Aprendizaje automático para la clasificación de cáncer de pulmón

Autor: Velázquez De Burnay, Babih
Institución: Universidad de A Coruña (UDC)
Localización: Universidad de A Coruña
Contacto: babih.velazquez@udc.es

Tutor: Puente Castro, Alejandro
Institución: Universidad de A Coruña (UDC)
Localización: Universidad de A Coruña
Contacto: a.puentec@udc.es

Tutor: Munteanu, Cristian Robert
Institución: Universidad de A Coruña (UDC)
Localización: A Coruña, España
Contacto: c.munteanu@udc.es

***Resumen*— La detección temprana del cáncer de células no pequeñas de pulmón (NSCLC) es uno de los grandes retos en la medicina clínica actual. Este estudio propone utilizar técnicas de aprendizaje automático para lograr este objetivo, utilizando datos RNA-Seq de Plaquetas Educadas por el Tumor (TEPs). Para ello, se identificaron muestras atípicas, se seleccionaron características relevantes para la detección del cáncer y se construyeron 11 modelos de aprendizaje automático con el fin de seleccionar el modelo con mayor acierto e interpretabilidad.**

Palabras Clave*— *Aprendizaje automático, cáncer, diagnóstico precoz, NSCLC.

I. Introducción

El cáncer ocupa la segunda posición como causa de fallecimiento en países desarrollados y en aquellas en vías de desarrollo, precedido únicamente por cardiopatías y enfermedades infecciosas [1]. En cuanto al cáncer de pulmón de células no pequeñas (NSCLC por sus siglas en inglés, *Non-small cell lung cancer*), representa aproximadamente el 13 % de todos los nuevos diagnósticos de cáncer. En la actualidad, este tipo de cáncer de pulmón destaca como una de las enfermedades con mayor tasa de mortalidad cada año [2]. Los tipos más comunes de NSCLC [3] son:

- **Carcinoma de Células Escamosas**: Constituye entre el 25 % y el 30 % de los casos de NSCLC y suele encontrarse cerca de los bronquios, en el centro de la cavidad torácica. Asociado comúnmente con la exposición al humo de tabaco.
- **Adenocarcinoma**: Representa el 40 % de todos los casos de NSCLC y generalmente se localiza en las regiones externas del pulmón. Existe una variante rara llamada carcinoma broncoalveolar (BAC) que se disemina a lo largo de todo el pulmón y tiene una incidencia creciente a nivel mundial, incluso en no fumadores.
- **Carcinoma de Células Grandes**: Conformando del 10 % al 15 % de los casos de NSCLC, es de crecimiento rápido y puede manifestarse en cualquier parte del pulmón.

El NSCLC está asociado a diversos factores de riesgo, siendo el más destacado el consumo de tabaco. El tabaquismo es el principal factor de riesgo del cáncer de pulmón y está relacionado con el 80-90% de estos cánceres [4]. La exposición a sustancias químicas como el amianto,

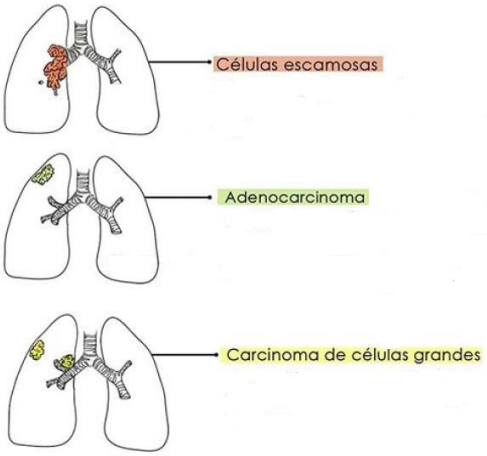

Fig. 1: Tipos de cáncer de NSCLC.

arsénico y níquel, así como al radón, un gas radiactivo liberado naturalmente a partir de ciertos tipos de suelos y rocas, también se asocian con un mayor riesgo [5]. La predisposición genética, exposición a agentes cancerígenos ambientales como la contaminación del aire, enfermedades pulmonares, radioterapia torácica previa y edad avanzada son factores adicionales que contribuyen al riesgo [6]. Es importante destacar que la presencia de estos factores no garantiza el desarrollo de NSCLC, ya que la interacción compleja de factores genéticos, ambientales y de estilo de vida contribuye a la variabilidad en la susceptibilidad de los individuos a este tipo de cáncer [7].

Desde el inicio del desarrollo de un cáncer de pulmón, hasta su diagnóstico mediante los métodos convencionales, puede transcurrir aproximadamente un año [8]. Por lo general, los primeros síntomas de esta enfermedad pueden ser confundidos fácilmente con problemas infecciosos o, en el caso de fumadores, con otras complicaciones asociadas al consumo de tabaco [9].

Las pruebas actuales para confirmar el diagnóstico de cáncer de pulmón se basan en técnicas que involucran el uso de radiación, como las radiografías, o métodos invasivos, como las punciones, además, estas metodologías suelen conllevar un alto coste [10]. En este contexto, se están llevando a cabo diversos estudios sobre un método no invasivo para la detección temprana del cáncer de pulmón: las biopsias líquidas. Este tipo de biopsias, obtenidas a partir de una muestra de sangre, analizan las plaquetas en busca de genes relacionados con el cáncer [11].

Aunque las plaquetas son generalmente conocidas por su función en la cicatrización de heridas, su interacción con las células tumorales permite la transferencia de biomoléculas, que a su vez estas sirven de biomarcador. Se dice que estas plaquetas están "entrenadas por el tumor" (TEP por sus siglas en inglés, "*tumor-educated blood platelets*"). La expresión genética de estos biomarcadores se puede cuantificar mediante técnicas como RNA-Seq o MicroArrays [12].

La detección temprana del cáncer de pulmón presenta desafíos significativos cuando se utilizan métodos clásicos, como radiografías o análisis bioquímicos [13]. Estos análisis a menudo vienen precedidos de la presencia de síntomas físicos, lo que puede llevar a diagnósticos tardíos y limitar las opciones de tratamiento efectivas. El diagnóstico precoz mediante técnicas mínimamente invasivas se considera crucial en el diagnóstico del cáncer. En este caso, mediante una simple muestra de sangre rutinaria, se podría realizar un diagnóstico de cáncer de pulmón a través de distintos algoritmos de aprendizaje automático [14]. Estos estudios son de gran relevancia en la lucha contra el cáncer, ya que la detección en etapas tempranas aumenta significativamente las tasas de supervivencia [15].

En contraste, la aplicación de inteligencia artificial (IA) [16] en el ámbito biomédico emerge como una alternativa prometedora, gracias a su capacidad para mejorar el diagnóstico y tratamiento de enfermedades mediante el análisis de grandes volúmenes de datos médicos. Al mismo tiempo, puede acelerar la investigación médica al identificar biomarcadores, optimizar procesos clínicos y predecir riesgos de enfermedades [17. La abundancia de casos para estudiar y la necesidad crítica de un diagnóstico precoz para mejorar las opciones de tratamiento hacen del NSCLC un problema relevante para la ciencia de datos. Los algoritmos de aprendizaje automático [18] poseen las siguientes características fundamentales: son capaces de aprender a partir de los datos, identificar patrones y tomar decisiones sin ser explícitamente programadas para cada tarea específica [19]. Estos algoritmos pueden analizar eficientemente patrones complejos en grandes conjuntos de datos, debido a su capacidad para procesar información rápidamente y encontrar relaciones ocultas en los datos. [20], permitiendo la identificación temprana de indicios sutiles de actividad génica asociada con el cáncer de pulmón [21]. La expresión génica se refiere a la forma en que los genes se activan o desactivan, y los cambios en estos patrones pueden indicar la presencia de células cancerosas [22].

Es por ello por lo que, el trabajo desarrollado en este estudio constituye una contribución en el ámbito de la detección precoz del cáncer de pulmón. La relevancia de este estudio radica en la aplicación de técnicas de inteligencia artificial y aprendizaje automático para identificar y seleccionar los genes más relevantes asociados con la detección temprana de esta enfermedad. A través de un enfoque, se busca construir un modelo robusto y eficiente capaz de clasificar el cáncer de pulmón con precisión, proporcionando así una herramienta de utilidad para la mejora de los protocolos de diagnóstico.

El contenido de las siguientes secciones queda resumido a continuación. En la Sección II se muestra el estado del arte actual. En la Sección III los fundamentos del estudio. En la Sección IV se detallan los modelos propuestos. En la Sección V se presentan y discuten los resultados. Finalmente, en la Sección VI se exponen las conclusiones obtenidas de este estudio y los futuros trabajos propuestos.

II. Estado de la Cuestión

A lo largo de las últimas dos décadas han surgido y se han empleado un gran número de técnicas para la detección del cáncer de pulmón. Si bien las más utilizadas son las pruebas histológicas, en la actualidad se están desarrollando modelos de aprendizaje automático y aprendizaje profundo para la ayuda en el diagnóstico de la enfermedad [23]. Existen distintas técnicas de abordaje al problema de la detección de cáncer de pulmón, el principal es mediante el análisis de radiografías. Este enfoque se ve realizado en el artículo publicado por Duan *et al.*, en 2020 de nombre "*Development of a machine leraning-based multimode diagnsosis system for lung cancer*" [24] donde se utilizan tres modelos de clasificación (árbol de decisión, neural net y máquinas de soporte de vectores) [25] para la predicción de la enfermedad.

Recientemente, se ha visto aumentado el uso de aprendizaje profundo para la detección de cáncer de pulmón a través de radiografías, en virtud de su eficacia a la hora de extraer características complejas, la captura de detalles a diferentes escalas y la capacidad de generalización del modelo. En el estudio publicado por Santos en 2023 [26] se utilizan técnicas de metaheurísticas y aprendizaje profundo para la detección de la enfermedad en su primer estadio, consiguiendo así mayores tasas de detección que los expertos radiólogos.

Sobre el mismo enfoque de análisis de radiografías, existe la posibilidad de unir modelos de aprendizaje automático y aprendizaje profundo, a estos sistemas se les conoce como sistemas híbridos. En un reciente estudio publicado por Vasavi y Sruthi en 2023 [27] se utiliza un sistema híbrido CNN-SVM, actuando la red convolucional (CNN) [28] como un extractor de características y SVM [29] como clasificador binario. Estudios similares basándose en sistemas híbridos como el de Vidhya y Mirnalinee publicado en 2022 [30] o la conferencia "*A hybrid feature based model development for computer aided diagnosis of lung cancer"* [31] abren un nuevo frente para la ayuda en el diagnóstico del cáncer de pulmón.

En este contexto, debido a las limitaciones existentes para realizar radiografías rutinarias a la población se dificulta la detección precoz de la enfermedad. A diferencia de los trabajos descritos, se plantea en este estudio conseguir un modelo altamente preciso para la ayuda en el diagnóstico precoz de NSCLC. El cambio para conseguir este objetivo es el uso de datos RNA-Seq extraídos de los pacientes mediante una muestra sanguínea. El aporte para el estado del arte existente consiste en la selección de técnicas de extracción de características adecuadas y la construcción de modelos de aprendizaje automático ajustados a este tipo de datos.

III. Fundamentos

Como se ha expuesto anteriormente, poder facilitar la detección temprana del cáncer de pulmón es uno de los retos actuales en el campo de la medicina clínica. Para explorar la posible existencia de genes en una biopsia líquida que muestren una expresión diferencial durante el crecimiento celular anormal, se emplean diversas herramientas y modelos. En particular, se analizan las TEP como indicadores

potenciales, con el objetivo de identificar patrones de expresión genética que puedan ser reveladores de un aumento en la actividad celular más allá de los límites normales. Para ello, en este estudio se va a realizar un estudio comparativo de distintas técnicas de procesado de datos de RNA-Seq, un análisis de distintas técnicas de selección de características y la optimización de hiperparámetros de diversos algoritmos de aprendizaje automático para maximizar el acierto en la clasificación de pacientes con cáncer, con el fin de desarrollar un sistema de ayuda a la toma de decisiones para el diagnóstico médico de la enfermedad.

A. Conjunto de Datos

Este estudio se basa en los datos del repositorio del NCBI, con el identificativo de acceso GSE89843 publicados en el artículo publicado por Best *et al*., en 2017 [11] en la revista Cell. A través del identificador tenemos acceso a dos matrices de datos. La primera matriz corresponde a datos de RNA-Seq de 4722 genes para cada uno de los 779 pacientes, de los cuales:

- 402 pertenecen a muestras recogidas de TEP con cáncer de pulmón.
- 377 pertenecen a muestras de pacientes sin cáncer de pulmón.

Y, por otra parte, la segunda matriz de datos recopila información de interés de los pacientes, como son el género, edad, si padece NSCLC, entre otras de menor relevancia.

Para el procesamiento de los datos, se propone seguir el siguiente flujo de trabajo:

- Preprocesado de los datos;
- Construcción y optimización de los hiperparámetros de distintos modelos de aprendizaje automático;
- Análisis de los resultados.

B. Preprocesado de los datos

El análisis exploratorio proporciona una comprensión de la naturaleza y la calidad de los datos antes de realizar análisis más avanzados. Un entendimiento detallado de los posibles valores de los elementos, la ausencia de datos en una variable o las relaciones entre variables facilita la selección de conjuntos de datos que posibiliten la extracción posterior de características relevantes en el modelo. Los conjuntos de datos en bruto requieren un preprocesamiento para su implementación en los modelos, permitiendo así una comprensión más profunda del problema y de las potenciales soluciones.

La normalización de los datos es una etapa esencial y crítica en este tipo de estudios. Al realizar análisis de RNA-Seq, las diferencias en eficiencia de secuenciación, las bibliotecas empleadas para su cuantificación y otras variables técnicas pueden introducir errores significativos [32]. La normalización mitiga estos sesgos al ajustar los datos, permitiendo así una interpretación más precisa de las diferencias biológicas reales.

En la evaluación de la calidad de las muestras, es crucial llevar a cabo una detección de valores atípicos. La detección correcta de estos valores es crucial para garantizar la integridad y validez de los análisis de expresión génica. La presencia de valores atípicos puede surgir de variaciones técnicas, contaminación de datos o incluso eventos biológicos inesperados, y su inclusión sin un tratamiento adecuado puede distorsionar significativamente los resultados y conclusiones del estudio.

Una vez superada la etapa de eliminación de muestras atípicas conviene reducir la dimensionalidad del conjunto de datos, con esa finalidad se procede a realizar la selección de características relevantes. En primer lugar, realizar la selección de características contribuye a mejorar la eficiencia computacional al disminuir la dimensionalidad del conjunto de datos, lo que resulta en tiempos de entrenamiento y evaluación de modelos más rápidos. Al mismo tiempo, evitamos generar un sobreajuste en los modelos de aprendizaje automático al simplificar la complejidad, promoviendo así su robustez y capacidad de generalización. Otro aspecto a considerar es la mejora de la interpretabilidad de los resultados, un conjunto de datos con menos características es más fácil de entender, facilitando la interpretación de los modelos. Finalmente, la eliminación de características redundantes o irrelevantes es otro beneficio clave, ya que se focaliza en las variables más influyentes para el problema en cuestión, evitando así la introducción de ruido.

C. Construcción y Optimización de Modelos

En la construcción de modelos es fundamental considerar la utilización de diversas familias de modelos y la optimización de sus hiperparámetros por diversas razones. En primer lugar, cada familia de modelos tiene sus propias características y supuestos subyacentes sobre la estructura de los datos. Al emplear diferentes familias de modelos, podemos explorar diferentes enfoques para abordar el problema, lo que nos permite capturar distintos tipos de relaciones entre las características y la variable objetivo. Otro aspecto importante es la reducción del sesgo de modelo. Esto ocurre cuando un modelo no puede capturar la verdadera relación entre las características y la variable objetivo debido a suposiciones incorrectas generadas en la fase de entrenamiento. Al utilizar una variedad de modelos, mitigamos este riesgo al poder seleccionar modelos con mejor rendimiento. La interpretación y explicabilidad también son consideraciones clave. Cada familia de modelos proporciona diferentes grados de interpretabilidad y explicabilidad. Hay veces que una fácil interpretabilidad del modelo puedo abrir nuevas líneas de investigación. Esto es debido a que podemos ver que variables asume el modelo como más importantes a la hora de la toma de la decisión.

De igual forma, al optimizar los hiperparámetros de cada modelo, mejoramos su capacidad de generalización a nuevos datos, evitando así el sobreajuste y garantizando un mejor rendimiento en diferentes escenarios.

D. Análisis de los resultados

Realizar una comparación múltiple de modelos nos permite seleccionar el mejor modelo para nuestro problema específico. Al evaluar y comparar múltiples modelos, podemos identificar cuál de ellos ofrece el mejor rendimiento predictivo. Además, nos permite establecer un punto de referencia o estándar de rendimiento. Esto desarrolla el estado del arte, siendo una información valiosa para futuros proyectos que trabajen con características similares a las nuestras, permitiendo una mejor toma de decisiones de selección de modelos y una mayor eficacia en el desarrollo de nuevos modelos.

IV. Modelos Propuestos

En esta sección se abordará el desarrollo y evaluación de los modelos propuestos en este estudio. Se describirá el proceso seguido, desde el preprocesado de los datos hasta la construcción de los modelos de aprendizaje automático, así como la exposición detallada de las métricas de rendimiento empleadas para evaluar la bondad de los modelos.

A. Preprocesado de los datos

Se procede en primer lugar a realizar un análisis de exploratorio de los datos. Se exponen las características clínicas de las muestras, esto es, la edad de los pacientes, su género y la proporción de afectados por NSCLC.

Acto seguido, se ejecuta la estandarización de los datos. La estandarización implica la transformación de los datos para que sigan una distribución normal con una media de cero y una desviación estándar de uno. Con la siguiente descripción matemática:

$$z = \frac{x - \mu}{\sigma}$$

Donde: z es el valor estandarizado, x es el valor original, μ es la media de la distribución y σ es la desviación estándar de la distribución.

Esto facilitará la detección de señales biológicas que podrían ocultarse debido al a diferencia de escalas y potencia la capacidad de realizar análisis diferenciales precisos.

Para la detección de valores atípicos se emplean cuatro técnicas, cada una con distintas propiedades y características. Con este fin, se aplican cuatro técnicas de detección:

1) **Teorema de Chebyshev**: Utiliza la desigualdad de Chebyshev [33] para identificar valores atípicos basándose en la relación entre la media y la desviación estándar. Una de sus mayores ventajas es que durante el proceso de cálculo no necesitamos asunciones sobre la distribución de los datos. Descrito formalmente:

$$P(|X - \mu| \leq k\sigma) \geq \left(1 - \frac{1}{k^2}\right)$$

Donde: X representa los datos. μ es la media de los datos. σ es la desviación estándar de los datos. k el número de desviaciones estándar respecto a la media.

2) **Isolation-Based Anomaly Detection:** Se basa en la construcción de árboles de decisión, asignando un valor de anomalía en función de la profundidad en la que se encuentra una observación en los [34]. Destaca por su capacidad para manejar grandes conjuntos de datos y su eficacia en la identificación de valores atípicos en dimensiones múltiples. Descrito formalmente:

$$S(x) = 2^{-\frac{E(h(x))}{c(n)}}$$

Donde: x representa el dato que se está evaluando. E(h(x)) es la profundidad esperada x en un árbol de decisión aleatorio. c(n) es una constante que se calcula utilizando la media de las profundidades esperadas de los árboles aleatorios.

3) **Local Outlier Factor (LOF***):* Evalúa la densidad local alrededor de cada observación y compara esta densidad con la de sus vecinos, identificando así valores que tienen una densidad significativamente más baja que sus vecinos [35]. Descrito formalmente:

$$LOF(x) = \frac{\sum_{y \in N(x)} \frac{dist(x, y)}{dist(y)}}{|N(x)|}$$

Donde: *x* es el punto de datos para el que se está calculando la puntuación de anomalía. *N*(*x*) es el conjunto de puntos de datos vecinos de *x*. dist(*x*,*y*) es la distancia máxima entre *x* e *y*, o la distancia entre *x* e *y si y es mayor que la distancia máxima entre x* y sus vecinos más cercanos. *N*(*x*) es el número de vecinos de *x*.

4) **Support Vector Data Description:** Utiliza una máquina de soporte vectorial (SVM) para construir un límite que encapsula la mayoría de los datos, identificando como atípicos aquellos que quedan fuera de este límite. Además, esta técnica es especialmente útil cuando se tiene un conjunto de datos predominantemente no atípico. Descrito formalmente:

$$\min_{w, \xi, \rho} \quad \frac{1}{2}|w|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i$$

Sujeto a las siguientes restricciones:

$$w^T \phi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1,2,\dots,n$$

Donde: w es el vector de peso, ξ es la variable de holgura. ρ es la distancia desde el hiperplano al origen en el espacio de características. ν es el parámetro de regularización que controla el número esperado de valores atípicos en el conjunto de entrenamiento.

De igual forma, para la selección de características se aplican cuatro técnicas distintas:

1) **Selección de Características basado en la varianza:** Se basa en métodos estadísticos para evaluar la relevancia de cada característica individual. Utiliza métricas como la puntuación F de ANOVA [36] para clasificar y seleccionar las mejores características en función de su relación con la variable objetivo. Descrito formalmente:

$$Seleccion(k) = \arg\max_{X} \quad F(X)$$

Donde: k es el número de características seleccionadas. X representa un subconjunto de características. F(X) es el valor estadístico F de ANOVA.

2) **Modified Variance Reduction Score (mVRS):** mVRS se centra en la modificación de la métrica de reducción de varianza para evaluar la importancia de las características [37]. Busca preservar la varianza de las características más relevantes y reducir la de las menos relevantes. Descrito formalmente:

$$mVRS(X) = \frac{1}{|S|}\sum_{i \in S} \frac{\sigma_i}{\sigma_{max}}$$

Donde: X es el conjunto de características. S es un subconjunto de características. σ_i es la varianza de la característica i. σ_{max} es la varianza máxima entre todas las características en el conjunto de datos.

3) **Random Forest Feature Selection:** Random Forest utiliza múltiples árboles de decisión y evalúa la importancia

de cada característica mediante la disminución en la precisión del modelo al omitir dicha característica [38].

4) **Double Radial Basis Function Kernel:** Es un método de selección de características basado en kernels [39]. Utiliza dos kernels de función de base radial (RBF) para evaluar la similitud entre instancias y seleccionar aquellas características que maximizan la separabilidad entre las clases en un conjunto de datos. Descrito formalmente:

$$K\ (x_i,\ x_j) = \alpha\, K_1(x_i,\ x_j;\ \gamma_1) + (1 - \alpha)\, K_2(x_i,\ x_j;\ \gamma_2)$$

Donde: $x_i,\ x_j$ son vectores de características de dos puntos de datos en el espacio de entrada. K_1 y K_2 son los kernels RBF con parámetros de escala γ_1 y γ_2 respectivamente. α es un peso que se aplica al primer kernel.

B. Construcción de modelos

En esta etapa, se procede a la construcción y optimización de los hiperparámetros de los modelos de aprendizaje automático**.** Es importante considerar una variedad de modelos para evaluar su rendimiento y determinar cuál se adapta mejor a los datos y a los objetivos del problema. Cada modelo tiene sus propias características, lo que los hace adecuados para diferentes tipos de problemas y conjuntos de datos**.** Es por ello que se emplean 11 modelos distintos para evaluar su rendimiento, conociendo así que modelo se adapta mejor a las características de nuestro problema. Los modelos empleados son: k-Nearest Neighbors (kNN), Proceso Gaussiano, Árbol de decisión, Random Forest, Neural Net, AdaBoost, Naive Bayes, SVM, Gradient Boosting, Regresión logística y Bagging. Estos modelos se agrupan por familias:

1) **Lazy Learning:**

Es un enfoque en el aprendizaje automático donde el modelo no generaliza desde un conjunto de datos de entrenamiento hasta que se presenta una instancia de prueba [40].

A esta familia pertenece kNN. kNN es uno de los clasificadores más populares y fáciles de entender, es un método de clasificación no paramétrico el cual dado una nueva instancia busca en los datos del conjunto de entrenamiento aquellas muestras que tengan las variables más parecidas para asignarle la clase más frecuente [41].

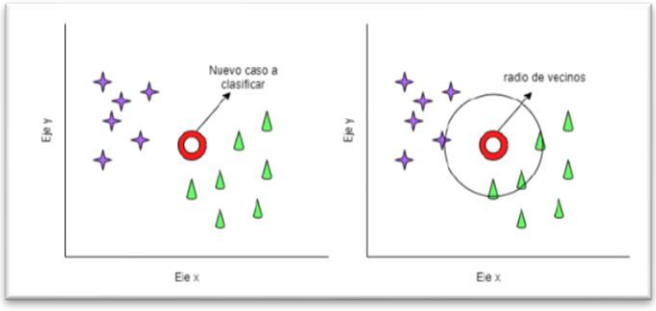

Fig. 2: Clasificador kNN.

2) **Modelos basados en árboles de decisión:**

Los modelos basados en árboles de decisión construyen una estructura de árbol que representa decisiones sobre cómo dividir los datos en diferentes grupos. Cada nodo en el árbol representa una pregunta sobre una característica, y las ramas representan posibles respuestas a esa pregunta [42]. Estos modelos son fáciles de interpretar y al mismo tiempo son capaces de capturar relaciones no lineales en los datos. Un árbol de decisión consiste en dividir iterativamente el conjunto de datos en subconjuntos más pequeños basados en características específicas, con el objetivo de crear un modelo predictivo en forma de árbol. Este proceso se repite hasta alcanzar una condición de parada, como la pureza de los subconjuntos o la profundidad máxima del árbol.

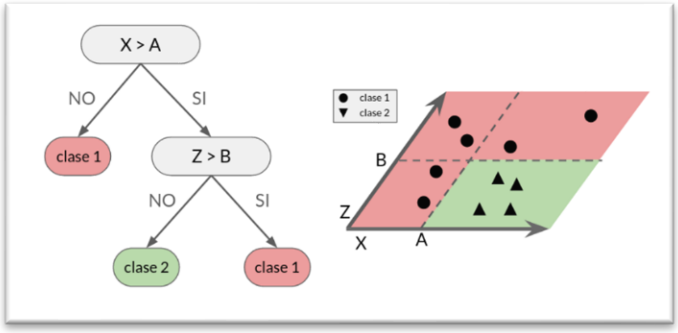

Fig. 3: Árbol de decisión

3) **Ensembles**:

Los modelos de tipo ensemble combinan múltiples modelos individuales para mejorar la precisión y la robustez del modelo final [43]. Estos modelos se basan en la idea de que la combinación de múltiples modelos puede proporcionar una mejor generalización y reducir el sesgo y la varianza del modelo. Dentro de esta familia tenemos:

Random Forest: Es un ensemble basado en modelos de árboles de decisión. Se construyen múltiples árboles de decisión independientes, cada uno entrenado con un subconjunto aleatorio de características y datos de entrenamiento. Luego, las predicciones de cada árbol se promedian para obtener la predicción final [44].

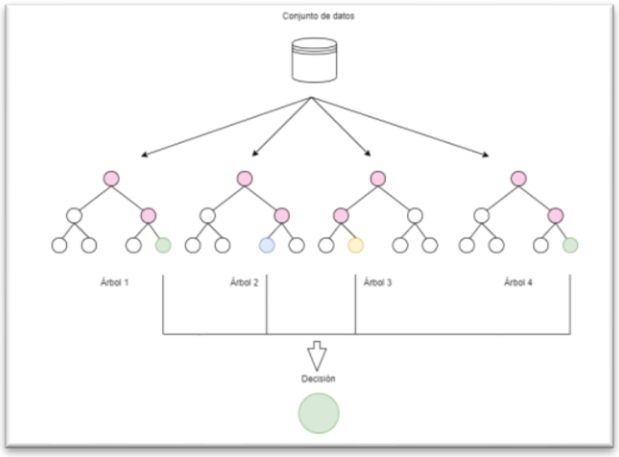

Fig. 4: Clasificador Random Forest.

Bagging: Este modelo se basa en entrenar múltiples modelos independientes utilizando diferentes subconjuntos de datos generados mediante muestreo con reemplazo del conjunto de entrenamiento original [45]. Las predicciones de cada modelo se votan para obtener la predicción final.

Gradient Boosting: Construye modelos de manera secuencial, donde cada modelo se enfoca en corregir los errores de los modelos anteriores. Se entrenan modelos “débiles” sucesivos, y cada uno se ajusta para minimizar la

función de pérdida en el gradiente descendente del error residual del modelo anterior [46].

AdaBoost: El último modelo de tipo ensemble, este construye una secuencia de modelos de manera adaptativa, asignando pesos a cada instancia de entrenamiento y ajustando los pesos en cada iteración para enfocarse en las instancias mal clasificadas por los modelos anteriores. Los modelos débiles se combinan ponderadamente para obtener la predicción final, donde los modelos con mejor rendimiento tienen más peso en la predicción final [47]

4) **Redes Neuronales:**

Consisten en capas de neuronas artificiales interconectadas, donde cada neurona realiza operaciones matemáticas en las entradas recibidas. Cada neurona toma entradas, realiza una combinación lineal de estas entradas ponderadas por los pesos, agrega un sesgo, y luego aplica una función de activación no lineal para producir una salida [48]. A este modelo se le conoce como Neural Net.

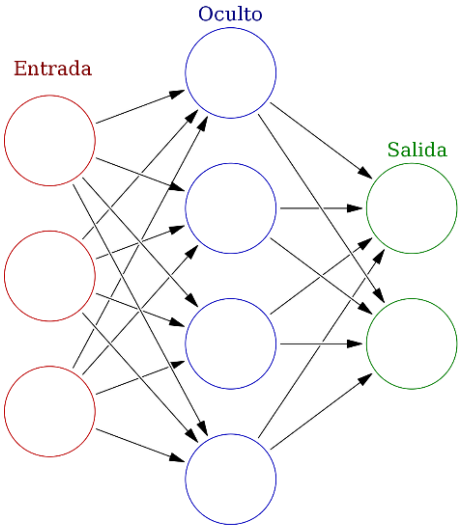

Fig. 5: Modelo Neural Net.

5) **Modelos Basados en Kernel:**

Los modelos basados en kernel transforman los datos de entrada en un espacio de características de mayor dimensión utilizando funciones de kernel [49]. Esto permite modelar relaciones no lineales entre las variables de entrada y salida. Dentro de esta familia tenemos a las máquinas de soporte de vectores (SVM).

El funcionamiento de las SVM consiste en encontrar un hiperplano de separación entre las instancias de dos categorías de datos distintas [50]. Su peculiaridad reside en que, de todos los posibles hiperplanos separadores, escoge aquel que tiene un margen de separación máximo. Este margen de separación se calcula como la distancia máxima entre las instancias que se encuentran en la frontera. El hiperplano o conjunto de hiperplanos generados poseen una alta dimensionalidad. Esta dimensionalidad se genera con las funciones kernel mediante transformaciones matemáticas no lineales, pueden ser de tipo polinómico, sigmoidal, lineal y RBF [51]. Estas propiedades hacen que SVM sea un modelo útil en datos de alta dimensionalidad debido a su capacidad para encontrar límites de decisión óptimos en espacios de características complejos.

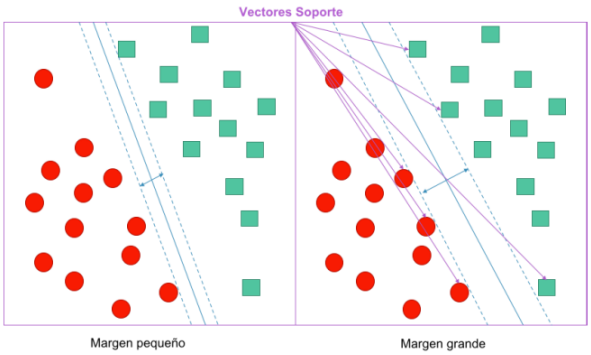

Fig. 6: Modelos SVM.

6) **Modelos estadísticos:**

Estos modelos tienen una base estadística sólida y se utilizan para modelar relaciones entre variables en función de la distribución de los datos observados. Dentro de esta familia se encuentran:

Proceso Gaussiano: Se utiliza para modelar la distribución conjunta de las variables de entrada y salida como una distribución gaussiana multivariante [52]. Es una técnica basada en la regresión que proporciona una estimación de la incertidumbre asociada con las predicciones.

Regresión Logística: Es un modelo de regresión utilizado para predecir la probabilidad de que una instancia pertenezca a una clase particular [53]. Aunque el nombre contiene "regresión", se utiliza principalmente para problemas de clasificación binaria o multinomial. Utiliza la función logística para modelar la relación entre las características y las probabilidades de pertenencia a una clase.

Naive Bayes: Es un modelo probabilístico basado en el teorema de Bayes [54]. Utiliza la regla de Bayes para calcular la probabilidad condicional de una clase dada una instancia de entrada, asumiendo que las características son independientes entre sí dada la clase.

La división del conjunto de datos es del 80% para entrenamiento y 20% para prueba. Este proceso se repite 25 veces para evaluar el rendimiento de cada modelo. Esta práctica permite entrenar el modelo en una parte de los datos y luego probar su rendimiento en una porción separada de datos no vistos previamente. Esta separación proporciona una visión precisa de cómo el modelo generaliza a nuevos datos. Esto es debido a que, reservar una parte de los datos como conjunto de prueba permite verificar si el modelo está aprendiendo patrones genuinos o simplemente memorizando los datos de entrenamiento. Si el rendimiento en el conjunto de prueba es significativamente peor que en el de entrenamiento, puede sugerir un posible sobreajuste.

Repetir este proceso múltiples veces brinda una evaluación más robusta del rendimiento de los modelos, permitiendo estimar mejor la variabilidad del rendimiento y obtener una evaluación más fiable. Esto es especialmente importante para comprender la confiabilidad de las métricas de rendimiento y reducir cualquier sesgo que pueda surgir de realizar una única división de los datos.

Para cada uno de los modelos descritos se realiza una búsqueda exhaustiva de la mejor combinación de hiperparámetros que maximice la capacidad predictiva del modelo. Con este fin se utiliza la función GridSearchCV perteneciente a la biblioteca scikit-learn de Python [55].

C. *Métricas de Rendimiento*

Para la correcta medición e interpretación de los resultados se emplean las siguientes métricas de rendimiento, basadas en los valores de la matriz de confusión [56] del Cuadro I.

CUADRO I.

MATRIZ DE CONFUSIÓN PARA UN PROBLEMA BINARIO

Estimado / Realidad	**POSITIVO**	**NEGATIVO**
POSITIVO	Verdadero Positivo (TP)	Falso Negativo (FN)
NEGATIVO	Falso Positivo (FP)	Verdaderos Negativos (TN)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\,\%$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\,\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\,\%$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Además, se realizan dos mediciones complementarias. Área bajo la curva ROC ya que es útil para evaluar la capacidad discriminativa de un modelo en problemas de clasificación binaria [57]. Representa la capacidad del modelo para distinguir entre clases positivas y negativas a través de diferentes umbrales de clasificación:

$$\text{Área bajo la Curva ROC} = \int_0^1 \text{TPR}(fpr)\; d(fpr)$$

Donde: TPR es la tasa de verdaderos positivos y fpr es la tasa de falsos positivos.

Y el coeficiente de Kappa de Cohen, ya que es una métrica de evaluación de concordancia entre dos sistemas de clasificación [58]. Se utiliza comúnmente en problemas de clasificación donde se necesita medir el grado de acuerdo más allá de lo que podría ocurrir al azar. Es decir, es de vital importancia en ayuda al diagnóstico médico.

$$\text{Kappa} = \frac{\text{p}_\text{o} - \text{p}_\text{e}}{1 - \text{p}_\text{e}}$$

Donde: p_o es el acuerdo observado relativo entre los observadores y p_e es la probabilidad hipotética de acuerdo por azar.

Para realizar una correcta comparación múltiple del rendimiento de los modelos es importante utilizar una métrica consistente y representativa. Se selecciona el F1-Score al combinar la precisión y el recall en una sola medida, lo que lo hace adecuado para evaluar la capacidad de un modelo para clasificar correctamente las instancias positivas y negativas.

Sin embargo, el uso del F1-Score por sí solo puede no ser suficiente para garantizar una comparación completa y significativa del rendimiento de los modelos. Por lo tanto, es apropiado realizar pruebas estadísticas adicionales para respaldar nuestras conclusiones sobre la superioridad de un modelo sobre otro.

El orden sugerido de las pruebas estadísticas (test de Shapiro, Levene, ANOVA y Tukey) es apropiado por las siguientes razones:

Test de Shapiro: Antes de realizar cualquier análisis estadístico paramétrico, como ANOVA, es importante verificar si los datos siguen una distribución normal. Este test se encarga de realizar dicha verificación [59]. Si los datos no siguen una distribución normal, puede ser más apropiado utilizar pruebas estadísticas no paramétricas.

Test de Levene: El test de Levene se utiliza para evaluar la igualdad de varianzas entre diferentes grupos de datos [60]. Es importante verificar si las varianzas de los F1-Score entre los diferentes modelos son homogéneas. Si las varianzas son significativamente diferentes, los resultados del ANOVA pueden no ser confiables.

ANOVA (Análisis de la Varianza): El ANOVA se utiliza para comparar las medias de múltiples grupos de datos y determinar si hay diferencias significativas entre ellos [61]. En este contexto de comparar el rendimiento de múltiples modelos, el ANOVA nos permite determinar si hay diferencias significativas en los resultados de F1-Score entre los diferentes modelos.

Tukey: Si el ANOVA indica que hay diferencias significativas entre los grupos, se utiliza el test de Tukey para realizar comparaciones múltiples entre todos los pares de grupos y determinar cuáles son significativamente diferentes entre sí [61]. Esto nos permite identificar qué modelos tienen un rendimiento significativamente mejor que otros.

Para cada uno de los test se presentan las siguientes hipótesis para los resultados de F1-Score de los modelos, a un nivel de significancia de α = Shapiro → Hipótesis nula (H_0): Los datos siguen una distribución normal. Hipótesis alternativa (H_a): Los datos no siguen una distribución normal.

Levene → H_0: La varianza es la misma en todos los grupos. H_a: Al menos una de las varianzas de los grupos es diferente de las demás.

ANOVA → H_0: Las medias de todos los grupos son iguales. H_a: Los datos no siguen una distribución normal. Tukey → H_0: No hay diferencias significativas entre las medias de los grupos. H_a: Al menos una diferencia entre las medias de los grupos es significativa.

V. RESULTADOS

Se realiza en primer lugar el análisis exploratorio de los datos, como se aprecia en la Fig. 7 los pacientes poseen un amplio rango de edad, siendo la media de edad de 54 años y 3 meses, mientras que la mediana de edad es de 55 años. Para cada cohorte de edad la distribución de los géneros se considera equilibrada. Del conjunto global de datos el 52.27% son mujeres y el 47.73% son hombres. De los cuales, como se ve recogido en la Fig. 9 para ambos géneros la distribución de muestras de pacientes de NSCLC y personas sanas están cercanos al 50%. Por lo que nos encontramos ante un conjunto de datos equilibrado.

Para evaluar el rendimiento de las técnicas de exclusión de muestras atípicas y de selección de características se escoge al modelo SVM, por sus propiedades detalladas anteriormente y su capacidad para tratar datos de alta dimensionalidad. Como se muestra en el Cuadro II y en la

Fig. 8 no existen diferencias significativas en el rendimiento de los distintos conjuntos de datos generados a partir de la exclusión de muestras con valores atípicos. Para cada una de las cuatro hipótesis propuestas no se rechaza la hipótesis nula, obteniendo valores alejados al nivel de significancia. En consecuencia, se prescinde de la eliminación de muestras atípicas y se continua el flujo de trabajo con el conjunto de datos original.

Para las distintas técnicas de selección de características aplicadas se presentan los resultados en el Cuadro III. Todas las técnicas cumplen las hipótesis de normalidad (Shapiro) y de homocedasticidad (Levene). Sin embargo, a la hora de realizar el test de ANOVA, se obtiene un p-valor = 0.026, siendo este valor es inferior al umbral de significancia. Por ello, se aplica el test de Tukey con la finalidad de hallar qué modelos presentan diferencias significativas entre sí, mostrado en la Fig. 10. Se observa como el conjunto de datos seleccionado por Random Forest es significativamente mejor que el seleccionado por mVRS, al obtener un p-valor de 0.014, inferior al nivel de significancia empleado. Aunque, no se aprecian diferencias significativas entre los otros modelos. Aun así, se opta por escoger el conjunto de datos seleccionado por Random Forest para su implementación en los 11 modelos, en virtud de reducir el coste computacional y aumentar la interpretabilidad de los modelos al reducir el número de genes de 4722 a 300.

Para apreciar la selección realizada por Random Forest se procede a visualizar un mapa de calor [62] de la expresión génica (Fig. 11). Respecto al mapa de calor, se seleccionan las 50 muestras con mayor varianza para facilitar la comprensión de este. Algunas muestras tienen niveles de sobreexpresión con valores cercanos a 10 (máximo valor de la escala) para un pequeño grupo genes, mientras que la gran mayoría tienen valores de expresión negativos. No se pueden generar conclusiones a partir de estas gráficas, por lo que se procede a evaluar la bondad de los modelos.

Tras visualizar las características de las variables, se implementan los datos en los modelos propuestos. Los valores de los conjuntos de entrenamiento y prueba se encuentran recogidos en el Cuadro IV. En siete de los modelos descritos previamente (AdaBoost, Gradient Boosting, Regresión Logística, Nearest Neighbors, Neural Net, Random Forest y SVM) se obtienen valores de rendimiento predictivo superiores a 0.85 para la métrica F1-Score. Sin embargo, el modelo Neural Net destaca en cuanto a robustez, solo dicho modelo es capaz de conseguir valores superiores a 0.90 para todas las métricas de rendimiento a excepción de Kappa. Neural Net demuestra una capacidad excepcional para clasificar correctamente las instancias en el conjunto de datos. Particularmente notable es el valor de Kappa obtenido, el cual se sitúa en 0.826 ± 0.032. Dicho valor indica que el modelo está logrando una concordancia sustancial entre las predicciones y las clases reales, más allá de lo que se esperaría si las predicciones se hicieran al azar. El modelo está proporcionando resultados significativos y confiables que van más allá de la casualidad, lo que refuerza la confianza en su capacidad para realizar predicciones precisas y útiles en la clasificación de datos. Además, es importante destacar que Neural Net exhibe la menor desviación típica entre los modelos evaluados, por lo que indica una mayor consistencia en el rendimiento del modelo en diferentes escenarios o conjuntos de datos.

Para complementar el rendimiento de los modelos se someten los resultados a las cuatro hipótesis planteadas en la Sección IV. Todos los modelos cumplen las hipótesis nulas de normalidad y homocedasticidad. En contraposición, al realizar el test de ANOVA se obtiene un p-valor = 1.19×10^{-9}, rechazándose así la hipótesis nula y aceptando la alternativa. Se realiza el test de Tukey para hallar las diferencias significativas entre los modelos, mostrado en la Fig. 13.

Aunque Neural Net sea el modelo con mayor rendimiento de F1-Score, no se aprecian diferencias significativas con los modelos AdaBoost, Gradient Boosting, Random Forest, Regresión Logística y SVM. En esta tesitura se propone escoger el modelo que facilite una mayor interpretabilidad de las decisiones predictivas. En consecuencia, el modelo de Regresión Logística se toma como el modelo con mayor capacidad combinativa de predictibilidad e interpretabilidad. Este modelo proporciona coeficientes para cada característica que indican la dirección y magnitud de su efecto en la variable de salida. Estos coeficientes son fáciles de interpretar y proporcionan información sobre qué características son más importantes para la predicción, pudiendo abrir nuevas líneas de investigación basadas en los genes con mayor relevancia a la hora de la toma de decisión.

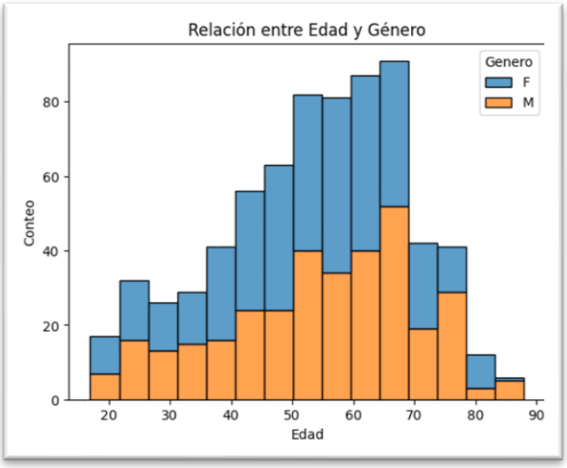

Fig. 7: Gráfico de barras apiladas. Conteo de edades discriminado por géneros.

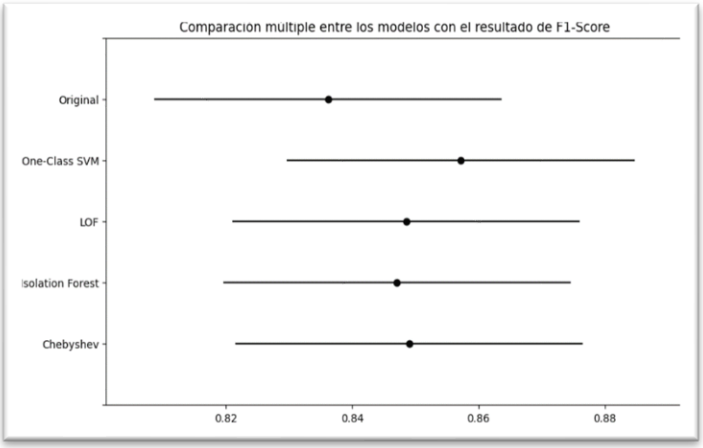

Fig. 8: Comparación múltiple de la bondad de los modelos. Basado en F1 Score para las distintas técnicas de detección de muestras atípicas.

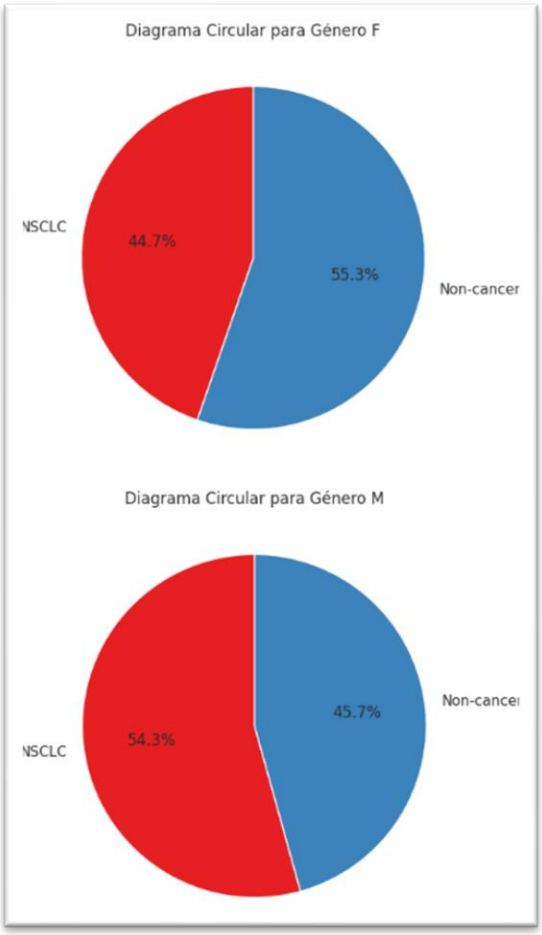

Fig. 9: Diagrama circular que muestra el porcentaje de pacientes NSCLC y sanos por género.

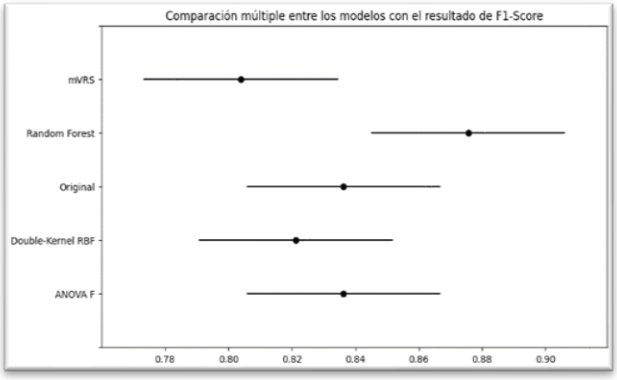

Fig. 10: Comparación múltiple de la bondad de los modelos. Basado en F1 Score para las distintas técnicas de selección de características.

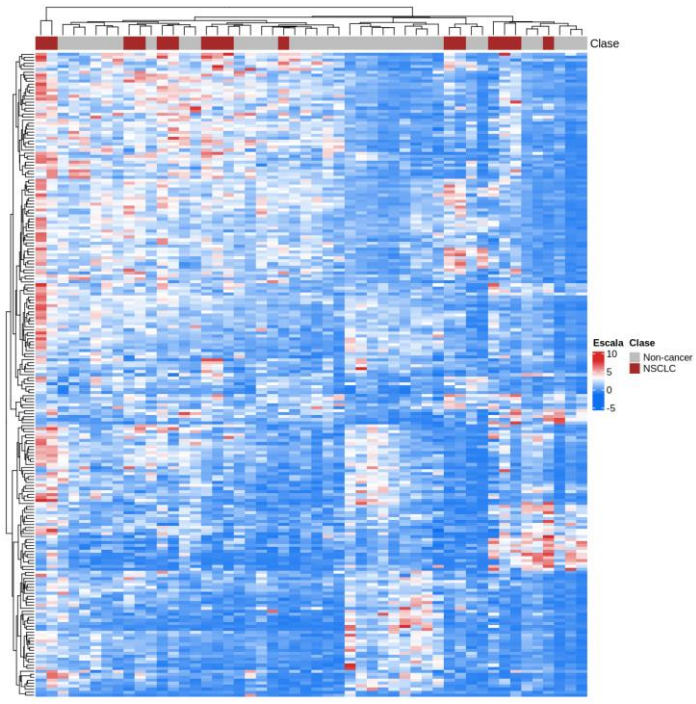

Fig. 11: Mapa de calor de los genes seleccionados por la técnica de selección de características Random Forest para las 50 muestras con mayor varianza.

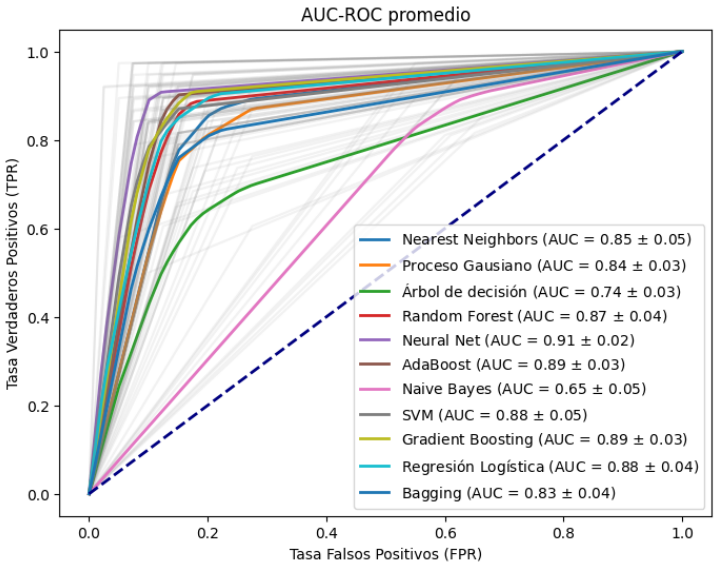

Fig. 12: Gráfica del área bajo la curva ROC.

CUADRO II.

RENDIMIENTO DEL CONJUNTO PRUEBA Y EVALUACIÓN DE HIPÓTESIS PARA LOS DISTINTOS CONJUNTOS DE DATOS DESPUÉS DE REALIZAR DETECCIÓN DE MUESTRAS ATÍPICAS.

SVM	**Nº Muestras Atípicos Encontrados**	**F1 Score**	**Shapiro** $\alpha = 0.05$	**Levene** $\alpha = 0.05$	**ANOVA** $\alpha = 0.05$	**Tukey** $\alpha = 0.05$
Conjunto completo	-	0.836 ± 0.044	$H_0 \rightarrow$ ✓	$H_0 \rightarrow$ ✓	$H_0 \rightarrow$ ✓	$H_0 \rightarrow$ ✓
Chebyshev	67	0.849 ± 0.038				
One-Class SVM	77	**0.857 ± 0.052**				
LOF	78	0.849 ± 0.024				
Isolation Forest	78	0.847 ± 0.042				

CUADRO III.

RENDIMIENTO DEL CONJUNTO PRUEBA Y EVALUACIÓN DE HIPÓTESIS PARA LOS DISTINTOS CONJUNTOS DE DATOS DESPUÉS DE REALIZAR TÉCNICAS DE SELECCIÓN DE CARACTERÍSTICAS.

SVM	**Nº Características Seleccionadas**	**F1 Score**	**Shapiro** $\alpha = 0.05$	**Levene** $\alpha = 0.05$	**ANOVA** $\alpha = 0.05$	**Tukey** $\alpha = 0.05$
Conjunto completo	4722	0.836 ± 0.044	$H_0 \rightarrow$ ✓	$H_0 \rightarrow$ ✓	$H_0 \rightarrow$ ✗ $H_a \rightarrow$ ✓ p-valor = 0.026 < α	$H_0 \rightarrow$ ✗ $H_a \rightarrow$ ✓ p-adj = 0.0158 entre mVRS y Random Forest
ANOVA F	195	0.850 ± 0.047				
Random Forest	300	**0.869 ± 0.049**				
mVRS	465	0.804 ± 0.047				
Double RBF	330	0.821 ± 0.044				

CUADRO IV.

RENDIMIENTOS DEL CONJUNTO PRUEBA PARA LOS DISTINTOS MODELOS DE APRENDIZAJE AUTOMÁTICO.

	Accuracy	**Precision**	**Recall**	**F1 Score**	**Kappa**	**AUC ROC**
AdaBoost	0.879 ± 0.026	0.867 ± 0.036	0.883 ± 0.038	0.874 ± 0.030	0.785 ± 0.052	0.890 ± 0.026
Árbol de Decisión	0.730 ± 0.036	0.739 ± 0.049	0.682 ± 0.072	0.706 ± 0.045	0.457 ± 0.074	0.739 ± 0.032
Bagging	0.840 ± 0.047	0.849 ± 0.031	0.810 ± 0.056	0.829 ± 0.034	0.678 ± 0.056	0.834 ± 0.038
Proceso Gausiano	0.840 ± 0.036	0.819 ± 0.039	0.857 ± 0.035	0.836 ± 0.027	0.628 ± 0.072	0.840 ± 0.030
Gradient Boosting	0.891 ± 0.040	0.879 ± 0.040	0.897 ± 0.065	0.887 ± 0.044	0.782 ± 0.052	0.892 ± 0.034
Regresión Logística	0.882 ± 0.021	0.855 ± 0.025	0.908 ± 0.021	**0.880 ± 0.022**	0.764 ± 0.041	0.884 ± 0.040
Naive Bayes	0.653 ± 0.041	0.590 ± 0.045	0.905 ± 0.038	0.713 ± 0.047	0.319 ± 0.081	0.653 ± 0.048
Nearest Neighbors	0.730 ± 0.036	0.853 ± 0.049	0.822 ± 0.072	0.851 ± 0.045	0.706 ± 0.074	0.854 ± 0.045
Neural Net	0.913 ± 0.019	0.906 ± 0.026	0.915 ± 0.023	**0.909 ± 0.021**	0.826 ± 0.032	0.907 ± 0.017
Random Forest	0.873 ± 0.025	0.861 ± 0.024	0.879 ± 0.028	0.869 ± 0.024	0.746 ± 0.047	0.874 ± 0.036
SVM	0.881 ± 0.023	0.888 ± 0.026	0.870 ± 0.024	0.875 ± 0.021	0.761 ± 0.048	0.881 ± 0.045

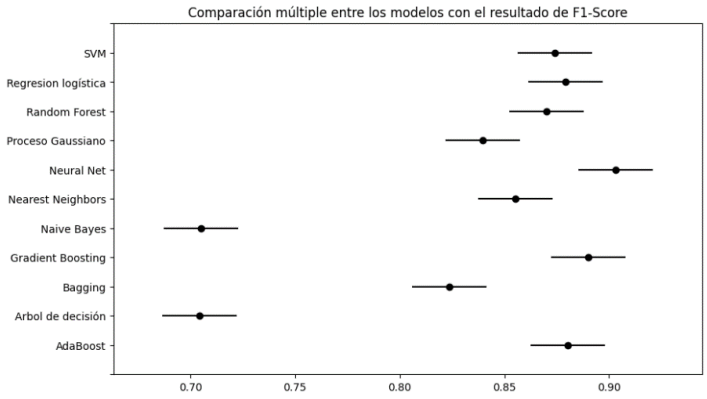

Fig. 13: Comparación múltiple de la bondad de los modelos. Basado en F1 Score para las distintas técnicas de selección de características.

VI. Conclusiones y Futuros Trabajos

En este estudio se ha propuesto un enfoque basado en técnicas de aprendizaje automático para la detección precoz del cáncer de pulmón de células no pequeñas utilizando datos RNA-Seq de plaquetas entrenadas por el tumor.

Se inició el estudio con la detección de muestras atípicas en el conjunto de datos original, aplicando diversas técnicas de análisis. A pesar de los esfuerzos, ninguna técnica demostró mejorar significativamente los resultados, lo que subraya la complejidad inherente de los datos y la necesidad de estrategias más avanzadas. De igual forma, aunque se observó una ligera mejora en los resultados, ninguna técnica de selección de características demostró mejorar significativamente el desempeño en la detección del NSCLC. Sin embargo, se optó por utilizar la técnica de Random Forest para este propósito, dado su potencial para reducir la dimensionalidad de los datos y aumentar la explicabilidad de los modelos. Esta decisión fue fundamental para identificar las variables más relevantes en la detección del NSCLC, lo que permitió mejorar la eficiencia computacional de los modelos y proporcionar una comprensión más clara de las características subyacentes asociadas con la enfermedad. Resaltando de nuevo, la complejidad de los datos.

Se evaluaron exhaustivamente 11 modelos diferentes de aprendizaje automático para determinar cuál ofrecía el mejor rendimiento en la detección del NSCLC. Entre estos modelos, destacó Neural Net por su precisión y sensibilidad superiores. Sin embargo, debido a la necesidad de interpretabilidad en entornos clínicos, se optó por seleccionar la Regresión Logística como el modelo final, con resultados para la métrica de F1-Score de 0.88. Aunque no mostró diferencias significativas en comparación con Neural Net, la Regresión Logística ofrece una comprensión más clara de las relaciones entre las variables y una interpretación más sencilla de los resultados, lo que la convierte en una opción más adecuada para su implementación clínica.

Estos resultados resaltan el potencial de las técnicas de aprendizaje automático para la detección temprana del cáncer de pulmón de células no pequeñas. La combinación de estrategias de preprocesamiento de datos, selección de características y evaluación de modelos ha permitido identificar un enfoque efectivo y práctico para la detección precoz de esta enfermedad, con importantes implicaciones para la práctica clínica y la investigación oncológica.

El enfoque adoptado en este estudio, que incluye la detección de muestras atípicas, la selección de características y la evaluación de múltiples modelos de aprendizaje automático, puede servir como un marco de trabajo robusto para futuras investigaciones en el campo de la oncología basado en otras enfermedades. Los resultados obtenidos no solo contribuyen al desarrollo de métodos más eficaces para la detección del NSCLC, sino que también destacan la importancia de considerar estrategias integrales que aborden la complejidad y heterogeneidad de los datos biológicos.

Desde una perspectiva clínica, la implementación de modelos de aprendizaje automático basados en datos RNA-Seq de TEPs podría integrarse fácilmente en la práctica médica habitual como una herramienta de apoyo al diagnóstico. Esto podría agilizar el proceso de detección y diagnóstico del NSCLC, permitiendo una intervención más rápida y precisa.

Mientras que, en el ámbito de la investigación oncológica, los resultados de estudios similares al presentado pueden inspirar nuevas investigaciones y colaboraciones para explorar aún más el potencial de los datos RNA-Seq de TEPs y las técnicas de aprendizaje automático en la identificación de biomarcadores y dianas terapéuticas para el NSCLC y otras enfermedades relacionadas con el cáncer. La combinación de datos de alta calidad con algoritmos de aprendizaje automático avanzados tiene el potencial de revolucionar nuestra comprensión y tratamiento del cáncer, abriendo nuevas vías para la personalización de la medicina y la mejora de los resultados clínicos.

VII. Referencias

[1] H. Sung et al., «Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries», CA: A Cancer Journal For Clinicians, vol. 71, n.o 3, pp. 209-249, feb. 2021, doi: 10.3322/caac.21660.

[2] F. Nasim, B. F. Sabath, y G. A. Eapen, «Lung Cancer», *Medical Clinics Of North America*, vol. 103, n.º 3, pp. 463-473, 2019, doi: 10.1016/j.mcna.2018.12.006.

[3] C. Barrionuevo y D. Dueñas, «Clasificación actual del carcinoma de pulmón. Consideraciones histológicas, inmunofenotípicas, moleculares y clínicas», *Horizonte Médico*, vol. 19, n.º 4, pp. 74-83, dic. 2019, doi: 10.24265/horizmed.2019.v19n4.11.

[4] J. Norum y C. Nieder, «Tobacco smoking and cessation and PD-L1 inhibitors in non-small cell lung cancer (NSCLC): a review of the literature», *ESMO Open*, vol. 3, n.º 6, p. e000406, ene. 2018, doi: 10.1136/esmoopen-2018-000406.

[5] M. Riudavets, M. G. De Herreros, B. Besse, y L. Mezquita, «Radon and lung cancer: current trends and future perspectives», *Cancers*, vol. 14, n.º 13, p. 3142, jun. 2022, doi: 10.3390/cancers14133142.

[6] Y. Alduais, H. Zhang, F. Fan, J. Chen, y B. Chen, «Non-small cell lung Cancer (NSCLC): A review of risk factors, diagnosis, and treatment», *Medicine*, vol. 102, n.º 8, p. e32899, feb. 2023, doi: 10.1097/md.0000000000032899.

[7] M. B. Schabath y M. L. Coté, «Cancer progress and priorities: Lung cancer», *Cancer Epidemiology, Biomarkers & Prevention*, vol. 28, n.º 10, pp. 1563-1579, oct. 2019, doi: 10.1158/1055-9965.epi-19-0221.

[8] S. Knight, P. Crosbie, H. Balata, J. Chudziak, T. Hussell, y C. Dive, «Progress and prospects of early detection in lung cancer», *Open Biology*, vol. 7, n.º 9, p. 170070, sep. 2017, doi: 10.1098/rsob.170070.

[9] D. R. Brenner, J. McLaughlin, y R. J. Hung, «Previous Lung Diseases and Lung Cancer Risk: A Systematic Review and Meta-Analysis», *PLOS ONE*, vol. 6, n.º 3, p. e17479, mar. 2011, doi: 10.1371/journal.pone.0017479.

[10] J. Corral *et al.*, «Estimation of lung cancer diagnosis and treatment costs based on a patient-level analysis in Catalonia (Spain)», *BMC Health Services Research*, vol. 15, n.º 1, feb. 2015, doi: 10.1186/s12913-015-0725-3.

[11] M. G. Best, S. G. J. G. I. 'T Veld, N. Sol, y T. Würdinger, «RNA sequencing and swarm intelligence–enhanced classification algorithm development for blood-based disease diagnostics using spliced blood platelet RNA», *Nature Protocols*, vol. 14, n.º 4, pp. 1206-1234, mar. 2019, doi: 10.1038/s41596-019-0139-5.

[12] W. Zhang *et al.*, «Comparison of RNA-seq and microarray-based models for clinical endpoint prediction», *Genome Biology*, vol. 16, n.º 1, jun. 2015, doi: 10.1186/s13059-015-0694-1.

[13] G. A. Silvestri *et al.*, «Methods for staging non-small cell lung cancer», *Chest*, vol. 143, n.º 5, pp. e211S-e250S, may 2013, doi: 10.1378/chest.12-2355.

[14] M. D. Podolsky, A. Барчук, V. I. Kuznetcov, N. Gusarova, V. S. Gaidukov, y S. A. Tarakanov, «Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels», *Asian Pacific Journal Of Cancer Prevention*, vol. 17, n.º 2, pp. 835-838, mar. 2016, doi: 10.7314/apjcp.2016.17.2.835.

[15] J. C. Nesbitt, J. B. Putnam, G. L. Walsh, J. A. Roth, y C. F. Mountain, «Survival in early-stage non-small cell lung cancer», *The Annals Of Thoracic Surgery*, vol. 60, n.º 2, pp. 466-472, ago. 1995, doi: 10.1016/0003-4975(95)00169-l.

[16] P. Hamet y J. Tremblay, «Artificial intelligence in medicine», *Metabolism*, vol. 69, pp. S36-S40, abr. 2017, doi: 10.1016/j.metabol.2017.01.011.

[17] D. Patel, Y. T. Shah, N. Thakkar, K. N. Shah, y M. Shah, «Implementation of artificial intelligence techniques for cancer detection», *Augmented Human Research*, vol. 5, n.º 1, nov. 2019, doi: 10.1007/s41133-019-0024-3.

[18] M. Á. Vega, L. M. Q. Mora, y M. V. C. Badilla, «Inteligencia artificial y aprendizaje automático en medicina», *Revista Médica Sinergía*, vol. 5, n.º 8, p. e557, ago. 2020, doi: 10.31434/rms.v5i8.557.

[19] D. Kreuzberger, N. Kuehl, y S. Hirschl, «Machine Learning Operations (MLOps): Overview, Definition, and Architecture», *IEEE Access*, vol. 11, pp. 31866-31879, ene. 2023, doi: 10.1109/access.2023.3262138.

[20] K. Y. Ngiam y I. W. Khor, «Big data and machine learning algorithms for health-care delivery», *The Lancet Oncology*, vol. 20, n.º 5, pp. e262-e273, may 2019, doi: 10.1016/s1470-2045(19)30149-4.

[21] A. Oulas, G. Minadakis, M. Zachariou, K. Sokratous, M. M. Bourdakou, y G. M. Spyrou, «Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches», *Briefings In Bioinformatics*, vol. 20, n.º 3, pp. 806-824, 2017, doi: 10.1093/bib/bbx151.

[22] A. Sánchez-Palencia *et al.*, «Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer», *International Journal Of Cancer*, vol. 129, n.º 2, pp. 355-364, nov. 2010, doi: 10.1002/ijc.25704.

[23] T. Inage, T. Nakajima, I. Yoshino, y K. Yasufuku, «Early lung cancer detection», *Clinics In Chest Medicine*, vol. 39, n.º 1, pp. 45-55, mar. 2018, doi: 10.1016/j.ccm.2017.10.003.

[24] S. Duan *et al.*, «Development of a machine learning-based multimode diagnosis system for lung cancer», *Aging*, vol. 12, n.º 10, pp. 9840-9854, may 2020, doi: 10.18632/aging.103249.

[25] M. Bachute y J. M. Subhedar, «Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms», *Machine Learning With Applications*, vol. 6, p. 100164, dic. 2021, doi: 10.1016/j.mlwa.2021.100164.

[26] D. F. Santos, «Tackling Lung Cancer: Advanced Image Analysis and Deep Learning for Early Detection», *TechRxiv*, jun. 2023, doi: 10.36227/techrxiv.23537685.v1.

[27] Ch. Vasavi y N. Sruthi, «Detection of lung cancer using optimized SVM-CNN model», *International Journal For Research In Applied Science And Engineering Technology*, vol. 11, n.º 6, pp. 4608-4613, jun. 2023, doi: 10.22214/ijraset.2023.54496.

[28] K. O'Shea y R. R. Nash, «An Introduction to Convolutional Neural Networks», *arXiv (Cornell University)*, nov. 2015, doi: 10.48550/arxiv.1511.08458.

[29] D. Pisner y D. M. Schnyer, «Support vector machine», en *Elsevier eBooks*, 2020, pp. 101-121. doi: 10.1016/b978-0-12-815739-8.00006-7.

[30] R. G. Vidhya y T. T. Mirnalinee, «Hybrid Optimized Learning for lung cancer classification», *Intelligent Automation And Soft Computing*, vol. 34, n.º 2, pp. 911-925, ene. 2022, doi: 10.32604/iasc.2022.025060.

[31] B. Kaur, B. Goyal and A. Dogra, "A Hybrid Feature Based Model Development for Computer Aided Diagnosis of Lung Cancer," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 1031-1036.

[32] D. Singh y B. Singh, «Investigating the impact of data normalization on classification performance», *Applied Soft Computing*, vol. 97, p. 105524, dic. 2020, doi: 10.1016/j.asoc.2019.105524.

[33] B. G. Amidan, T. A. Ferryman, y S. K. Cooley, «Data outlier detection using the Chebyshev theorem», *IEEE*, ene. 2005, doi: 10.1109/aero.2005.1559688.

[34] F. T. Liu, K. M. Ting, y Z. Zhou, «Isolation Forest», *IEE*, dic. 2008, doi: 10.1109/icdm.2008.17.

[35] M. Breunig, H. Kriegel, R. T. Ng, y J. Sander, «LOF: identifying density-based local outliers», *ACM SIGMOD*, may 2000, doi: 10.1145/342009.335388.

[36] L. St»Hle y S. Wold, «Analysis of variance (ANOVA)», *Chemometrics And Intelligent Laboratory Systems*, vol. 6, n.º 4, pp. 259-272, nov. 1989, doi: 10.1016/0169-7439(89)80095-4.

[37] O. Gascuel, «Data Model and Classification by Trees: The Minimum Variance Reduction (MVR) Method», *Journal Of Classification*, vol. 17, n.º 1, pp. 67-99, ene. 2000, doi: 10.1007/s003570000005.

[38] A. Cutler, D. R. Cutler, y J. R. Stevens, «Random forests», en *Springer eBooks*, 2012, pp. 157-175. doi: 10.1007/978-1-4419-9326-7_5.

[39] K. Thurnhofer-Hemsi, E. López-Rubio, M. A. Molina-Cabello, y K. Najarian, «Radial basis function kernel optimization for Support Vector Machine classifiers», *arXiv (Cornell University)*, jul. 2020, doi: 10.48550/arxiv.2007.08233.

[40] M.-L. Zhang y Z. Zhou, «ML-KNN: A lazy learning approach to multi-label learning», *Pattern Recognition*, vol. 40, n.º 7, pp. 2038-2048, jul. 2007, doi: 10.1016/j.patcog.2006.12.019.

[41] G. Guo, H. Wang, D. A. Bell, Y. Bi, y K. Greer, «KNN Model-Based Approach in Classification», en *Lecture Notes in Computer Science*, 2003, pp. 986-996. doi: 10.1007/978-3-540-39964-3_62.

[42] S. Kotsiantis, «Decision trees: a recent overview», *Artificial Intelligence Review*, vol. 39, n.º 4, pp. 261-283, jun. 2011, doi: 10.1007/s10462-011-9272-4.

[43] L. Rokach, «Ensemble-based classifiers», *Artificial Intelligence Review*, vol. 33, n.º 1-2, pp. 1-39, nov. 2009, doi: 10.1007/s10462-009-9124-7.

[44] A. Cutler, D. R. Cutler, y J. R. Stevens, «Random forests», en *Springer eBooks*, 2012, pp. 157-175. doi: 10.1007/978-1-4419-9326-7_5.

[45] D. W. Opitz y R. Maclin, «Popular Ensemble Methods: an Empirical study», *Journal Of Artificial Intelligence Research*, vol. 11, pp. 169-198, ago. 1999, doi: 10.1613/jair.614.

[46] C. Bentéjac, A. Csörgő, y G. Martínez-Muñoz, «A comparative analysis of gradient boosting algorithms», *Artificial Intelligence Review*, vol. 54, n.º 3, pp. 1937-1967, ago. 2020, doi: 10.1007/s10462-020-09896-5.

[47] Y. Cao, Q. Miao, J. Liu, y L. Gao, «Advance and Prospects of AdaBoost Algorithm», *Acta Automatica Sinica*, vol. 39, n.º 6, pp. 745-758, jun. 2013, doi: 10.1016/s1874-1029(13)60052-x.

[48] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, y G. Monfardini, «The Graph Neural Network model», *IEEE Transactions On Neural Networks*, vol. 20, n.º 1, pp. 61-80, ene. 2009, doi: 10.1109/tnn.2008.2005605.

[49] S. Suthaharan, «Support Vector machine», en *Integrated series on information systems*, 2016, pp. 207-235. doi: 10.1007/978-1-4899-7641-3_9.

[50] N. K. Jong y P. Stone, «Kernel-Based Models for Reinforcement Learning», *Texas University*, ene. 2006, [En línea]. Disponible en: https://www.cs.utexas.edu/~pstone/Papers/bib2html-links/ICML06-nick.pdf

[51] A. Patle y D. Chouhan, «SVM kernel functions for classification», *IEEE*, ene. 2013, doi: 10.1109/icadte.2013.6524743.

[52] M. Seeger, «GAUSSIAN PROCESSES FOR MACHINE LEARNING», *International Journal Of Neural Systems*, vol. 14, n.º 02, pp. 69-106, abr. 2004, doi: 10.1142/s0129065704001899.

[53] M. P. LaValley, «Logistic regression», *Circulation*, vol. 117, n.º 18, pp. 2395-2399, may 2008, doi: 10.1161/circulationaha.106.682658.

[54] I. Rish, «An empirical study of the naive Bayes classifier», *IJCAI*, ene. 2001, [En línea]. Disponible en: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf

[55] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *HAL (Le Centre Pour la Communication Scientifique Directe)*, oct. 2011, [En línea]. Disponible en: https://hal.inria.fr/hal-00650905

[56] R. Susmaga, «Confusion Matrix Visualization», en *Springer eBooks*, 2004, pp. 107-116. doi: 10.1007/978-3-540-39985-8_12.

[57] S. H. Park, J. M. Goo, y C. H. Jo, «Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists», *Korean Journal Of Radiology*, vol. 5, n.º 1, p. 11, ene. 2004, doi: 10.3348/kjr.2004.5.1.11.

[58] S. M. Vieira, U. Kaymak, y J. M. Sousa, «Cohen's kappa coefficient as a performance measure for feature selection», *International Conference On Fuzzy Systems*, jul. 2010, doi: 10.1109/fuzzy.2010.5584447.

[59] Z. Hanusz, J. Tarasinska, y W. Zieliński, «Shapiro–Wilk Test with Known Mean», *DOAJ (DOAJ: Directory Of Open Access Journals)*, feb. 2016, doi: 10.57805/revstat.v14i1.180.

[60] J. L. Gastwirth, Y. R. Gel, y W. Miao, «The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice», *Statistical Science*, vol. 24, n.º 3, ago. 2009, doi: 10.1214/09-sts301.

[61] J. Jaccard, M. A. Becker, y G. Wood, «Pairwise multiple comparison procedures: A review.», *Psychological Bulletin*, vol. 96, n.º 3, pp. 589-596, nov. 1984, doi: 10.1037/0033-2909.96.3.589.

[62] L. Wilkinson y M. Friendly, «The History of the Cluster Heat Map», *The American Statistician*, vol. 63, n.º 2, pp. 179-184, may 2009, doi: 10.1198/tas.2009.0033.