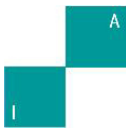
Inteligencia Artificial, 28(75) (2025), 260-280
doi: 10.4114/intartif.vol28iss75pp260-280



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

# Diagnóstico de cáncer de mama usando el tamaño del efecto $d$ de Cohen como selector de características

Nicolás Martín Masino[1,A], Antonio Quintero-Rincón[1,2,A]

[1]Laboratorio de Ciencia de Datos e Inteligencia Artificial, Departamento de Ciencia de Datos, Universidad Católica Argentina (UCA), Argentina
[2]Departamento de Informática, Universidad Católica Argentina (UCA), Argentina
[A]nicolasmasino@uca.edu.ar, antonioquintero@uca.edu.ar

**A**bstract: Breast cancer is a tumor that begins to grow in the milk ducts or lobules and can become lethal if treatment is not administered in time. According to the World Health Organization (WHO), there were approximately 2.3 million cases of breast cancer in 2020. Furthermore, breast cancer can affect anyone, particularly women over 50 years old. Therefore, it is crucial to have early diagnostic techniques. We propose a novel method based on Cohen's $d$ for feature selection in this context. Cohen's $d$ is a statistical concept that quantifies the strength of the relationship between two populations on a numeric scale. The central idea is to utilize Cohen's $d$ effect size as a feature selector to reduce the dimensionality of the data and enhance the predictors through a Machine Learning (ML) classifier model for diagnosing breast cancer. For experimental purposes, the Breast Cancer Wisconsin database was used. This proposed feature selector is compared with two classical methods: Learning Vector Quantization (LVQ) and Recursive Feature Elimination (RFE). A random evaluation of the features of each selector was conducted 100 times using a Support Vector Machine (SVM) classifier, resulting in the following average outcomes: Cohen's $d$ based feature selector showed 0.96 sensitivity and 0.97 specificity, RFE based feature selector exhibited 0.95 sensitivity and 0.98 specificity, and LVQ based feature selector demonstrated 0.91 sensitivity and 0.96 specificity. These promising results indicate that the proposed methodology utilizing Cohen's $d$ may be a valuable feature selector and sheds light on the long-standing research into breast cancer detection.

**Keywords**: Breast cancer, Effect Size, Cohen's d, Feature Selection.

**R**esumen: El cáncer de mama es un tumor maligno que comienza a desarrollarse dentro de los conductos galactóforos o de los lobulillos que producen leche del seno, lo que resulta mortal si no se recibe tratamiento a tiempo. En consonancia con la Organización Mundial de la Salud (OMS), en el año 2020 se diagnosticaron cerca de 2,3 millones de mujeres con cáncer mamario. Asimismo, el tumor puede afectar a cualquier individuo indiscriminadamente, aunque la tasa suele ser mayor en sujetos femeninos mayores de 50 años. Con este escenario mundial resulta imprescindible contar con estrategias de detección temprana de cáncer de mama. Bajo dicho lema, se propone usar el tamaño del efecto $d$ de Cohen como selector de características para ser aplicado en un modelo de clasificación de Machine Learning (ML). El objetivo es reducir la dimensionalidad de los datos y así optimizar los predictores para diagnosticar el cáncer de mama. $d$ de Cohen mide la fuerza de la relación entre dos poblaciones en una escala numérica. Este selector propuesto se compara con dos métodos clásicos: cuantificación de vectores de aprendizaje (LVQ: Learning Vector Quantization) y eliminación recursiva de características (RFE: Recursive Feature Elimination). Para fines de experimentación se trabajo con las base de datos Breast Cancer Wisconsin. La evaluación aleatoria de las características de cada selector, se realizó 100 veces a través de un clasificador de máquina de vectores de soporte (SVM: Support Vector Machine), obteniéndose en promedio, los siguientes

**ISSN: 1137-3601 (print), 1988-3064 (on-line)**

resultados: una sensibilidad de 0.91 y una especificidad de 0.96 para el modelo basado en LVQ, una sensibilidad de 0.96 y una especificidad de 0.97 empleando el método propuesto $d$ de Cohen, contra una sensibilidad de 0.95 y una especificidad de 0.98 utilizando RFE. Estos resultados prometedores sugieren que la metodología propuesta es potencialmente útil como selector de características y abren una luz en la larga investigación en la detección de cáncer de mama.

**Palabras Clave**: Cáncer de mama, Tamaño de efecto, $d$ de Cohen, Selección de características.

# 1. Introducción

Dada su prevalencia, el cáncer de mama es el tumor más frecuente y mortal. Según la Organización Mundial de la Salud, de los 2,3 millones de mujeres diagnosticadas con cáncer de mama en 2020, alrededor de 685.000 fallecieron debido a dicha afección [1]. El cáncer de mama se categoriza de acuerdo al alcance y propagación que tiene dentro del cuerpo, y su tratamiento varía de acuerdo al tipo de neoplasia y su dispersión. Los métodos tradicionales para detectar el cáncer de mama a nivel mundial, se centran en la prueba de triple evaluación, la cual se forma mediante la combinación de tres exámenes médicos: la evaluación clínica de mama, las imágenes radiológicas (mamografía y/o ecografía) y una evaluación patológica por aspiración con aguja fina (FNA) o biopsia con aguja gruesa (CNB)[2]. El cáncer de mama a un paciente se le diagnostica positivamente, si al menos una de estas tres pruebas indica malignidad. Por el contrario, se le diagnostica negativamente si las tres pruebas indican una afección mamaria benigna. Este trabajo se centrará en la biopsia por aspiración con aguja fina (FNA: Fine-Needle Aspiration), la cual es una técnica potente y amigable donde se extrae, aspirando por medio de una jeringa, una pequeña cantidad de tejido o líquido de la región de interés [3]. Dicho procedimiento es ampliamente utilizado para el muestreo de diferentes tipos de tumores [4].

Las biopsias de mama mediante la técnica de FNA, se clasifican en una de las siguientes cinco categorías [5]. C1: Insatisfactoria. C2: Benigna. C3: Atípica/Indeterminada, favorece Benigna. C4: Sospechosa, favorece Maligna. C5: Maligna. Las categorías C1, C2 y C5 suelen ser sencillas y no suelen plantear dificultades a los patólogos, mientras que las categorías C3 y C4 muestran características malignas distintivas, las cuales se deben analizar con sumo detalle. Precisamente, encontrar unas características óptimas que ayuden a llegar a los diagnósticos correctos es el objetivo de este trabajo. Para fines de experimentación, se consideró la base de datos (BD) *Diagnostic Wisconsin Breast Cancer*, la cual es de libre acceso en el repositorio de ML de la Universidad de California, Irvine [6]. Treinta características de esta BD, describen las propiedades de los núcleos celulares presentes en una imagen digitalizada de una muestra mamaria extraída mediante FNA. Todas ellas relacionadas con la textura, el área y la suavidad, las cuales están catalogadas en dos clases: $M$ para las muestras Malignas y $B$ para las muestras Benignas. Con el objetivo de encontrar las características óptimas de esta BD, se propone usar el tamaño de efecto $d$ de Cohen como reductor y selector de características, las cuales serán posteriormente empleadas en el diagnóstico del cáncer de mama. Antes de introducir el tamaño de efecto $d$ de Cohen, es necesario definir en qué consiste la selección de características.

La selección de características (SC) es un conjunto de estrategias imprescindibles en ML, aplicado al preprocesamiento de datos, especialmente en datos de alta dimensión [7]. Su objetivo es determinar un subconjunto de variables predictoras de acuerdo a un criterio, el cual permita seleccionar las características más relevantes que expliquen la problemática que se quiere modelar. De esta manera, se logra reducir tanto la dimensionalidad de los datos como la complejidad del modelo de ML; mejora el tiempo de entrenamiento de los algoritmos, reduce el sobreajuste y hace más fácil la interpretación del modelo [8]. La SC se pueden agrupar en 3 métodos: de Filtro, Envolventes e Integrados. Los métodos de filtro se usan por lo general en la etapa de preprocesamiento de los datos y son independientes del modelo de ML usado. Las características son seleccionadas a partir de criterios estadísticos, a los cuales se les asigna un valor numérico dado por su correlación con la variable objetivo. Estos modelos tienen buen tiempo de computación, una complejidad baja y son robustos al sobreajuste. Los Métodos Envolventes entrenan un modelo de ML utilizando diferentes tamaños y combinaciones de subconjuntos de variables. Estas se van añadiendo o eliminando del subconjunto hasta llegar a un resultado óptimo. No obstante, son más complejos, tienen un tiempo de ejecución alto y pueden tender al sobreajuste si el número de observaciones es insuficiente. Por otro lado, los métodos Integrados combinan las dos técnicas anteriores,

optimizando las ventajas de ambos, es decir, se generan subconjuntos basados en decisiones estadísticas y se entrena un algoritmo simultáneamente. Para más detalles de los métodos de SC, consultar [9, 10].

Dentro de cada uno de estos métodos, hay un gran número de técnicas de selección de características. Es por ello, que elegir una estrategia específica depende mucho del contexto y puede llegar a ser una tarea compleja. Por lo tanto, es esencial conocer la dimensión y el tamaño de los datos, así como el costo computacional requerido y aceptado.

En el campo de diagnóstico médico, las técnicas mayormente utilizadas se basan en: correlación, filtro basado en consistencia, separación de clases, ganancia de información, Relief, RFE (Recursive Feature Elimination) y regularización Lasso [11]. Más específicamente, en el contexto de la identificación de características de núcleos celulares relevantes para el diagnóstico de cáncer de mama, existen varios estudios en la literatura. Por ejemplo, en [12] se propusieron tres métodos diferentes, basados en la mínima-redundancia máxima-relevancia, suma de rangos de Wilcoxon y RF (Random Forest); Otros artífices emplearon selección secuencial en avance (Sequential Forward Selection) [13], o incluso mediante un método combinado usando Máquinas de Vectores de Soporte (SVM) basado en F-score como en el caso de [14]. Otro estudio del mismo campo comparó RF con otros selectores, como SVM-RF, RRF (Regularized Random Forest), SBS (Sequential Backward Selection) y VarSelRF (Iterative Random Forest) para reducir las características nucleares y clasificar leucocitos [15]. Por otro lado, en [16] plantearon un modelo ensamble de selección de características de filtro llamado EnSNR, basado en entropía y en la Relación Señal-Ruido (SNR: Signal-Noise Relation).

La SC define la relevancia de los datos y depende del objetivo final del análisis. La idea general de este trabajo es, diseñar un buen estimador de cáncer a partir de características significativas de núcleos celulares presentes en una imagen digitalizada de un muestra mamaria obtenida con FNA. Este enfoque presenta las siguientes propiedades:

- Al reducir la dimensión del problema, se puede mejorar la exactitud y precisión de un modelo predictivo.
- Cada característica puede resultar costosa de medir, por lo tanto, la predicción puede ser computacionalmente costosa en entornos de alta dimensiones.
- Reducir el número de funciones puede proporcionar información adicional sobre el funcionamiento del modelo.

En el presente trabajo se propone usar el tamaño de efecto $d$ de Cohen como reductor y selector de características, con el objetivo de diagnosticar el cáncer de mama. El tamaño de efecto mide la magnitud de asociación entre dos poblaciones de interés a través de una escala numérica. Este valor estadístico cobró mayor importancia en los últimos años en investigaciones de meta-análisis, tales como psicológicas, educacionales y conductuales [17, 18, 19, 20, 21, 22, 23]. El tamaño de efecto es una familia de estadísticos que se dividen en mediciones estandarizadas de la diferencia entre dos grupos, y en medidas de correlación. En ambos casos se pueden estimar tanto de forma paramétrica como no-paramétrica [24, 25].

Una particularidad del efecto $d$ de Cohen es que, se puede destinar para medir el grado de significancia que tiene un fenómeno en estudio sobre una población determinada, o evaluar el efecto que tiene un tratamiento sobre una población. Por ejemplo, comparando la población que recibe un tratamiento, contra una población de control que no lo recibe. De esta manera se permite ver que los efectos no sean producto del azar y además, permite dar un complemento al valor-$p$, el cual es utilizado esencialmente en investigaciones médicas y psicológicas como soporte para las hipótesis, ya que permite rechazar o aceptar la hipótesis nula. Por el contrario, el tamaño del efecto permite conocer que tanto rechazar o aceptar dicha hipótesis, además de otros beneficios que trae su implementación [26]. En la Sección 2.2.1 se ahondará más en el tamaño del efecto $d$ de Cohen.

En el contexto del presente trabajo, el tamaño de efecto dado por $d$ de Cohen, brinda una manera de cuantificar la significancia práctica que tiene una característica. Aquellas características con mayor efecto serán las seleccionadas, y las de menor efecto serán descartadas. Note que este método se ajusta dentro de los métodos de filtro de SC. En este sentido, resulta intuitivo pensar que una característica debe comportarse de manera distinta ante una clase o la otra. De hecho, mientras más explique a la variable objetivo, se tendrán más rangos numéricos que respondan diferente a la clase M o la clase B. En este contexto, se plantea la siguiente hipótesis a resolver: *Las medias para la clase de muestra Maligna (M)*

*y la clase de muestra Benigna (B) no son iguales, por lo tanto, será coherente pensar que una misma característica puede tomar distribuciones divergentes para estas dos clases.* A manera de ilustración, la Figura 1 ejemplifica esta premisa en el mejor de los casos.

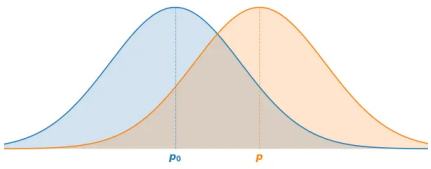



Figura 1: Comportamiento de una variable para dos grupos o clases distintas, color azul para la clase benigna y color amarillo para la clase maligna. Observe como las medias $p_0$ y $p$, se pueden diferenciar claramente entre sí.

El objetivo de este trabajo, es usar el tamaño del efecto $d$ de Cohen como un selector de características de baja complejidad computacional, para ser implementado en un clasificador basado en técnicas de ML y así tener una herramienta de diagnóstico de cáncer de mama en imágenes de núcleos celulares. Es importante destacar que la idea del trabajo no es competir con los clásicos métodos de selector de características, sumamente conocidos en la comunidad científica; sino comparar la efectividad de $d$ de Cohen como selector de características, con el objetivo de proponer y evaluar una nueva alterativa para realizar este procedimiento.

El artículo está organizado de la siguiente manera. La Sección 2 presenta la metodología propuesta, donde se introduce el conjunto de datos experimentales y el protocolo médico usado (Sección 2.1), junto con tres selectores de características; el selector propuesto basado en el tamaño del efecto $d$ de Cohen y dos selectores clásicos como la Cuantización de Vectores de Aprendizaje (LVQ) y la Eliminación Recursiva de Características (RFE) (Sección 2.2), luego, en la sección 2.3 se introduce el modelo de clasificación a través de la Máquina de Vectores de Soporte (SVM). La experimentación se presenta en la Sección 3 y se discuten los resultados. Finalmente, en la Sección 4 se presentan las conclusiones y los trabajos futuros.

## 2. Metodología

Actualmente, las técnicas de ML y los selectores de características son novedosos y eficaces, ya que permiten, bajo reglas lógicas y expresiones matemáticas, llegar a resultados certeros al tener un mejor procesamiento de los datos [27]. Naji et al. [28] afirma que los modelos de ML pueden contribuir al diagnóstico temprano de cáncer mamario, y Shanbehzadeh et al. [29] entiende que los algoritmos de ML pueden diagnosticar rápidamente el cáncer de forma efectiva.

En contraposición, los médicos se valen de las herramientas de ML como una ayuda para la toma de decisiones importantes. Gracias a la selección de características, estas sirven como métodos de simplificación de los trabajos en el área de diagnóstico, en primer lugar por disminuir la complejidad de los datos, y en segundo lugar por realizar trabajos automatizados. Asimismo, es importante destacar que la implementación de estas estrategias significa una reducción en los costos computacionales y una mayor versatilidad en el uso de la información. El estudio del diagnóstico de mama basada en modelos de ML es un tema actual de investigación en el avance y el desarrollo de las afecciones cancerosas, debido a la significativa capacidad de detectar cambios importantes o patrones en conjuntos de datos complejos y con mucho ruido [16, 30, 27, 28, 31, 32, 33, 29, 34, 35]. La Metodología propuesta se resume en tres grandes etapas, ver Figura 2. En la primera etapa (color verde) se introduce el protocolo médico, donde a la imagen de la biopsia de interés, se le extraen 10 características de los núcleos celulares. A estas características, se les hacen diversas mediciones estadísticas, como la media, el valor extremo y la desviación estándar, formando 30 características totales. En la segunda etapa (color cyan) se propone el selector de características usando el tamaño del efecto $d$ de Cohen, y se compara con dos selectores clásicos de la literatura, LVQ y RFE:

1. Análisis del tamaño del efecto a partir de la $d$ de Cohen, para cuantificar la diferencia de las distribuciones de una misma variable para las clases de muestras benignas ($B$) y malignas (M).

Este método es el principal aporte a esta investigación.

2. LVQ divide en regiones una misma variable, definiendo un vector que caracteriza las clases de muestras, como benignas ($B$) o malignas (M). Este método se usa para contrastar nuestro método propuesto.

3. RFE permite probar todas las distintas combinaciones de las variables y entregar un subconjunto funcional de las mismas para poder ser usadas eficientemente en la separación de clases. Este método se usa para contrastar nuestro método propuesto.

Finalmente, en la tercera etapa (Color azul), se construye un plano de separación usando el selector de características dado por cada método. Note que cada grupo de características es la entrada a un clasificador SVM para detectar si una muestra de biopsia tiene presencia de cáncer o no. A continuación se introducen todos los métodos usados en este trabajo.

## 2.1. Base de Datos y software

Para fines de experimentación se trabajó con la base de datos *Breast Cancer Wisconsin*, de la universidad de California, Irvine [6], la cual usa en el protocolo médico explicado en la Sección del Apéndice 6. En esta base de datos se extraen pequeñas muestras de la biopsia de mama mediante aspiración con aguja fina (FNA). Posteriormente, se digitalizan las muestras en imágenes, para poder procesarlas con modelos de contornos activos conocidos como serpientes. El objetivo de este modelo de contorno activo, es extraer las siguientes 10 características iniciales de los núcleos celulares: *radius, perimeter, symmetry, concavity, area, compactness, smoothness, texture, concave points y fractal dimension*. Para cada una de estas características se calculó la media (*mean*), desviación estándar (*se*) y máximo valor (*worst*) considerando cada muestra extraída, obteniéndose un total de 30 características por imagen, llamadas: *radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, fractal dimension mean, radius se, texture se, perimeter se, area se, smoothness se, compactness se, concavity se, concave points se, symmetry se, fractal dimension se, radius worst, texture worst, perimeter worst, area worst, smoothness worst, compactness worst, concavity worst, concave points worst, symmetry worst, fractal dimension worst.* Observe que estas características corresponden a mediciones sobre los núcleos celulares. El dataset contiene 569 observaciones divididas en 357 casos benignos (clase B) y 212 casos malignos (clase M). Para mayor detalles visite [36].

La implementación de los siguientes métodos fueron hechas en *RStudio 2023.09.1+494, Desert Sunflower Release.*

## 2.2. Selector de características

### 2.2.1. Análisis del tamaño del efecto d de Cohen

El tamaño del efecto se refiere a una forma de cuantificar la magnitud de la diferencia entre dos grupos. Corresponde a una parte especial en la estadística descriptiva, donde se estima la magnitud de una relación, independientemente de si este vínculo aparente en los datos, refleja o no, una verdadera relación en la población [37]. Aún así, está íntimamente asociado con el *p-value*, de hecho, el tamaño del efecto surge de medir la relevancia de la diferencia estadísticamente significativa [38]. Para estos tipos de hallazgos, los cuales conllevan el rechazo de la hipótesis nula, pueden ser irrelevante cuando se tiene magnitudes bajas.

En la práctica, el tamaño del efecto es particularmente prominente en la investigación social y médica para hacer comparaciones entre dos distribuciones, una de control y otra de tratamiento, y de esta manera, evaluar que tan grande es la diferencia o el impacto de un tratamiento sobre la muestra de control [37]. En este trabajo se propone usar el tamaño del efecto, dado por $d$ de Cohen para medir la magnitud de la diferencia entre dos grupos de variables, permitiendo determinar cuáles características, en realidad, son las que poseen mayor importancia.

Básicamente, existen tres formas de medir el tamaño de efecto entre dos distribuciones [37, 38, 39], las cuales se describen a continuación:

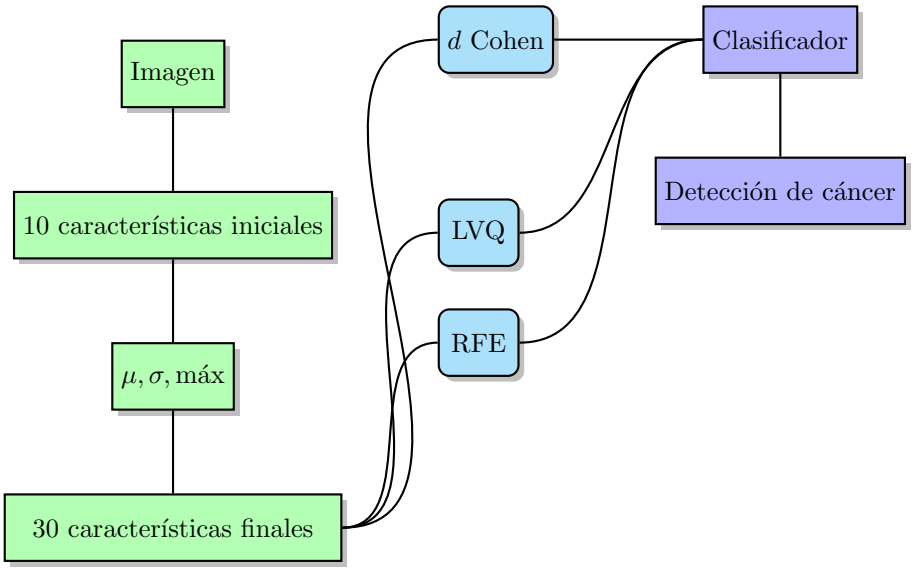



Figura 2: Diagrama de bloques que ilustra la metodología propuesta. El color verde hace referencia al protocolo médico. $\mu, \sigma$, máx corresponden a la media, la desviación estándar y el valor extremo respectivamente. El color cyan corresponde a los selectores de características estudiados. El color azul corresponde al clasificador para detectar cáncer.

- Diferencia de medias estandarizadas: Es de particular interés cuando se pretende estudiar la diferencia de las medias de dos grupos. La formula más popular que se emplea en estos casos es la $d$ de Cohen, ver Ecuación (1).
- Coeficiente de correlación: Intrínseco para calcular la relación cuantitativa y lineal de dos variables. En estos casos el tamaño del efecto se calcula con el coeficiente de correlación de Pearson.
- Razón de probabilidad: Los bien conocidos *odds ratio* son de interés cuando se pretende estudiar las probabilidades de éxito en un grupo de tratamiento, en relación con las probabilidades de éxito en un grupo de control.

En este trabajo se usa la diferencia de medias estandarizadas, ya que es de interés medir la diferencia de las medias que toma cada variable para las dos clases en estudio, clase Maligna (M) y clase Benigna (B), como se ilustró en la Figura 1.

La $d$ de Cohen puede ser definida de forma general, de la siguiente manera:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{PooledSD} \tag{1}$$

Siendo $\overline{x}_1$ y $\overline{x}_2$ las medias correspondientes para el primer y segundo grupo respectivamente y $PooledSD$ la desviación estándar combinada de los dos grupos. No obstante, para los casos en el que los tamaños de los grupos sean iguales, la $d$ de Cohen se define de la siguiente manera:

$$d = \frac{\overline{D}}{SD(D)} \tag{2}$$

Donde $\overline{D}$ es la media de las diferencias, definida como $D_i = x_{1,i} - x_{2,i}$, la diferencia entre el valor $i$ del primer grupo y del segundo grupo, y $SD(D)$ es la desviación estándar de las diferencias. Para los demás casos en el que se cuenta con tamaños de muestras desiguales en los grupos, se emplea la Ecuación (1) expresando la desviación típica combinada como:

$$PooledSD = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}} \tag{3}$$

Donde $n_1$ y $n_2$ es el tamaño de la muestra para el grupo 1 y el grupo 2 respectivamente [25].

Las interpretaciones de los valores $d$ de Cohen dependen del tamaño de la diferencia entre los dos grupos. Por ejemplo, un valor $d = 3$ representa que la diferencia entre la media del grupo 1 y del grupo 2, es de tres desviaciones estándares. Cohen [25] propone la siguiente regla de decisión de valores tipificados que depende de cada caso práctico en el que se esté aplicando:

- Valores $d$ inferiores a 0,2: Nula existencia de alguna diferencia.
- Valores $d$ entre 0,2 y 0,5: Serán entendidos como una diferencia pequeña.
- Valores $d$ que oscilan entre 0,5 y 0,8: Pueden significar una diferencia moderada.
- Valores $d$ superiores a 0,8: Señalan una diferencia o tamaño de efecto grande.

En este trabajo, se propone usar el tamaño del efecto a través de los $d$ de Cohen con los valores más grandes, es decir, los pesos que maximizan la $d$ de Cohen para cada clase. Estos pesos serán entonces los candidatos como selectores de características. A manera de ilustración, la Figura 3 muestra 3 ejemplos de esta idea. Observe que para un valor $d = 0{,}2$, el cual es pequeño, las clases están superpuestas y la maximización es pequeña, para un valor $d = 0{,}5$, el cual es medio, las clases empiezan a diferenciarse y su maximización es media; y para un valor $d = 0{,}8$, el cual es grande, las clases se diferencian más claramente y por lo tanto, su maximización es grande.

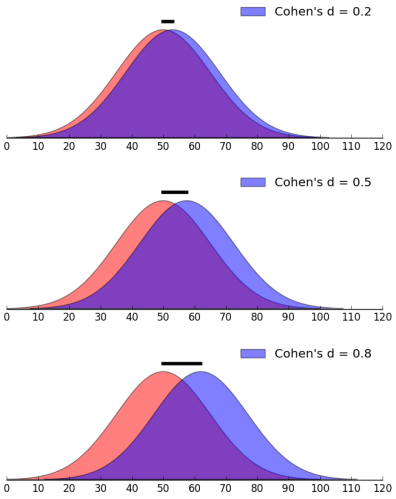



Figura 3: Tamaño del efecto para diferentes $d$ de Cohen para dos clases. Observe que a mayor valor $d$, mayor es la diferenciación entre las clases.

El objetivo es usar $d$ de Cohen como un estimador del tamaño de efecto que funcione como un selector y reductor de características. Los resultados obtenidos, se contrastan con dos métodos computacionales clásicos más complejos que se introducen en las secciones 2.2.2 y 2.2.3. El selector de características propuesto basado en $d$ de Cohen se puede resumir en el Algoritmo 1. Note que para cada variable en el conjunto de características se extra el grupo de datos pertenecientes a la clase M ($k = 1$) y a la clase B ($k = 2$), computando finalmente el calculo de la Ecuación (1) con la Ecuación (3). En este sentido se obtiene un valor $d$ para cada característica, seleccionando en última instancia aquellas que cumplen con $d > 0{,}8$. Consulte [25, 26, 17] para un tratamiento completo de las propiedades matemáticas del tamaño de efecto $d$ de Cohen.

**Algorithm 1** Método propuesto basado en $d$ de Cohen

**Require:** $S = \{X_1, X_2, X_3, \ldots, X_n\}$
  **for** $X_i \in S$ **do** ▷ Para cada característica $X_i$ del conjunto de características $S$ repetir:
    Sea $M_i$ los datos pertenecientes a la característica $X_i$ con clase $k = 1$.
    Sea $B_i$ los datos pertenecientes a la característica $X_i$ con clase $k = 2$.
    Calcular la diferencia de medias estandarizadas de $M_i$ y $B_i$ mediante la Ecuación (1)
  **end for**
    **return** $S' = \{X_1^{'}, X_2^{'}, X_3^{'}, \ldots, X_n^{'}\}$ con $d > 0{,}8$ ▷ Subconjunto final de características con valores $d$ de Cohen mayores a 0.8

### 2.2.2. Aprendizaje por Cuantificación Vectorial (LVQ)

LVQ es una red neuronal dedicada en parte a la reducción de dimensionalidad [40]. A grandes rasgos, LVQ propuesta por Kohonen et al. [41], se enfoca en dividir el espacio de características en regiones que corresponden a las clases, asignando pesos a los nodos de la red. Los pesos se ajustan de acuerdo a medidas de distancia, ver Ecuación (4), como es en el caso de un clasificador k-NN, permitiendo, a través de un aprendizaje competitivo, clasificar los datos basado en la cercanía con la muestra de entrenamiento. LVQ reajusta los pesos midiendo la longitud entre el vector de entrada $x$ y el vector prototipo $c$ mediante la Ecuación (4). Siendo $c$ el vector que representa a las regiones divididas.

$$dist(x, c) = \sum_{i=1}^{n} (x_i - c_i)^2 \tag{4}$$

Donde $n$ es el numero de características y *dist* es la distancia euclidiana.

Posteriormente, LVQ hace uso de estos pesos para asignarle una clase a los datos de entrada [41, 42]. LVQ es una red de dos capas, una capa de datos de entrada y otra capa de datos de salida, como se ilustra en la Figura 4. La idea principal de LVQ es emplear los puntos de entrenamiento $x_i$ para atraer los vectores prototipos $C$ a la clase correcta y repeler otros prototipos, reduciendo a su vez, la tasa de aprendizaje $\epsilon$ en cada iteración hasta que tienda a cero. El Algoritmo 2 describe el funcionamiento del algoritmo de LVQ [43].

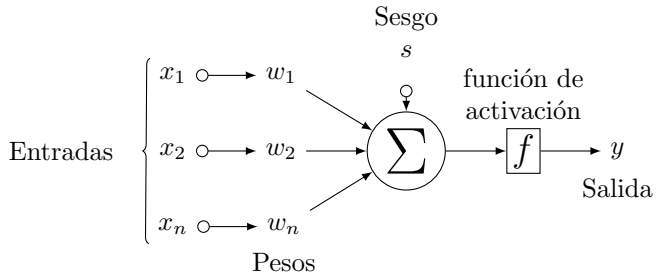



Figura 4: Diagrama de la arquitectura LVQ. Las entradas $x$ corresponden a las características, el sesgo desplaza el resultado obtenido a través de la función de activación, la cual es lineal para LVQ, la salida $y$ puede ser $y = 1$ para la clase $M$ o $y = 0$ para la clase $B$

### 2.2.3. Eliminación Recursiva de Características (RFE)

RFE es un método envolvente de selección de características, es decir, elimina aquellos atributos redundantes y de poca correlación con una variable objetivo. El proceso de eliminación no causa aumentos de error o ruido en una etapa de entrenamiento, dejando de esta manera, las *características* más independientes y fuertes que permitan mejorar la capacidad del modelo de generalizar los datos [44]. La operatividad del método está basado en un procedimiento iterativo de eliminación de características recursiva, a esto se lo conoce comúnmente como selección en retroceso (*Backward Selection*). Esta técnica

**Algorithm 2** Aprendizaje por cuantificación vectorial

Elegir $C$ prototipos iniciales para cada clase: $m_1^{(k)}, m_2^{(k)}, ..., m_R^{(k)}$, con $k = 1, 2, ..., K$. Estos prototipos pueden seleccionarse aleatoriamente a partir de los puntos de entrenamiento de cada clase.

**while** $\epsilon > tol$ **do** ▷ Reducir la tasa de aprendizaje $\epsilon$ en cada iteración hasta que tienda a cero.

Seleccionar aleatoriamente un punto de entrenamiento $x_i$.

Encontrar el prototipo más cercano $m_j^{(k)}$ a $x_i$.

Si $g_i = k$ (es decir, pertenecen a la misma clase), actualizar el prototipo moviéndolo hacia el punto de entrenamiento:

$$m_j^{(k)} \leftarrow m_j^{(k)} + \epsilon(x_i - m_j^{(k)}) \tag{5}$$

Si $g_i \neq k$ (es decir, pertenecen a clases diferentes), actualizar el prototipo moviéndolo lejos del punto de entrenamiento:

$$m_j^{(k)} \leftarrow m_j^{(k)} - \epsilon(x_i - m_j^{(k)}) \tag{6}$$

**end while**

implica la construcción de un modelo de aprendizaje automático considerando el total de las características, las cuales clasifica de acuerdo con su importancia al eliminar las características de bajo beneficio, reconstruye el modelo y recalcula la importancia nuevamente, hasta llegar a un tamaño de subconjunto de atributos óptimo. RFE utiliza el modelo de RF, por su particularidad de evaluar la importancia de los predictores basándose en el decrecimiento de la impureza Gini en los nodos [45]. Los Algoritmos 3 y 4 describen el funcionamiento de RFE [9, 43, 46]. Sea $S$ el conjunto de características de dimensión $\sigma$. La meta es reducir $S$ a una dimensión óptima $\sigma_0$, de tal manera que conserve las características más relevantes de los datos al reducir de dimensionalidad. De esta forma RFE entrena y evalúa un modelo RF, primero con todo $S$ y luego con cada subconjunto de característica encontrado $S_p$ de dimensión $\hat{\sigma_0}$ hasta encontrar recursivamente, aquel subconjunto con una dimensión óptima. Posteriormente se emplea el decrecimiento de la impureza Gini, ver Ecuación (7) para medir la importancia de cada característica. RF entrena $B$ arboles de decisión con una muestra de bootstrap de tamaño $N$ con $m$ características seleccionadas aleatoriamente y punto de corte, dado por la moda, que corresponde a la predicción más frecuente entre los árboles. La decisión final se toma por una votación del promedio de la predicción de los árboles.

**Algorithm 3** Eliminación recursiva de características en RF

**Require:** $S = \{X_1, X_2, X_3, \ldots, X_\sigma\}$ ▷ Conjunto de características

**while** $\hat{\sigma_0} \neq \sigma_0$ **do**

Entrenar un RF inicialmente con $S$ y luego recursivamente con cada subconjunto de características $S_p,\ 1 \leq p \leq \sigma$ usando el Algoritmo (4)

Obtener el coeficiente de Gini para cada característica $p$:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2 \tag{7}$$

▷ Donde $n$ es la cantidad de clases y $p_i$ la proporción de casos de la clase $i$ en el nodo

Remover las características con el menor coeficiente de Gini

Calcular la exactitud del subconjunto de características $S_p$

**end while**

**return** Subconjunto $S_p$ de dimensión $\sigma_0$

**Algorithm 4** Random Forest

**for** $b = 1$ hasta $B$ **do**
  Obtener la muestra de bootstrap $Z^*$ de tamaño $N$ del conjunto de entrenamiento
  **while** $n_i^b \neq n_{min}$ **do** ▷ Para cada nodo terminal $n_i$ del árbol $T_b$ realizar los siguientes pasos hasta alcanzar el tamaño del nodo mínimo $n_{min}$
    Seleccionar $m$ variables aleatorias entre $p$ características
    Seleccionar la mejor variable y punto de corte del subconjunto $m$
    Dividir el nodo en dos nodos hijos
  **end while**
**end for**
  **return** Ensamble de arboles $\{T_b\}_1^B$

## 2.3. Modelo de clasificación

Una vez hecho la selección de característica, prosigue la etapa de modelado de los datos a partir de la clásica máquina de vectores de soporte, la cual se introduce a continuación.

### 2.3.1. Máquina de Vectores de Soporte (SVM)

Las SVM están basadas en la idea de minimización del riesgo estructural [47], sintetizados en la extracción de un hiperplano de separación equidistante a los datos más cercanos de ambas clases. De esta forma, se intenta lograr maximizar un margen entre el hiperplano y los puntos de datos. En el momento de construir el hiperplano, se consideran aquellos puntos que caen en la frontera del margen, lo que se conoce como vectores de soporte [48]. El hiperplano aludido no es único, sino que se pueden hallar infinitos hiperplanos que causen una separación en los datos. Encontrar el hiperplano que sea mejor para el modelo representa un problema de optimización [49]. La Figura 5, ilustra el caso del hiperplano óptimo. El hiperplano debe causar un margen máximo entre él mismo y los vectores de soporte, se puede

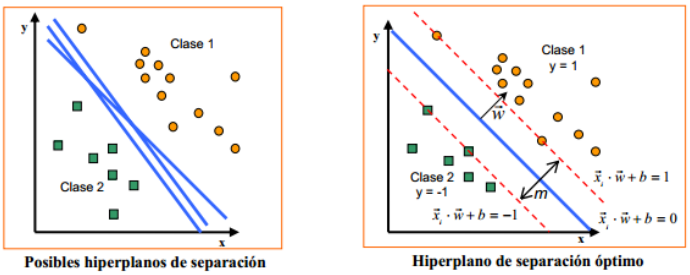



Figura 5: Hiperplano de separación en un espacio bidimensional

visualizar que teniendo un conjunto de datos de $n$ observaciones linealmente separables en un espacio bidimensional en pares $(x_1, x_2)$ con una variable tipo objetivo de dos niveles, $y_i \in \{+1, -1\}$, entonces el hiperplano quedará definido como:

$$\vec{w} \cdot \vec{x}_i + b = 0 \tag{8}$$

Obteniéndose consecuentemente la expresión (9) como fórmula general de la clasificación de un SVM.

$$\begin{cases} y_i = 1, & si \quad \vec{w} \cdot \vec{x}_i + b > 0 \\ y_i = -1, & si \quad \vec{w} \cdot \vec{x}_i + b < 0, \quad i = 1, \ldots, n \end{cases} \tag{9}$$

Donde $\vec{w} \in \Re^p$ es ortogonal al hiperplano y $b \in \Re$ para el supuesto caso que se diera que $\vec{x} \in \Re^p$. Para el ejemplo mencionado $\vec{w} \in \Re^2$. Cabe destacar que a los efectos de *kernalización* de los datos, $\vec{z_i} = \phi(\vec{x_i})$

son los puntos correspondientes al mapeo con una función $\phi$ a un espacio de dimensión $Z$. Por lo tanto, el hiperplano de separación queda redefinido de la siguiente manera [47]:

$$\vec{w} \cdot \vec{z_i} + b = 0 \tag{10}$$

Ahora bien, el concepto de margen está intrínsecamente relacionado con el poder de generalización del modelo, ya que un hiperplano con límites muy acotados tenderá a un sobreajuste (*overfitting*) de los datos [48]. En la práctica, cuando se modela datos con valores atípicos (*outliers*), el margen puede quedar muy estrecho, lo que trae problemas al clasificar nuevos datos. Una solución a este conflicto es permitir que algunos puntos del entrenamiento violen el margen, consiguiendo que este sea más ancho y haya una mayor separación entre los vectores de soporte [48], esto se conoce como margen suave (*soft margin*). Mientras que en un margen duro (*hard margin*), como el que se ilustra en la Figura 5, la clasificación es perfecta y no hay violaciones. El margen suave es medido basándose en un parámetro de regularización $C$ (costo) [50]. El ajuste del mismo genera un balance entre la maximización del margen y el número de violaciones a la clasificación. Valores de $C$ muy grandes producen margenes más estrechos y por lo tanto más sobreajuste, mientras que valores de $C$ muy pequeños incitan margenes más amplios a costa de mayores errores [48]. En estos casos, la ecuación (9) queda incompleta y debe ser incorporada la corrección del costo, que a los efectos del presente trabajo queda excluida su definición. Para un tratamiento integral de SVM remitimos al lector a [49, 51, 43].

## 3. Resultados y discusiones

En el presente documento se trabajó con 357 casos pertenecientes a la clase de muestra benigna (B) y 212 casos pertenecientes a la clase de muestra Maligna (M). 569 observaciones en total. Recordar que las 30 características finales, expuestas en la sección 2.1, surgen de la medición de la media, valor extremo y desviación estándar de las 10 características nucleares iniciales de cada muestra de biopsia. A continuación se analizan los tres métodos de selección de características de estas 30 características.

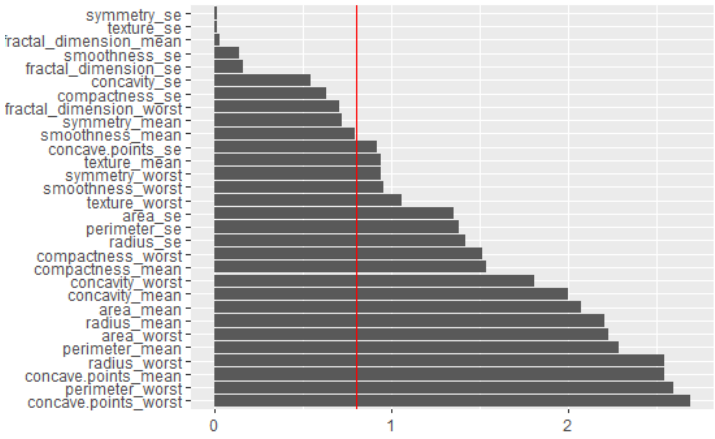
1. Para el análisis de características $d$ de Cohen se empleó la ecuación (1) considerando la desviación típica de la ecuación (3). Recordar que el objetivo es discriminar los casos malignos (M) y benignos (B) a partir de la cuantificación de las características de la sección 2.1. Entonces la idea es, crear una tabla de selección de pesos en orden ascendente, con los valores $d$ de Cohen que son más grandes. En este caso los valores tipificados $d > 0{,}8$ serán los candidatos representativos para tomar como selector de características.

   La Figura 6(a) muestra estos valores de $d$ de Cohen para cada característica. Note que la linea roja representa el corte en 0,8, las características superiores a esta serán las seleccionadas. Se puede observar que *concave points worst*, *perimeter worst*, *concave points mean* y *radius worst* son las características que poseen mayores valores de $d$ de Cohen. Precisamente, la Figura 7(a) muestra el comportamiento de la variable *concave points worst* para muestras benignas (B) y aquellas con cáncer (M). En este ejemplo se tiene una media 0,074 para el primer nivel de la variable objetivo (B), y una media de 0,182 para el segundo nivel (M), tumor mamario, con rangos de 0 a 0,175 y de 0,028 a 0,291 respectivamente, lo que integra a *concave points worst* como un buen detector de cáncer de mamas. Por otro lado, en la Figura 7(b) se observa que hay una superposición entre ambas distribuciones, la de muestras malignas y la de muestras benignas, lo cual hace que no sea un buen candidato para ser un estimador de cáncer. Con el nuevo subconjunto de características seleccionados, se remuestrearon los datos cien veces de forma aleatoria para medir la calidad de las predicciones.

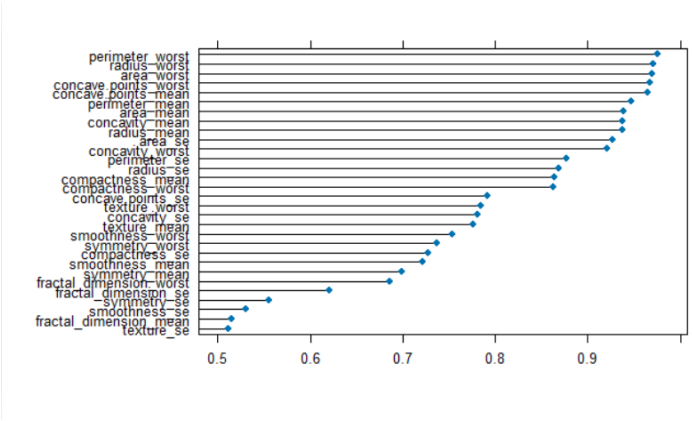
2. Para el método de LVQ se incorporó validación cruzada con 10 carpetas (*10-fold Cross Validation*) repetidos en 10 instancias. En la Figura 6(b) se visualiza los resultados de esta metodología. Se puede observar que *perimeter worst*, *radius worst*, *area worst*, *concave points worst*, *concave points mean* y *perimeter mean* fueron de las características seleccionadas de mayor interés.

   Observe que las características seleccionadas por el método LVQ, *concave points worst*, *perimeter worst*, *concave points mean* y *radius worst* son similares a las encontradas en el análisis de $d$ de Cohen, ver Figura 6(a).

3. La Figura 8 expone los resultados obtenidos al aplicar el selector de características utilizando RFE. Se puede observar la exactitud del modelo para diferentes consideraciones de tamaños de subconjuntos de características. Por ejemplo, el punto sólido más óptimo, son 12 variables que pasan a ser la selección de características más representativas. Están compuestas por: *area worst*, *concave points worst*, *perimeter worst*, *radius worst*, *texture worst*, *concave points mean*, *area se*, *texture mean*, *concavity worst*, *smoothness worst*, *concavity mean* y *area mean*.




(a) Pesos de cada característica $d$ de Cohen. La linea roja corresponde a un valor $d = 0{,}8$

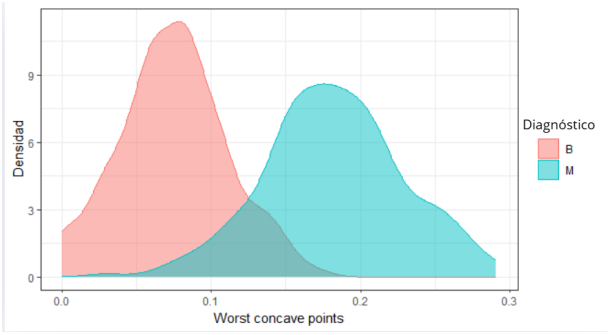



(b) Pesos de cada característica LVQ

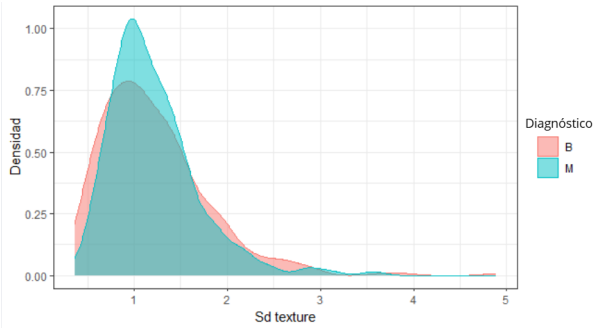
Figura 6: (a) Valor de $d$ de Cohen para cada característica. (b) Importancia de las características con LVQ.

En el Cuadro 1, se expresa las variables pertenecientes al subconjunto de 12 características de RFE. En la columna de *Puntuación general* se dispone de los respectivos índices de importancia para cada atributo. Cabe aclarar que la partición de la tabla no es más que para facilitar la lectura, y donde se ordena de forma decreciente los índices de izquierda a derecha. Es más, de esta labor se puede extraer otro conjunto de 6 predictores que mejores valores obtuvieron. *area worst*, *concave points worst*, *perimeter worst*, *radius worst*, *texture worst* y *concave points mean*.

La selección de las características más significativas, para los distintos métodos, obtuvieron diferentes resultados muy interesantes. Aún así, es importante destacar que aquellos resultados no son más que ligeras modificaciones, y que en todos, la mayoría de las características que ofrecen mejores rendimientos se encuentran presentes. Note que si se compara el Cuadro 1 con la Figura 6(b), el conjunto tomado por LVQ: *perimeter worst*, *radius worst*, *area worst*, *concave points worst*, *concave points mean* y *perimeter mean*, presenta los mismos elementos que la selección llevada a cabo por




(a)




(b)

Figura 7: (a) Densidad de Concave Points Worst, (b) Densidad de Texture Standard Deviation. Ambas para las clases de muestras benignas (B) y malignas (M).

RFE, a excepción de *perimeter mean*. Sin embargo, y algo que es valioso resaltar, estas características mencionadas forman parte de las seis más importantes del RFE que se habló en un principio. En este sentido, el paralelismo entre $d$ de Cohen y LVQ es notable. Adicionalmente, se puede observar lo siguiente:

- Entre las metodologías $d$ Cohen y LVQ, se tiene que, la mayor correlación que puede presentarse entre la selección de características. Retomando la Figura 6(b), la característica *texture se* se encuentra en el fondo de la lista con una importancia apenas superior a 0,5; véase Figura 6(a), lo cual, también ocurre en términos de $d$ de Cohen, puesto que sus distribuciones para los grupos $B$, muestras benignas, y el grupo $M$, muestras con cáncer, son cuasi idénticas, ver Figura 7(b).
- Basado en los resultados con esta base de datos, se pudo evidenciar un desempeño similar entre RFE, LVQ y $d$ de Cohen.

Para evaluar el comportamiento de las características predictivas de cada selector, se usó el modelo SVM. Se realizó una partición en los datos a razón de 70 % para un conjunto de entrenamiento, y un 30 % restante para un conjunto de testeo. La motivación de usar SVM, radica en que es una técnica óptima para resolver problemas de clasificación de este estilo. Diferentes estudios dan soporte y evidencian buenos resultados, Collazo et al. [52] realiza una comparación entre RF, redes neuronales y SVM, de lo que en sus investigaciones concluye de igual manera, que dichos métodos de aprendizaje son herramientas muy útiles, no obstante, SVM logró destacarse entre las tres en la tarea de diagnosticar cáncer de mama, y que incluso, se recomienda su utilización en la detección temprana de la neoplasia. Asimismo, Núñez et al. [53] aplica los mismos modelos que el trabajo anterior analiza, a cuatro bases de datos diferentes, a lo que arriba que SVM obtuvo mejores rendimientos en promedio. En [28, 54] se compararon los modelos de RF, K-NN, regresión logística, árboles de decisión, SVM y Naïve Bayes (NB) con la base de datos *Breast Cancer Wisconsin*, donde

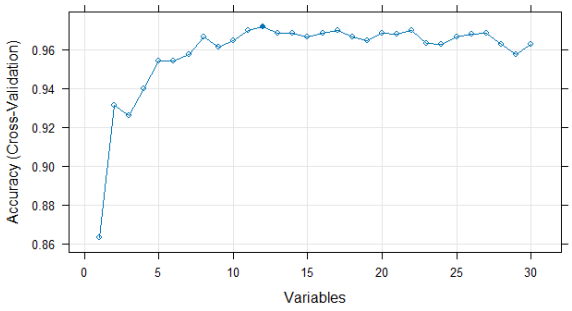



Figura 8: Subconjunto de las 12 características óptimas usando RFE.

SVM tuvo la menor tasa de error, 97.2 % y 97.13 % de accuracy respectivamente. Salod y Singh [2] revisaron la literatura referente a la detección de cáncer mamario usando técnicas de ML en el periodo de 2015 a 2019. Concluyen que el modelo de ensamble de ANN con SVM obtuvo el mejor accuracy, cercano al 100 %; el modelo RF con SVM logró el 97 % de accuracy, mientras que NB junto con clústers con tabla pivot presentó un accuracy del 99 %.

Cuadro 1: Importancia de las 12 características por RFE

| **Conjunto de 12 características** | | | |
|---|---|---|---|
| **Características** | **Puntuación general** | **Características** | **Puntuación general** |
| *Area worst* | 13.720 | *Concave points worst* | 13.558 |
| *Perimeter worst* | 12.815 | *Radius worst* | 12.812 |
| *Texture worst* | 10.736 | *Concave points mean* | 10.353 |
| *Area se* | 9.169 | *Texture mean* | 9.028 |
| *Concavity worst* | 9.009 | *Concavity mean* | 8.038 |
| *Smoothness worst* | 8.026 | *Area mean* | 7.695 |

Para medir la capacidad de identificación de cáncer mamario en el conjunto de datos de testeo, se usó la sensibilidad (TPR) y especificidad (TNR) en los tres selectores de características estudiados: $d$ Cohen, LVQ y RFE. Como es bien sabido, la sensibilidad está dada por $TPR = \frac{TP}{TP+FN}$, donde $TP$ son los casos que el modelo predijo con exactitud muestras con cáncer, y $FN$ son los casos que el modelo predijo una muestra benigna, cuando en realidad se trataba de malignidad; mientras que la especificidad esta dada por $TNR = \frac{TN}{TN+FP}$, donde $TN$ son aquellos casos que se clasificaron correctamente muestras benignas, y $FP$ los falsos positivos. Finalmente, El Cuadro 3 muestra los resultados para cada una de las características seleccionadas, usando el modelo clasificador SVM através del kernel RBF (*Radial Basis Function*). Precisamente, los kernels RBF y lineal en la clasificación de las clases de muestras benignas (B) y malignas (M), mostraron los mejores resultados frente a otros tipos de kernels, tales como el sigmoideo y el polinómico. El uso de SVM se respalda en diversos trabajos que han obtenido un buen rendimiento en la detección de neoplasia [36, 55, 56].

Cabe resaltar, que si bien en [57] se puede encontrar que $d$ de Cohen se usó como selector de características en el dataset Breast Cancer Wisconsin, no se proporcionó una investigación metodológica, y no tiene el animo de evaluar los distintos tamaños de efecto como el presente trabajo. Aún así, la propuesta resulta ser novedosa y puede brindar un punto de inicio usando el lenguaje de programación Python. De hecho, [57] utilizó una formula del efecto $d$ de Cohen ligeramente diferente a la utilizada en el presente trabajo para seleccionar características. En su desarrollo la diferencia de las medias estandarizadas se usó mediante la ecuación (1), pero la desviación combinada de la ecuación (3) fue definida sin la corrección de Bessel. El hecho no resulta significativo pues los valores estimados de $d$ de Cohen son similares. Posteriormente realiza algunas observaciones en las características y calcula el valor-$p$ para cada una de ellas. Finalmente selecciona aquellas características con un $d$ de Cohen superiora 0,4 y un $p \leq 0{,}05$, en el

subconjunto final quedan expresadas 25 características de 30. Lo cual es destacable en comparación con este trabajo, donde se propone seleccionar aquellas características superiores a un $d = 0{,}8$, obteniendo 20 características de 30.

En el presente trabajo se realizo un remuestreo aleatorio de los datos para asegurar que los resultados obtenidos no se debieran al azar, un remuestreo de cien instancias para ser precisos, como se aclaro anteriormente. Sin embargo, en el trabajo llevado a cabo por [57] no se evidencia ni se indica alguna técnica de remuestreo utilizada que permita asegurar la efectividad del modelo, por lo que no se puede hacer un comparación objetiva sobre el funcionamiento de ambos modelos. De esta manera con la propuesta de $d$ de Cohen usando SVM se obtuvo un AUC de $0{,}9952 \pm 0{,}003$, mientras que [57] obtuvo un AUC de 0,9814 usando el clasificador *Light Gradient-Boosting Machine* (LGBM) [58, 59]. El Cuadro 2 muestra las métricas evaluadas para $d$ de Cohen usando un clasificador SVM.

| Selector | ACC | TPR | TNR | F1-Score | AUC |
|---|---|---|---|---|---|
| $d$ de Cohen | $0{,}9724 \pm 0{,}011$ | $0{,}9611 \pm 0{,}024$ | $0{,}9790 \pm 0{,}012$ | $0{,}9781 \pm 0{,}008$ | $0{,}9952 \pm 0{,}003$ |
| LVQ | $0{,}9486 \pm 0{,}015$ | $0{,}9183 \pm 0{,}032$ | $0{,}9682 \pm 0{,}020$ | $0{,}9595 \pm 0{,}012$ | $0{,}9843 \pm 0{,}008$ |
| RFE | $0{,}9715 \pm 0{,}011$ | $0{,}9506 \pm 0{,}024$ | $0{,}9838 \pm 0{,}011$ | $0{,}9775 \pm 0{,}009$ | $0{,}9927 \pm 0{,}005$ |

Cuadro 2: Métricas del clasificador SVM según cada selector, remuestreando los datos cien veces de forma aleatoria para medir la calidad de las predicciones. ACC = Exactitud, TPR = Razón de Verdaderos Positivos, TNR = Razón de Verdaderos Negativos, AUC = Área Bajo la Curva.

Otro tema de interés es la complejidad computacional en términos de *Big* $\mathcal{O}$, $d$ de Cohen tiene una complejidad lineal, LVQ tiene una complejidad exponencial y RFE tiene una complejidad recursiva. El Cuadro 3 muestra un resumen de los predictores característicos seleccionados por cada selector, junto con su complejidad computacional.

| Selector | Predictores característicos | Complejidad |
|---|---|---|
| $d$ de Cohen | *radius mean*, *texture mean*, *perimeter mean*, *area mean*, *compactness mean*, *concavity mean*, *concave points mean*, *radius se*, *perimeter se*, *area se*, *concave points se*, *radius worst*, *texture worst*, *perimeter worst*, *area worst*, *smoothness worst*, *compactness worst*, *concavity worst*, *concave points worst* y *symmetry worst* | $\mathcal{O}(n)$ |
| LVQ | *perimeter worst*, *radius worst*, *area worst*, *concave points worst*, *concave points mean* y *perimeter mean* | $\mathcal{O}(2^n)$ |
| RFE | *area worst*, *concave points worst*, *perimeter worst*, *radius worst*, *texture worst*, *concave points mean*, *area se*, *texture mean*, *concavity worst*, *smoothness worst*, *concavity mean* y *area mean* | $\mathcal{O}(n \log(n))$ |

Cuadro 3: Predictores propuestos para cada selector. $\mathcal{O}$ corresponde a la complejidad del selector en términos de Big $\mathcal{O}$.

# 4. Conclusiones y trabajo futuros

El presente trabajo propone usar el tamaño del efecto $d$ de Cohen como un selector de características de los datos. El clasificador SVM usando las características estimadas con $d$ de Cohen, pudo identificar el 96 % de los casos de muestras con cáncer y el 97 % de las muestras benignas. Estos resultados sugieren que la metodología propuesta, puesta en un ambiente clínico, permite el diagnóstico de la neoplasia de mamas y la existencia de células incipientes de cáncer en pacientes con estas patologías. Este trabajo mostró que las características *area worst* y *concave points worst* ilustraron la clasificación de las muestras, brindando muy buenos puntajes de importancia bajo distintos métodos de selección de características, lo cual permite detectar el cáncer incipiente en imágenes de núcleos celulares. Asimismo, se demostró que $d$ de Cohen es potencialmente útil como selector de características, siendo beneficioso para dar una primera aproximación a la importancia de las variables en un problema de clasificación dicotómico.

El mayor limitante del método propuesto, es el supuesto de que los datos siguen una distribución normal. Asimismo, $d$ de Cohen no resulta ser un selector robusto y su uso esta limitado a un espacio bidimensional, siendo así, incapaz de determinar la importancia de las características en un modelo de clasificación y detección multiclase.

La mayor fortaleza radica en que, al tener $d$ de Cohen una baja complejidad computacional, hace posible que sea fácil y rápido de implementar en sistemas embebidos que usan algoritmos de ML para diagnóstico médico. Otro buen punto a la hora de utilizar $d$ de Cohen como selector de características, es que permite dar una explicación lógica que proviene únicamente de los datos, sin recurrir a una algoritmia compleja.

Como trabajos futuros, se planea implementar el tamaño del efecto $d$ de Cohen, basado en sus valores tipificados como un clasificador de células cancerígenas, junto con sus intervalos de confianza [60, 61]. También se puede combinar con técnicas de procesamiento de imágenes para el diagnóstico de neoplasia mamaria, como las utilizadas por [36], para analizar biopsias y detectar cáncer incipiente o para un diagnóstico temprano. El trabajo actual, si bien se presentó $d$ de Cohen como selector de características, se puede combinar con otros tamaños de efecto junto con otras metodologías del estado del arte, para dar un alternativa más robusta. Finalmente, evaluar otras bases de datos, permitirá tener una idea más amplia de cómo ajustar $d$ de Cohen adaptativamente a nuevos datos.

## 5. Disponibilidad del software

El software utilizado en este estudio está disponible en la plataforma de desarrollo colaborativo `https://github.com/Nicolas-Masino/Effect-size`, accessed: 2024-24-03.

## 6. Apéndice

En esta sección se introduce el protocolo médico usado para extraer las 10 Características de los núcleos celulares. Este protocolo se puede encontrar en el trabajo de investigación "*Nuclear Feature Extraction For Breast Cancer Diagnosis*", realizada por *W.N. Street, W.H. Wolberg y O.L. Mangasarian* [36]. En el contexto médico, se realizó una biopsia completa de la masa del seno mediante una tradicional técnica de cirugía no invasiva, llamada aspiración con aguja fina (FNA), cuyo objetivo es poder extraer una pequeña muestra del tejido. Luego de la biopsia se emplearon técnicas de procesamiento de imagen para la identificación de los rasgos de las células. Se examinó individualmente cada una de ellas obteniendo sus características como el tamaño, la textura y otras constituciones específicas que se detallan más adelante.

El material aspirado fue colocado y teñido en un portaobjetos de vidrio. La imagen para el análisis digital fue generado por una cámara de video a color JVC TK-1070U montada sobre un microscopio Olympus con un 63X objetivo y 2.5X ocular. La imagen fue digitalizada por un tablero de captura de marco de color ComputerEyes/RT (Digital Vision, Inc., Dedham MA 02026), con una resolución de $512 \times 480$ pixels, con 8 bits por píxel archivo Targa (TGA). Se desarrolló una interfaz gráfica de usuario, para que una persona externa contorne manualmente con el mouse el límite de cada célula. La interfaz fue elaborada usando X Window System (X11) y el Athena Widget Set en un DECstation 3100. Este trabajo manual se tomó como una primera aproximación de los límites de la célula, que luego fue optimizado con un modelo de contorno activo conocido como *serpiente*, el cual se adapta a la forma exacta de los núcleos. Esto permite un análisis preciso y automatizado del tamaño, la forma y la textura de los núcleos. El sistema de diagnóstico de visión por computador, extrajo diez características distintas de los modelos de contorno activo de los núcleos celulares, las cuales se introducen a continuación, tal como se detallan en [36] y como se usan en este trabajo:

1. *Radius*: Promedio de las distancias del epicentro de la célula hasta el perímetro
2. *Perimeter*: Total de distancias entre los puntos del perímetro.
3. *Area*: Cantidad de los píxeles del interior de las células, sumado a la mitad de los píxeles del perímetro
4. *Texture*: Varianza de los valores de una escala de grises.

5. *Compactness*: Medida que disminuye cuando la célula tiene una forma circular y aumenta cuando es irregular. Sin embargo, suele sesgarse si la célula es alargada o pequeña, la compactación tiende a aumentar sin significar que haya probabilidades de que se esté tratando de cáncer. Básicamente, es la combinación del perímetro y el área de la célula, se calcula como: $Compactness = Perimeter^2/Area$.

6. *Smoothness*: Curvatura de las células medidas como la diferencia entre, el largo de una línea radial y la media del largo de las líneas que la rodean. En la Figura 9(a) se ejemplifica el cálculo de la curvatura.

7. *Concavity*: Capta las irregularidades de la forma de los núcleos celulares. Se trazan cuerdas que unen distintos puntos no adyacentes dentro de la misma frontera de la célula, y se determinan las partes del límite que difiere con el interior de cada cuerda. En este sentido, se mide la magnitud de las concavidades de la célula. Como el parámetro no informa de buena manera si las cuerdas son largas, se utilizan cuerdas pequeñas para captar concavidades pequeñas, ver Figura 9(b).

(a) Smoothness (b) Convavity (c) Symmetry

Figura 9: (a) Smoothness: Curvatura de una célula. (b) Convavity: Cuerdas utilizadas para calcular la concavidad. (c) Symmetry: Segmentos de la simetría de una célula.

8. *Concave points*: Medida del número de concavidades que posee la célula.

9. *Symmetry*: Se calcula de la siguiente manera: se traza un eje central, de tal forma que divida la célula a la mitad, y se miden las diferencias del largo de una línea perpendicular de un lado del eje, con las líneas ortogonales del lado opuesto, hasta la frontera del núcleo. En la Figura 9(c) se visualiza la metodología para estimar la simetría.
Puede suceder que el eje corte el límite del núcleo debido a una concavidad, en este caso, el parámetro no informaría la simetría de la célula.

10. *Fractal dimension*: Describe el perímetro de la célula, utilizando técnicas de aproximación de línea costera para generar rectas, con largos cada vez mayores, que se adapten a los límites del núcleo, Ver Figura 10.

Figura 10: Secuencia de medidas de la dimensión fractal

Finalmente a cada una de estas diez características iniciales, se les estimó el valor medio, el valor extremo (*worst*) y la desviación estándar, logrando tener de esta manera, un total de treinta características de análisis, las cuales fueron introducidas en la sección 2.1.

# Referencias

[1] WHO launches new roadmap on breast cancer. `https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer`, 2023. Accessed: 2025-03-24.

[2] Zakia Salod and Yashik Singh. A five-year (2015 to 2019) analysis of studies focused on breast cancer prediction using machine learning: A systematic review and bibliometric analysis. *Journal of Public Health Research*, 9(1):65–75, 2020. doi: 10.4081/jphr.2020.1772.

[3] Edneia Tani, Nelson Fuentes-Martinez, and Lambert Skoog. A review of the use of fine-needle aspiration biopsy of mammary tumors for diagnosis and research. *Acta Cytologica*, 61(4-5):305–315, 2017. doi: 10.1159/000477373.

[4] Bo Franzén, Gert Auer, and Rolf Lewensohn. Minimally invasive biopsy-based diagnostics in support of precision cancer medicine. *Molecular Oncology*, 18(11):2612–2628, 2024. doi: 10.1002/1878-0261.13640.

[5] Nina S. Shabb, Fouad I. Boulos, and Fadi W. Abdul-Karim. Indeterminate and Erroneous Fine-Needle Aspirates of Breast with Focus on the 'True Gray Zone': A Review. *Acta Cytologica*, 57(4): 316–331, 2013. doi: 10.1159/000351159.

[6] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. doi: https://doi.org/10.24432/C5DW2B.

[7] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[8] Shulin Wang y Sheng Yang Jie Cai, Jiawei Luo. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2017.11.077. URL `https://www.sciencedirect.com/science/article/pii/S0925231218302911`.

[9] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh. *Feature Extraction Foundations and Applications*. Springer, 2006.

[10] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. *Feature selection for High-Dimensional Data (Artificial Intelligence: Foundations, Theory, and Algorithms)*. Springer, 2015.

[11] Beatriz Remeseiro and Veronica Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112:103375, 2019. doi: 10.1016/j.compbiomed.2019.103375.

[12] Cheng; Lu, David; Romo-Bucheli, Xiangxue; Wang, Andrew; Janowczyk, Shridar; Ganesan, Hannah; Gilmore, David; Rimm, and Anant Madabhushi. Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers. *Laboratory Investigation*, 98(11):1438–1448, 2018. doi: 10.1038/s41374-018-0095-7.

[13] Alireza Osareh and Bita Shadgar. Machine learning techniques to diagnose breast cancer. In *2010 5th International Symposium on Health Informatics and Bioinformatics*, pages 114–120, 2010. doi: 10.1109/HIBIT.2010.5478895.

[14] Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2, Part 2):3240–3247, 2009. doi: 10.1016/j.eswa.2008.01.009.

[15] Mukesh; Saraswat and K. V. Arya. Feature selection and classification of leukocytes using Random Forest. *Medical & Biological Engineering & Computing*, 52(12):1041–1052, 2014. doi: 10.1007/s11517-014-1200-8.

[16] Supoj Hengpraprohm and Suwimol Jungjit. Ensemble feature selection for breast cancer classification using microarray data. *Inteligencia Artificial*, 23(65):100–114, 2020. doi: 10.4114/intartif.vol23iss65pp100-114.

[17] Geoff Cumming and Robert Calin-Jageman. *Introduction to the New Statistics. Estimation, Open Science, and Beyond*. Routledge, 2024.

[18] S. Olejnik and J. Algina. Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25(3):241–286, 2000. doi: 10.1006/ceps.2000.1040.

[19] Denis Cousineau. Approximating the distribution of Cohen's d in within-subject designs. *Quantitative methods for psychology*, 16(4):418–421, 2020. doi: 10.20982/tqmp.16.4.p418.

[20] Luis Miguel Sanchez-Loyo, Jesua Guzmán-González, Saúl Ramírez De los Santos, and Franco Sánchez-García. Tamaño del efecto para distribuciones no paramétricas: Concordancia entre medidas para la robustez de análisis en ciencias de la conducta. *Instrumentación, Estadística y Lógica*, 1(14):43–52, 2023.

[21] Sareh Panjeh, Anders Nordahl-Hansen, and Hugo Cogo-Moreira. Establishing new cutoffs for Cohen's d. an application using known effect sizes from trials for improving sleep quality on composite mental health. *International Journal of Methods in Psychiatric Research*, 32(3), 2023. doi: 10.1002/mpr.1969.

[22] Maria del Carmen Carcelen-Fraile, Noelia del Pino Deniz-Ramirez, Jessica Sabina-Campos, Agustin Aibar-Almazan, Yulieth Rivas-Campo, Ana Maria Gonzalez-Martin, and Yolanda Castellote-Caballero. Exercise and nutrition in the mental health of the older adult population: A randomized controlled clinical trial. *Nutrients*, 16(11), 2024. doi: 10.3390/nu16111741.

[23] J. Vaske J. Jerry, Beaman Jay, and A. Miller A. Craig. Practical application of a minimal important percent difference formulation of Cohen's d. *Human Dimensions of Wildlife*, 29(3):269–283, 2024. doi: 10.1080/10871209.2023.2233544.

[24] Lee A Becker. Effect size (ES). University of Colorado springs. `https://www.uv.es/~friasnav/EffectSizeBecker.pdf`, 2000. Accesed: 2025-03-24.

[25] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013. doi: 10.4324/9780203771587.

[26] Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in psychology*, 4:863, 2013. doi: 10.3389/fpsyg.2013.00863.

[27] Javier Pérez Córdova. Técnicas de machine learning aplicadas a la búsqueda de biomarcadores de cáncer de mama. Máster Universitario en Ciencia de Datos, Universitat oberta de Catalunya, 2021. Available at `https://openaccess.uoc.edu/bitstream/10609/127711/6/javipercorTFM0121memoria.pdf`.

[28] Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, and Olivier Debauche. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191:487–492, 2021. doi: 10.1016/j.procs.2021.07.062.

[29] Mostafa Shanbehzadeh, Hadi Kazemi-Arpanahi, Mohammad Bolbolian Ghalibaf, and Azam Orooji. Performance evaluation of machine learning for breast cancer diagnosis: A case study. *Informatics in Medicine Unlocked*, 31:101009, 2022. doi: 10.1016/j.imu.2022.101009.

[30] Freire Hidalgo and Jimmy Mauricio. Machine learning y clustering para detección de cáncer de mama a través de imágenes de mamografía, 2021. Available at `https://dspace.utpl.edu.ec/handle/20.500.11962/29121`.

[31] Jiande Wu and Chindo Hicks. Breast cancer type classification using machine learning. *Journal of Personalized Medicine*, 11(2), 2021. doi: 10.3390/jpm11020061.

[32] Amin Rezaeipanah, Rahmad Syah, Siswi Wulandari, and A Arbansyah. Design of ensemble classifier model based on MLP neural network for breast cancer diagnosis. *Inteligencia Artificial*, 24(67): 147–156, 2021. doi: 10.4114/intartif.vol24iss67pp147-156.

[33] Jhelly Núñez. Investigación sobre el método de predicción del cáncer de mama basado en aprendizaje automático. *Revista de investigación de Sistemas e Informática*, 15:5–12, 12 2022. doi: 10.15381/ri-si.v15i2.23402.

[34] Ghassan Ahmad Ismaeel. Machine learning to diagnose breast cancer. *Przeglad Elektrotechniczny*, 99(1):10–12, 2023. doi: 10.15199/48.2023.01.02.

[35] Inayatul Haq, Tehseen Mazhar, Hinna Hafeez, Najib Ullah, Fatma Mallek, and Habib Hamam. Exploring machine learning classifiers for breast cancer classification. *KSII Transactions on Internet and Information Systems*, 18(4):860–880, 2024. doi: 10.3837/tiis.2024.04.003.

[36] W. Nick Street, W.H. Wolberg, and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 1905:861–870, 1993. doi: 10.1117/12.148698.

[37] Dolores Frías-navarro, Juan Llobell, and Fernando García. Tamaño del efecto del tratamiento y significación estadística. *Psicothema*, 12(2):236–240, 2000.

[38] Mario Enrique Rendón-Macías, Irma Susana Zarco-Villavicencio, and Miguel ángel Villasís-Keever. Métodos estadísticos para el análisis del tamaño del efecto. *Revista Alergia*, 68(2):128,136, 2021.

[39] Mario Luis Iovaldi. Effect size. *Revista Argentina de Cirugía*, 115(3):217–219, 2023.

[40] Alain Guerrero Enamorado and Daimerys Ceballos Gastell. Una evaluaciõn del algoritmo LVQ en una colección de texto. *Revista Cubana de Ciencias Informáticas*, 10(4):154–170, 2016.

[41] Teuvo Kohonen. *Learning Vector Quantization*, pages 245–261. Springer Berlin Heidelberg, 2001. doi: 10.1007/978-3-642-56927-2_6.

[42] Jason Brownlee. Learning vector quantization for machine learning. `https://machinelearningmastery.com/learning-vector-quantization-for-machine-learning/`, 2020. Accesed: 2025-03-25.

[43] Robert Tibshirani y Jerome Friedman Trevor Hastie. *The Elements of Statistical Learning*. Springer, 2017.

[44] Puneet Misra and Arun Yadav. Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal on Emerging Technologies*, 11(3):659–665, 2020.

[45] Burcu F. Darst, Kristen C. Malecki, and Corinne D. Engelman. Using recursive feature elimination in Random Forest to account for correlated variables in high dimensional data. *BMC Genet*, 19(1): 2–6, 2017.

[46] Qi Chen, Zhaopeng Meng, Xinyi Liu, Qianguo Jin, and Ran Su. Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes*, 9(6), 2018. doi: 10.3390/genes9060301.

[47] Gustavo Betancourt. Las máquinas de soporte vectorial (SVMs). *Scientia Et Technica*, XI(27): 67–72, 01 2005. ISSN 0122-1701.

[48] Enrique Carmona Suárez. Tutorial sobre máquinas de vectores soporte (SVM). *ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia, Madrid*, pages 1–27, 2014.

[49] Bernhard Schlkopf and Alexander J. Smola. *Learning with Kernels*. The MIT Press, 2002.

[50] Chih-wei Hsu, Chih-chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. `https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`, 2003. Accesed: 2025-03-25.

[51] Lutz H. Hamel. *Knowledge Discovery with Support Vector Machines*. Wiley, 2011.

[52] Nelson del Castillo Collazo. Predicción en el diagnóstico de tumores de cáncer de mama empleando métodos de clasificación. *Revista de Investigacion en Tecnologias de la Informacion*, 8(15):96–104, 2020.

[53] Pérez Núñez and Jhelly Reynaluz. Investigación sobre el método de predicción del cáncer de mama basado en aprendizaje automático. *Revista de investigación de Sistemas e Informática*, 15(2):5–12, 2022. doi: 10.15381/risi.v15i2.23402.

[54] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83:1064–1069, 2016. doi: 10.1016/j.procs.2016.04.224.

[55] Cheng-Lung Huang, Hung-Chang Liao, and Mu-Chen Chen. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Systems with Applications*, 34(1):578–587, 2008. doi: 10.1016/j.eswa.2006.09.041.

[56] Nishit Kaul, Majid Zaman, Waseem Jeelani Bakshi, Sameer Kaul, Bharti Bhat, and Sheikh Amir Fayaz. Analytical study of breast cancer and treatment techniques. *Procedia Computer Science*, 235: 578–587, 2024. doi: 10.1016/j.procs.2024.04.057.

[57] Feature selection with Cohen effect size. `https://www.kaggle.com/code/cast42/feature-selection-with-Cohen-effect-size/notebook`, 2018. Accesed: 2025-03-25.

[58] Guolin ke. LightGBM. AI for science. `https://github.com/guolinke`, 2024. Accessed: 2025-03-25.

[59] Microsoft LightGBM. `https://github.com/Microsoft/LightGBM`, 2024. Accessed: 2025-03-25.

[60] Alexander Bowring, Fabian J.E. Telschow, Armin Schwartzman, and Thomas E. Nichols. Confidence sets for Cohen's d effect size images. *NeuroImage*, 226:117477, 2021. doi: 10.1016/j.neuroimage.2020.117477.

[61] Nicolás Masino and Antonio Quintero-Rincón. Effect sizes as a statistical feature-selector-based learning to detect breast cancer. In *2024 IEEE Biennial Congress of Argentina (ARGENCON)*, pages 1–7, 2024. doi: 10.1109/ARGENCON62399.2024.10735908.