Clasificador Basado En *Principal Component Analisys (PCA)* Para Predecir La Presencia De Cáncer De Mama A Partir De Información Extraída De Los Núcleos Celulares De Masa Mamaria Con El Método De Aspiración De Aguja Fina (FNA)

Sebastián Rogeles Sánchez y Yuliana Andrea Pérez Guerrero

Trabajo de Grado para optar por el título de Ingeniero Mecánico

Director:
Jabid Eduardo Quiroga Méndez
Doctor en Ingeniería Civil

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería Mecánica
Bucaramanga
2022

Agradecimientos

Agradecemos a Dios, a nuestras familias y amigos por acompañarnos durante nuestro viaje universitario. A todos aquellos que de una u otra forma fueron de ayuda dándonos guía o consejo.

Un agradecimiento especial a nuestro director de trabajo de grado por dedicarnos su tiempo, experiencia y conocimientos para la realización de este proyecto.

Índice general

Índice de figuras

Índice de tablas

Resumen

Título: Clasificador Basado En *Principal Component Analisys (PCA)* Para Predecir La Presencia De Cáncer De Mama A Partir De Información Extraída De Los Núcleos Celulares De Masa Mamaria Con El Método De Aspiración De Aguja Fina (FNA) *

Autores: Sebastián Rogeles Sánchez † y Yuliana Andrea Pérez Guerrero ‡

Palabras Clave: Cáncer de mama, *Data driven,* Inteligencia artificial, *Machine Learning,* PCA.

Descripción: Este trabajo usó PCA como herramienta para reducir las dimensiones de una base de datos de la Universidad de Wisconsin del Departamento de Medicina y Ciencias Computacionales, donde se recoge información de características cuantificables de núcleos celulares de 569 personas y 32 atributos tales como: radio, textura, perímetro, área, simetría, concavidad, suavidad, compacidad, puntos cóncavos y dimensión fractal para desarrollar un clasificador el cual estime con mayor efectividad la presencia de Cáncer de Mama en el paciente usando los datos obtenidos de los núcleos de sus células mamarias contribuyendo al diagnóstico oportuno. La aplicación de PCA a esta información permitió desarrollar un clasificador óptimo que estima la presencia de cáncer de mama en sus etapas iniciales. Este clasificador hará un procesamiento de datos usando PCA para luego aplicar distintos clasificadores de Machine Learning, los resultados de estos serán utilizados en matrices de confusión para posteriormente ser contrastadas y así con el debido análisis e interpretación se corroboren los resultados y se examinen con los parámetros de rendimiento. A pesar de haber arrojado excelentes resultados la aplicación de PCA se comparó sin el uso de esta técnica la cual dio unos parámetros de rendimiento mayores, por lo que se concluye que es mejor incluir toda la base datos en el entrenamiento del clasificador puesto que prima la exactitud al momento de detectar cáncer.

*: Trabajo de grado

†: Facultad de Ingenierías Fisicomécanicas. Director: PhD. Jabid Eduardo Quiroga Méndez

‡: Facultad de Ingenierías Fisicomécanicas. Director: PhD. Jabid Eduardo Quiroga Méndez

Abstract

Title: Principal Component Analysis (PCA) Based Classifier For Predicting The Presence Of Breast Cancer From Information Extracted From Breast Mass Cell Nuclei Using Fine Needle Aspiration (FNA) Method [§]

Authors: Sebastián Rogeles Sánchez[¶] y Yuliana Andrea Pérez Guerrero [‖]

Keywords: Artificial Intelligence, Breast Cancer, Data driven, Machine Learning, PCA.

Description: This thesis used PCA as a tool to reduce the dimensions of a database from the University of Wisconsin Department of Medicine and Computer Science, where quantifiable feature information is collected from cell nuclei of 569 individuals and 32 attributes such as: radius, texture, perimeter, area, symmetry, concavity, smoothness, compactness, concave points and fractal dimension to develop a classifier which more effectively estimates the presence of Breast Cancer in the patient using the data obtained from the nuclei of their breast cells contributing to timely diagnosis. The application of PCA to this information allowed the development of an optimal classifier that estimates the presence of breast cancer in its early stages. This classifier will do a data processing using PCA to then apply different Machine Learning classifiers, the results of these will be used in confusion matrices to later be contrasted and thus with the proper analysis and interpretation the results will be corroborated and examined with the performance parameters. In spite of having yielded excellent results, the application of PCA was compared without the use of this technique, which gave higher performance parameters, so it is concluded that it is better to include the entire database in the training of the classifier, since accuracy at the time of detecting cancer is more important.

[§]: Bachelor Thesis
[¶]: Facultad de Ingenierías Fisicomécanicas. Director: PhD. Jabid Eduardo Quiroga Méndez
[‖]: Facultad de Ingenierías Fisicomécanicas. Director: PhD. Jabid Eduardo Quiroga Méndez

Introducción

El creciente poder de la tecnología, representado con la Ley de Moore, ofrece una capacidad de almacenamiento y de procesamiento inimaginable permitiendo en diversos escenarios utilizar todos los datos disponibles sin tener que hacer previamente agregaciones que disminuyan la cantidad de información, es por tanto que la ciencia de datos tenga tantas aplicaciones y alcances, permitiendo que se aprovechen las distintas técnicas para inferir y calcular modelos a partir de herramientas que identifican correctamente las dinámicas subyacentes que predominan en cualquier fenómeno. Una de estas herramientas es el Análisis De Componentes Principales o PCA por sus siglas en inglés, y es uno de los usos centrales de la Descomposición De Valores Singulares (SVD), adicionalmente PCA ofrece un sistema de coordenadas jerárquico para representar datos correlacionados de alta dimensión dentro de la ciencia basada en datos o *Data-Driven*.

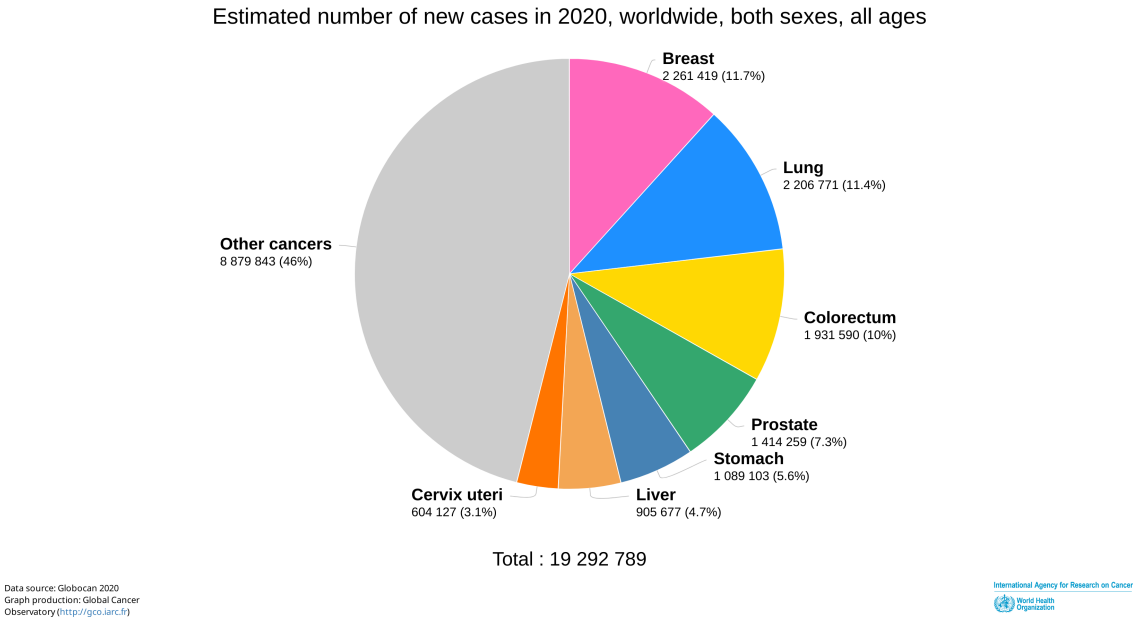
En consecuencia PCA es aplicable a un sin fin de comportamientos y fenómenos, en este caso se usa para la contribución al diagnóstico oportuno de cáncer de mama usando una base de datos de la Universidad de Wisconsin del Departamento de Medicina y Ciencias Computacionales, donde se recoge información de características cuantificables de núcleos celulares de 569 personas y 32 atributos tales como: radio, textura, perímetro, área, simetría, concavidad, suavidad, compacidad, puntos cóncavos y dimensión fractal (Wolberg, Street & Mangasarian, 1995). La aplicación de PCA a esta información permitirá desarrollar un clasificador óptimo que estime la presencia de cáncer de mama en sus etapas iniciales. Este clasificador hará un preprocesamiento de datos usando PCA para luego aplicar distintos clasificadores de *Machine Learning*, los resultados de estos serán utilizados en matrices de confusión para posteriormente ser contrastadas y así con el debido análisis e interpretación se corroboren los resultados y se examinen con los parámetros de rendimiento. Si bien esto puede parecer una aplicación inesperada para la Ingeniería Mecánica, la creciente tasa de casos de cáncer de mama en la región pone sobre la mesa un reto fascinante: ¿Hay algo que pueda hacerse al respecto usando lo que la academia comparte en la formación del pregrado de Ingeniería Mecánica? ¿Hasta dónde puede llegar la interdisciplinaridad del ejercicio del ingeniero mecánico? El *Data-Driven* no es desconocido en esta rama de la ingeniería, la ciencia de datos es ampliamente utilizada en la gestión de vanguardia al punto que actualmente muchas empresas afirman practicarla de una forma u otra, además permite la expansión a temas que antes resultaban inaccesibles en el campo de la investigación.

1. Planteamiento del Problema

Según la Organización Mundial de la Salud - OMS, la principal causa de muerte con aproximadamente 10 millones de defunciones al año es el cáncer, entre todos los tipos más común es el de mama, en el que se detectaron 2,26 millones de casos nuevos a nivel global y que afecta en un 99 % a las mujeres (OMS, 2020).

Figura 1

Nuevos casos en el mundo en 2020.



Cifras verdaderamente alarmantes, aproximadamente la cuarta parte de los nuevos casos corresponden a América y además se prevé que para el 2040 el número de mujeres diagnosticadas aumentará en un 39 % (Organización Panamericana de la Salud – OPS, 2020).

En Colombia las cifras muestran que el cáncer de mama ocupa el primer lugar en prevalencia con el 13.7 % lo que correspondió a 15.509 casos nuevos y 4.411 decesos (OMS, 2020). Bucaramanga y su área metropolitana es donde más se presentan casos de Cáncer de Mama (Alcaldía de Bucaramanga, 2021), esto sucede en gran medida por un diagnóstico tardío en los pacientes.

Figura 2

Nuevos casos en el LATAM y Norte America en 2020.

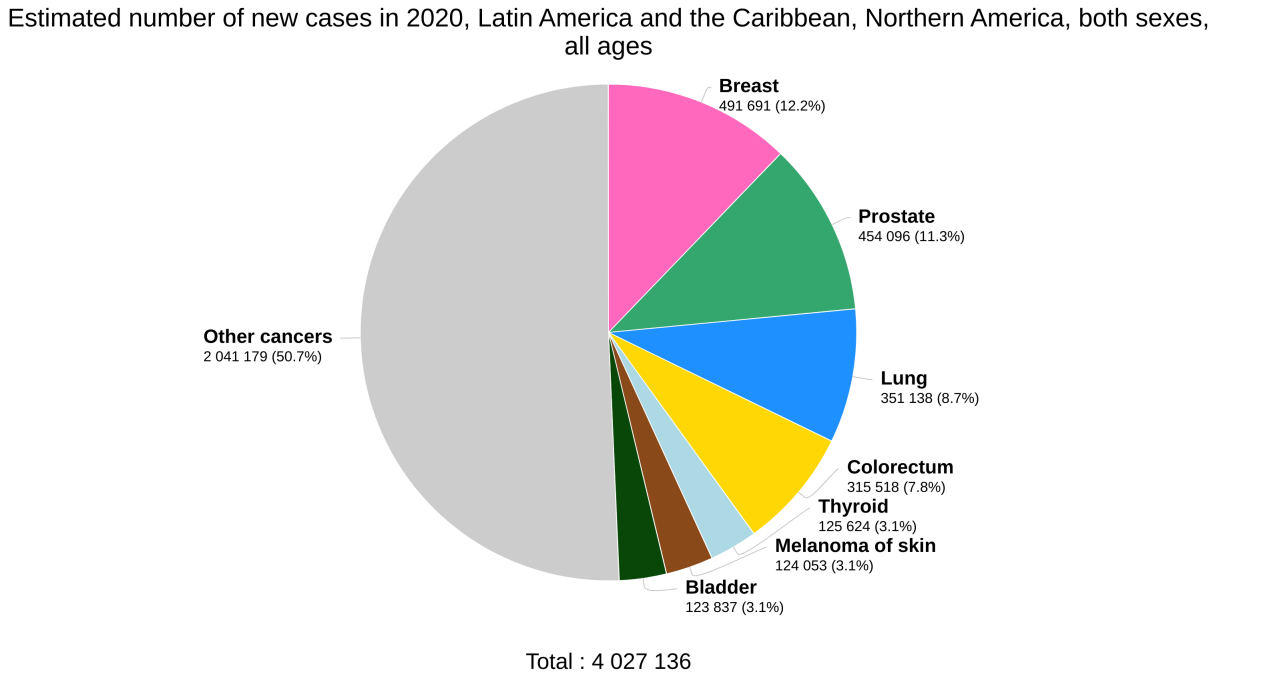
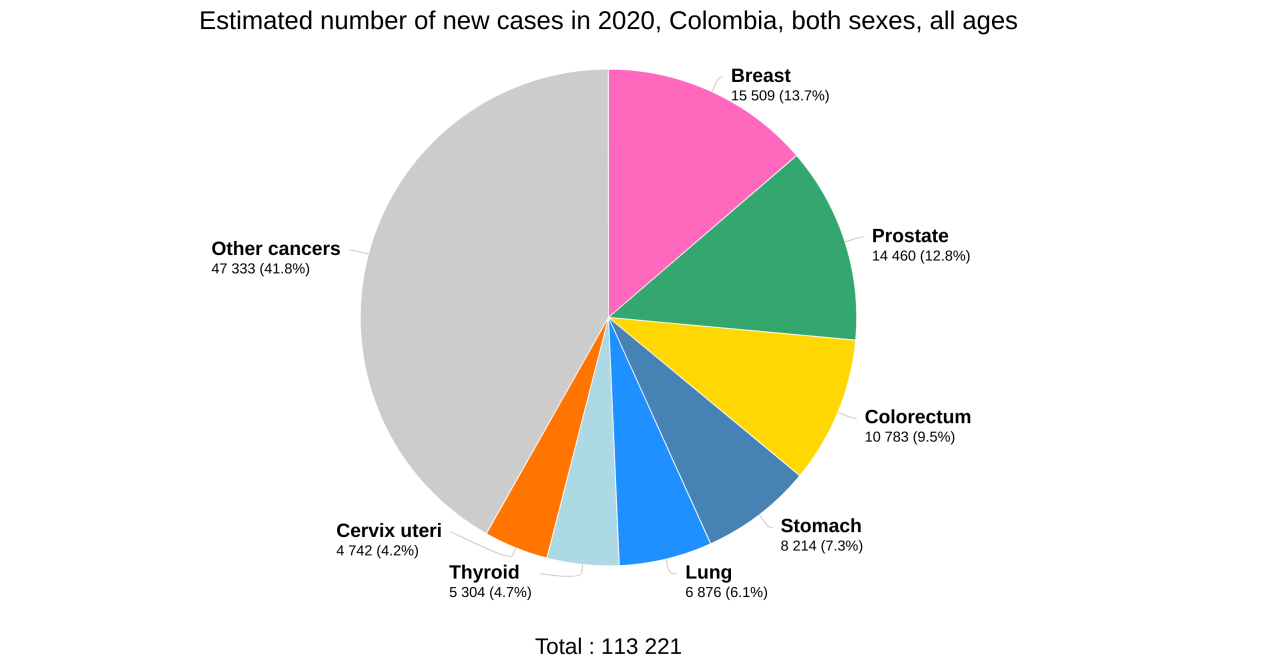


Figura 3

Nuevos casos en Colombia en 2020.



Ante esta problemática y la creciente tasa de casos positivos, se propone el desarrollo de un clasificador, haciendo uso de herramientas tecnológicas, puntualmente Inteligencia Artificial - IA aportando a la implementación de nuevas técnicas para la detección temprana del cáncer de mama en el campo de la biomédica. Se hace hincapié donde se refleja la tendencia a nivel mundial, que busca implementar la digitalización como medio de transición al hacer los procesos de forma eficiente, que, a su vez, sería el caso del proyecto; Desarrollo de un clasificador basado en Principal Component Analisys – PCA en ambiente Matlab para predicción de cáncer de mama, para contribuir a la investigación del diagnóstico oportuno a través de nuevas técnicas y aplicaciones con el fin de reducir los casos avanzados donde la enfermedad ya no es tratable. La Ingeniería Biomédica es una disciplina relativamente nueva en la que se aplican los criterios de ingeniería con un fuerte conocimiento del Tratamiento Computacional de la Información a las ciencias de la vida con sólidos fundamentos en Biología y Medicina (Ingeniería Biomédica, 2022). Teniendo en cuenta todos estos factores previamente mencionados, desde la Universidad Industrial de Santander – UIS, se plantea desarrollar un clasificador basado en Principal Component Analisys – PCA en ambiente Matlab para la predicción de Cáncer de Mama validando el desempeño y capacidad de discriminación del clasificador usando la tasa de positivos reales (TPR) y la tasa de falsos negativos (FNR).

2. Objetivos

2.1. Objetivo general

Desarrollar un clasificador basado en PCA para predecir presencia de cáncer de mama a partir de las características extraídas de los núcleos celulares mamarios, tales como: Radio, Textura, Perímetro, Área, Simetría, Concavidad, Suavidad, Compacidad, Puntos Cóncavos y Dimensión Fractal.

2.2. Objetivos específicos

- Realizar una búsqueda de literatura en el área (Investigación bibliográfica) para seleccionar la *data* que servirá como muestra para el propósito del proyecto, la cual tiene múltiples dimensiones que corresponden a valores discretos y etiquetas para las características analizadas.
- Procesar la base de datos *"Breast Cancer Wisconsin (Diagnostic) Data set"* con PCA para eliminar la redundancia y organizar la información de manera jerárquica de tal forma que se puedan escoger las dimensiones en el dominio de PCA (rasgos) que permitan un mejor desempeño del clasificador.
- Evaluar diferentes clasificadores tales como: Árboles de Decisión, Naive Bayes, Máquina de Soporte Vectorial (SVM), Regresión Logística, de Ensamble y Redes Neuronales; en los datos post-procesados por PCA y determinar el mejor desempeño vía Matriz de Confusión.
- Evaluar el desempeño (predicciones) del clasificador ante datos no usados en su entrenamiento, determinando su Exactitud, Precisión, Sensibilidad/Exhaustividad, Especificidad y BER para luego comparar con los resultados obtenidos en el entrenamiento del *data set*.

3. Justificación

Este proyecto contribuye a la línea de investigación biomédica de la Escuela de Ingeniería Mecánica de la Universidad Industrial de Santander y su aplicación específica es el diagnóstico de Cáncer de Mama en etapas tempranas usando información de los núcleos celulares almacenada en una tabla de datos con características cuantificadas como: radio, textura, perímetro, área, simetría, concavidad, suavidad, compacidad, puntos cóncavos y dimensión fractal; con las cuales se pretende encontrar un patrón de correlación en el diagnóstico de cáncer maligno gracias al estudio de datos. En este sentido, la relación de los datos que se obtiene del comportamiento del cáncer es difícil de establecer, generalmente el análisis basado en datos *(Data-Driven)* busca establecer patrones ocultos en los mismos que permitan clasificar o estimar alguna variable o estado. Por lo tanto, el valor de los sistemas basados en datos es obtener la información de estos sin pretender entender la naturaleza que los generó, esta es una de las grandes ventajas del *Data-Driven*, al final el sistema clasificará sin importar que se trate de una turbina, cáncer de mama o una población. La razón de usar *Data-Driven* y la herramienta PCA es por la trasversalidad que estas ofrecen, pues son ampliamente utilizadas en el análisis exploratorio de datos y en la construcción de modelos predictivos con aplicaciones muy variadas, desde la identificación de notas musicales (Tobón González, & Cortés Osorio, 2018), la clasificación de pacientes en proceso de intubación (González Acevedo, Arizmendi Pereira, & Giraldo-Giraldo, 2015), la evaluación de deformaciones en tuberías (Reyes Combariza, Acuña Hernández, Villamizar Mejía, & Sandoval Cáceres, 2011), o incluso para la relación de compuestos volátiles precursores de aroma y notas sensoriales en variedades del cacao (Gualdrón Zambrano, López Giraldo, Palencia Blanco & Guarín Henao, 2017).

Esta aplicación fue seleccionada para el proyecto debido a que las cifras en Colombia muestran que el cáncer de mama ocupa el primer lugar en prevalencia con el 13.7 %, lo que correspondió a 15.509 casos nuevos y 4.411 decesos (Organización Mundial de la Salud, 2020). Bucaramanga y su área metropolitana es donde más se presentan casos de cáncer de mama (Alcaldía de Bucaramanga, 2021), esto sucede mayormente por el diagnóstico tardío en los pacientes. El Análisis de Componentes Principales (PCA) permite manejar y clasificar enormes cantidades de datos, reduciendo las dimensiones de los datos, determinando proyecciones de datos que produzcan variables significativas para el posterior desarrollo de un clasificador sin perder información, esta tesis responde a la búsqueda de nuevas técnicas para la investigación del diagnóstico oportuno usando PCA como herramienta para reducir las dimensiones de la información permitiendo que un clasificador estime con mayor efectividad la presencia de Cáncer de Mama en el paciente usando los datos obtenidos de los núcleos de sus células mamarias, lo anterior además de sintonizar con el campo investigativo de vanguardia también pertenece al ejercicio de la Ingeniería Mecánica que integralmente contribuye al

mejoramiento de la calidad de vida, proporcionando soluciones óptimas mediante el uso eficiente de los recursos que se tienen a disposición (UIS, 2022).

Los ingenieros mecánicos de la Universidad Industrial de Santander deben conocer y manejar este tipo de herramientas, el siguiente nivel de la gestión que hoy se enseña es la gestión basada en datos la cual, y según un estudio realizado por el *Massachusetts Institute of Technology* (MIT) presenta una productividad entre el 5 % y 6 % mayor a las organizaciones que no cuentan con una cultura de datos (Brynjolfsson, Hitt & Kim, 2011) también registra una reducción de costos cerca al 10 % y un aumento de hasta 11 % en productividad como se evidencia en un estudio de Avanade (Avanade 2019), la automatización de procesos de datos y reportes permite contar con información inmediata y tomar decisiones con *data* "fresca" además de filtrar con mayor facilidad los indicadores que resulten más relevantes mostrando patrones de comportamiento y dejando atrás los cálculos manuales y los extensos reportes de información solicitada, en cambio la cultura *Data-Driven* permite visualizar las fuentes de datos actualizadas que pueden ser representadas en gráficas interactivas destinando más tiempo a la evaluación de la información pertinente y planteamiento de estrategias en lugar destinar tiempo y esfuerzo en el diseño de un plan de acción.

4. Marco Teórico

4.1. Antecedentes

4.1.1. Antecedentes internacionales. Katherine Michelle Gancino Chacha en sus tesis "Enfoque de la Teoría de Juegos en Detección de Cáncer de Mama, Asistido por un Algoritmo Clasificador" Universidad de las Fuerzas Armadas, 2021, tuvo como objetivo: Desarrollar un algoritmo clasificador con teoría de juegos aplicado a la detección de cáncer de mama.

Nataly Marlene Díaz Bernilla en su tesis ."Análisis Comparativo de Clasificadores para la Detección de Subtipos de Cáncer" Universidad Señor de Sipán. Pimentel, 2021 tuvo como objetivo: Comparar clasificadores para la detección de subtipos de cáncer.

Rosana Pirchio en su artículo Çlasificación de Cáncer de Mama con Técnicas de Análisis de la Componente Principal-Kernel PCA, Algoritmos de Máquina de Vectores de Soporte y Regresión Logística" Universidad Tecnológica Nacional. Pimentel, 2022. Concluyó lo siguiente: los resultados de las métricas mejoraron utilizando PCA y kPCA (Pirchio, 2022).

4.1.2. Antecedentes nacionales. Freddy Armando Rodríguez Quintero en sus tesis "Pronóstico de cáncer de mama benigno y maligno: comparación de nueve métodos de clasificación usando R" Universidad Nacional de Colombia. Manizales, 2019, tuvo como objetivo:

> Aplicar y comparar los nueve métodos de clasificación de aprendizaje automático propuestos, al set de datos de cáncer de mama creada por el Dr. William H. Wolberg en el hospital de la Universidad de Wisconsin EE. UU., para encontrar un modelo de predicción con alta capacidad de precisión en la clasificación (Rodríguez, 2019).

"Modelo en *Machine Learning* para el Diagnóstico del Cancer de Mama" (Jorge Millán y Jaime Robles, 2020). En esta tesis de posgrado se desarrollaron 6 diferentes tipos de clasificadores en Phyton, además de una aplicación web en la cual se puedan cargar datos nuevos para predecir presencia de cáncer de mama; además, realizo la clasificación de manera correcta un 100 % del total de los datos destinados para la verificación teniendo una precisión superior a los demás modelos seleccionados.

Juan Camilo Peña Vahos en sus tesis "Despliegue de un modelo de clasificación de tumores de cáncer de mama Reto Kaggle" Universidad de Antioquia. Medellín, 2022, tuvo como objetivo: reducir la necesidad de realizar biopsias para detectar si el cáncer de mama es beningno o maligno.

4.1.3. Antecedentes regionales. A nivel regional es muy poca la información nueva encontrada a pesar de Bucaramanga ser la ciudad donde más se registran casos de cáncer de mama.

"Supervivencia a 5 Años de las Mujeres con Cáncer de Mama de Bucaramanga y su Área Metropolitana según el Estadio Clínico" (Sonia Osma, 2012). En esta tesis de posgrado se hizo un estudio exhaustivo para dar a conocer de manera más clara esta problemática que afecta a Bucaramanga y su área metropolitana.

Eliana Ximena González Morales en sus tesis "Tamización de Pacientes con Sospecha de Cáncer de Mama Mediante Redes Neuronales Artificiales Utilizando Variables Clínicas" Universidad Industrial de Santander. Bucaramanga, 2013, tuvo como objetivo: Selección y validación de un modelo basado en RNA que tamizara a los pacientes sospechosos de cáncer para colaborar en el diagnóstico objetivo a los médicos.

David Enrique Arenas Queeman en sus tesis "Análisis Preliminar de la Termografía como Herramienta para Estimar las Propiedades Térmicas de Estructuras Laminares a Partir de un Modelo Unidimensional de Transferencia de Calor Enfocado a La Detección de Cáncer de Mama" Universidad Industrial de Santander. Bucaramanga, 2019, tuvo como objetivo: evaluar la viabilidad de la termografía como herramienta para estimar las propiedades térmicas de estructuras laminares considerando un modelo de transferencia de calor unidimensional enfocado a la detección de cáncer de mama.

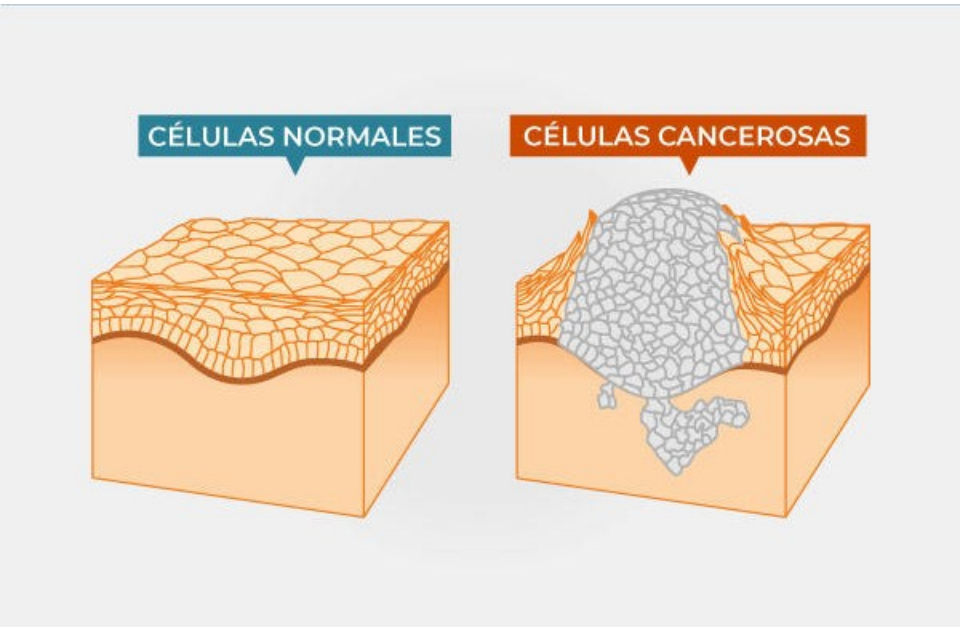
4.2. Definiciones

4.2.1. Cáncer. El *National Cancer Institute* (NCI) de los Estados Unidos define el cáncer como la enfermedad donde sucede la multiplicación sin control de algunas células y la diseminación en otras partes del cuerpo.

En el cáncer las células dañadas o anormales se forman y se multiplican cuando no deberían esto pueded desencadenar la formacion de tumores,estos pueden ser o no cancerosos, los cancerosos se diseminan en los tejidos cercanos e incluso viajan a otras partes del cuerpo para formar tumores en un proceso llamado metástasis (NCI, 2021)

Figura 4

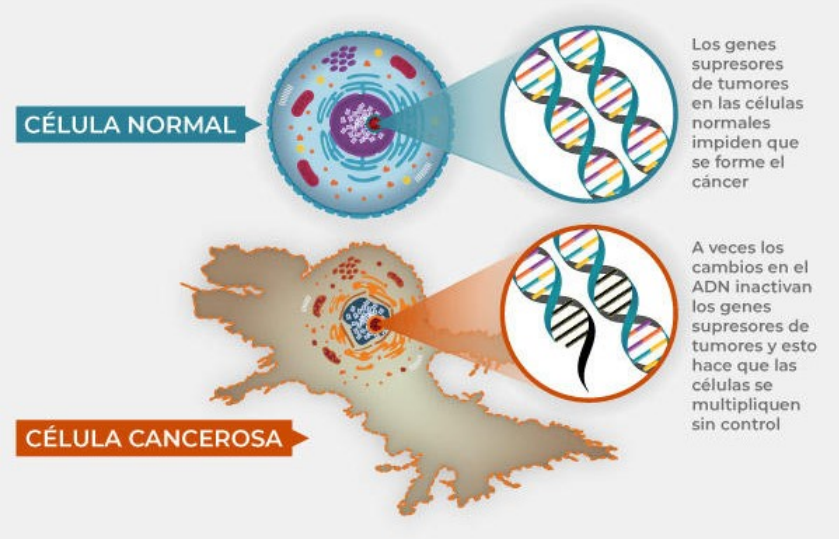
Como se forma el cáncer.



Cuando el cáncer es de origen genético, los genes encargados de controlar el funcionamientode las células se conviertan en oncogenes estos hacen que las células dañadas se multipliquen sin control.

Figura 5

Células cancerosas vs. células normales.

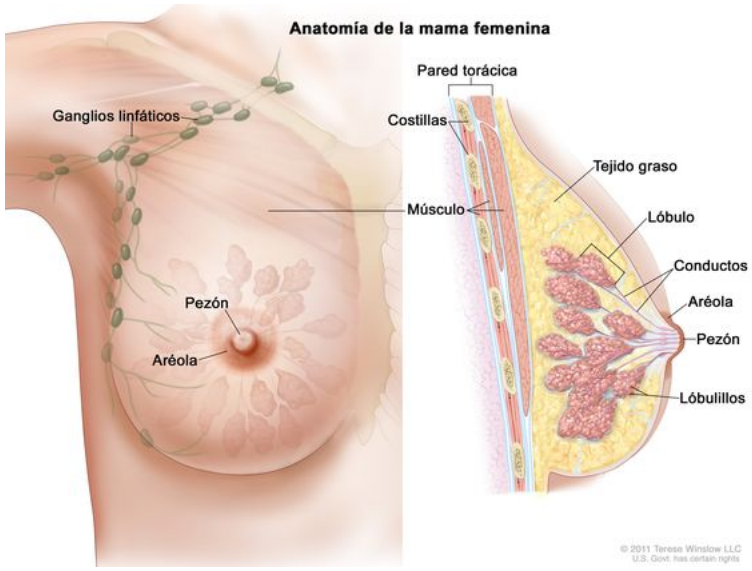


4.2.1.1. Cáncer de mama. Este cáncer puede presentarse en mujeres y hombre, si bien es poco frecuente en estos últimos. El cáncer de mama se forma en los tejidos de la mama y se conoces los siguientes tipos:

- Carcino intraductal
- Carcinoma ducta invasivo
- Carcinoma lobulillar invasivo
- Inflamatorio
- Triple negativo
- Angiosarcoma
- Enfermedad de Paget del pezón
- Tumores filodes

Figura 6

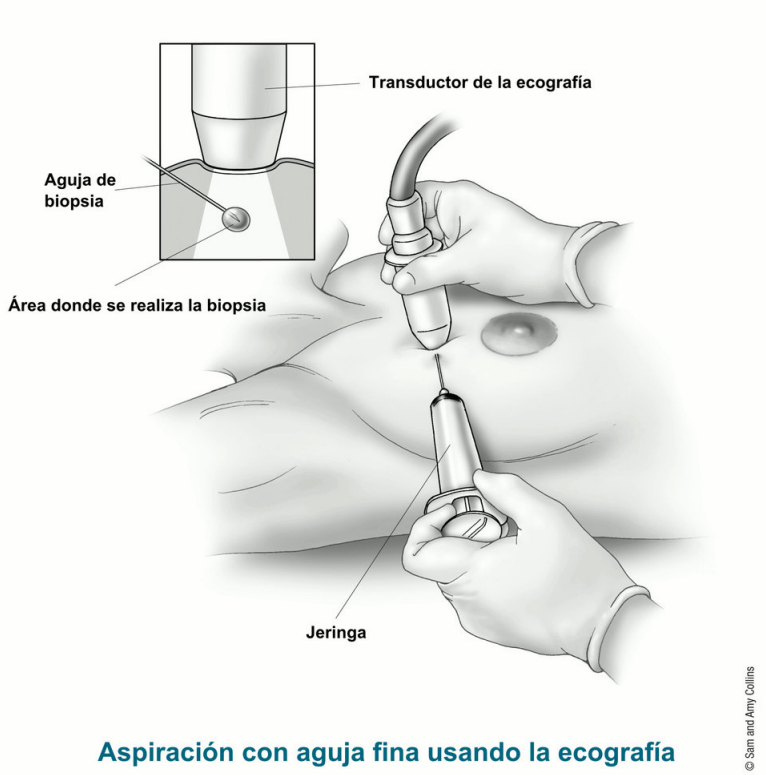
Anatomía de la mama femenina.



4.2.1.2. Método de aguja fina. El método de aguja fina o FNA por sus siglas en inglés es una biopsia por aspiración con aguja fina, se trata de la extracción del tejido o liquido en la zona del cuerpo que causa sospecha para examinar si contiene células cancerosas en la que se utiliza una aguja hueca muy fina adherida a una jeringa (American Cancer Society, 2019).

Figura 7

Método de aguja fina.



(a) Adaptado de Biopsia por aspiración con aguja fina [Fotografía], American Cancer Society, 2019 (https://www.cancer.org).

4.2.2. Inteligencia Artificial. En el diccionario de *Oxford Languages* la inteligencia artificial se puntualiza como un programa computacional diseñado para realizar operaciones que se consideran propias de la inteligencia humana, como el autoaprendizaje.

La Inteligencia Artificial – IA, tiene como precursor a Alan Turing matemático e informático teórico británico, reconocido mundialmente por descifrar códigos nazis en la segunda guerra mundial. En 1956 se formalizó el termino de Inteligencia Artificial como un nuevo area de investigación científica en una disertación en Darmouth que fue organizada por John McCarthy, Marvin Minsky, Claude Shannon Y Nathaniel Rochester. Una idea arraigada en el estudio de la IA es que el pensamiento es una forma de computación no exclusiva de los seres humanos o biológicos, incluso existe la hipótesis de que la inteligencia humana es posible de replicar en máquinas digitales.

Existen dos paradigmas de investigación en IA, uno se refiere a la IA simbólica y el otro a la IA conexionista. la primera modela la mente humana como una computadora procesadora de símbolos, sus algoritmos

hacen parte de un conjunto de métodos basados en la representar¿ción el conocimiento procedimental que tienen los seres humanos de forma explícita, a traves del uso de símbolos y reglas en programas informáticos. Ahora bien, la IA conexionista se basa en modelar la biología del cerebro que está compuesto por redes neuronales biológicas.

4.2.3. *Data driven.* *Data-Driven* (impulsado/basado en datos) se refiere a que el progreso o desarrollo en una actividad está impulsada en datos y no en la intuición o experiencia personal. Puede referirse a distintas disciplinas dependiendo de su aplicación, como lo es en este caso *Data-Driven Science* o simplemente *Data-Science.*

Cuando una empresa emplea un enfoque *"Data-Driven"* significa que toma decisiones estratégicas basadas en el análisis y la interpretación de datos

4.2.3.1. Data-driven science. Aunque el terminoa *Data Science* ha estado presente desde los últimos 30 años, es en los años 70 cuando el término empezó a usarse para referirse a los métodos de procesamiento de datos.

> *Data-science* o ciencia de datos es una disciplina científica centrada en el análisis de grandes fuentes de datos para extraer información, comprender la realidad y descubrir patrones con los que tomar decisiones. La ciencia de datos combina herramientas matemáticas, estadísticas e informáticas para la optimización en la toma de decisiones. Para el 2021 la ciencia de datos se separa del *big data* y se proclama como una disciplina independiente, pues a diferencia del *big data* la ciencia de datos proporciona un increíble potencial de rendimiento y utiliza modelos inteligentes que aprenden de sí mismos como sucede en *Machine Learning,* junto con métodos estadísticos para entrenar los ordenadores (Universidad Complutense de Madrid, 2021).

4.2.3.2. Reducción dimensional. La alta dimensionalidad es un desafío común en el procesamiento de datos de sistemas complejos. Estos sistemas pueden involucrar grandes conjuntos de datos medidos, incluidos audio, imágenes o video. Es muchos sistemas naturales se observa que los datos exhiben un patrón dominante que puede estar caracterizado por un atractor de baja dimensión.

4.2.3.3. Descomposición de valores singulares (SVD). SVD proporciona una descomposición de matriz numéricamente estable que se puede utilizar para una variedad de propósitos y se garantiza su existencia, se usa para obtener aproximaciones de bajo rango a matrices y realizar pseudo inversas de matrices no cuadradas para encontrar solución a un sistema de ecuaciones **Ax=b**. Otro uso importante de SVD es como algoritmo subyacente del PCA. Esta es una técnica *Data-Driven* puesto que los patrones se descubren exclusivamente a partir de los datos, sin echar mano de conocimiento experto o intuición, las aplicaciones de SVD van mucho más allá de la reducción de dimensionalidad de datos de alta dimensión, además de calcular pseudo inversas, proporciona soluciones a ecuaciones matriciales

sub-determinadas o sobre-determinadas, para eliminar el ruido de los conjuntos de datos, la caracterización de la geometría de entrada y salida de un mapa lineal entre espacios vectoriales.

Generalmente, es de interés analizar un conjunto de datos grandes $X \in \mathbb{C}^{nxm}$:

X=

$$\begin{bmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_m \\ | & | & & | \end{bmatrix} \tag{4.1}$$

Las columnas de $X_k \in \mathbb{C}^n$ pueden ser mediciones de simulaciones o experimentos. Los vectores de columna también pueden representar el estado de un sistema físico que está evolucionando en el tiempo. El índice k es una etiqueta que indica el K^{th} conjunto distinto de medidas, **X** consistirá en una serie temporal de datos, y **X**= $X(k \triangle t.)$ A menudo, la *dimensión de estado* **n** es muy grande, del orden de millones o miles de millones de grados de libertad.

Las columnas a menudo se denominan instantáneas y m es el número de instantáneas en **X**. Para muchos sistemas $n \gg m$, resulta en una matriz alta y delgada, en contraposición a una matriz baja y ancha cuando $n \ll m$.

$X \in \mathbb{C}^{nxm}$:

$$X = U\Sigma V^* \tag{4.2}$$

Donde $U \in \mathbb{C}^{nxm}$ y $V \in \mathbb{C}^{nxm}$ son matrices con columnas ortonormales, y $\Sigma \in \mathbb{R}^{nxm}$ es una matriz con entradas reales, no negativas en la diagonal y ceros fuera de la diagonal. Aquí * denota la transpuesta conjugada compleja. Cuando $n \geq m$, la matriz Σ tiene como máximo m elementos distintos de cero en la diagonal, y puede ser escrito como:

$$\Sigma = \begin{bmatrix} \widehat{\Sigma} \\ 0 \end{bmatrix} \tag{4.3}$$

Por lo tanto, es posible representar X exactamente usando la economía SVD:

$$X = U\Sigma V^* = \begin{bmatrix} \widehat{U} & \widehat{U^{\perp}} \end{bmatrix} \begin{bmatrix} \widehat{\Sigma} \\ 0 \end{bmatrix} V^* = \widehat{U}\widehat{\Sigma}V^* \tag{4.4}$$

Las columnas de $\widehat{U^{\perp}}$ abarcan un espacio vectorial que es complementario y ortogonal al que abarca $\widehat{U}$. Las columnas de **U** se denominan *vectores singulares izquierdos* de **X** y las columnas de **V** son *vectores singulares derechos*. Los elementos diagonales de $\widehat{\Sigma} \in \mathbb{C}^{nxm}$ se denominan *valores singulares* y están ordenados de mayor a menor. El rango de **X** es igual al número de valores singulares distintos de cero (Kutz, 2013).

5. Desarrollo del Clasificador

5.1. Muestra

La muestra usada para la finalidad debe ceñirse a las siguientes características:

- Tipos de datos: Multivariable.
- Tarea predeterminada: Clasificación.
- Tipo de atributos: Real.
- Número de instancias: Cantidad mínima aceptable 150.
- Número de atributos: Límite inferior 20.
- Costo de acceso: Ninguno.

Tabla 1

Data sets: evaluados

	Nombre del conjunto
1	Breast Cancer
2	Breast Cancer Wisconsin (Original)
3	Breast Cancer Wisconsin (Prognostic)
4	Breast Cancer Wisconsin (Diagnostic)

Tabla 2

Data sets: características

Conjunto de datos	**1**	**2**	**3**	**4**
Tipos de datos	Multivariable	Multivariable	Multivariable	Multivariable
Tarea predeterminada	Clasificación	Clasificación	Clasificación, regresión	Clasificación
Tipo de atributos	Categórico	Entero	Real	Real
Instancias	286	699	198	569
Atributos	9	10	34	32
Costo	$ 0	$ 0	$ 0	$ 0

Los conjuntos de datos mostrados en la Tabla 6.1, fueron tomados de *MACHINE LEARNING REPOSITORY UCI.*

Los conjuntos de datos 1 y 2 no cumplen con los criterios de Tipo y número de atributo en consecuencia, se desestiman.

Tanto el conjunto 3 y 4 cumplen con los criterios establecidos, sin embargo, el conjunto 4 es el más lejano a las cantidades mínimas de instancias y atributos, es por esto finalmente que se elige el conjunto 3 de datos para la muestra.

Características de la muestra elegida:

- *Data set* multivariable.
- Tarea predeterminada: clasificador.
- Datos de origen real.
- 569 participantes.

- 32 atributos 2 de los cuales son de identificación para los participantes, dejando solo 30 para el análisis.

- La información de los atributos corresponde a Radio, Textura, Perímetro, Área, Simetría, Concavidad, Suavidad, Compacidad, Puntos Cóncavos y Dimensión Fractal de los núcleos celulares mamarios.

5.2. División de la base de datos

Para la adecuada realización de modelos predictivos se deben tener en cuenta los siguientes conjuntos de datos (Vaquerizo, 2010):

- Conjunto de datos de entrenamiento: datos con los cuales se entrenan los modelos.

- Conjunto de datos de validación: datos usados en el entrenamiento para determinar la exactitud del modelo.

- Conjunto de datos de prueba: datos que nos ofrecen la exactitud real del clasificador previamente validado.

La base de datos se divide en una proporción 80/20 de la cual el 80 % es para entrenamiento y validación y el 20 % para prueba.

5.3. Procesamiento de los datos con PCA y curva ROC

Figura 8

Clasificador: código.

```
obs=readmatrix("Entrenamiento y verificación","Range",'C1:AF455'); %datos
grp=readcell("Entrenamiento y verificación","Range",'B1:B455'); %etiquetas

[U,S,V] = svd(obs,'econ');

figure
plot(cumsum(diag(S))./sum(diag(S)),'k-o','LineWidth',1.5)
set(gca,'FontSize',13), axis tight, grid on
set(gcf,'Position',[100 100 600 250])

figure, hold on
for i=1:size(obs,1)
    a = V(:,1)'*obs(i,:)';
    b = V(:,2)'*obs(i,:)';
    c = V(:,3)'*obs(i,:)';

    if(grp{i}=='M')
        plot3(a,b,c,'rx','LineWidth',2);
    else
        plot3(a,b,c,'bo','LineWidth',2);

    end
    table(i,1)=a;
    table(i,2)=b;
    table(i,3)=c;

    results(i,1)=grp{i};
end
view(85,25), grid on, set(gca,'FontSize',13)
```

Figura 9

Curva ROC.

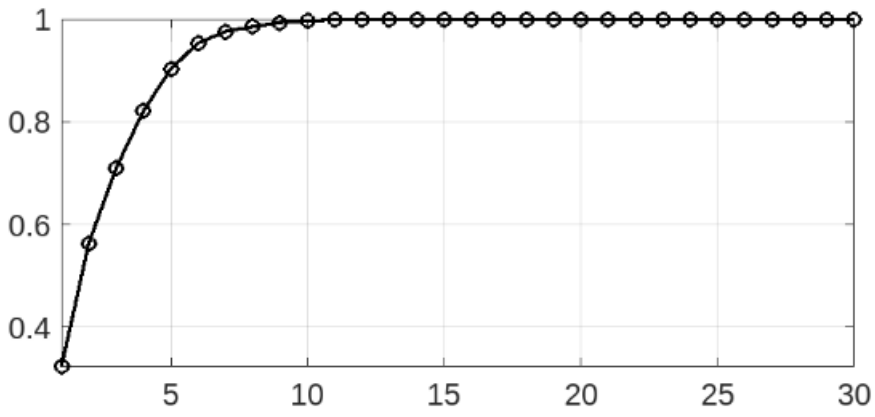
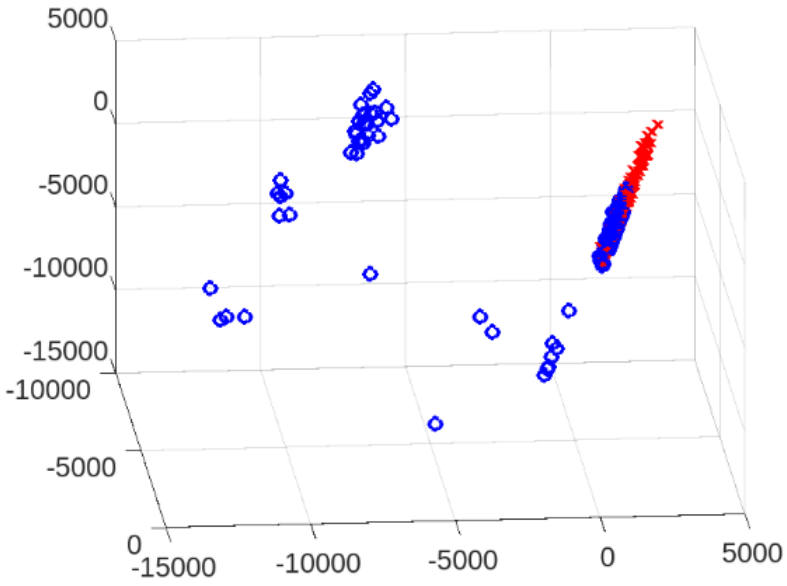


Figura 10

3 dimensiones en PCA.



De acuerdo con la curva ROC (Característica Operativa del Receptor), se puede identificar la fiabilidad en el área bajo la curva (AUC por sus siglas en inglés) puesto que esta proporciona una medida agregada del rendimiento en todos los umbrales de clasificación posibles (Clasificación: Curva ROC y AUC — Machine Learning — Google Developers, 2022). Según Hernández, F. Lemus Pineda los valores referenciados para la interpretación del AUC se muestran a continuación:

- Valores entre 0.5 y 0.6 implican que el modelo ajustado no es el adecuado.
- Valores entre 0.6 y 0.75 implican que el modelo ajustado tiene una tasa de clasificación regular.
- Valores entre 0.75 y 0.9 implican que el modelo ajustado tiene una buena tasa de clasificación.
- Valores entre 0.9 y 0.97 implican que el modelo ajustado tiene una tasa de clasificación muy buena.
- Valores entre 0.97 y 1 implican que el modelo ajustado tiene una excelente tasa de clasificación.

Por tal motivo se escoge para la búsqueda de un clasificador adecuado la menor cantidad de dimensiones PCA que cumplan para la tasa de clasificación buena, muy buena y excelente.

- 4D PCA con AUC=0.82142
- 5D PCA con AUC=0.902983
- 7D PCA con AUC=0.976026

5.4. Implementación de los clasificadores

Se usarán los diferentes clasificadores del apartado de *Machine Learning* en Matlab, los cuales son:

- Árbol de decisión:
  - *Fine tree.*
  - *Medium tree.*
  - *Coarse tree.*
- Análisis discriminante:
  - Discriminante lineal.
  - Discriminante cuadrática.
- Regresión logística.
- Naive Bayes:
  - *Gaussian Naive Bayes.*
  - *Karnel Naive Bayes.*
- Máquina de soporte vectorial (SVM):
  - SVM Lineal.
  - SVM Cuadrática.
  - SVM Cúbica.
  - *Fine Gaussian SMV.*
  - *Medium Gaussian SVM.*
  - *Coarse Gaussian SVM.*
  - *Medium tree.*
  - *Coarse tree.*
- Aproximación Kernel:
  - Kernel SVM.
  - Regresión logística Kernel.

- Ensamble:
  - *Boosted trees.*
  - *Bagged trees.*
  - Subespacio discriminante.
  - Subespacio KNN.
  - Boosted trees.
  - Regresión logística Kernel.
- Redes neuronales:
  - Angosta.
  - Mediana.
  - Grande.
  - Bicapa.
  - Tricapa.

6. Resultados

Los resultados fueron obtenidos para la cantidad de dimensiones en PCA seleccionadas en las etapas de entrenamiento y prueba. En la siguiente tabla se enlistan los clasificadores usados.

Clasificadores	
# de clasificador	Tipo
1	Fine Tree
2	Medium Tree
3	Coarse Tree
4	Linear Discriminant
5	Quadratic Discriminant
6	Logistic Regression
7	Gaussian Naive Bayes
8	Kernel Naive Bayes
9	Linear SVM
10	Quadratic SVM
11	Cubic SVM
12	Fine Gaussian SVM
13	Medium Gaussian SVM
14	Coarse Gaussian SVM
15	Fine KNN
16	Medium KNN
17	Coarse KNN
18	Cosine KNN
19	Cubic KNN
20	Weighted KNN
21	Boosted Trees
22	Bagged Trees
23	Subspace Discriminant
24	Subspace KNN
25	RUSBoosted Trees

Sigue en la página siguiente.

Clasificadores	
# de clasificador	Tipo
26	Narrow Neural Network
27	Medium Neural Network
28	Wide Neural Network
29	Bilayered Neural Network
30	Trilayered Neural Network
31	SVM Kernel
32	Logistic Regression Kernel

6.1. Resultados: etapa de entrenamiento

Tabla 4

Entrenamiento: 4D usando PCA

Clas.	Exactitud(%)	Precisión(%)	Sensibilidad(%)	Especificidad(%)	BER
1	84,61538462	86,44067797	89,47368421	76,47058824	0,1702786378
2	84,61538462	86,44067797	89,47368421	76,47058824	0,1702786378
3	89,01098901	87,3015873	96,49122807	76,47058824	0,1351909185
4	80,21978022	76,71232877	98,24561404	50	0,2587719298
5	45,05494505	**100**	12,28070175	**100**	0,4385964912
6	85,71428571	82,35294118	98,24561404	64,70588235	0,1852425181
7	45,05494505	**100**	12,28070175	**100**	0,4385964912
8	85,71428571	82,35294118	98,24561404	64,70588235	0,1852425181
9	83,51648352	80	98,24561404	58,82352941	0,2146542828
10	87,91208791	84,84848485	98,24561404	70,58823529	0,1558307534
11	78,02197802	74,02597403	**100**	41,17647059	0,2941176471
12	81,31868132	78,57142857	96,49122807	55,88235294	0,2381320949
13	81,31868132	77,77777778	98,24561404	52,94117647	0,2440660475
14	78,02197802	74,66666667	98,24561404	44,11764706	0,2881836945
15	78,02197802	81,3559322	84,21052632	67,64705882	0,2407120743
16	81,31868132	80	94,91525424	58,82352941	0,2313060818
17	79,12087912	76,38888889	96,49122807	50	0,2675438596
18	80,21978022	79,10447761	92,98245614	58,82352941	0,2409700722
19	80,21978022	79,10447761	92,98245614	58,82352941	0,2409700722
20	81,31868132	80,3030303	92,98245614	61,76470588	0,2262641899
21	85,71428571	85,48387097	92,98245614	73,52941176	0,1674406605
22	85,71428571	85,48387097	92,98245614	73,52941176	0,1674406605
23	78,02197802	74,02597403	**100**	41,17647059	0,2941176471
24	82,41758242	80,59701493	94,73684211	61,76470588	0,2174922601
25	87,91208791	85,9375	96,49122807	73,52941176	0,1498968008
26	89,01098901	86,15384615	98,24561404	73,52941176	0,141124871
27	89,01098901	86,15384615	98,24561404	73,52941176	0,141124871
28	87,91208791	87,09677419	94,73684211	76,47058824	0,1439628483
29	87,91208791	87,09677419	94,73684211	76,47058824	0,1439628483
30	**90,10989011**	90	94,73684211	82,35294118	**0,1145510836**
31	76,92307692	75,71428571	92,98245614	50	0,2850877193
32	70,32967033	69,23076923	94,73684211	29,41176471	0,3792569659

Tabla 5

Entrenamiento: 5D usando PCA

Clas.	Exactitud(%)	Precisión(%)	Sensibilidad(%)	Especificidad(%)	BER
1	87,91208791	91,07142857	89,47368421	85,29411765	0,1261609907
2	89,01098901	91,22807018	91,22807018	85,29411765	0,1173890609
3	90,10989011	91,37931034	92,98245614	85,29411765	0,1086171311
4	75,82417582	73,97260274	94,73684211	44,11764706	0,3057275542
5	48,35164835	91,66666667	19,29824561	**97,05882353**	0,4182146543
6	90,10989011	91,37931034	92,98245614	85,29411765	0,1086171311
7	48,35164835	91,66666667	19,29824561	**97,05882353**	0,4182146543
8	87,91208791	88,33333333	92,98245614	79,41176471	0,1380288958
9	80,21978022	78,26086957	94,73684211	55,88235294	0,2469040248
10	89,01098901	88,52459016	94,73684211	79,41176471	0,1292569659
11	85,71428571	87,93103448	89,47368421	79,41176471	0,1555727554
12	82,41758242	81,53846154	92,98245614	64,70588235	0,2115583075
13	78,02197802	76,05633803	94,73684211	50	0,2763157895
14	73,62637363	71,42857143	**96,49122807**	35,29411765	0,3410732714
15	81,31868132	85,71428571	84,21052632	76,47058824	0,1965944272
16	83,51648352	82,8125	92,98245614	67,64705882	0,1968524252
17	70,32967033	69,73684211	92,98245614	32,35294118	0,3733230134
18	82,41758242	82,53968254	91,22807018	67,64705882	0,205624355
19	82,41758242	82,53968254	91,22807018	67,64705882	0,205624355
20	82,41758242	84,74576271	87,71929825	73,52941176	0,1937564499
21	89,01098901	92,72727273	89,47368421	88,23529412	0,1114551084
22	87,91208791	89,65517241	91,22807018	82,35294118	0,1320949432
23	72,52747253	70,51282051	**96,49122807**	32,35294118	0,3557791538
24	81,31868132	85,71428571	84,21052632	76,47058824	0,1965944272
25	87,91208791	91,07142857	89,47368421	85,29411765	0,1261609907
26	86,81318681	89,47368421	89,47368421	82,35294118	0,1408668731
27	83,51648352	87,5	85,96491228	79,41176471	0,1731166151
28	86,81318681	89,47368421	89,47368421	82,35294118	0,1408668731
29	86,81318681	89,47368421	89,47368421	82,35294118	0,1408668731
30	**92,30769231**	**93,10344828**	94,73684211	88,23529412	**0,08513931889**
31	76,92307692	80	84,21052632	64,70588235	0,2554179567
32	75,82417582	79,66101695	82,45614035	64,70588235	0,2641898865

Tabla 6

Entrenamiento: 7D usando PCA

Clas.	**Exactitud(%)**	**Precisión(%)**	**Sensibilidad(%)**	**Especificidad(%)**	**BER**
1	87,91208791	89,65517241	91,22807018	82,35294118	0,1320949432
2	87,91208791	89,65517241	91,22807018	82,35294118	0,1320949432
3	92,30769231	90,32258065	98,24561404	82,35294118	0,09700722394
4	86,81318681	83,58208955	98,24561404	67,64705882	0,1705366357
5	48,35164835	**100**	17,54385965	**100**	0,4122807018
6	**95,6043956**	94,91525424	98,24561404	91,17647059	**0,05288957688**
7	49,45054945	**100**	19,29824561	**100**	0,4035087719
8	89,01098901	88,52459016	94,73684211	79,41176471	0,1292569659
9	92,30769231	90,32258065	98,24561404	82,35294118	0,09700722394
10	91,20879121	88,88888889	98,24561404	79,41176471	0,1117131063
11	86,81318681	86,8852459	92,98245614	76,47058824	0,1527347781
12	81,31868132	80,3030303	92,98245614	61,76470588	0,2262641899
13	87,91208791	84,84848485	98,24561404	70,58823529	0,1558307534
14	85,71428571	81,42857143	**100**	61,76470588	0,1911764706
15	86,81318681	86,8852459	92,98245614	76,47058824	0,1527347781
16	89,01098901	87,3015873	96,49122807	76,47058824	0,1351909185
17	80,21978022	76	**100**	47,05882353	0,2647058824
18	89,01098901	86,15384615	69,13580247	10	0,6043209877
19	89,01098901	87,3015873	96,49122807	76,47058824	0,1351909185
20	87,91208791	85,9375	96,49122807	73,52941176	0,1498968008
21	85,71428571	89,28571429	87,71929825	82,35294118	0,1496388029
22	89,01098901	89,83050847	92,98245614	82,35294118	0,1233230134
23	83,51648352	79,16666667	**100**	55,88235294	0,2205882353
24	87,91208791	88,33333333	92,98245614	79,41176471	0,1380288958
25	87,91208791	89,65517241	91,22807018	82,35294118	0,1320949432
26	94,50549451	93,33333333	98,24561404	88,23529412	0,06759545924
27	89,01098901	91,22807018	91,22807018	85,29411765	0,1173890609
28	91,20879121	91,52542373	94,73684211	85,29411765	0,09984520124
29	94,50549451	94,82758621	96,49122807	91,17647059	0,06166150671
30	87,91208791	88,33333333	92,98245614	79,41176471	0,1380288958
31	84,61538462	82,08955224	96,49122807	64,70588235	0,1940144479
32	83,51648352	80,88235294	96,49122807	61,76470588	0,2087203302

Tabla 7

Entrenamiento: sin PCA

Clas.	**Exactitud(%)**	**Precisión(%)**	**Sensibilidad(%)**	**Especificidad(%)**	**BER**
1	94,50549451	91,93548387	**100**	85,29411765	0,07352941176
2	94,50549451	91,93548387	**100**	85,29411765	0,07352941176
3	95,6043956	96,49122807	96,49122807	94,11764706	0,04695562436
4	93,40659341	93,22033898	96,49122807	88,23529412	0,07636738906
5	93,40659341	**98,11320755**	91,22807018	**97,05882353**	0,05856553148
6	91,20879121	92,98245614	92,98245614	88,23529412	0,09391124871
7	91,20879121	91,52542373	94,73684211	85,29411765	0,09984520124
8	92,30769231	94,64285714	92,98245614	91,17647059	0,07920536636
9	96,7032967	95	**100**	91,17647059	0,04411764706
10	95,6043956	94,91525424	98,24561404	91,17647059	0,05288957688
11	95,6043956	94,91525424	98,24561404	91,17647059	0,05288957688
12	79,12087912	75	**100**	44,11764706	0,2794117647
13	94,50549451	94,82758621	96,49122807	91,17647059	0,06166150671
14	93,40659341	93,22033898	96,49122807	88,23529412	0,07636738906
15	93,40659341	93,22033898	96,49122807	88,23529412	0,07636738906
16	94,50549451	91,93548387	**100**	85,29411765	0,07352941176
17	87,91208791	85,9375	96,49122807	73,52941176	0,1498968008
18	94,50549451	93,33333333	98,24561404	88,23529412	0,06759545924
19	91,20879121	90,16393443	96,49122807	82,35294118	0,1057791538
20	94,50549451	91,93548387	**100**	85,29411765	0,07352941176
21	62,63736264	62,63736264	**100**	0	0,5
22	**97,8021978**	96,61016949	**100**	94,11764706	**0,02941176471**
23	94,50549451	91,93548387	**100**	85,29411765	0,07352941176
24	91,20879121	88,88888889	98,24561404	79,41176471	0,1117131063
25	95,6043956	96,49122807	96,49122807	94,11764706	0,04695562436
26	93,40659341	94,73684211	94,73684211	91,17647059	0,07043343653
27	96,7032967	96,55172414	98,24561404	94,11764706	0,03818369453
28	95,6043956	94,91525424	98,24561404	91,17647059	0,05288957688
29	96,7032967	96,55172414	98,24561404	94,11764706	0,03818369453
30	96,7032967	96,55172414	98,24561404	94,11764706	0,03818369453
31	79,12087912	85,18518519	80,70175439	76,47058824	0,2141382869
32	83,51648352	86,20689655	87,71929825	76,47058824	0,1790505676

Tabla 8

Tiempo de entrenamiento en los diferentes clasificadores.

Clasificador	Tiempo de entrenamiento (segundos)			
	4-D	5-D	7-D	No PCA
Fine Tree	16,086046	4,342412	17,087087	12,422414
Medium Tree	21,211357	17,626232	24,928996	25,885016
Coarse Tree	9,079733	32,989466	13,410278	12,949551
Linear Discriminant	8,634354	6,998354	10,806013	10,650496
Quadratic Discriminant	5,657363	4,546236	10,523047	8,288804
Logistic Regression	11,692838	2,971073	7,446222	6,824349
Gaussian Naive Bayes	28,99427	2,61157	4,850534	5,248492
Kernel Naive Bayes	12,671609	9,090672	15,304169	4,903892
Linear SVM	10,120437	8,955482	31,615227	14,66289
Quadratic SVM	8,882679	13,488264	11,815701	32,7651
Cubic SVM	14,655321	16,727393	10,107601	14,080439
Fine Gaussian SVM	4,947428	12,120438	9,365503	13,307337
Medium Gaussian SVM	3,337899	8,529961	8,610226	12,626469
Coarse Gaussian SVM	2,540109	6,252116	6,549399	10,259206
Fine KNN	14,410083	4,000706	4,179223	8,895068
Medium KNN	37,194254	2,860052	3,095647	6,740082
Coarse KNN	18,509428	9,104084	13,829238	4,319475
Cosine KNN	16,327574	8,960784	19,015456	3,236105
Cubic KNN	13,87604	22,475321	8,345609	15,234348
Weighted KNN	10,655423	10,792182	6,071856	30,520815
Boosted Trees	8,316178	8,50385	6,141685	10,013715
Bagged Trees	6,950763	6,999662	6,051551	8,505668
Subspace Discriminant	4,542345	4,741639	5,797932	6,259505
Subspace KNN	3,639253	3,61447	9,957502	5,002595
RUSBoosted Trees	20,438531	12,885853	4,688838	2,399181
Narrow Neural Network	36,886359	12,720238	3,338968	1,207819
Medium Neural Network	10,82782	28,393487	6,726477	12,800957
Wide Neural Network	8,557529	10,649095	6,619729	16,122108
Bilayered Neural Network	6,171637	7,903767	15,434695	6,443714
Trilayered Neural Network	3,873389	5,378045	4,436774	8,736941
SVM Kernel	3,490083	3,705019	3,328935	2,666719
Logistic Regression Kernel	13,202711	12,976711	6,714337	5,333903

Los tiempos de entrenamiento son cortos y serán desestimados en la validación de los clasificadores debido a que esta característica es irrelevante para la selección del clasificador.

6.2. Resultados: etapa de prueba

Tabla 9

Prueba: 4 dimensiones en PCA

Clas.	Exactitud(%)	Precisión(%)	Sensibilidad(%)	Especificidad(%)	BER
1	85,96491228	88,88888889	88,88888889	80,95238095	0,1507936508
2	87,71929825	89,18918919	91,66666667	80,95238095	0,1369047619
3	91,22807018	93,05555556	93,05555556	88,0952381	0,09424603175
4	83,33333333	81,92771084	94,44444444	64,28571429	0,2063492063
5	49,12280702	93,75	20,83333333	**97,61904762**	0,4077380952
6	86,84210526	86,07594937	94,44444444	73,80952381	0,1587301587
7	48,24561404	88,23529412	20,83333333	95,23809524	0,4196428571
8	91,22807018	93,05555556	93,05555556	88,0952381	0,09424603175
9	87,71929825	86,25	95,83333333	73,80952381	0,1517857143
10	91,22807018	89,74358974	**97,22222222**	80,95238095	0,1091269841
11	68,42105263	73,07692308	79,16666667	50	0,3541666667
12	86,84210526	89,04109589	90,27777778	80,95238095	0,1438492063
13	85,96491228	85	94,44444444	71,42857143	0,1706349206
14	79,8245614	77,52808989	95,83333333	52,38095238	0,2589285714
15	85,0877193	89,85507246	86,11111111	83,33333333	0,1527777778
16	86,84210526	88	91,66666667	78,57142857	0,1488095238
17	82,45614035	86,11111111	86,11111111	76,19047619	0,1884920635
18	88,59649123	91,54929577	90,27777778	85,71428571	0,1200396825
19	87,71929825	89,18918919	91,66666667	80,95238095	0,1369047619
20	85,0877193	88,73239437	87,5	80,95238095	0,1577380952
21	85,0877193	91,04477612	84,72222222	85,71428571	0,1478174603
22	84,21052632	92,1875	81,94444444	88,0952381	0,1498015873
23	78,07017544	75,2688172	**97,22222222**	45,23809524	0,2876984127
24	85,0877193	88,73239437	87,5	80,95238095	0,1577380952
25	85,0877193	89,85507246	86,11111111	83,33333333	0,1527777778
26	92,10526316	95,65217391	91,66666667	92,85714286	0,07738095238
27	88,59649123	96,82539683	84,72222222	95,23809524	0,1001984127
28	88,59649123	95,38461538	86,11111111	92,85714286	0,1051587302
29	90,35087719	94,20289855	90,27777778	90,47619048	0,09623015873
30	**92,98245614**	**97,05882353**	91,66666667	95,23809524	**0,06547619048**
31	78,94736842	84,28571429	81,94444444	73,80952381	0,2212301587
32	87,71929825	87,17948718	94,44444444	76,19047619	0,1468253968

Tabla 10

Prueba: 5 dimensiones en PCA

Clas.	Exactitud(%)	Precisión(%)	Sensibilidad(%)	Especificidad(%)	BER
1	87,71929825	92,64705882	87,5	88,0952381	0,1220238095
2	89,47368421	92,85714286	90,27777778	88,0952381	0,1081349206
3	**91,22807018**	93,05555556	93,05555556	88,0952381	**0,09424603175**
4	83,33333333	81,92771084	94,44444444	64,28571429	0,2063492063
5	49,12280702	93,75	20,83333333	**97,61904762**	0,4077380952
6	90,35087719	94,20289855	90,27777778	90,47619048	0,09623015873
7	49,12280702	88,88888889	22,22222222	95,23809524	0,4126984127
8	89,47368421	90,54054054	93,05555556	83,33333333	0,1180555556
9	87,71929825	86,25	95,83333333	73,80952381	0,1517857143
10	**91,22807018**	90,78947368	95,83333333	83,33333333	0,1041666667
11	90,35087719	90,66666667	94,44444444	83,33333333	0,1111111111
12	87,71929825	86,25	95,83333333	73,80952381	0,1517857143
13	86,84210526	85,18518519	95,83333333	71,42857143	0,1636904762
14	84,21052632	80,68181818	**98,61111111**	59,52380952	0,2093253968
15	82,45614035	90,625	80,55555556	85,71428571	0,1686507937
16	86,84210526	86,07594937	94,44444444	73,80952381	0,1587301587
17	78,07017544	78,31325301	90,27777778	57,14285714	0,2628968254
18	85,96491228	84,14634146	95,83333333	69,04761905	0,1755952381
19	85,96491228	85	94,44444444	71,42857143	0,1706349206
20	85,96491228	88,88888889	88,88888889	80,95238095	0,1507936508
21	84,21052632	89,70588235	84,72222222	83,33333333	0,1597222222
22	86,84210526	92,53731343	86,11111111	88,0952381	0,128968254
23	80,70175439	77,17391304	**98,61111111**	50	0,2569444444
24	85,0877193	89,85507246	86,11111111	83,33333333	0,1527777778
25	86,84210526	93,84615385	84,72222222	90,47619048	0,1240079365
26	89,47368421	92,85714286	90,27777778	88,0952381	0,1081349206
27	89,47368421	**95,45454545**	87,5	92,85714286	0,09821428571
28	88,59649123	94,02985075	87,5	90,47619048	0,1101190476
29	89,47368421	94,11764706	88,88888889	90,47619048	0,1031746032
30	90,35087719	94,20289855	90,27777778	90,47619048	0,09623015873
31	79,8245614	81,01265823	88,88888889	64,28571429	0,2341269841
32	71,05263158	73,4939759	84,72222222	47,61904762	0,3382936508

Tabla 11

Prueba: 7 dimensiones en PCA

Clas.	Exactitud(%)	Precisión(%)	Sensibilidad(%)	Especificidad(%)	BER
1	87,71929825	93,93939394	86,11111111	90,47619048	0,1170634921
2	90,35087719	92,95774648	91,66666667	88,0952381	0,1011904762
3	91,22807018	94,28571429	91,66666667	90,47619048	0,08928571429
4	89,47368421	86,58536585	**98,61111111**	73,80952381	0,1378968254
5	50	**100**	20,83333333	**100**	0,3958333333
6	**93,85964912**	95,77464789	94,44444444	92,85714286	**0,06349206349**
7	50,87719298	**100**	22,22222222	**100**	0,3888888889
8	92,10526316	90,90909091	97,22222222	83,33333333	0,09722222222
9	92,98245614	92,10526316	97,22222222	85,71428571	0,08531746032
10	92,10526316	90,90909091	97,22222222	83,33333333	0,09722222222
11	92,98245614	92,10526316	97,22222222	85,71428571	0,08531746032
12	92,10526316	95,71428571	93,05555556	92,68292683	0,07130758808
13	92,98245614	91,02564103	**98,61111111**	83,33333333	0,09027777778
14	88,59649123	85,54216867	**98,61111111**	71,42857143	0,1498015873
15	87,71929825	98,33333333	81,94444444	97,61904762	0,1021825397
16	90,35087719	89,61038961	95,83333333	80,95238095	0,1160714286
17	85,96491228	83,33333333	97,22222222	66,66666667	0,1805555556
18	90,35087719	89,61038961	95,83333333	80,95238095	0,1160714286
19	90,35087719	89,61038961	95,83333333	80,95238095	0,1160714286
20	92,10526316	95,65217391	91,66666667	92,85714286	0,07738095238
21	88,59649123	94,02985075	87,5	90,47619048	0,1101190476
22	89,47368421	94,02985075	87,5	90,47619048	0,1101190476
23	87,71929825	84,52380952	**98,61111111**	69,04761905	0,1617063492
24	90,35087719	92,95774648	91,66666667	88,0952381	0,1011904762
25	87,71929825	95,3125	84,72222222	92,85714286	0,1121031746
26	91,22807018	95,58823529	90,27777778	92,85714286	0,08432539683
27	90,35087719	96,92307692	87,5	95,23809524	0,08630952381
28	92,10526316	97,01492537	90,27777778	95,23809524	0,07242063492
29	85,96491228	95,16129032	81,94444444	92,85714286	0,1259920635
30	91,22807018	94,28571429	91,66666667	90,47619048	0,08928571429
31	89,47368421	88,46153846	95,83333333	78,57142857	0,1279761905
32	87,71929825	89,18918919	91,66666667	80,95238095	0,1369047619

Tabla 12

Prueba: Sin uso PCA

Clas.	Exactitud(%)	Precisión(%)	Sensibilidad(%)	Especificidad(%)	BER
1	87,71929825	95,3125	84,72222222	92,85714286	0,1121031746
2	87,71929825	95,3125	84,72222222	92,85714286	0,1121031746
3	91,22807018	93,05555556	93,05555556	88,0952381	0,09424603175
4	**98,24561404**	98,61111111	98,61111111	97,61904762	**0,01884920635**
5	72,80701754	**100**	56,94444444	**100**	0,2152777778
6	**98,24561404**	98,61111111	98,61111111	97,61904762	**0,01884920635**
7	94,73684211	97,14285714	94,44444444	95,23809524	0,05158730159
8	84,21052632	93,5483871	80,55555556	90,47619048	0,1448412698
9	**98,24561404**	98,61111111	98,61111111	97,61904762	**0,01884920635**
10	97,36842105	98,5915493	97,22222222	97,61904762	0,02579365079
11	97,36842105	98,5915493	97,22222222	97,61904762	0,02579365079
12	79,8245614	76,34408602	98,61111111	47,61904762	0,2688492063
13	97,36842105	98,5915493	97,22222222	97,61904762	0,02579365079
14	96,49122807	95,94594595	98,61111111	92,85714286	0,04265873016
15	92,98245614	95,71428571	93,05555556	92,85714286	0,07043650794
16	94,73684211	94,59459459	97,22222222	42,85714286	0,2996031746
17	94,73684211	92,30769231	**100**	85,71428571	0,07142857143
18	93,85964912	95,77464789	94,44444444	92,85714286	0,06349206349
19	95,61403509	95,89041096	97,22222222	92,85714286	0,0496031746
20	94,73684211	95,83333333	95,83333333	92,85714286	0,05654761905
21	63,15789474	63,71681416	**100**	0	0,5
22	97,36842105	97,26027397	98,61111111	95,23809524	0,03075396825
23	96,49122807	95,94594595	98,61111111	92,85714286	0,04265873016
24	92,98245614	94,44444444	94,44444444	90,47619048	0,0753968254
25	93,85964912	**100**	90,27777778	**100**	0,04861111111
26	96,49122807	98,57142857	95,83333333	97,61904762	0,03273809524
27	96,49122807	97,22222222	97,22222222	95,23809524	0,0376984127
28	96,49122807	98,57142857	95,83333333	97,61904762	0,03273809524
29	95,61403509	98,55072464	94,44444444	97,61904762	0,03968253968
30	94,73684211	98,52941176	93,05555556	97,61904762	0,04662698413
31	90,35087719	89,61038961	95,83333333	80,95238095	0,1160714286
32	89,47368421	89,47368421	94,44444444	80,95238095	0,123015873

6.3. Resumen

En la siguiente tabla se aprecia el mejor clasificador de acuerdo con la cantidad de dimensiones PCA.

Tabla 13

Resumen de los resultados

Dimensiones PCA	Clasificador	Tipo	Exactitud (%)	BER
4	Trilayered Neural Network	Validación	90,10989011	0,1145510836
	Trilayered Neural Network	Prueba	92,98245614	0,06547619048
5	Trilayered Neural Network	Validación	92,30769231	0,08513931889
	Coarse Tree	Prueba	91,22807018	0,09424603175
7	Logistic Regression	Validación	95,6043956	0,05288957688
	Logistic Regression	Prueba	93,85964912	0,06349206349
Sin PCA	Bagged Trees	Validación	97,8021978	0,02941176471
	Linear Discriminant	Prueba	**98,24561404**	**0,01884920635**
	Logistic Regression	Prueba	**98,24561404**	**0,01884920635**
	Linear SVM	Prueba	**98,24561404**	**0,01884920635**

Figura 11

Matriz de confusión de los clasificadores óptimos.

7. Conclusiones y recomendaciones

- Todos los clasificadores que se encuentran en el resumen de resultados a excepción del de 5 dimensiones PCA, superan al clasificador de Syantra (compañía privada de atención médica de precisión que está cambiando la forma en que se detecta y trata el cáncer) que tiene una exactitud de 92.2 % (Syantra, 2022).
- El mejor clasificador de cada grupo tiene un alto índice de exactitud y un bajo índice de BER.
- Los mejores modelos clasificadores son de la categoría sin PCA: Linear Discriminant, Logistic Regression y Linear SVM, todos los 3 con una exactitud de 98,2456140 y un BER de 0,01884920635
- Se recomienda realizar una base de datos de mujeres santandereanas para implementar el clasificador puesto que se considera exitoso el proyecto.

Referencias Bibliográficas

Abeliuk, A., y Gutiérrez, C. (2021). *Historia y evolución de la inteligencia artificial* (Vol. 21).

Ambrocio Ríos, J. A. (2018). *Ecología de levaduras asociadas a la taberna, bebida extraída de la palma de coyol (acrocomia aculeata (jacq.) lodd. ex mart).*

Arenas, D. (2019). *Análisis preliminar de la termografía como herramienta para estimar las propiedades térmicas de estructuras laminares a partir de un modelo unidimensional de transferencia de calor enfocado a la detección de cáncer de mama.*

Avanade. (2020). *Las buenas transformaciones digitales son future ready.* Descargado de `https://www.avanade.com/-/media/asset/research/future-ready-now-research-spanish.pdf`

Biopsia por aspiración con aguja fina. (2019). American Cancer Society. Descargado de `https://amp.cancer.org/es/cancer/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/biopsia-del-seno/biopsia-del-seno-por-aspiracion-con-aguja-fina.html` (Acceso: 18-07-2022)

Brunton, S., y Kutz, J. (2019). *Data-driven science and engineering.* doi: 10.1017/9781108380690

Brynjolfsson, H. L., E., y Kim, H. (2011). *Strength in numbers: How does data-driven decisionmaking affect firm performance?* Descargado de `https://deliverypdf.ssrn.com`

Cancer today. (2011). Descargado de `https://gco.iarc.fr/`

Clasificación: Curva roc y auc — machine learning. (2022). Google Developers. Descargado de `https://developers.google.com` (Acceso: 08-10-2022)

Colina, C. L., y López-Roldán, P. (1991). El análisis de componentes principales: Aplicación al análisis de datos secundarios. *Revista de psicología*, 31-63.

Cuadrado, A. M. (2019). *Utilización del machine learning en la industria 4.0.*

Cáncer. (2022). Descargado de `https://www.who.int/es/news-room/fact-sheets/`

detail/cancer

Cáncer de mama. (2022). Organización Panamericana de la Salud. Descargado de https://www.paho.org/es/temas/cancer-mama

Detectar el cáncer de mama a tiempo salva vidas. (2022). Alcaldía de Bucaramanga. Descargado de https://www.bucaramanga.gov.co/noticias/detectar-el-cancer-de-mama-a-tiempo-salva-vidas/ (Acceso: 12-07-2022)

Diccionario de cáncer del nci. instituto nacional del cáncer. (2022). Descargado de https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/cancer-de-mama (Acceso: 18-07-2022)

Díaz, N. (2021). *Análisis comparativo de clasificadores para la detección de subtipos de cáncer.*

Forbes. (2020). *Google cloud brandvoice: How the world became data-driven, and what's next.* Descargado de https://www.forbes.com/sites/googlecloud/2020/05/20/how-the-world-became-data-driven-and-whats-next/ (Acceso: 18-07-2022)

Gancino, K. (2021). *Enfoque de la teoría de juegos en detección de cáncer de mama* (Inf. Téc.). Universidad de las Fuerzas Armadas. Descargado de http://repositorio.espe.edu.ec/bitstream/21000/23825/1/T-ESPE-044311.pdf

Kutz, N. (2013). *Data-driven modelling and scientific computation* (First ed.). Oxford.

Osma, S. (2012). *Supervivencia a 5 años de las mujeres con cáncer de mama de bucaramanga y su Área metropolitana según el estadio clínico* (Tesis de Master no publicada). Universidad Industrial de Santander.

Peña, J. (2022). *Despliegue de un modelo de clasificación de tumores de cáncer de mama, reto kaggle* (Tesis de Master no publicada). Universidad de Antioquia.

Pirchio, R. (2022). *Clasificación de cáncer de mama con técnicas de análisis de la componente principal-kernel pca, algoritmos de máquina de vectores de soporte y regresión logística.* Medisur.

Robles, B., y Millán, J. (2022). *Modelo en machine learning para el diagnóstico del cáncer*

de mama (Tesis de Master no publicada). Universidad Distrital Francisco José De Caldas.

Rodríguez, F. (2019). *Pronóstico de cáncer de mama benigno y maligno: comparación de nueve métodos de clasificación usando r.*

Rojas, E. M. (2020). *Machine learning: Análisis de lenguajes de programación y herramientas para desarrollo* (Vol. 28).

Sampieri, R. H., Collado, C. F., y Lucio, M. B. (2014). *Metodología de la investigación* (6th ed.). McGraw - Hill.

Syantra. (2022). Descargado de `https://www.syantra.com/for-patients` (Acceso: 08-10-2022)

Universidad industrial de santander. (2022). Descargado de `https://uis.edu.co/ffm-pre-mecanica-es/` (Acceso: 12-07-2022)

Vaquerizo. (2010). *Entrenamiento, validación y test - análisis y decisión.* Descargado de `https://analisisydecision.es/entrenamiento-validacion-y-test/` (Acceso: 12-07-2022)

¿qué es data science? (2021). Universidad Complutense de Madrid. Descargado de `https://www.masterdatascienceucm.com/que-es-data-science/` (Acceso: 18-07-2022)

¿qué es el cáncer? (2021). Instituto Nacional del Cáncer. Descargado de `https://www.cancer.gov/espanol/cancer/naturaleza/que-es` (Acceso: 18-07-2022)

¿qué es la ingeniería biomédica? (2021). Ingeniería Biomédica. Descargado de `https://www.ingenieriabiomedica.org/queesingenieriabiomedica` (Acceso: 12-07-2022)