nivel celular

Las células que adquieren un perfil cancerígeno pueden aparecer en cualquier tipo de tejido, más sin embargo los órganos que poseen líneas celulares con tiempos de reemplazo más cortos suelen ser mayormente afectados. Debido a que las células exponen el material genético al medio con mayor frecuencia, lo cual incrementa la tasa de mutaciones durante los tres procesos celulares anteriormente mencionados [9] [13].

La célula cancerígena se considera como una anomalía, la cual desencadena un conjunto de nuevas células incapaces de reconocer su propio límite natural, y en casos más complejos pueden diseminarse a áreas del cuerpo donde no pertenecen. Para que un tejido adquiera un perfil benigno o maligno depende principalmente de dos factores, el primero es la cantidad de mutaciones que las células posean, y el segundo es el nivel de supresión de mecanismos de regulación oncogénica [13] [14].

Las células que adquieren una cantidad significativa de mutaciones, pierden la capacidad de regular la expresión génica de manera endógena. Para este caso existen dos variantes que permiten determinar el comportamiento del cáncer; si la célula pierde el control sobre su propia homeostasis, pero los mecanismos de regulación celular del sistema inmune pueden reducir la tasa de crecimiento anormal, el cáncer se denomina benigno, más sin embargo si todos los mecanismos biológicos pierden el control sobre este grupo de células

el cáncer adquiere la clasificación de maligno [10] [13]. En la Figura 4 se puede encontrar las características principales que tiene una célula con perfil cancerígeno.

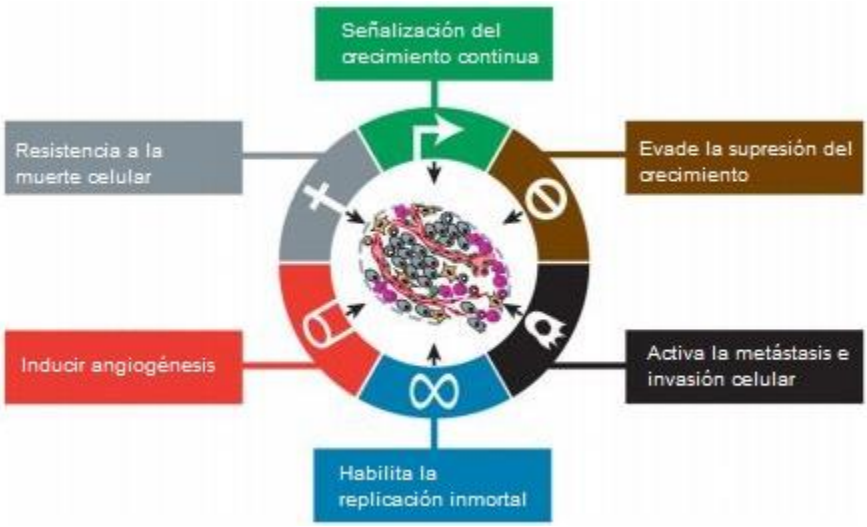

Figura 4. Características del cáncer [14].

1.1.3 El sistema inmune y el cáncer

El sistema inmune se puede agrupar en dos estrategias o subsistemas principales de defensa, el primero es el sistema inmune innato, el cual presenta barreras biológicas para impedir que agentes externos penetren el dominio celular, más sin embargo no posee medios de respuesta frente a patógenos que traspasen sus barreras o aquellas que sean originadas en el interior del cuerpo. Pero, por el contrario, el segundo subsistema conocido como sistema inmune adaptativo posee unidades especializadas en analizar, eliminar e inmunizar al cuerpo de las anomalías que perturben la homeostasis celular. [10] [13].

Uno de los agentes más efectivos del sistema inmune adaptativo son las células citotóxicas tipo T. Esta línea celular está compuesta de los agentes más mortíferos que hay entre los linfocitos, debido a que se pegan a la pared celular de la célula mutada y por medio de uniones *gap* introducen enzimas tóxicas que degradan el núcleo celular, promueve la lisis de la membrana y la descomposición de organelos. En términos de funcionalidad se considera depuración de agentes peligrosos para el organismo [14]. Pese a que el cuerpo humano posee un método muy efectivo para la eliminación de agentes extraños, algunas enfermedades como lo es el cáncer pueden proliferar sin ser detectadas debido a diversas razones [9] [13].

Para la correcta activación de los agentes citotóxicos se debe cumplir un proceso previo, el cual consiste en que: las proteínas mutantes dentro del citosol son degradadas por los lisosomas, desnaturalizando sus secuencias para extraer los epítopos de las proteínas. Los epítopos actúan como una bandera que indica la presencia de agentes nocivos al interior de la célula. Posteriormente estas secuencias peptídicas se transportan al retículo endoplasmático, en donde se unen con un complejo de mayor histocompatibilidad clase I (MHC I) y son transportados al medio extracelular. Una vez migrados los epítopos y el MHC una estructura de las células linfoides conocida como el receptor de células T (TCR) se une a este complejo, provocando que un catalizador de linfocitos citotóxicos se active y

promueva así su diferenciación [13]. La secuencia de activación de las células linfoides se aprecia en la Figura 5.

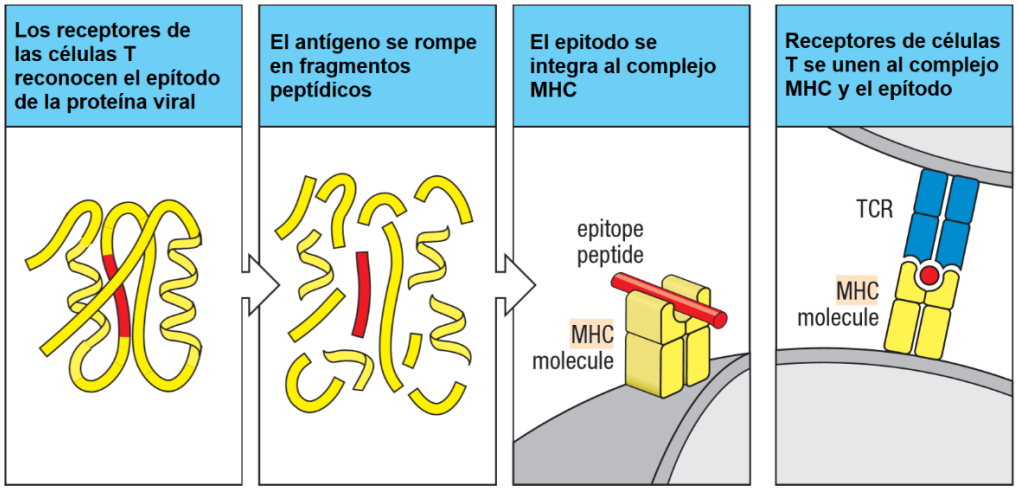

Figura 5. Secuencia de operación del MHC (K. Murphy, 2017).

Por lo tanto, para la correcta eliminación de las células cancerígenas depende del nivel que esté presente al interior de la célula. Cuando se posee un tejido con un número pequeño de mutaciones, el sistema inmune procede a eliminar el número excedente de células de la especie que hace parte del tumor benigno. En caso contrario, si se inhibe la actividad de lisis de los epítopos, el sistema inmune no tiene ningún tipo de control sobre las células que se consideran malignas [10] [14].

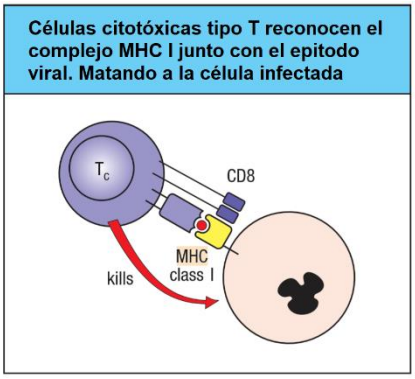

Figura 6. Complejo MHC I (K. Murphy, 2017).

El sistema inmune posee 2 complejos de mayor histocompatibilidad diferentes, ambos tienen la capacidad de activar la respuesta de los linfocitos T citotóxicos, pero el catalizador químico y la estrategia de reconocimiento en cada uno es diferente. Mientras que el MHC I está presente en la membrana de la mayoría de células nucleadas, y utiliza como catalizador un complejo CD8. El MHC 2 solo está presente en las células linfoides y tiene como catalizador al complejo CD4 [13].

A partir de la premisa anterior se puede realizar una correlación de cómo los complejos HMC I pueden convertirse en un biomarcador útil para la detección de nuevas especies cancerígenas, debido a que este tipo de patología inhabilita la formación adecuada de este

complejo [15] [16]. Para ello es vital realizar un análisis comparativo de las propiedades fisicoquímicas en las especies cancerígenas y no cancerígenas. En la actualidad una de las líneas de investigación que realiza este tipo de estudios bajo un entorno de simulación es la bioinformática, la cual permite analizar el comportamiento de las estructuras celulares [17] [16].

Identificar un biomarcador potencial es un hecho bastante significativo para un estudio bioinformático. A partir de este elemento se pueden realizar diversos entornos experimentales para analizar el comportamiento biológico del problema. Para contextualizar cómo sería una herramienta de soporte en los procesos de detección del cáncer, es útil conocer la trayectoria y los desarrollos en materia de bioinformática, con el fin de enriquecer el trabajo y aprender de los planteamientos de diversos investigadores.

1.2 Estado del Arte

1.2.1 Historia de la Bioinformática

La bioinformática ha sido implementada en los últimos 50 años como una herramienta para el análisis de la información biológica. En los años 70 el impacto de la bioinformática marcó un antes y un después en la forma en que se analizaban y leían las secuencias de ADN. Desde ese entonces se empezaron a construir lenguajes específicos para interpretar y leer los datos que contenían las secuencias de ADN, y como esta se traducía al ARN, y este a las proteínas respectivamente, más sin embargo la mayor limitación en la secuenciación completa del genoma eran las unidades de procesamiento de esa época, impidiendo procesar cientos de miles de datos que se obtenían. El primer gran avance para este campo de estudio llegó a partir de los años 90, en donde se contaba con unidades GNU que permitían implementar códigos más robustos, almacenar información biológica más grande y analizar variables más complejas. Esta década también se denominó como la era de la secuenciación, en donde el estudio primordial era culminar el mapa del genoma humano que se había buscado desde los años 50 [18].

En la primera década de este siglo la investigación en bioinformática había rendido grandes avances con la culminación de la secuenciación del genoma humano, más sin embargo este logro desencadenó la segunda era de la secuenciación debido a que se buscaba obtener el mapa genético de la mayor cantidad de especies vivas en el planeta. Pero no fue hasta el 2005 que las academias más importantes del mundo desarrollaron un estándar de base que debe poseer una secuencia. Con todo el desarrollo de los últimos años en la actual década, la bioinformática ha apuntado a nuevos horizontes: con el uso de las secuenciaciones que se han realizado en los últimos 30 años se pretende diseñar modelos que simulen el comportamiento de sistemas biológicos, con el fin de crear tratamientos, aplicaciones y desarrollos que mejoren la calidad de vida de las personas [18].

Las innovaciones modernas en materia de bioinformática se pueden agrupar en seis categorías o líneas de investigación, las cuales son: genómica, proteómica, micro arreglos, sistemas biológicos y simulación, evolución y otras aplicaciones. El objetivo común de la mayoría de estas líneas de investigación es predecir el comportamiento de una unidad funcional biológica en un entorno *in silico*, ya sea a nivel proteico o génico y trasladar los experimentos a entornos *in vitro* e *in vivo* para corroborar las predicciones [19].

1.2.2 Bioinformática y cáncer

Las investigaciones en cáncer desde la línea de bioinformática son tan variadas como las líneas celulares mismas que pueden expresar esta enfermedad. Muchos de los desarrollos en materia de esta enfermedad se realizan desde las líneas de genómica, proteómica y simulación biológica [20]. Estos tres campos basan sus estudios en las estructuras que intervienen activamente en las vías metabólicas, la señalización celular y los residuos que dejan. Existen otras dos líneas de investigación que tienen un aporte significativo en investigaciones en cáncer. Desde la línea de micro arreglos y procesamiento de imágenes biomédicas se anexa información de las células, basándose en información bidimensional y tridimensional [21] [18].

Al considerar lo numeroso que son las investigaciones en cáncer y las diferentes líneas que existen para generar conocimiento se puede agrupar todos estos desarrollos bajo dos grandes principios, los cuales son: desarrollar métodos para la detección temprana de cáncer y generar nuevos tratamientos oncológicos a partir de herramientas de predicción informáticas [22] [15].

Desde la línea de detección temprana de cáncer, el objetivo primordial ha sido la detección de nuevos biomarcadores que indiquen la presencia de esta enfermedad en cualquier etapa de su gestación, ya sea a partir de los residuos que generan las células, malformaciones en los dominios de las proteínas, anormalidades en las tasas de gasto metabólico de organelos específicos y entre otros biomarcadores [15] [23].

En paralelo al crecimiento de la bioinformática, otra ciencia de la computación ha generado grandes avances en modelos y algoritmos estadísticos, con la capacidad de procesar grandes bancos de información, agruparlos y convertir la información en conocimiento. Es en este punto donde las técnicas de aprendizaje automático ingresan como herramientas para el análisis de la información biológica [20] [24].

1.3 Aprendizaje Automático y la bioinformática

El aprendizaje automático o *Machine Learning* (ML) es una ciencia que estudia algoritmos y modelos estadísticos para el desarrollo de funciones específicas. Una manera menos convencional de referirse a este estudio es; la manera de aproximar los métodos que los seres vivos utilizan para generar conocimiento por medio de información previa, y emularlo en sistemas digitales. Los algoritmos de aprendizaje automático son ampliamente utilizados debido a que pueden procesar una gran cantidad de información en N-dimensiones, y pueden reconstruir grupos para tareas de clasificación o generar regresiones numéricas para tareas de predicción [25] [26].

El uso del aprendizaje automático en la bioinformática tuvo un impacto significativo al inicio del siglo XXI, debido a que estaban en la segunda era de la secuenciación y tanto el número de repositorios biológicos como los tamaños de los mismos estaban en aumento [19].

El crecimiento desmedido de los autores en bioinformática provocó problemas en este campo. El primero era el manejo eficiente de los datos, debido a que se reportaban cada vez más variables biológicas de una misma especie y no se podía analizar toda la información emergente de manera homogénea, y el segundo era la redundancia de los trabajos de investigación para las diferentes especies [18]. Para la solución de ambos

problemas se empezaron a implementar sistemas neurales para procesar las primeras secuencias genómicas de repositorios saturados y redundantes, con el objetivo de disminuir la dimensionalidad de los datos, mientras se penalizaba la redundancia de las muestras [19] [18].

Posteriormente el aprendizaje automático generó tendencia en las otras líneas de la bioinformática. En la proteómica se utilizó para clarificar el nivel de unión de biomarcadores con los dominios de enzimas catalizadoras o predecir la incidencia de un compuesto en la supervivencia celular de algunas bacterias extremófilas a partir de la espectrometría de masas [27] [28]. En campos como la evolución y simulación de sistemas se adecuaron algoritmos genéticos para predecir la incidencia de los factores que más peso tenían en la supervivencia de las especies y cómo pueden evolucionar con los datos topográficos actuales [29].

Muchas de las aplicaciones modernas del aprendizaje automático en la bioinformática se han enfocado en desarrollos para tratar o analizar las enfermedades desde diferentes enfoques. Una línea de investigación que surgió a partir de la bioinformática es la inmunoinformática, la cual utiliza los mismos enfoques metodológicos, pero centra sus estudios en los organismos del sistema inmune, con el objetivo de analizar porque fallan los sistemas de defensa del cuerpo o desarrollar tratamientos que puedan activar respuestas específicas de estos sistemas para responder de manera autónoma frente a algunas enfermedades [18] [30].

1.3.1 Inmunoinformática y Aprendizaje Automático.

El enfoque tradicional de la inmunoinformática busca predecir: el efecto de un estímulo en términos del número de células inmunes que se proliferan en el tiempo, la tasa de reconocimiento y activación de las células B y T o generar modelos evolutivos que predicen la herencia de información de una generación a otra de células B [31].

Desde la línea de proteómica se han desarrollado diferentes estudios con epítopos y complejos MHC para la creación de aplicaciones inmunoinformáticas para tratamiento y detección de diferentes enfermedades, debido a que estos dos son agentes que determinan la respuesta del sistema inmune frente a un antígeno [30].

Los experimentos con epítopos son de los más comunes en la práctica de la de inmunoinformática. Muchos de estos experimentos buscan identificar la longitud óptima de un receptor MHC específico, con el fin de caracterizar el comportamiento de los compuestos encargados de la detección de antígenos [32]. Otro uso común de los epítopos es para la predicción de las tasas de respuesta de las células T y B frente a la unión de estos compuestos con un receptor de células linfoides [33].

Para el desarrollo de tratamientos inmuno-específicos, la inmunoinformática realiza procesos de regresiones para determinar el grado de unión entre un complejo MHC disfuncional y una célula linfoide con la intervención de un catalizador [34]. El análisis de catalizadores como elementos para mejorar la tasa de unión entre complejos inmunes es uno de los enfoques más modernos que tiene la inmunoinformática. El objetivo primario de esta línea de investigación es el diseño de tratamientos contra enfermedades inmunosupresoras de manera personalizada, con el fin de predecir la dosis óptima de

compuesto que mantenga la integridad celular, mientras que reactiva la reactiva enzimática de las células linfoides [35].

El uso que se le da a los antígenos en la inmunoinformática se centra en predecir el perfil de las patologías a nivel celular y apoyar en funciones de diagnóstico, debido a que enfermedades de origen viral o bacteriano mutan de manera exógena los procesos de expresión celular, mientras que enfermedades autoinmunes lo realizan a partir de la acumulación de mutaciones [36].

La primera aplicación que se especializó en predecir antígenos cancerígenos, bacterianos y virales fue VaxiJen. Este algoritmo del año 2007 implementaba un modelo de trasformación de auto covarianza cruzada para predecir el perfil de antígeno a partir de las propiedades fisicoquímicas de la proteína, y tenía una exactitud del 80% [37]. La segunda gran aplicación en la predicción de antígenos es TIminer de 2017, a diferencia de su antecesor este programa solo se enfoca en la predicción de antígenos tumorales y de los dominios inmunogénicos dentro del antígeno para predecir componentes en vacunas, este programa en su fase de predicción de antígenos reporto una exactitud del 89 % [38].

Con referencia en el estado del arte y al marco teórico, el presente trabajo propone una comparación de modelos de aprendizaje automático para la detección de células cancerígenas a partir de los antígenos del complejo MHC I. Utilizando protocolos de extracción de características físico-químicas de las proteínas y un proceso comparativo de las medidas de desempeño en la fase de validación y prueba de los modelos. Con este procedimiento se pretende determinar cuál modelo de aprendizaje automático presenta el mejor desempeño en la predicción de antígenos cancerígenos, utilizando propiedades fisicoquímicas como marcadores de entrada.

2. OBJETIVOS

2.1. General

Comparar el funcionamiento de tres algoritmos de aprendizaje automático para la detección de antígenos cancerígenos en secuencias peptídicas, empleando medidas de sensibilidad, especificidad y curva ROC.

2.2. Específicos

a) Determinar qué características fisicoquímicas de los aminoácidos que conforman la cadena peptídica de los antígenos bajo estudio son relevantes para la detección de cáncer partir del complejo MHC I.
b) Entrenar modelos los modelos de bosques aleatorios, máquinas de soporte vectorial y redes neuronales para la detección de antígenos extraídos de secuencias peptídicas en organismos sanos y cancerígenos.
c) Analizar los modelos propuestos a partir de sus medidas de sensibilidad, especificidad y curva ROC, comparando sus prestaciones y complejidad en cuanto a su tamaño.

3. METODOLOGÍA

En la Figura 7 se encuentra un diagrama de bloques con el procedimiento experimental simplificado. En esta figura se observan las tres fases del proyecto, desde el proceso de adquisición de datos hasta las pruebas de validación de los modelos entrenados.

Figura 7. Procedimiento experimental.

Para organizar el desarrollo experimental del proyecto de grado la metodología se divide en tres fases:

- Construcción de la base de datos.
- Selección y Entrenamiento de los modelos.
- Análisis del desempeño.

3.1. Construcción de la base de datos

La base de datos está compuesta por un conjunto de secuencias peptídicas en formato FASTA del complejo MHC I de diferentes líneas celulares. La primera parte de los datos se toma del repositorio "TANTIGEN", el cual es una base de datos especializada en antígenos y epítopos cancerígenos. El repositorio posee un compendio de 900 proteínas registradas y 4.000 unidades peptídicas secuenciadas en experimentos *in vivo* e *in vitro* [39]. Para la selección de la muestra de trabajo se toma una única secuencia peptídica, con una secuenciación completa por cada tipo de proteína registrada, descartando de ese modo las isoformas y las secuenciaciones parciales. El objetivo de estos criterios es evitar sobre ajustes a ciertas especies celulares que se sobreexpresan en el cáncer, permitiendo así que los modelos se generalicen de una manera más adecuada [23].

La segunda sección de la base de datos está conformada por proteínas de células no cancerígenas del repositorio IEDB. Esta base de datos contiene un registro de epítopos, de los complejos MHC I y II y porcentaje de enlace celular entre ambas estructuras peptídicas; dichas proteínas fueron secuenciadas en experimentos *in vivo* e *in vitro* para más de 9.615 especies. Los antígenos del complejo MHC I en humanos ascienden a 10.000 muestras secuenciadas, de un total de 586 clases celulares diferentes. A este conjunto de datos se le aplican los mismos criterios que al anterior, excluyendo a las secuencias parciales y tomando una única isoforma por cada tipo celular. De este conjunto de datos se tomó una muestra arbitraria de 256 proteínas [23] [40].

Una vez construida la base de datos con los 1.156 antígenos se procede a extraer las características de las secuencias. Para este paso se utilizan dos protocolos experimentales predefinidos, los cuales son: "*Standardized criteria for statistical analysis of inmunoglobulina V-REGION and amino acid properties*" [41] y "*Amino acid side chain parameters for correlation studies in biology and pharmacology*" [42]. El primer protocolo describe que los atributos estáticos de los aminoácidos basándose en el conteo de grupos funcionales, la polaridad y la categoría según el tamaño de cada aminoácido existente en la secuencia peptídica. El segundo protocolo extrae las características de las proteínas a partir de la ponderación total de todos los donantes de hidrógeno, la carga y el coeficiente de hidrofobicidad de cada aminoácido.

3.1.1. Descripción de las características Físico-Químicas

Las características que se consideran a partir de ambos protocolos son:

- Índice de Hidropatía: Representa la energía libre de transferencia del aminoácido sometido a un medio de baja constante dieléctrica al agua.
    - Hidrofóbico
    - Hidrofílico
    - Neutro
- Carga: Es el valor energético que queda en la cadena lateral, una vez se haya formado una secuencia peptídica.
    - Cuando los aminoácidos son expuestos a un pH fisiológico, dos adquieren carga negativa: ácido aspártico (Asp, D) y ácido glutámico (Glu, E) (cadenas laterales ácidas), y tres tienen carga positiva: lisina (Lys, K), arginina (Arg, R) e histidina (His, H) (cadenas laterales básicas).

- Número donantes de Hidrógenos: Capacidad de recibir o aportar un grupo hidroxilo para formar puentes de hidrógeno.
    - No donantes: No forma puentes de hidrógeno
    - Donante simple: Forma 1 puente de hidrógeno
    - Donante doble: Forma 2 puente de hidrógeno
    - Donante fuerte: Forma 4 puente de hidrógeno
- Composición por grupos funcionales: Los aminoácidos que constituyen las proteínas poseen propiedades diferentes debido sus grupos funcionales, los cuales se dividen en:
    - Alifático
    - Aromático
    - Sulfuro
    - Hidroxilo
    - Básico
    - Acido
    - Amida
- Polaridad: Capacidad de separación de las cargas eléctricas en la misma molécula.
    - Polar
    - Apolar
- Volumen: Es la unidad de espacio que ocupa una proteína en una solución. Esta unidad se coloca en Angstroms cúbicos $(\dot{A}^3)$, sin embargo, el estándar permite agrupar a los aminoácidos en 5 categorías.
    - Diminuto
    - Pequeño
    - Medio
    - Grande
    - Muy grande

En la Figura 8, se encuentran los coeficientes de hidropatía para los 20 aminoácidos esenciales, clasificando las especies que poseen un comportamiento hidrofóbico, neutro e hidrofílico.

I	V	L	F	C	M	A	W	G	T	S	Y	P	H	N	D	Q	E	K	R
4,5	4,2	3,8	2,8	2,5	1,9	1,8	-0,9	-0,4	-0,7	-0,8	-1,3	-1,6	-3,2	-3,5	-3,5	-3,5	-3,5	-3,9	-4,5
Hidrofóbico								**Neutro**						**Hidrofílico**					

Figura 8. Coeficientes de hidropatía por aminoácido [43].

La Figura 9 contiene el agrupamiento general de los 20 aminoácidos con respecto a las características de carga, composición de grupos funcionales, polaridad y volumen.

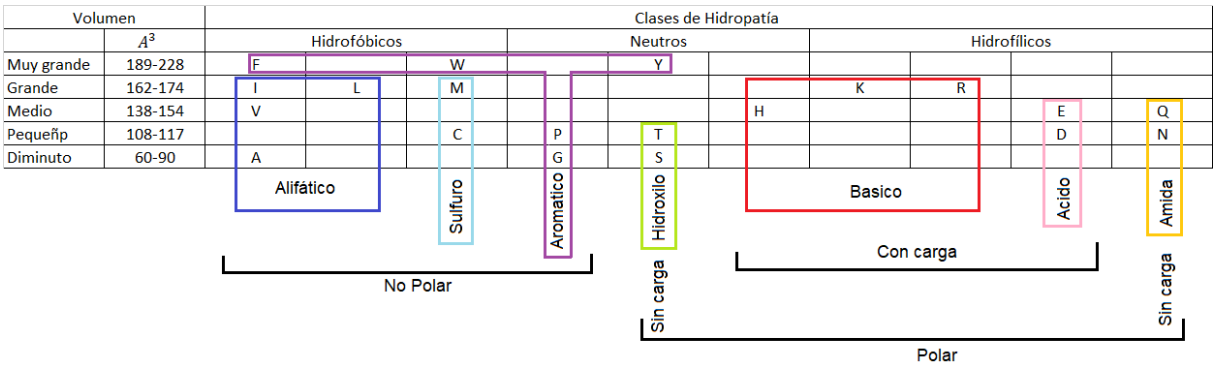

Figura 9. Características generales de los 20 aminoácidos [43].

La Tabla 1 muestra el número total de donantes y aceptores de hidrógeno que posee cada uno de los 20 aminoácidos principales.

Aminoácidos	*Donantes de hidrógeno*
Ala	0
Arg	4
Asn	2
Asp	1
Cys	0
Gln	2
Glu	1
Gly	0
His	1
He	0
Leu	0
Lys	2
Met	0
Phe	0
Pro	0
Ser	1
Thr	1
Trp	1
Tyr	1
Val	0

Tabla 1.Número de donantes de hidrógeno por aminoácido [42].

La segunda etapa del preprocesamiento de la matriz de características se basa en el escalamiento y normalización de los datos. El proceso de escalamiento consiste en una transformación lineal para estandarizar las características de las proteínas basándose en la longitud de las secuencias, con el objetivo de poder comparar los niveles de energía de las proteínas cortas o inferiores a los 50 aminoácidos, con las grandes o superiores a los

250 aminoácidos [23]. El proceso de normalización se basa en otra transformación lineal que permite escalar los datos para disminuir los traslapes entre las distribuciones de probabilidad natural de los datos [44].

3.1.2. Escalamiento de características

Para el procedimiento de escalamiento se considera la frecuencia relativa de cada uno de los 20 aminoácidos principales. Estas variables se utilizan para el cálculo de la frecuencia acumulada de los grupos: polaridad, grupos funcionales y volumen. Para las variables de número de donantes de hidrógeno, carga y coeficiente de hidrofobicidad, a estas frecuencias se les multiplica un coeficiente respectivo por aminoácidos y característica evaluada [23]. Esto se aprecia en las ecuaciones 1 y 2, las cuales representan la frecuencia relativa por aminoácido y la frecuencia relativa con un coeficiente especifico respectivamente.

$$\left(h[aa] = \frac{f[aa]}{N}\right) \qquad (1)$$

$$\left(h_i[aa] = \frac{f[aa]}{N} * C_{i[aa]}\right) \qquad (2)$$

En la ecuación 1 y 2 los términos $f[aa]$ y N representan la cantidad de un aminoácido y la longitud de la secuencia peptídica respectivamente. Los términos $h[aa]$ y $h_i[aa]$ representan la frecuencia relativa por aminoácido, con el condicionamiento de que h_i es multiplicado por el respectivo coeficiente de la característica. Una vez culminado el cálculo de las frecuencias se procede a la sumatoria de características, dependiendo de la variable como se aprecia en la ecuación 3 [23].

$$H[val] = \sum_{x=aa}^{n} h[aa] \qquad (3)$$

Este proceso impide que las distribuciones de probabilidad en los datos sean funciones que dependan de la longitud de las proteínas [23].

3.1.3. Normalización de los datos

El proceso de normalización se realiza a partir de la desviación estándar de las muestras, como se observa en ecuación número 4. Esta técnica escala los datos homogéneamente distribuidos con valores negativos a todo valor bruto que tenga una puntuación menor a la media, mientras que los positivos son de aquellos puntajes superiores a la media [44].

$$\frac{X - \bar{X}}{S} \qquad (4)$$

El uso de procesos de normalización basado en la desviación de las variables permite aproximar a una distribución normal, con el objetivo de que la distribución de probabilidad sea asimétrica debido al orden de magnitud de los atributos de entrada [45].

Culminado los procesos de escalamiento y desviación se obtiene una matriz de 1.156 muestras con 34 características, las cuales se aprecian en la Tabla 2. En esta tabla se puede apreciar; la frecuencia relativa (F.R) de los 20 aminoácidos principales, la frecuencia

acumulada (F.A) de los distintos grupos fisicoquímicos y los coeficientes netos de tres características químicas.

CARACTERSITICAS	***ETIQUETA***
F.R Alanina	*f_a*
F.R Cisteína	*f_c*
F.R Aspartato	*f_d*
F.R Ácido glutámico	*f_e*
F.R Felanina	*f_f*
F.R Glicina	*f_g*
F.R Histidina	*f_h*
F.R Isoleucina	*f_i*
F.R Lisina	*f_k*
F.R Leucina	*f_l*
F.R Metionina	*f_m*
F.R Asparagina	*f_n*
F.R Prolina	*f_p*
F.R Glutamina	*f_q*
F.R Arginina	*f_r*
F.R Serina	*f_s*
F.R Treonina	*f_t*
F.R Valina	*f_v*
F.R Triptófano	*f_w*
F.R Tirosina	*f_y*
Carga Neta	*CaNet*
Donantes de Hidrógeno	*DoHi*
Coeficiente Hidrofóbico	*CoHi*
F.A Alifático	*Ali*
F.A Aromático	*Aro*
F.A Sulfuro	*Sul*
F.A Hidroxilo	*Hdrx*
F.A Amida	*Amid*
F.A Ácidos	*Acd*
F.A Aminoácidos Diminutos	*Tiny*
F.A Aminoácidos Pequeños	*Small*
F.A Aminoácidos Grandes	*Lrg*
F.A Polares	*Plr*
F.A Apolares	*Aplr*

Tabla 2. Características de estudio.

3.2. Selección de los modelos

Para la selección de uno de los modelos se tomó como trabajo guía al artículo: "*TTAgP 1.0: A computational tool for the specific prediction of tumor T-cell antigens*", el cual presentaba un código libre para la predicción de neo antígenos cancerígenos implementado como modelo base de clasificación en un algoritmo de bosques aleatorios o Random Forest [23]. La selección de los modelos de máquina de soporte vectorial y redes neuronales son el resultado de una revisión bibliográfica realizada en la base de datos SCOPUS, para determinar cuáles son los modelos más implementados en aplicaciones de proteómica e

inmunoinformática. Para esta revisión se selecciona una ventana de 5 años para considerar únicamente artículos modernos. Anexo a estas consideraciones, los tres modelos seleccionados tienen su respectiva modalidad de clasificación en la toolbox de aprendizaje automático de Python (scikit-learn), lo cual es acorde con la aplicación de predicción de células cancerígenas a partir de antígenos.

3.2.1. Modelo de Bosques aleatorios (RF)

El modelo de bosque aleatorio se puede desglosar primeramente de una manera empírica. Un bosque es el agrupamiento aleatorio de un conjunto de árboles, por lo tanto, para entender el concepto de "bosque" es necesario comprender que es un árbol.

3.2.1.1. Árbol de decisión

El modelo de árbol de decisión se basa en una secuencia de preguntas significativas a partir de las características de entrada. Acorde a esto su topología se asemeja a la de un árbol invertido, su pregunta inicial se considera el tronco del árbol o nodo raíz, posteriormente cada ramificación representa más preguntas o nodos de prueba, con el fin de culminar en nodos puros u hojas que nos permite interpretar la respuesta del problema. Este proceso permite dividir de manera rectangular el espacio vectorial cada vez que un dato pasa por una pregunta, formando así regiones de decisión fácil de interpretar incluso sin conocimientos a priori del problema [46]. Una representación gráfica de los planos que el modelo puede crear durante su entrenamiento, se visualiza en la Figura 10.

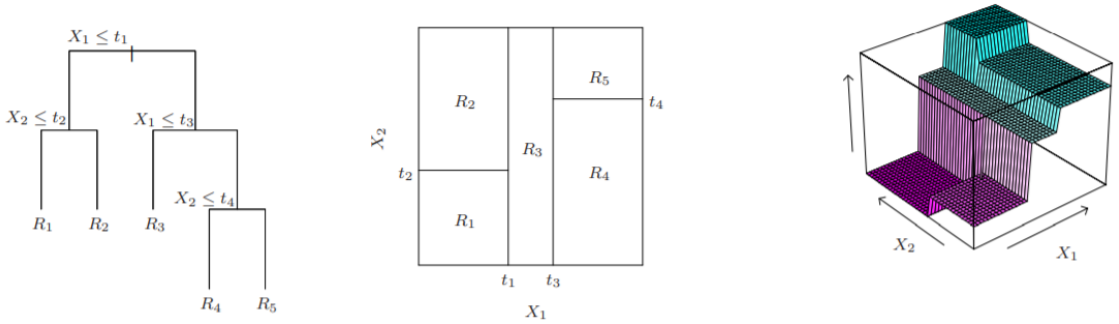

Figura 10. Espacio general de un árbol de decisión [46].

La secuencia de corte o de asignación de preguntas tienen como base la estructura de un árbol, por lo tanto, debe tener el mayor espesor en su base, y este caso en particular la cantidad de información extraída en los primeros niveles de corte debe ser la mayor [46]. Para asegurar que el modelo extrae de manera correcta la información, se toma como el nivel de información que puede aportar las características como nodo. Esta extracción se modela en la ecuación 5.

$$IG(D_P) = I(D_P) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \qquad (5)$$

La ecuación 5 muestra el cálculo de la información (IG) a partir de la impureza del padre $I(D_P)$ y la sumatoria de las impurezas de los hijos $I(D_j)$, multiplicado por la razón de datos entre el nodo hijo y el nodo padre (N_j/N_P). Con esta ecuación el modelo puede determinar qué nodo debe estar en la sección superior del árbol por medio de un proceso iterativo, pero para ello se requiere determinar medidas de impureza acorde a la aplicación [47].

Como se tiene un problema de clasificación, las funciones de impureza óptima para un árbol de decisión son: la entropía cruzada y el índice de Gini. El método de entropía mide la calidad de las distribuciones, mientras que el de Gini minimiza el error de clasificación al disminuir la varianza de las muestras. Estos dos modelos matemáticos están representados en las ecuaciones 6 y 7 respectivamente [46].

$$II_H(t) = -\sum_{i=1}^{c} p(i/t) \log_2 p(i/t) \qquad (6)$$

$$I_G(t) = \sum_{i=1}^{c} p(i/t)\,(1 - p(i/t)) \qquad (7)$$

Al iterar de manera finita el modelo converge generando respuestas de diferentes profundidades (Ver Figura 10). Esto indica que la calidad de los datos puede conducir a la respuesta óptima dependiendo del nivel de información que cada corte haya generado en el espacio vectorial. Uno de los problemas de permitir que un árbol se extienda indefinidamente es el sobreajuste que se puede generar de los datos, impidiéndole al modelo reproducir la distribución natural de los datos [46].

3.2.1.2. Bosque aleatorio

Implementar y entrenar un árbol de decisión es teórica y computacionalmente simple, sin embargo, al ser el único clasificador estocástico corre el riesgo de presentar altos niveles de varianza y sesgo en sus predicciones. Por este motivo implementar un bosque aleatorio se convierte en una opción sólida, debido a que tiene todas las ventajas matemáticas de un único árbol, pero con el objetivo que los árboles extras puedan suplir las debilidades de la clasificación entre ellos [46] [48].

Para la construcción de un bosque aleatorio se toma un conjunto de muestras $n \in N$, donde: N es la totalidad de la población y n su muestra. Posteriormente se entrena el árbol y cuando termina de crecer se crea un nuevo conjunto de entrenamiento $n_{\tau+1} \neq n_{\tau}$, y se prosigue con el entrenamiento del nuevo árbol. El entrenamiento culmina cuando todos los árboles generan una predicción, y esta se pondera para dar el resultado final [46] [48].

3.2.2. Modelo de Redes neuronales artificiales (ANN)

El modelo de redes neurales es un modelo inspirado en la arquitectura del cerebro, debido a que pondera los valores de las entradas junto a los pesos sinápticos que se le otorga a las neuronas, pero en términos de funcionalidad son un modelo no lineal de regresión [49].

3.2.2.1. Función *Feed-forward*

El modelo de la red neuronal se basa en la combinación lineal de parámetros (ecuación 8), la cual tiene que pasar a través de una función de activación no lineal (ecuación 9).

$$a_j = \sum_{i=1}^{D} {w_{ji}}^{(1)} x_i + {w_{jo}}^{(1)} \qquad (8)$$

$$z_j = h(a_i) \qquad (9)$$

Debido a que las redes neuronales tienen diversas entradas, todas estas características tienen que ser ponderadas en las diversas neuronas de las capas ocultas, este proceso convierte el espacio de entradas en una sumatoria de parámetros adaptativos que están diseñados para realzar las propiedades de mayor peso entre las características [25].

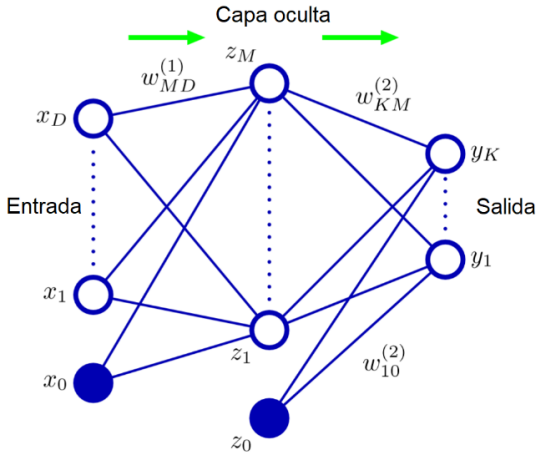

Figura 11. Estructura genérica de una red neuronal [25].

Para construir adecuadamente una red neural, no es computacionalmente complejo. Los pesos sinápticos de la red son asignados de manera aleatoria con valores pequeños, y con estos se calcula un estado inicial aleatorio (Ecuación 10), donde la salida y_k, es usada para el cálculo del error de a partir de los pesos w (Ecuación 11).

$$y_k = \sum_{j=0}^{M} w_{kj} z_j \qquad (10)$$

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \|y_k - t_k\|^2 \qquad (11)$$

3.2.2.2. Aprendizaje del modelo

Una vez se tienen todos los valores de las funciones de activación en la red, es necesario calcular el gradiente del error de cada uno de los pesos sinápticos. Debido a que el modelo tiene un comportamiento regresivo, la ecuación óptima para la actualización de los pesos se considera la ecuación 12, la cual calcula el peso sináptico de la siguiente época ($w_j{}^{\tau+1}$) a partir del peso actual (w_j^{τ}) menos el gradiente del error por la tasa de aprendizaje ($\eta \nabla E$).

$$w_j{}^{\tau+1} = w_j^{\tau} - \eta \nabla E\left(w_j^{\tau}\right) \qquad (12)$$

3.2.2.3. Propagación reversa

Como el proceso de optimización empieza siendo completamente aleatorio, la ecuación 10 no presenta buenos resultados en las tareas de clasificación o regresión, pero a partir del cálculo del error en la ecuación 11 y el proceso de optimización de los pesos sinápticos de

la ecuación 12, se entrenan las neuronas para construir el hiper-plano de decisión óptimo. Por lo tanto, es necesario implementar el método de propagación inversa para asignar porciones de error total del modelo a las capas de neuronas anteriores [25].

El cálculo del error en la capa de salida (δ_k) es equivalente a la diferencia entre el resultado teórico (t_k) y el generado por el modelo (y_k), debido a que no tiene función de activación la salida de las redes neurales.

$$\delta_k = y_k - t_k \tag{13}$$

Para calcular el error en las capas anteriores (δ_j) es necesario recurrir al valor de la función de activación de la neurona objetivo y multiplicarlo por la sumatoria de error de los errores de la capa siguiente (Ecuación 14). Una vez se tiene el valor del error se calcula el gradiente por medio de la multiplicación del error con el valor de entrada de la neurona (Ecuación 15)

$$\delta_j = \left(1 - {z_j}^2\right) \sum_{k=1}^{K} w_{kj}\delta_k \tag{14}$$

$$\frac{dE}{dw_{ji}^{(n)}} = \delta_j x_i \tag{15}$$

3.2.3. Modelo de Máquina de soporte vectorial (SVM)

El principio de una máquina de soporte vectorial consiste en diseñar un hiper plano óptimo que separe categorías binarias. Más sin embargo, el funcionamiento matemático de este modelo se describe como una categoría de una red neuronal *feedforward* [49].

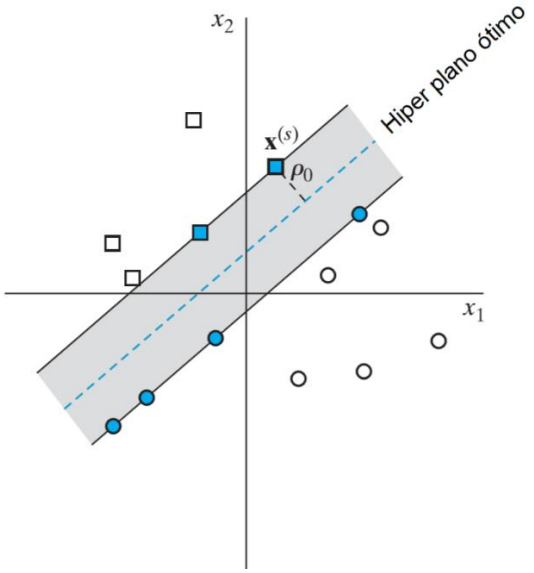

Figura 12. Hiper plano de separación óptimo (Haykin, 2009).

Lo novedoso de la máquina de soporte vectorial es que la diferencia de una red neuronal es el concepto de proyección de vectores. Para este modelo se tiene un vector de soporte

que delimita la mínima distancia de separación entre las muestras y el vector de entrada. El objetivo es encontrar el plano que permite maximizar el margen de los vectores de soporte. La función general describe el plano de separación que se expuso en la ecuación 8, pero en la ecuación 16 se observa lo diferente del modelo ante una red neural. El vector de pesos w, el de entrada x y el parámetro de *bias* b se comportan igual, pero en vez de ser una red ponderativa donde se la suma de las funciones de activación no lineales generan el espacio de decisión, aquí se tiene un único plano, donde su respuesta depende de la clasificación objetivo d_i [49].

$$g(x) = w_0^T x + b = \mp 1\,;\;\; d_i = \mp 1 \qquad (16)$$

Para optimizar este modelo es necesario calcular la distancia de separación r entre el vector de entrada x y el espacio de decisión $g(X)$. Esta distancia también es conocida como la proyección del vector de entrada sobre el plano.

$$r = \frac{g(x)}{\|w_0\|} = \begin{Bmatrix} \dfrac{1}{\|w_0\|} \\ \dfrac{-1}{\|w_0\|} \end{Bmatrix}; \begin{matrix} d = 1 \\ d = -1 \end{matrix} \qquad (17)$$

El concepto de proyección se asocia al plano perpendicular que une dos planos objetivos por medio de la distancia más corta. Al considerar esta premisa y la ecuación 17 nos aseguramos que al maximizar r, se maximicen los márgenes de separación entre clases [49]. Una interpretación geométrica de esta afirmación se encuentra en la Figura 13.

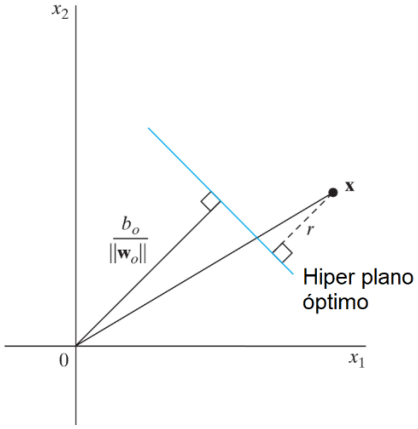

Figura 13. Representación de la proyección en 2D [49].

3.2.3.1. Entrenamiento para máquinas de soporte vectorial.

El problema de optimización para este modelo es un poco más complejo debido a que está planteado bajo un modelo de optimización convexa; lo cual genera que la optimización tenga que hacerse a partir de dos premisas [49].

La primera premisa es la optimización de la función de costo, la cual se implementa a partir de los multiplicadores de Lagrange. En la ecuación 18 se observa como la minimización del costo $w^T w$ queda representada como relación del espacio óptimo bajo los coeficientes b, α.

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i [b_i (w_0^T x + b) - 1] \qquad (18)$$

El uso de multiplicadores coloca ciertas restricciones en el uso de coeficientes, si no se selecciona con cuidado la función de costo podría volverse inestable y nunca converger. Para solucionar esto se restringen los coeficientes de la ecuación 18 con estas dos restricciones:

$$\frac{dJ(w, b, \alpha)}{dw} = 0$$

$$\frac{dJ(w, b, \alpha)}{db} = 0$$

La segunda premisa es el cálculo del mínimo valor de los coeficientes de Lagrange, el cual es un problema ya solucionado con funciones de optimización de segundo orden. Con la ecuación 19 se resuelve el cálculo de los coeficientes w para toda variable de entrada x_i, mientras que con la ecuación 20 se obtiene el valor mínimo del Lagrangeano que optimiza la ecuación 18, por medio de un proceso interactivo.

$$w_0 = \sum_{i=1}^{Ns} \alpha, d_i x_i \qquad (19)$$

$$b_0 = 1 - w_0^T X = \sum_{i=1}^{Ns} \alpha, d_i x_i \qquad (20)$$

3.3. Método de Entrenamiento

Para la fase de entrenamiento se propone construir dos escenarios para el entrenamiento y prueba de los modelos. Para el primer entorno de trabajo se toma la totalidad de los datos procesados en el apéndice 3.1.3, para este modo de trabajo la cantidad de muestras por categoría no está balanceada. En el segundo escenario se realiza un sub muestreo sobre la base de datos procesada, con el propósito de igualar el número de muestras patológicas al de muestras sanas. El objetivo tras esta división de entornos es comprobar que el desempeño de los modelos no presenta un sesgo de acuerdo a la cantidad de datos que tiene el grupo de entrada, evitando así la paradoja de la exactitud [50].

Los datos de ambos entornos se separan en porcentajes del 0.9 y 0.1 para el conjunto de validación y el de prueba, respectivamente. El proceso de entrenamiento-validación se realiza de manera simultánea a partir de las funciones del paquete de aprendizaje automático de Python. Este paquete permite el ajuste de parámetros de optimización acorde a una grilla de exploración, y el resultado del entrenamiento es el mismo de la validación cruzada [51].

El objetivo de implementar directamente una validación cruzada a 10 *folds* es agrupar el conjunto de validación en 10 subconjuntos diferentes. Una vez construido estos subconjuntos se entrena al modelo con los primeros 9 y se prueba con el último. El dato de dicha prueba se guarda y se prosigue a entrenar nuevamente el modelo cambiando el subconjunto de prueba sin permitir su repetición. Dicho proceso genera 10 pruebas con datos independientes los cuales se ponderan para obtener el resultado de la validación. La ventaja de trabajar directamente sobre la validación es que se descarta que los conjuntos de entrada tengan incidencia directa en el proceso de clasificación, debido al azar en el cálculo de las regiones de separación de los modelos [25] [52].

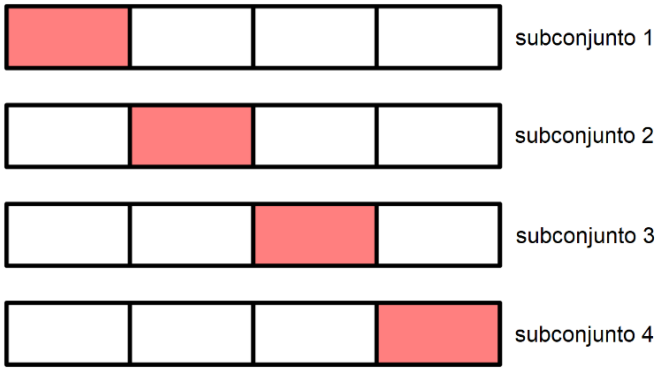

Figura 14. Esquema general de los datos en una validación cruzada [25].

El uso de validación cruzada con grilla puede generar inconvenientes logísticos debido a que los parámetros son ajustados directamente por el programador, más sin embargo se implementa una estrategia de grillas fijas y aleatorias con el fin de encontrar el conjunto de hiper parámetros óptimos para cada uno de los modelos. En esta estrategia se realizan numerosas exploraciones cortas. El objetivo de esta estrategia es acotar el hiperplano de decisión con diversas mallas, para aproximarse al mínimo global con distintas aproximaciones. Una vez explorados los diferentes espacios de decisión se prosigue con el diseño de mallas aleatorias, estas deben estar contenidas en las mallas fijas que presentaron el mejor desempeño con el fin de explorar los hiper espacios óptimos [53].

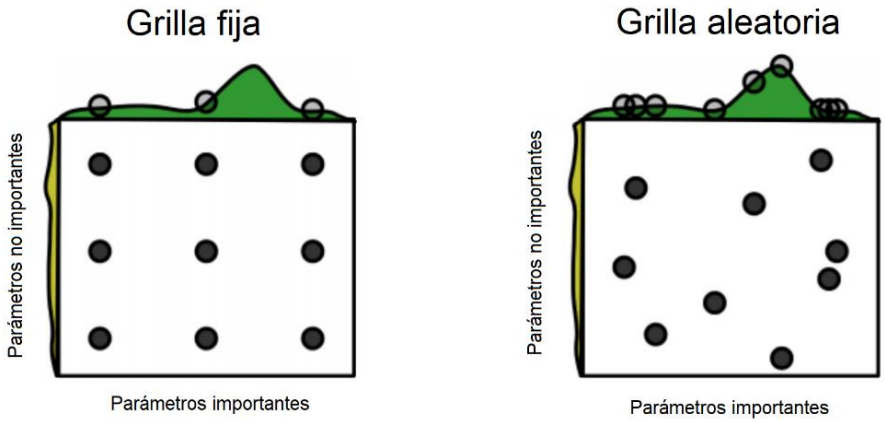

Figura 15. Modelos con grillas fijas y aleatorias [53] .

3.3.1.1. Parámetros de optimización

Basándose en el apartado 3.2, los hiper parámetros que se seleccionaron para el proceso de optimización se encuentran en la Tabla 3.

Modelo de bosque aleatorio	Número de estimadores: Número de árboles de decisión creados.
	Profundidad máxima: Longitud máxima que pueden tener las ramas.
	Mínimo número para división: El número mínimo de muestras que debe tener un nodo de prueba para dividirse.
	Mínimo número para hojas: El número mínimo de muestras para considerar un nodo una hoja pura.
Máquina de soporte vectorial	Factor C: Parámetro de regularización.
	Kernel: Selecciona el modelo Kernel a implementar.
	Gamma: Coeficiente máximo que pueden tomar las funciones Kernel.
Red neuronal artificial	Tamaño de la capa oculta: Número de neuronas en la primera capa oculta.
	Tipo de activación: Función de activación.
	Tasa de aprendizaje inicial: Valor inicial de la tasa de aprendizaje.
	Tasa de aprendizaje: Tasa de aprendizaje con valor fijo o adaptativo.

Tabla 3. Parámetros de optimización.

3.4. Análisis del desempeño

3.4.1. Medidas de desempeño en Test

Una vez entrenado los modelos se prueba su desempeño a partir de la: Sensibilidad (Sen), especificidad (Esp), desviación heurística o suma de producto (SP) y exactitud (Ex). Las medidas de sensibilidad y especificidad miden la capacidad de un estimador para predecir adecuadamente casos positivos y negativos respectivamente, mientras que la exactitud es el desempeño general de un estimador basado en la totalidad de las predicciones adecuadas [54]. La suma de producto o desviación heurística es una medida de desempeño que balancea de manera suave la relación que existe entre la sensibilidad y especificidad del modelo [55]. Dichas métricas se pueden apreciar de la ecuación 21 a la 24.

$$Ex = \frac{TN + TP}{TN + TP + FN + FP} \tag{21}$$

$$Sen = \frac{TP}{TP + FN} \tag{22}$$

$$Esp = \frac{TN}{TN + FP} \tag{23}$$

$$SP = \sqrt{\left(\frac{Sen + Esp}{2}\right) \cdot \sqrt{Sen \cdot Esp}} \tag{24}$$

En las ecuaciones anteriores TP, TN, FP y FN representan respectivamente el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos que genera un estimador [54]. Para analizar estadísticamente estas medidas se utiliza la validación cruzada para determinar su desviación, para posteriormente analizar el clasificador en el conjunto de prueba y comparar su desempeño [23].

3.4.2. Curva ROC

La curva de característica operativa del receptor (ROC) permite medir el desempeño de un modelo de predicción a partir de la probabilidad de que la muestra esté bien clasificada. Este cálculo se hace a partir de las funciones de densidad de probabilidad para la clasificación positiva y negativa de las muestras; a medida que aumente el traslape entre las curvas, las muestras tienen menor probabilidad de quedar bien clasificadas. En caso de que el traslape de curvas sea del 50%, la selección de categoría se considera un proceso de azar [30].

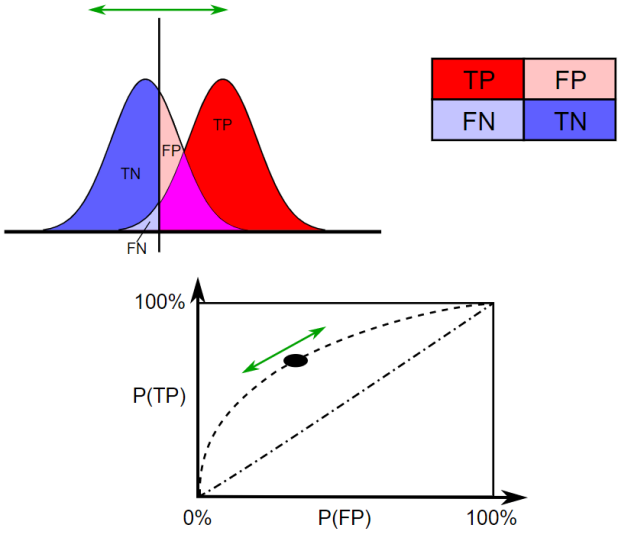

Figura 16. Curva ROC [56].

El proceso del cálculo de las funciones de probabilidad se realiza con la toolbox aprendizaje automático de python, la cual permite decidir si se retorna únicamente el valor de la clasificación o si se acompaña con el valor de densidad de probabilidad [57].

3.4.3. Análisis de relevancia de las características

Para la selección de características más relevantes se realiza una modificación al conjunto de datos de prueba. Esta consiste en asignarle a una característica el valor de su media con el objetivo de que los modelos no tengan la capacidad de diferenciar la tendencia de dicha característica, generando así un posible error de clasificación, en caso de que la variable tenga mucha relevancia en el proceso de predicción. Al realizar el procedimiento anterior en todas las características se analizan las medidas de desempeño en cada caso para determinar si la característica tiene relevancia o no en el proceso de clasificación [58] [59].

El último procedimiento de esta etapa experimental es entrenar nuevamente los modelos con las características principales y repetir los procesos de los apartados 3.4.1 y 3.4.2 para comparar todos los resultados.

4. RESULTADOS

En esta sección se muestran los resultados obtenidos en los diferentes experimentos que se hicieron con las bases de datos.

En la Tabla 4 se visualizan las medidas de desempeño para los 3 modelos de aprendizaje automático entrenados previamente con el método de grilla aleatoria en el grupo de prueba. Los modelos pertenecientes al entorno (N) representan al escenario de entrenamiento con datos desbalanceados entre clases, mientras que los modelos con el entorno (H) son aquellos que son entrenados en un escenario balanceado. Los resultados resaltados en negrillas indican el desempeño de los mejores modelos en cada entorno.

Modelo	***Sen***	***Esp***	***SumPro***	***Ex***
MLP_N	1.0	0.896	0.948	0.952
MLP_H	1.0	0.773	0.883	0.907
RF_N	**1.0**	**1.0**	**1.0**	**1.0**
RF_H	**1.0**	**0.955**	**0.978**	**0.981**
SVM_N	1.0	0.897	0.948	0.952
SVM_H	1.0	0.864	0.931	0.943

Tabla 4. Desempeño de los modelos en el conjunto de prueba, por medio de las métricas de desempeño.

De manera experimental el parámetro de desempeño que arroja los mejores resultados en la fase de exploración cambia de acuerdo al entorno de entrenamiento. Para el entorno de datos balanceados la exactitud se convirtió en el mejor parámetro de búsqueda, mientras que en el entorno de datos desbalanceados la suma de productos arrojó los mejores resultados en la exploración.

Como se puede observar en la Figura 17 y Figura 18, el desempeño de los mejores modelos de cada entorno en la fase de validación concuerda con los resultados del grupo de prueba, indicando que el resultado no es mejor que el azar.

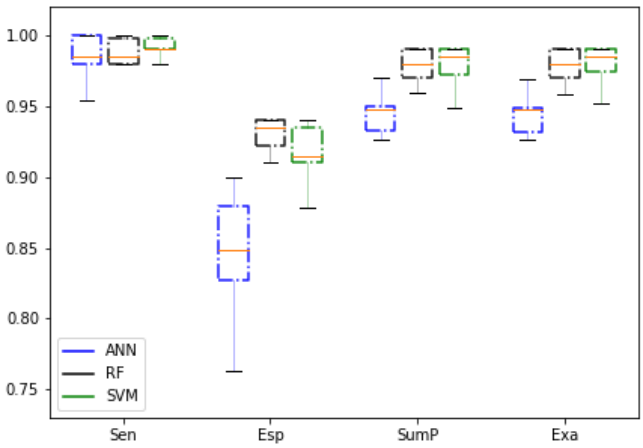

Figura 17. Desempeños de los modelos en la fase de validación para el entorno balanceado.

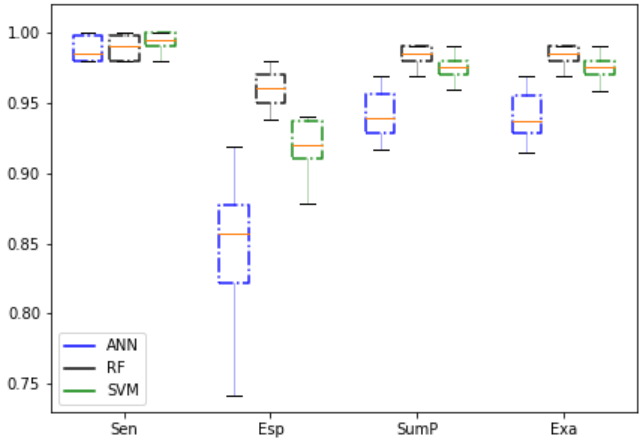

Figura 18. Desempeños de los modelos en la fase de validación para el entorno desbalanceado.

En la Figura 19 y en la Figura 20 se presentan las curvas ROC, realizadas a partir del conjunto de datos de prueba para los dos entornos de entrenamiento, los cuales muestran que el desempeño de los modelos en términos de la separación de las curvas de probabilidad es alto.

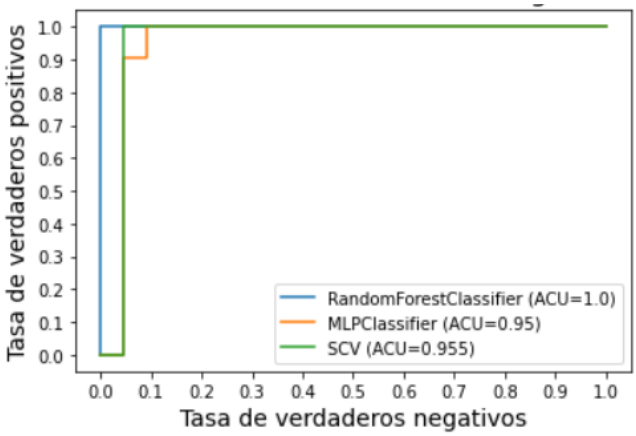

Figura 19. Curvas ROC para los modelos con el grupo de datos balanceados.

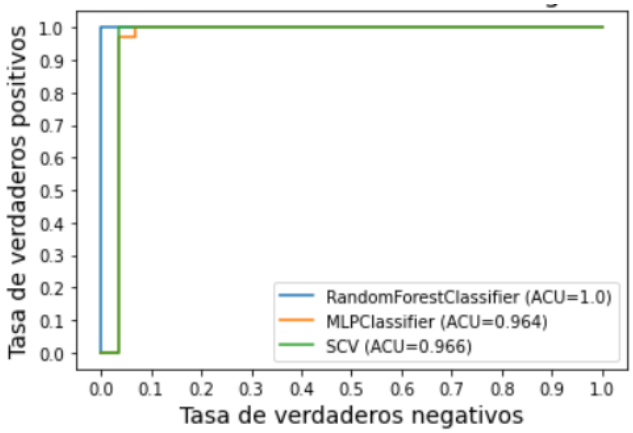

Figura 20. Curvas ROC para los modelos con el grupo de datos desbalanceados.

4.1. Análisis de Relevancia de Características

En la Tabla 5 se mide el desempeño de los seis modelos entrenados cuando se promedia una de sus características de entrada en el conjunto de prueba. Los resultados son evaluados en términos de la exactitud y la suma de productos para los entornos balanceado y desbalanceado respectivamente. Los parámetros resaltados en negrilla se ajustan a los cambios más significativos en el desempeño de los modelos.

Característica	Ex RF	Ex MLP	Ex SVM	SP RF	SP MLP	SP SVM
f_a	0.981	0.887	0.925	1.000	0.930	0.948
f_c	0.981	0.906	0.962	1.000	0.948	0.948
f_d	0.981	0.925	0.981	1.000	0.948	0.948
f_e	0.981	0.906	0.962	1.000	0.948	0.948
f_f	0.981	0.925	0.943	1.000	0.930	0.948
f_g	0.981	0.887	0.943	1.000	0.948	0.948
f_h	0.981	0.906	0.943	1.000	0.948	0.948
f_i	1.000	0.925	0.943	1.000	0.948	0.948
f_k	0.981	0.906	0.925	1.000	0.948	0.948
f_l	0.981	0.925	0.962	1.000	0.930	0.948
f_m	0.981	0.906	0.943	1.000	0.948	0.948
f_n	0.981	0.906	0.943	1.000	0.948	0.948
f_p	0.981	0.906	0.943	1.000	0.930	0.948
f_q	0.981	0.906	0.981	1.000	0.948	0.948
f_r	0.981	0.887	0.943	1.000	0.930	0.948
f_s	0.981	0.906	0.943	1.000	0.930	0.948
f_t	0.981	0.906	0.943	1.000	0.930	0.948
f_v	0.981	0.906	0.943	1.000	0.948	0.948
f_w	0.981	0.925	0.943	1.000	0.930	0.948
f_y	1.000	0.925	0.943	1.000	0.930	0.948
CaNet	**0.623**	0.906	0.962	**0.700**	0.948	0.965
DoHi	**0.396**	**0.415**	**0.415**	**0.000**	**0.000**	**0.000**
CoHi	1.000	0.925	0.962	1.000	0.965	0.965
Ali	0.981	0.925	0.943	1.000	0.912	0.948
Aro	0.981	0.906	0.943	1.000	0.948	0.948
Sul	0.981	0.925	0.962	1.000	0.930	0.948
Hdrx	0.981	0.925	0.962	1.000	0.930	0.948
Amid	1.000	0.906	0.906	1.000	0.948	0.948
Acd	0.981	0.887	0.943	1.000	0.948	0.948
Tiny	0.981	0.906	0.943	1.000	0.930	0.948
Small	0.981	0.887	0.943	1.000	0.930	0.948
Lrg	0.981	0.887	0.943	1.000	0.933	0.948
Plr	0.981	0.906	0.962	1.000	0.948	0.948
Aplr	1.000	0.925	0.943	0.983	0.948	0.948

Tabla 5. Desempeño de los modelos en el grupo de prueba, promediando una de sus características.

Al ponderar las características se encuentra que el número de donantes de hidrógeno afecta a los tres modelos de aprendizaje automático de una manera evidente, por lo tanto, se asume que es una característica de alta importancia para el proceso de clasificación. La segunda característica que se asume de alta importancia es la de carga neta, pese a que solo afecta de manera relevante al modelo de bosques aleatorios. La diferencia entre los desempeños es muy alta y por lo tanto selecciona como importante.

Para probar de manera experimental la importancia de estas características se realiza el reentrenamiento y prueba con los mismos grupos de datos, pero considerando únicamente las características 2 principales. Como se puede observar en la Tabla 6, el desempeño de todos los modelos presenta buenos valores, y si se compara con los resultados de la Tabla 4 se nota un comportamiento ligeramente superior en las medidas de desempeño.

Modelo	***Sen***	***Esp***	***SumPro***	***Ex***
MLP_N	1.0	0.931	0.965	0.967
MLP_H	1.0	0.863	0.931	0.943
RF_N	**1.0**	**1.0**	**1.0**	**1.0**
RF_H	**1.0**	**1.0**	**1.0**	**1.0**
SVM_N	1.0	0.931	0.965	0.967
SVM_H	1.0	0.954	0.977	0.981

Tabla 6. Desempeño de los modelos en el conjunto de prueba con únicamente dos características.

Para corroborar la fiabilidad de los modelos con únicamente dos características se presentan las gráficas de las medidas de desempeño y su desviación en la fase de validación, como se aprecia en la Figura 21 y la Figura 22.

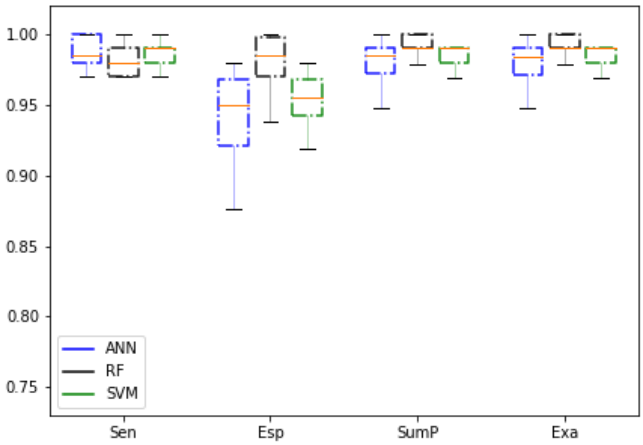

Figura 21. Desempeños de los modelos en la fase de validación para el entorno balanceado únicamente con las características relevantes.

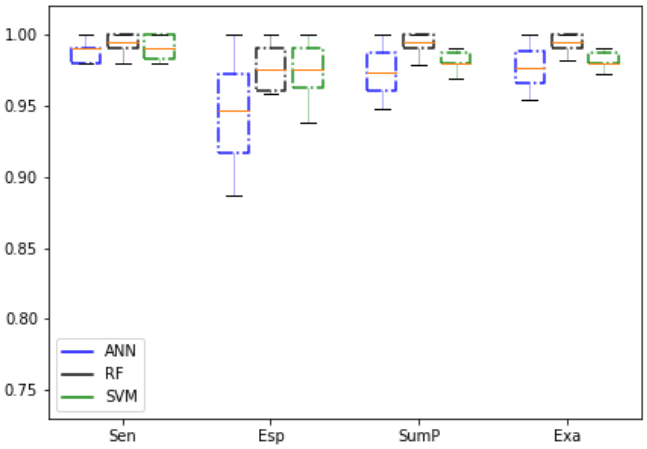

Figura 22. Desempeños de los modelos en la fase de validación para el entorno desbalanceado únicamente con las características relevantes.

En la Figura 23 y en la Figura 24 se presentan las curvas ROC, realizadas a partir del conjunto de datos de prueba para los dos entornos de entrenamiento con únicamente las características principales.

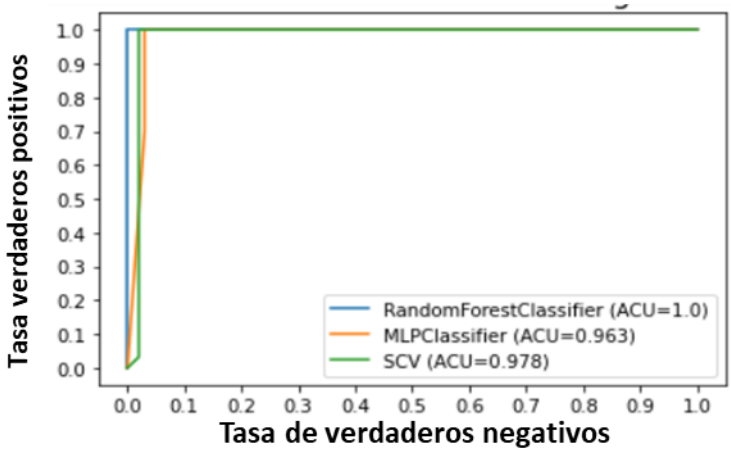

Figura 23. Curvas ROC para los modelos con el grupo de datos balanceados y solo dos características.

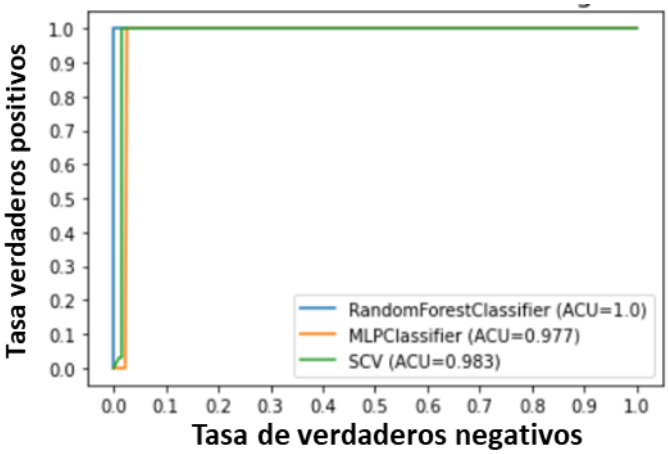

Figura 24. Curvas ROC para los modelos con el grupo de datos desbalanceados y solo dos características.

En la Tabla 7 se observa las medidas de desempeño que se utilizaron para exploración en los entornos balanceado y desbalanceado. Dicha tendencia muestra que el proceso de clasificación fue superior para los modelos entrenados únicamente con las características más relevantes.

Modelo	***SumPro_2C***	***SumPro_CC***	**∆ *SumPro***	***Ex_2C***	***Ex_CC***	**∆ *Ex***
MLP_N	0.965	0.947	**0.017**	0.967	0.951	**0.016**
MLP_H	0.930	0.882	**0.047**	0.943	0.905	**0.037**
RF_N	1.0	1.0	**0.0**	1.0	1.0	**0.0**
RF_H	1.0	0.977	**0.022**	1.0	0.981	**0.018**
SVM_N	0.965	0.947	**0.017**	0.967	0.951	**0.016**
SVM_H	0.977	0.930	**0.046**	0.981	0.943	**0.037**

Tabla 7. Comparación del desempeño de los modelos en términos de sus medidas de exploración con la totalidad y únicamente 2 de sus características.

5. DISCUSIÓN

El desempeño final de los modelos en la etapa de prueba y validación es acorde a lo esperado. En la literatura en el año 2019 se lanzó un modelo de detección de antígenos cancerígenos que tenía un desempeño alto a comparación de otras aproximaciones comerciales, como lo serian VaxiJen y TIminer. Este algoritmo presenta una sensibilidad, especificidad y exactitud del 0.89, 0.92 y 0.9 respectivamente en el conjunto de validación [23]. Mientras que los modelos entrenados con el criterio de la grilla aleatoria y depurados con la fase de relevancia obtuvieron medias de 0.97, 0.95, 0.96 en las mismas meticas en la fase de validación [53].

Si se compara los resultados en la fase de validación y prueba de este trabajo con los reportados en las aplicaciones de detección de péptidos VaxiJen y TIminer, se puede asegurar que todos los modelos de clasificación entrenados son capaces de predecir de manera adecuada si el antígeno es cancerígeno o no. Al tomar en cuenta que el desempeño del enfoque balanceado resultó inferior, se asume que al tener un menor número de muestras no permitió generalizar el espacio de decisión más óptimo, lo cual provocó que su tasa de predicción bajara [25].

Al analizar el desempeño de los modelos en la fase de validación con todas las características (Figura 21 y Figura 22) se observa que las muestras no cancerígenas son un poco más complejas de clasificar que las muestras patológicas. Esto indica que algunas muestras no patológicas presentan características fisicoquímicas similares a las cancerígenas. A partir de esta consideración y al observar la tendencia de los clasificadores se asume que los bordes rígidos de los bosques aleatorios permiten delimitar el espacio de decisión entre las muestras de una manera estricta, impidiendo que las muestras que estén cerca del borde de decisión afecten de manera significativa al modelo [46]. Para el caso de las máquinas de soporte vectorial y las redes neuronales al ser modelos que suavizan los espacios de decisión puede incrementar el error de clasificación para las muestras no patológicas [30] [34].

Al basar el desempeño final de un modelo de clasificación en términos de las medidas de estadísticas de predicción, se está desconociendo el estado estocástico de los modelos. Para contrarrestar esta inquietud se presenta la curva de característica operativa del receptor (ROC). Esta curva al medir el desempeño en términos de la probabilidad de que las muestras estén bien clasificadas, permite anexar un indicativo que de los algoritmos. Basándose en esta premisa y al considerar que el área bajo la curva de todos los modelos es superior al 90%, se considera que el aprendizaje de los modelos fue óptimo, y descartando así una incidencia del azar en estos modelos [30] [60].

Cuando se pondera una de las características de la matriz de entrada, la sección en la que cae dicha muestra en el espacio de decisión se ve afectada dependiendo del peso de la característica. Si la muestra al ponderarse no cambia de manera significativa la medida de desempeño final, como sucedió con la mayoría de los casos de la Tabla 5, significa que las características ingresadas no aportan mucha información o inclusive se convierten en ruido para el modelo [58]. A diferencia de las variables de carga neta y número de donantes de hidrógeno, una vez fueron ponderadas la tendencia de las medidas de desempeño disminuyó dramáticamente, por lo tanto, se corrobora que el mayor peso para la predicción recae únicamente en 2 de las 34 categorías. Basándose en este resultado, se reentrenan

los modelos con las características principales, con el fin de corroborar experimentalmente que la depuración de dichas características mejora o empeora el desempeño [59].

Como se puede observar en la Tabla 7, la acción de reentrenar los modelos con las características relevantes permitió a los modelos de ambos entornos presentar un incremento en sus métricas de desempeño para la fase de prueba. Como se debe validar el desempeño de estos modelos de menor dimensionalidad se analiza la media y su desviación en el conjunto de validación; los resultados consignados en la Figura 21 y la Figura 22 muestran que el desempeño de los modelos en la etapa de validación cruzada es homogéneo, indicando que la reducción de dimensionalidad de los registros no solo simplifico los modelos, sino que redujo la dispersión de los datos de acuerdo a su media [53]. Por último, al analizar el desempeño de las curvas ROC en la Figura 23 y Figura 24 se encuentra que el proceso de separación de las curvas de probabilidad presenta una métrica más alta, por lo tanto de descarta el azar del desempeño de estos modelos [30]. Todos estos resultados permiten afirmar que la mayoría de características aportaban ruido a los modelos, y problema presentaba una menor dimensionalidad de la que se esperaba al inicio de este trabajo [59]

6. RECOMENDACIONES Y TRABAJOS FUTUROS

Para complementar este trabajo se podría hacer desde varias perspectivas. El primer método sería incrementar el tamaño de la base de datos con el objetivo de generar modelos más robustos, pero manteniendo la restricción de las isoformas para evitar el sobreajuste a una línea celular especifica. El segundo método para complementar este trabajo es implementar técnicas de aprendizaje no supervisado, con el fin de utilizar aproximaciones heurísticas sin guía para identificar de una manera más empírica los patrones que diferencian los antígenos cancerígenos a los de cualquier otra especie. Para este procedimiento se proponen utilizar modelos de agrupamiento (clustering), mapas autoorganizados (SOM), modelos de k-medias (k-means), o inclusive comparaciones de estos tres, para determinar qué tipo de proceso heurístico es más óptimo para agrupar este tipo de muestras.

Un segundo enfoque para este tipo de experimentos es realizar un análisis parecido a las especies del complejo MHC II, con el fin de determinar qué condiciones fisicoquímicas son relevantes para la diferenciación de células sanas y patológicas. Este procedimiento no es excluyente de líneas celulares con cáncer. Otro posible trabajo sería analizar los epítopos provenientes de enfermedades de origen viral, bacterial, parasital, alérgicos y entre otros con el fin de conocer un perfil más amplio de las enfermedades y sus efectos sobre las células del cuerpo.

7. CONCLUSIONES

El desempeño general de los modelos presenta una buena tasa de predicción para los modelos con los dos entornos de entrenamiento. Pero el factor que ayudó de manera más positiva fue la correcta selección de medidas de desempeño. El modelo con el entorno desbalanceado presentaba en un inicio la paradoja de la exactitud, pero al implementar la suma de productos bajo un criterio experimental se empezó a encontrar los parámetros de optimización en un menor número de exploraciones. Otro de los factores que influyó de manera positiva en el desempeño de los algoritmos fue el número de muestras, debido a que, al entrenar con más datos, la media general en todos los modelos subió un poco. La medida de desempeño que fue inferior en todos los modelos fue la especificidad, indicando así que algunas proteínas de la categoría no cancerígena tienen propiedades fisicoquímicas parecidas al grupo de las cancerígenas.

El modelo de bosques aleatorios entre todos los modelos presento los mejores indicadores de desempeño tanto en validación como prueba, junto al hecho que su área bajo la curva ROC es del 100%, indicando así que la probabilidad de que el modelo haya clasificado bien las muestras del conjunto de prueba es alta.

Por medio del análisis de relevancia se encontró que el número de donantes de hidrógeno es la característica principal que separa los grupos, pero el uso de la carga neta le permite al modelo de bosque aleatorio diferenciar de manera casi perfecta las regiones de separación. Este hecho es un indicativo de que los procesos de expresión de antígenos en patologías relacionadas con cáncer, no tienen un patrón definido de las especies que pueden formar puentes de hidrógeno.

Acorde al alcance definido por los objetivos de este proyecto, se concluye que cada objetivo experimental fue culminado, basándose en los resultados presentados en este documento.

Referencias

[1] OMS, «Cáncer,» 12 septiembre 2018. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/cancerH.

[2] ASCO, «American Society of Clinical Oncology,» *Journal of Clinical Oncology,* pp. 212-222, 20 Febrero 2017.

[3] DANE, «Boletín Técnico,» 20 Diciembre 2019. [En línea]. Available: https://www.dane.gov.co/files/investigaciones/poblacion/bt_estadisticasvitales_IIItrim_2019pr-20-diciembre-2019.pdf.

[4] MinSalud, «PLAN DECENAL PARA EL CONTROL EN COLOMBIA,» 17 Marzo 2012. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/IA/INCA/plan-nacional-control-cancer.pdf.

[5] MinSalud, «Cáncer,» Enero 25 2020. [En línea]. Available: https://www.minsalud.gov.co/salud/publica/PENT/Paginas/Prevenciondel-cancer.aspx.

[6] A. L. L. T. G. Óscar, «Costos directos de la atención del cáncer,» *Cancerol,* vol. XX, nº 2, pp. 52-60, 2016.

[7] J. Greening, «The peptidome comes of age: Mass spectrometry-based characterization of the circulating cancer peptidome,» *Enzymes,* vol. 42, pp. 27-64, 2017.

[8] S. Lou, «Automated detection of radiology reports that require follow-up imaging using natural language processing feature engineering and machine learning classification,» *Journal of Digital Imaging,* vol. 33, nº 1, pp. 131-136, 2020.

[9] G. M. H. R. E. &. W. N. Cooper, La célula: Geoffrey M. Cooper y Robert E. Hausman, Madrid: Madrid: Marbán, 2014.

[10] B. S. D. Lodish, Biología Celular y Molecular, 5 ed., Madrid: Panamericana, 2016, pp. 590-630.

[11] S. I. Fox, FISIOLOGÍA HUMANA, vol. XII, New York: Pierce College, 2011, pp. 50-90.

[12] V. M. Saikumar P., Apoptosis and Cell Death. In: Allen T., vol. II, Boston: Molecular Pathology Library, 2009.

[13] C. W. K. Murphy, ImmunBiology, 9 ed., vol. I, Bostom: Garlad Science, 2017, pp. 3-35.

[14] W. R. Hanahan D, «Review: Hallmarks of Cancer,» 11 Junio 2011. [En línea]. Available: http://ez.urosario.edu.co/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ed selp&AN=S0092867411001279&lang=es&site=eds-live&scope=site.

[15] S. Gortzak-Uzan, «A Proteome Resource of Ovarian Cancer Ascites: Integrated Proteomic and Bioinformatic Analyses To Identify Putative Biomarkers,» *Journal of Proteome Research,* vol. VII, nº 1, p. 339–351, 2013.

[16] H. Mattsson, «Improved pan-specific prediction of MHC class I peptide binding using a novel receptor clustering data partitioning strategy,» *HLA Inmune Response Genetics,* pp. 1-6, 2016.

[17] R. J. A. N. M. N. H. V. I. Jurtz, «An introduction to Deep learning on biological sequence data – Examples and solutions,» *Oxford University Press,* 2017.

[18] J. &. V. A. &. C. S. &. D. N. Gauthier, «A Brief History of Bioinformatics. Briefings in Bioinformatics,» pp. 11-34, 2018.

[19] B. C. R. S. C. B. J. G. I. I. Pedro Larrañaga, «Machine learning in bioinformatics,» p. 86–112, 2006.

[20] S. R. J. Y. H. J. L. H. J. L. E. J. L. Kim, «Bioinformatic and metabolomic analysis reveals miR-155 regulates thiamine level in breast cancer,» *Cancer Letters,* vol. II, nº 357, p. 488–497, 2015.

[21] G. B. V. C. D. F. &. S. S. Musumarra, «A Bioinformatic Approach to the Identification of Candidate Genes for the Development of New Cancer Diagnostics,» *Biological Chemistry,* vol. II, nº 384, pp. 391-398, 2004.

[22] C. G. B. T. Y. C. S. Zhou. J, «Genetic and bioinformatic analyses of the expression and function of PI3K regulatory subunit PIK3R3 in an Asian patient gastric cancer library,» *Medical Genomic,* vol. V, nº 1, pp. 5-12, 2012.

[23] J. F. H. B. L. &. F. J. G. Beltrán Lissabet, «TTAgP 1.0: A computational tool for the specific prediction of tumor T cell antigens,» *Computational Biology and Chemistry,* nº 83, 2019.

[24] D. Chicco, «Ten quick tips for machine learning in computational biology,» *BioMed Central,* vol. I, nº 10, pp. 1-17, 2017.

[25] C. Bishop, PATTERN RECOGNITION AND MACHINE LEARNING, Primera ed., 2006, pp. 5-15.

[26] O. Theobald, Machine Learning for Absolute Beginners, vol. 2, Independently Published, 2018, pp. 10-12.

[27] A. M. D. A. A. L. Swan, «Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology,» *OMICS : a Journal of Integrative Biology,* vol. XVII, nº 12, pp. 595-610, 2010.

[28] J. Ulintz, «Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches,» *Molecular & Cellular Proteomics,* vol. V, nº 3, pp. 97-509, 2005.

[29] S. Jeet, «Machine Learning Biogeographic Processes from Biotic Patterns: A New Trait-Dependent Dispersal and Diversification Model with Model Choice By Simulation-Trained Discriminant Analysis,» *Systematic Biology,* vol. 65, nº 3, pp. 525-55, 2016.

[30] S. Zhang, «PromPDD, a web-based tool for the prediction, deciphering and design of promiscuous peptides that bind to HLA class I molecules,» *Journal of Immunological Methods,* pp. 476-489, 2020.

[31] K. D. R. N. Tomar, «Immunoinformatics: an integrated scenario,» *British Society of Immunology,* vol. 131, nº 2, p. 153–168, 2010.

[32] K. D. H. Youngmahn, «Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction,» *BMC Bioinformatics,* 2017.

[33] E. Loan, «Building MHC Class II Epitope Predictor Using Machine Learning Approaches,» *Springer,* vol. 1268, pp. 67-73, 2015.

[34] T. Alvarez, «NNAlign-MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved t-cell epitode,» *Molecular and Cellular Proteomic,* vol. 18, nº 12, pp. 2459-2477, 2019.

[35] M. M. H. &. B. M. Nosrati, «Introducing of an integrated artificial neural network and chou's pseudo amino acid composition approach for computational epitope-mapping of crimean-congo haemorrhagic fever virus antigens,» *International Immunopharmacology,* nº 78, 2020.

[36] S. X. Z. Weilong, «Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes,» *Computacional Biology,* pp. 1-28, 2018.

[37] R. D. A.D. Irini, «VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines,» *BMC Bioinformatics,* pp. 22-26, 2007.

[38] F. F. W. J. T. Elias, «TIminer: NGS data mining pipeline for cancer immunology and immunotherapy,» *Bioinformatics,* vol. 33, nº 19, p. 3140–3141, 2017.

[39] L. T. S. L. H. Olsen, «TANTIGEN: a comprehensive database of tumor T cell antigens,» *Cancer Immunol Immunother,* nº 66, 2017.

[40] M. S. O. J. D. S. M. S. C. J. W. D. S. A. P. B. Vita R, «The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res,» *PudMed,* 2018.

[41] C. L. S. S. R. L. G. a. L. Pommié, «IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties,» *IMGT,* nº 17, pp. 17-32, 2004.

[42] J.-L. C. M. K. L. B. V. A. &. P. V. FAUCHÈRE, «Amino acid side chain parameters for correlation studies in biology and pharmacology,» *International Journal of Peptide and Protein Research,* nº 32(4), p. 269–278, 2009.

[43] ImMunoGeneTics, «Amino acids,» 20 04 2004. [En línea]. Available: http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html.

[44] Q. B. A. G. K. Maricel, «Optimization of a new score function for the detection of remote homologs,» *Proteins: Structure, Function, and Bioinformatics,* vol. XLI, nº 4, pp. 498-503, 2000.

[45] k. S.-H. T. H. D. G. Eleni, «Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie,» *BMC Bioinformatics,* vol. XV, nº 16, 2014.

[46] R. T. J. F. T. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Segunda ed., Stanford, California: Springer, 2008.

[47] G. Thomas, «Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms,» 5 1995. [En línea]. Available: http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/tr-bias.pdf.

[48] SciKit Learn, «sklearn.ensemble.RandomForestClassifier,» 2019. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[49] S. Haykin, Neural Network and Learning Machines, Tercera ed., Boston: Person, 2009.

[50] F. J. &. P.-M. C. Valverde-Albacete, «100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox.,» *Plos One,* vol. IX, nº 1, 2014.

[51] Scikit Learn, «GridSearchCV,» 2 Enero 2020. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV.

[52] C. C. H. S. M. S. Ciampi A., «Recursive Partition: A Versatile Method for Exploratory-Data Analysis in Biostatistics,» *Springer,* vol. XXXVII, 1999.

[53] B. Y. B. James, «Random Search for Hyper-Parameter Optimization,» *Journal of Machine Learning Research,* vol. XIII, pp. 281-305, 2012.

[54] T. S. H. T. A. ENDO, «Comparison of Seven Algorithms to Predict Breast Cancer Survival,» *International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association,* vol. XIII, nº 2, pp. 11-16, 2008.

[55] M. J. D. A. F.F. João, «PLEURAL TUBERCULOSIS DIAGNOSIS BASED ON ARTIFICIAL,» *Sociedade Brasileira de Inteligência Computacional,* 2011.

[56] Kakau, «ROC curves,» 17 06 2010. [En línea]. Available: http://creativecommons.org/licenses/by-sa/3.0/.

[57] SciKit Learn, «sklearn.neural_network.MLPClassifier,» 20 Enero 2020. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier.predict_proba.

[58] C. L. D. I. Seixas JM, «Relevance criteria for variable selection in classifier design. In: International conference on engineering applications of neural networks,» de *International conference on engineering applications of neural networks*, London, 1996.

[59] F. S. T. R. C. P. J. V. K. A. L. S. J. M. &. M. F. C. Aguiar, «Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro,» *Medical & biological engineering & computing,* vol. 54, nº 11, pp. 1751-1759, 2016.

[60] M. L. F. W. N.A Obuchowski, «ROC curves in clinical chemistry: uses, misuses, and possible solutions.,» *Clenecal Chemistry,* vol. L, nº 7, pp. 18-25, 2004.

[61] N. Tkachev, «Flexible data trimming improves performance of global machine learning methods in omics-based personalized oncology,» *International Journal of Molecular Sciences,* vol. XXI, nº 3, 2020.

[62] L. Breiman, «Random forests,» *Springer Netherlands,* vol. XLV, nº 1, pp. 5-32, 2001.