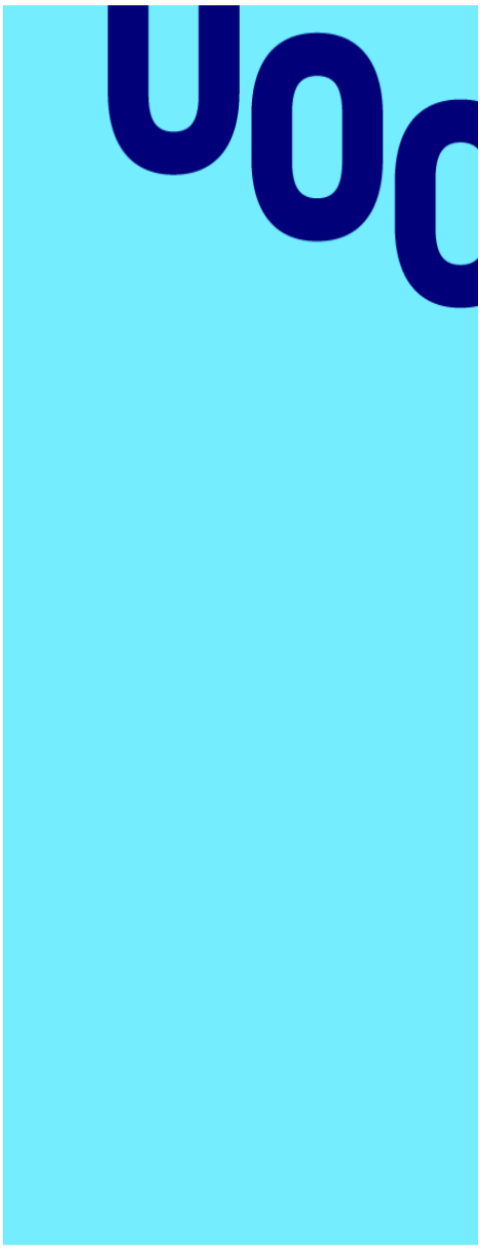
Aplicación de técnicas de Clustering para la segmentación de masas mamográficas



Universitat Oberta de Catalunya

Demetrio Muñoz Álvarez

Bioinformática Estadística y Aprendizaje Automático

Máster universitario Online de Bioinformática y Bioestadística (interuniversitario: UOC, UB)

Nombre del director/a de TF:

Romina Astrid Rebrij

18 de junio de 2024

Ficha Del Trabajo Final

Título del trabajo:	"Aplicación de técnicas de Clustering para la segmentación de masas mamográficas"
Nombre del autor/a:	Demetrio Muñoz Álvarez
Nombre del director/a de TF:	Romina Astrid Rebrij
Nombre del/de la PRA:	Romina Astrid Rebrij
Fecha de entrega:	18 de junio de 2024
Titulación o programa:	Máster universitario Online de Bioinformática y Bioestadística (interuniversitario: UOC, UB)
Área del trabajo final:	Bioinformática Estadística y Aprendizaje Automático
Idioma del trabajo:	Castellano
Palabras clave:	Cáncer de mama, Aprendizaje automático no supervisado, Agrupamiento (Clustering), Masas mamarias, Bi-rads.

Resumen del trabajo

Este trabajo de fin de máster explora los algoritmos de agrupación en el ámbito oncológico, concretamente relacionados con el cáncer de mama. Para ello, se utilizaron datos basados en imágenes mamográficas utilizando la clasificación BI-RADS para agrupar significativamente las lesiones tumorales y establecer el riesgo de malignidad con algoritmos de agrupación no supervisados (K-means, K-modes, agrupación jerárquica, entre otros) de manera automática, sin la intervención preliminar de un especialista. Tras obtener un modelo satisfactorio, se desarrolló una web interactiva de uso público que condensa todo el trabajo realizado.

Abstract

This master's project explores clustering algorithms in the field of oncology, specifically related to breast cancer. For this purpose, mammographic image-based data using the BI-RADS classification were used to meaningfully cluster tumor lesions and establish the risk of malignancy with unsupervised learning algorithms (K-means, K-modes, hierarchical clustering, among others) automatically, without preliminary intervention by a specialist. After obtaining an optimal model, an interactive website was developed for public use that condenses all the work done.

Índice general

Índice de figuras

Capítulo 1

1. Introducción

1.1. Contexto y Justificación del Trabajo

El cáncer de mama es el cáncer más diagnosticado en el mundo. En el año 2022 presentó una incidencia estimada del 11,7% entre todos los tipos de cáncer y fue el cáncer más común en la población femenina, con una tasa de detección del 24,5% y una mortalidad del 15,5% (Chhikara & Parang, 2023)[1]. Un diagnóstico y tratamiento tempranos pueden reducir de manera drástica la mortalidad de los pacientes de alto riesgo de cáncer de mama (Wang, 2017)[2]. De los factores asociados a este tipo de cáncer, la densidad mamográfica (cantidad relativa de tejido radio denso) es uno de los más importantes después del sexo, la edad, las mutaciones genéticas y los antecedentes familiares (Boyd et al., 1998)[3]. Según la literatura, las mujeres con una densidad mamográfica alta tienen un riesgo de cáncer de mama mayor que las mujeres con una densidad mamográfica baja (Boyd et al., 2010)[4].

Una de las formas de interpretar las mamografías es la puntuación estandarizada Bi-Rads (por sus siglas en inglés, Breast Imaging Reporting and Data Systemn) (Tabla 1), que categoriza de forma simple y entendible, según el especialista, el resultado de las imágenes de las mamografías, ecografías y resonancias magnéticas en una escala numérica dependiendo la gravedad de los resultados (American College of Radiology, 2003)[5].

TABLA 1. CLASIFICACIÓN BI-RADS. INFORMACIÓN EXTRAÍDA DE AIBAR ET AL., 2011[6].

Valor Bi-Rads	Descripción
0	No concluyente por lectura incompleta
1	Mama normal
2	Benigna (probabilidad de cáncer similar a la población general)
3	Hallazgos probablemente benignos. (< 2% de riesgo de malignidad)

4	Probablemente maligna (valor predictivo positivo para cáncer entre 29-34% hasta 70%)
5	Altamente sugerente de malignidad (valor predictivo positivo para cáncer superior al 70%)
6	Malignidad confirmada histológicamente

La implementación de nuevas tecnologías basadas en el aprendizaje automático es un objetivo principal en el ámbito de la salud, ya que mejoran los diagnósticos tempranos de las enfermedades, desplazando técnicas invasivas como la biopsia, reduciendo los tiempos de los tratamientos, disminuyendo los costes monetarios y aumentando la productividad en el ámbito medico (Topol, 2019)[7]. Sin embargo, los enfoques basados en el aprendizaje automático suelen tener una precisión baja y unos tiempos de cálculos elevados (Haq et al., 2021)[8]. Debido a esto, se están realizando esfuerzos para mejorar estas tecnologías con el desarrollo de algoritmos más avanzados, optimización de la integración de técnicas de computación en la nube y procesamientos paralelos.

En este contexto exploramos el agrupamiento no supervisado ("clustering") como una alternativa frente al aprendizaje supervisado para afrontar los desafíos del diagnóstico preliminar del cáncer de mama. A diferencia del aprendizaje supervisado, el cual requiere un volumen grande de datos etiquetados para el entrenamiento de los modelos (Roy, Meena, & Lim, 2022)[9], el aprendizaje no supervisado puede procesar, analizar y detectar patrones en los datos sin la necesidad de etiquetas (Calamuneri et al., 2017)[10]. Esto es interesante en el ámbito oncológico, ya que las características de los tumores provenientes de los datos de imágenes mamográficas pueden ser utilizados para identificar grupos de pacientes, determinar características relevantes y mejorar el análisis de la heterogeneidad de los tumores (O'Connor et al., 2015)[11].

Aunque la visión del aprendizaje no supervisado pueda ser prometedora presenta desafíos a la hora de su implementación en el campo de la medicina. Estos desafíos conllevan la construcción de modelos sólidos capaces de manejar una alta variabilidad de datos médicos y su implementación en el flujo de trabajos en los

centros de salud, así como tener en cuenta las implicaciones éticas que conllevan las nuevas tecnologías relacionadas con la inteligencia artificial. Además, se tiene que garantizar la interpretabilidad de estos modelos debido a que, en el ámbito médico, la comprensión y la confianza de los profesionales de la salud es fundamental para la implementación de nuevas herramientas. Estas nuevas tecnologías deben ser transparentes y sus resultados fáciles de interpretar por los especialistas, que son los últimos en tomar las decisiones basadas en estos resultados.

Hemos implementado el uso de modelos de agrupamiento no supervisado como una alternativa a la clasificación Bi-Rads, que, aunque ampliamente utilizada y útil para estandarizar la interpretación de imágenes mamográficas, tiene sus limitaciones debido al criterio del especialista y su sistema categórico. Estos modelos pueden mejorar o reemplazar en algunos casos esta clasificación al agrupar características tumorales sin necesidad de una clasificación previa. Al no restringir el riesgo y gravedad del cáncer a categorías, estos modelos pueden llegar a identificar nuevos grupos de riesgo precisos, recortar tiempos de procesamiento y proporcionar una evaluación más detallada, lo que conlleva un diagnóstico temprano y a tratamientos personalizados pudiendo reducir el uso de técnicas de diagnóstico invasivas.

Por lo tanto, se pretende conseguir la automatización de la clasificación de pacientes según el riesgo que presenten sus características tumorales obtenidas de imágenes mamográficas, sin la necesidad de una interpretación preliminar por parte del especialista.

1.2. Objetivos del trabajo

1.2.1. Objetivo general

- Implementar un modelo de agrupamiento no supervisado utilizando datos de masas mamográficas capaz de identificar lo mejor posible grupos de riesgo con características malignas.

1.2.2. Objetivos específicos

- Explorar diversas técnicas de agrupamiento no supervisado para determinar el mejor modelo de agrupamiento.
- Publicar el código del proyecto en un repositorio de GitHub de acceso público (https://github.com/DemetrioMunoz/TFM_Code).
- Desarrollar una web interactiva para el modelo seleccionado.

1.3. Impacto en sostenibilidad, ético-social y de diversidad

El presente trabajo tiene el potencial de impactar positivamente en el ámbito de la salud pública al mejorar la eficacia en el diagnóstico del cáncer de mama. Las principales implicaciones son la reducción de diagnósticos erróneos y pruebas invasivas como las biopsias. Además, puede llegar a mejorar la tasa de supervivencia de las pacientes por la detección temprana de este tipo de afección. Todo esto se traduce en una reducción de los recursos médicos y económicos necesarios en el proceso de diagnóstico del cáncer de mama.

Los Objetivos de Desarrollo Sostenible (**ODS**) (Naciones Unidas, 2023) [12] impactados positivamente con este trabajo son:

- **ODS 3 (Salud y Bienestar):** El proyecto contribuye a mejorar la salud y el bienestar de las mujeres al facilitar la detección temprana del cáncer de mama.
- **ODS 9 (Industria, Innovación e Infraestructura):** El proyecto fomenta la innovación en el sector de la salud mediante el desarrollo de nuevas tecnologías para el diagnóstico del cáncer de mama.
- **ODS 10 (Reducción de las Desigualdades):** El proyecto puede contribuir a reducir las desigualdades en el acceso a la atención médica al facilitar el diagnóstico temprano del cáncer de mama a mujeres de diferentes grupos socioeconómicos.

Sin embargo, existen aspectos negativos que debemos considerar, como el posible sesgo de los resultados. El modelo podría discriminar a ciertos grupos vulnerables

que tienen un difícil acceso a la sanidad pública. Para evitar esto, se deberían implementar medidas como:

- Realizar pruebas exhaustivas del modelo con diferentes grupos de población para asegurar que no hay sesgos en los resultados.
- Desarrollar estrategias para que el acceso a estos métodos diagnósticos sea equitativo para todos los grupos socioeconómicos y sociales.
- Establecer programas de formación para los profesionales de la salud sobre cómo utilizar estos métodos diagnósticos de forma responsable y ética.

Otro aspecto que considerar es la privacidad y seguridad de los datos sensibles. Al hacer uso de nuevas tecnologías, como la inteligencia artificial, se corre el riesgo de exponer información sensible. Para proteger la privacidad de los pacientes, se deben tomar las siguientes medidas:

- Implementar medidas de seguridad robustas para proteger los datos de los pacientes.
- Obtener el consentimiento informado de los pacientes antes de utilizar sus datos para cualquier propósito.
- Establecer políticas claras sobre cómo se recopilan, utilizan y almacenan los datos de los pacientes.

La introducción de herramientas de diagnóstico computacionales puede provocar la pérdida de puestos de trabajo intermedios entre paciente y médico. Para minimizar este impacto, se pueden tomar las siguientes medidas:

- Invertir en la formación de los profesionales de la salud para que puedan adaptarse a las nuevas tecnologías.
- Desarrollar nuevos programas de empleo en el sector de la tecnología médica.
- Ofrecer apoyo a los trabajadores que se vean afectados por la automatización con el objetivo de que puedan encontrar nuevos empleos.

En conclusión, el proyecto tiene el potencial de generar impactos positivos en la salud pública, la igualdad social y el desarrollo económico. Sin embargo, es importante tomar las medidas necesarias para minimizar los impactos negativos.

1.4. Enfoque y método seguido

Teniendo en cuanta los datos que manejamos en este trabajo, datos basados características tumorales provenientes de las imágenes mamográficas del estudio de Elter et al. (2017)[13] se optó por implementar modelos de aprendizaje no supervisados, concretamente modelos de agrupamiento que buscan identificar estructuras intrínsecas en los datos al agruparlos en conjuntos homogéneos o "clústeres" basados en similitudes entre las observaciones (Patel, 2019)[14]. Este enfoque se ha elegido frente al aprendizaje supervisado por varios motivos. El aprendizaje no supervisado no requiere que los datos estén previamente etiquetados, además permite identificar patrones y relaciones del conjunto de datos, dando una mejor comprensión de la heterogeneidad de las características tumorales. Al mismo tiempo, al no depender de las etiquetas para entrenar el modelo, el aprendizaje no supervisado puede adaptarse mejor a la variabilidad de los datos médicos, lo que puede conllevar una mejora en la precisión y la utilidad de los modelos para el diagnóstico, en nuestro caso, del cáncer de mama en el ámbito clínico.

1.5. Estado del Arte

Actualmente, las nuevas tecnologías están ganando terreno en diversos campos de interés. En el ámbito de la salud, el aprendizaje automático y la inteligencia artificial están asumiendo un papel cada vez más relevante (Jovel & Greiner, 2021)[15], ya que pueden detectar fases tempranas de afecciones, mejorar los diagnósticos, agilizar la atención a los pacientes y permitir tratamientos personalizados al manejar y procesar grandes volúmenes de datos.

Concretamente, en el campo oncológico del cáncer de mama, se han desarrollado y se están desarrollando métodos de diagnóstico no invasivos que utilizan técnicas de aprendizaje automático supervisado y no supervisado. En particular, las técnicas no

supervisadas, como el agrupamiento o “clustering”, permiten identificar patrones y relaciones ocultas en los datos médicos sin la necesidad de intervención humana. Este proceso puede traducirse en una detección temprana y precisa del cáncer de mama. Estudios recientes han demostrado que estas técnicas pueden analizar datos médicos, mejorando significativamente la precisión y la rapidez del diagnóstico (Thanoon et al., 2023; Jalloul et al., 2023)[16, 17]. Por último, el uso del aprendizaje automático reduce la necesidad de procedimientos invasivos, ya que proporciona información sobre la malignidad de las lesiones detectadas en imágenes radiográficas, como en el caso de las mamografías (Nasser & Yusof, 2023)[18].

1.6. Planificación del trabajo

1.6.1. Tareas y calendario

Diagrama de Gantt, ver Anexo I.

1.6.2. Hitos

TABLA 2. DESCRIPCIÓN DE LOS HITOS DEL PROYECTO.

Hito	Descripción	Fecha de inicio	Fecha de entrega
Entrega de la actividad P1	Definición y plan de trabajo. PEC1.	28/02/2024	19/03/2024
	Búsqueda de información sobre el tema del proyecto.		05/03/2024
	Selección del conjunto de datos y modelos de agrupamiento.		12/03/2024
Entrega de la actividad P2	Desarrollo del trabajo (Fase1). PEC2.	20/03/2024	23/04/2024

	Preprocesamiento del conjunto de datos.		29/03/2024
	Aplicación preliminar de los modelos de agrupamiento.		05/06/2024
	Creación de un repositorio público (GitHub).		19/04/2024
Entrega de la actividad P3	Desarrollo del trabajo (Fase2). PEC3.	24/04/2024	28/05/2024
	Selección de modelos de agrupamiento		05/06/2024
	Progreso de la memoria.		18/06/2024
	Desarrollo de la aplicación web interactiva.		17/06/2024
Entrega de la actividad P4	Cierre de la memoria y de la presentación. PEC4.	29/05/2024	18/06/2024
	Entrega de la memoria y presentación.		18/06/2024
Entrega de la actividad P5	Defensa pública. PEC5.	25/06/2024	05/07/2024

1.6.3. Análisis de riesgos

TABLA 3. DESCRIPCIÓN DEL ANÁLISIS DE RIESGOS DURANTE EL PROYECTO.

Descripción del riesgo	Severidad	Probabilidad	Mitigación
Funcionamientos no esperados de las librerías de Python	Alta	Leve	Uso de librerías alternativas
Incapacidad para implementar los modelos elegidos en los datos	Alta	Moderada	Revisión de los modelos, optimización de parámetros, entrenamiento adicional...
No conseguir el rendimiento objetivo	Alta	Moderada	Optimización de hiperparámetros, evaluación más exhaustiva, incremento del tamaño muestral y diagnóstico de errores
No conseguir reproducibilidad con otro conjunto de datos	Leve	Leve	Estandarizar o normalizar los datos, establecer semillas aleatorias, estudio se sensibilidad, comprobar la estabilidad...
Problemas con el repositorio	Moderada	Leve	Mantener el desarrollo y el control de versiones actualizado. Seguir las prácticas recomendadas de Github
No poder implementar la web interactiva	Moderada	Moderada	Investigación y aprendizaje. Explorar otros frameworks alternativos para desarrollo web
Complejidad técnica entre la relación de los modelos y la aplicación web interactiva	Alta	Moderada	Investigación y aprendizaje de herramientas especializadas y alternativas

1.7. Breve sumario de productos obtenidos

De este proyecto obtenemos los siguientes productos:

- **Plan de trabajo:** informe en formato PDF que contenga la justificación, los objetivos y el procedimiento del proyecto, desglosando las tareas con los tiempos estimados de realización para cada una de ellas. Incluido en la memoria.

- **Memoria:** documento en PDF que contenga una descripción detallada del proyecto, incluyendo la justificación, objetivos, metodología, resultados obtenidos, análisis de estos, conclusiones y posibles recomendaciones para futuros estudios. Además, se incluirá una sección dedicada al código implementado en el proyecto.
- **Repositorio público:** repositorio público (GitHub) de acceso libre al código desarrollado a lo largo del proyecto.
- **Página web interactiva:** desarrollada en Flask para el procesamiento y segmentación de datos basados en masas mamarias.
- **Presentación virtual:** PowerPoint y video que exponga de manera clara y concisa los aspectos clave del proyecto.

1.8. Breve descripción de los otros capítulos de la memoria

1.8.1. Material y métodos

Explicación de la base de datos utilizada, análisis exploratorio y técnicas de preprocesamientos de nuestros datos. Modelos de agrupamiento utilizados, parámetros e índices de validación para cada modelo. Características y diseño de la web interactiva.

1.8.2. Resultados

Explicación de los resultados obtenidos para cada modelo. Valoración de los índices de validación y justificación del número de clústeres obtenido.
Justificación del modelo/modelos no supervisados elegido para implementar en la web interactiva. Demo y explicación de uso de la web interactiva.

1.8.3. Conclusión y trabajos futuros

Conclusiones basadas en los resultados obtenidos. Aplicaciones futuras del trabajo realizado y posibles próximos pasos.

1.8.4. Glosario

Apartado para definir términos técnicos empleados durante el proyecto que puedan ser de ayudado para entender el tema tratado.

1.8.5. Bibliografía

Bibliografía usada y citada a lo largo del proyecto.

Capítulo 2

2. Material y Métodos

2.1. Conjunto de datos

El conjunto de datos seleccionado "mammograpich masses" contiene información sobre características tumorales extraídas de imágenes mamográficas. Este conjunto de datos proviene del estudio de Elter et al. (2017)[13] y ha sido descargado del repositorio "**data.wold**".

El trabajo de Elter et al. (2017)[13] evaluó dos sistemas de diagnóstico asistido por computadora (CAD) para predecir con precisión los resultados de biopsias de cáncer de mama y para facilitar la interpretación y toma de decisiones por parte de los radiólogos, destacando la importancia de la precisión y comprensibilidad de estos sistemas. Nuestro proyecto busca seguir un enfoque similar al hacer uso del aprendizaje automático para mejorar la capacidad predictiva y la interpretación de los resultados en el diagnóstico del cáncer de mama. Sin embargo, a diferencia del estudio original, que hizo uso de etiquetas para entrenar el modelo supervisado, nuestro enfoque utiliza técnicas de aprendizaje no supervisado. Este método no precisa de etiquetas para entrenar los modelos, lo que permite identificar patrones y relaciones ocultos en los datos.

El conjunto de datos "**mammograpich mases**" consta de 961 observaciones con 6 atributos, incluyendo la evaluación Bi-Rads, la edad del paciente, la forma de la masa, el margen, la densidad tumoral y la gravedad (diagnóstico) (Tabla 4).

TABLA 4. DESCRIPCIÓN DE VARIABLES, EQUIVALENCIAS Y TIPOS DE DATOS DEL CONJUNTO "MAMMOGRAPICH MASSES".

Variable	Descripción	Tipo de variable
Bi-Rads	**Clasificación Bi-Rads: 0 a 6**	**Ordinal**
Edad	**Edad del paciente en años**	**Continuo**

Forma	**Forma tumoral**	**Categórico**
	Redonda	1
	Ovalada	2
	Lobulada	3
	Irregular	4
Margen	**Margen tumoral**	**Categórico**
	Circunscrita	1
	Micro lobulada	2
	Oscurecida	3
	Mal definida	4
	Espiculada	5
Densidad	**Densidad tumoral**	**Ordinal**
	Alta	1
	Iso (densidad similar al tejido mamario circundante)	2
	Baja	3
	Contiene grasa	4
Severidad	**Diagnóstico**	**Binomial**
	Benigna	0
	Maligna	1

En un primer paso, se eliminaron los datos faltantes del conjunto de datos, descartando toda fila que tuviera un valor faltante en alguno de sus atributos. Reduciendo el número de observaciones de 961 a 830, con las cuales realizamos los análisis posteriores. Luego, comparamos la distribución de la edad (“age”), que es la única variable continua, con las demás variables categóricas para comprobar si la variable edad es distintiva entre categorías (Figura 1). La variable edad se conservó para realizar diferentes modelos de agrupamiento. Aunque en la clasificación de las lesiones mamarias la edad no es un factor directo, sí se puede relacionar con la densidad mamaria (Spak et al., 2017)[19]. Finalmente, se descartaron las variables de evaluación Bi-Rads ("score"), al ser comparable a nuestra variable objetivo, y la variable diagnóstico ("malignant"), que se usó como una forma de evaluar la calidad de los clústeres obtenidos.

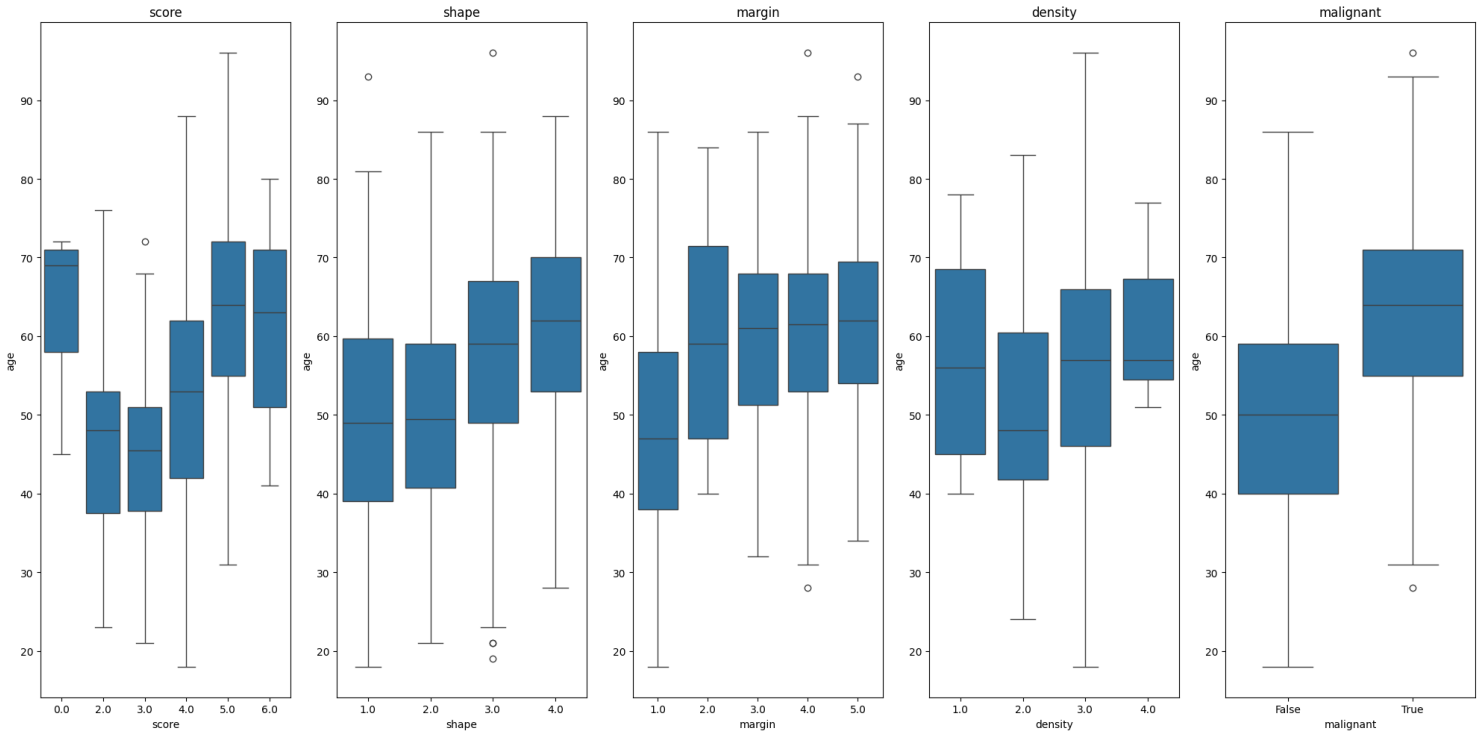


FIGURA 1. DISTRIBUCIÓN DE LA VARIABLE EDAD ("AGE") POR CADA VARIABLE DEL CONJUNTO DE DATOS.

El conjunto de datos resultante para implementar los modelos de agrupamiento se distribuye y se relaciona de la siguiente manera (Figura 2):

```
      age  shape  margin  density
0    45.0    2.0     4.0      3.0
1    69.0    4.0     5.0      3.0
2    58.0    4.0     4.0      3.0
3    72.0    4.0     3.0      3.0
4    71.0    4.0     4.0      3.0
..    ...    ...     ...      ...
954  80.0    4.0     5.0      3.0
955  51.0    4.0     4.0      3.0
956  71.0    4.0     4.0      3.0
957  68.0    4.0     3.0      3.0
958  72.0    4.0     3.0      3.0

[830 rows x 4 columns]
<class 'pandas.core.frame.DataFrame'>
Index: 830 entries, 0 to 958
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   age      830 non-null    float64
 1   shape    830 non-null    object
 2   margin   830 non-null    object
 3   density  830 non-null    object
dtypes: float64(1), object(3)
memory usage: 32.4+ KB
```

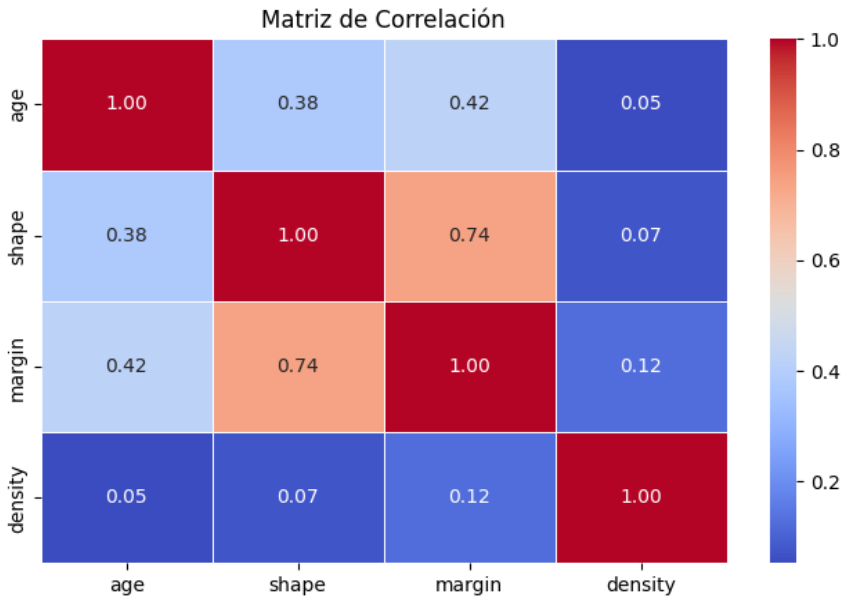


FIGURA 2. CONJUNTO DE DATOS RESULTANTE, TIPO DE VARIABLE Y CORRELACIÓN ENTRE VARIABLES.

En la correlación de las variables se observó que la forma y margen tumoral tienen una correlación de 0,75, mientras que con la edad y la densidad los valores de correlación son menores a 0,45.

La edad y la densidad mamaria, aunque son factores importantes en el riesgo del cáncer de mama, no son tan determinantes como las características de las lesiones tumorales en cuanto a la forma y el margen para la clasificación en el diagnóstico por mamografía. A medida que aumenta la edad, el riesgo de cáncer de mama también aumenta . Asimismo, una densidad mamaria elevada es un factor de riesgo que puede dificultar la detección de tumores en las mamografías debido a que el tejido denso puede ocultar las lesiones (Boyd et al., 2007; Edwards et al., 2014)[20, 21]. En cambio, las características específicas de las lesiones, forma y márgenes, tienen una relación directa con el riesgo de malignidad y, por lo tanto, son más determinantes en el diagnostico.

2.2. Herramientas y métodos de procesamiento de datos

2.1.1. Lenguaje de programación y paquetes

El código y la implementación de los algoritmos se desarrollaron en el lenguaje de programación Python utilizando el programa Visual Studio Code, aprovechando Jupyter Notebook (Microsoft, n.d.; Jupyter Project, n.d.)[22, 23].

Para el análisis de datos y la implantación de los modelos de agrupamientos se utilizaron los siguientes paquetes de Python:

- **Pandas** para la manipulación y análisis de datos estructurados en formato dataframes (The pandas development team, 2020)[24].
- **Seaborn y Matplotlib** para la visualización de datos y la generación de gráficos y figuras (Waskom, 2021; Hunter, 2007)[25, 26].
- **Scikit-learn** para el preprocesamiento y transformaciones de los datos. Además, junto con **kmodes y scikit-learn-extra** se implementaron algoritmos de clustering como KMeans, MeanShift, KModes,

KPrototypes, y KMedoids (Pedregosa et al., 2011; De Vos, n.d.; Lemaître, Nogueira & Aridas, 2020)[27, 28, 29].

- **SciPy** para realizar análisis de clúster jerárquico (Virtanen et al., 2020)[30].
- **UMAP** para la reducción de dimensionalidad (McInnes, Healy & Melville, 2018)[31].
- **Matrices de distancia** para utilizar algoritmos de agrupamiento y análisis de datos que requieren una medida de distancia.

2.1.2. Transformación de los datos

El conjunto de datos utilizado en este proyecto contiene datos mixtos, tanto variables categóricas como números. Debido a estas características se han usado diferentes estrategias para poder implementar los algoritmos de agrupación.

2.1.2.1. Codificación de variables categóricas (One-Hot Encoding)

La codificación de variables categorías “One-Hot Encoding” (Kumar et al., 2018)[32] se usó para convertir variables categóricas en un formato que puede ser proporcionado a algoritmos de aprendizaje automático para mejorar su rendimiento. Cada categoría de una variable categórica se convierte en una nueva columna binaria (0 o 1). Algunos modelos no manejan las variables categóricas de manera adecuada, por ello, se realizó la codificación de las variables categóricas para los modelos de agrupamiento basados en la distancia.

2.1.2.2. Reducción de dimensionalidad (UMAP)

Con las variables categóricas codificas (sin tener en cuenta la variable edad), se aplicó la técnica de reducción de dimensionalidad UMAP (Uniform Manifold Approximation and Projection). Esta técnica permitió facilitar la identificación y visualización de patrones en la estructura de nuestros datos. Reduciendo la dimensión a dos componentes principales, se identificó tres zonas en las que se agrupan

nuestros datos (Figura 3) que podrían representar diferentes subgrupos cada uno con características únicas o compartidas. Esta técnica en combinación con modelos de agrupamiento proporcionó una manera eficaz para analizar y agrupar las características de nuestros datos.

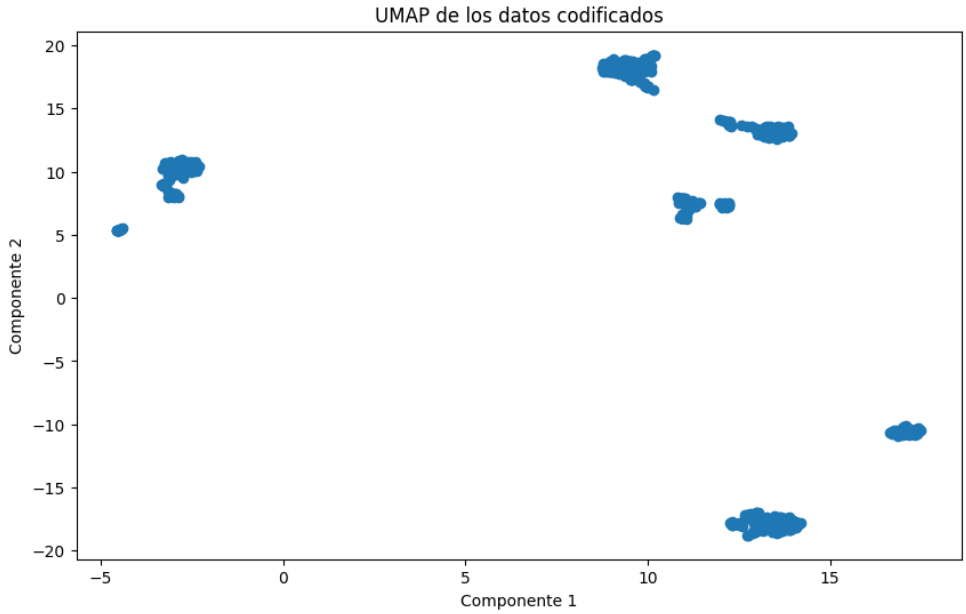


FIGURA 3. VISUALIZACIÓN DE LA ESTRUCTURA DE LOS DATOS CON LA REDUCCIÓN DE DIMENSIONALIDAD UMAP.

2.2.2.3. Discretización de variables

Una de las otras técnicas utilizadas para poder aplicar de manera óptima los modelos de agrupamiento fue discretizar las variables continuas para los modelos que solo manejen variables categóricas. Se transformo la variable continua edad (“age”) a grupos en intervalos de diez años.

2.2.2.4. Matriz de distancia

La última técnica utilizada para transformar los datos y aplicar los modelos de agrupamiento fue la utilización de matrices de distancia, concretamente se optó por usar la matriz de Gower (Gower, 1971)[33]. Al usar la matriz de Gower se puede calcular la similitud entre las observaciones considerando tanto las variables numéricas como las categóricas. Esta medida de distancia se usó en modelo donde se pueden

aplicar en modelo que permiten medidas de distancia precomputadas, como los métodos de agrupamiento jerárquicos.

2.3. Modelos de Agrupamiento (Clustering)

Los modelos de agrupamiento son técnicas para el análisis exploratorio de los datos en los que se pueden descubrir grupos, identificar distribuciones y patrones en los datos subyacentes. Estos modelos consisten en dividir un conjunto de datos dado en grupos (clústeres) de tal manera que las observaciones dentro del grupo sean similares entre sí que las observaciones de diferentes grupos (Guha et al., 1998)[34]. El proceso de agrupamiento se puede observar en la Figura 4.

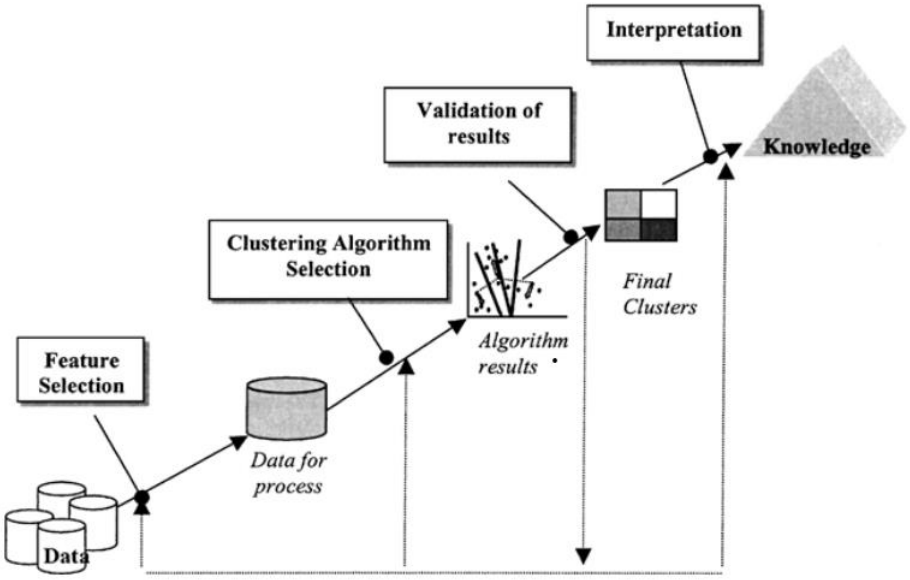


FIGURA 4. ETAPAS DEL PROCESO DE AGRUPACIÓN (FAYYAD ET AL., 1996)[35]

Los modelos implementados en este proyecto se muestran en la siguiente tabla:

TABLA 5. LISTADO DE MODELOS DE AGRUPAMIENTOS IMPLEMENTADOS.

Modelos de Agrupamiento
K-Means
K-Medoids
K-Modes

K-Prototypes

Hierarchical clustering

2.2.1. K-Means

El algoritmo K-Means (Sinaga & Yang, 2020)[36] es una técnica de agrupamiento de particiones (no jerárquico) con el objetivo de dividir un conjunto de observaciones en ***k*** grupos o clústeres de manera que las observaciones dentro de un mismo grupo sean más similares entre sí que con las observaciones de otros grupos.

Este algoritmo utiliza la distancia euclidiana como medida de similitud entre observaciones de los datos. Esta distancia entre dos puntos en un espacio euclidiano es la longitud del segmento de línea que une los dos puntos.

El algoritmo K-Means divide un conjunto de datos ***n*** con ***x*** observaciones en ***k*** clústeres ***C***, descritos por la media $\boldsymbol{u_j}$ de las muestras en el clúster. Esta media se le conoce como centroide, que, aunque no está incluido en el conjunto comparte el mismo espacio. Con lo cual, el algoritmo tiene como objetivo elegir los centroides que minimicen la inercia, siguiendo la ecuación:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right)$$

La inercia puede reconocerse como una medida del grado de coherencia interna de los clústeres.

En resumen, el algoritmo tiene tres pasos de acción. El primer paso asigna cada observación al centroide más cercano. El segundo paso crea nuevos centroides tomando el valor medio de todas las observaciones asignadas a cada centroide anterior. Se calcula la diferencia entre los centroides antiguos y los nuevos, y

se repiten estos dos últimos pasos hasta que este valor sea menor que un umbral.

2.2.2. K-Medoids

El algoritmo K-Medoids se basa en el algoritmo K-Means, pero con algunas diferencias (Arora & Varshney, 2016)[37]. K-Means minimiza la suma de cuadrados dentro del clúster, mientras que K-Medoids minimiza la suma de las distancias entre cada punto y el medoid del clúster. El medoid es un punto en los datos que tiene la menor distancia total a las otras observaciones de su clúster, es decir, las disimilitudes con las observaciones a su alrededor son mínimas. Esta característica permite cualquier métrica de distancia para el agrupamiento (por ejemplo, una matriz de similaridad o matriz de disimilaridad).

La disimilitud del medoid (***Ci***) y las observaciones (***Pi***) se calcula mediante:

$$\boldsymbol{E = |Pi - Ci|}$$

donde el coste del modelo se calcula mediante la ecuación:

$$\boldsymbol{c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|}$$

2.2.3. K-Modes

El algoritmo K-Modes está diseñado para agrupar datos categóricos (Huang, 1998)[38]. Define los clústeres basándose en el número de categorías coincidentes entre las observaciones de los datos, es decir, usa la moda en lugar de la media para actualizar los centroides y la disimilitud de Hamming para

calcular la distancia. Esto difiere del algoritmo K-Means, que agrupa datos numéricos basándose en la distancia euclidiana.

La distancia de Hamming (d_H) se calcula entre dos vectores categóricos x e y de longitud m:

$$d_H(x, y) = \sum_{i=1}^{m} \delta(x_i, y_i)$$

donde $\delta(x_i, y_i)$ es una función indicadora:

$$\delta(x_i, y_i) = \begin{cases} 0 \; si \; x_i = y_i \\ 1 \; si \; x_i \neq y_i \end{cases}$$

Con lo cual, cada observación se asigna al clúster cuyo "mode" (centroide) está más cercano en términos de Hamming:

$$C_i = \{ x_j \mid d_H(x_j, m_j) \leq d_H(x_j, m_k) \; para \; todos \; k = 1, 2, \ldots, K \}$$

donde m_i es el "mode" del clúster C_i y K es el número de clúster.

Después de asignar las observaciones a los clústeres, se recalculan los "modes" de cada clúster. El "mode" para cada atributo en un clúster es el valor más frecuente de ese atributo en el clúster:

$$m_{ij} = mode(x_{kj} | x_k \in C_i)$$

Donde m_{ij} es el "mode" del atributo j en el clúster i, y x_{kj} es el valor del atributo j para la observación k en el clúster i.

Por último, se repite la asignación y el cálculo del "mode" hasta que las observaciones no cambian de clúster.

2.2.4. K-Prototypes

El algoritmo K-Prototypes combina K-Modes y K-Means y es capaz de agrupar datos mixtos, tanto numéricos como categóricos (Huang, 1997)[39]. Se logra

gracias a que combina la distancia euclidiana para variables numéricas y una medida adecuada para variables categóricas. Este algoritmo calcula la distancia total entre dos observaciones $\boldsymbol{x}$ e $\boldsymbol{y}$, una con características numéricas $\boldsymbol{x_{num}}$ y otra con características categóricas $\boldsymbol{x_{cat}}$, la ecuación se define como:

$$\boldsymbol{d(x,y) = \sqrt{\sum_{j=1}^{P_{num}} (x_{num,j} - y_{num,j})^2} + \sum_{j=1}^{P_{cat}} \delta(x_{cat,j}, y_{cat,j})}$$

donde:

- $\boldsymbol{P_{num}}$ es el número de características numéricas.
- $\boldsymbol{P_{cat}}$ es el número de características categóricas.
- $\boldsymbol{x_{num,j}}$ e $\boldsymbol{y_{num,j}}$ valores de la característica numérica $\boldsymbol{j}$ para $\boldsymbol{x}$ e $\boldsymbol{y}$.
- $\boldsymbol{x_{cat,j}}$ e $\boldsymbol{y_{cat,j}}$ valores de la característica categórica $\boldsymbol{j}$ para $\boldsymbol{x}$ e $\boldsymbol{y}$.
- $\boldsymbol{\delta(x_{cat,j}, y_{cat,j})}$ es la distancia entre las características categóricas $\boldsymbol{j}$.

2.2.5. Hierarchical clustering

Los algoritmos jerárquicos construyen los clústeres mediante la fusión o división sucesiva de las observaciones (Nielsen, 2016)[40], formando una estructura jerárquica que se representa como un dendrograma. Podemos encontrarnos dos tipos: aglomerativos y divisivos. El método aglomerativo comienza con cada observación en su propio clúster y fusiona clústeres sucesivamente, mientras que el método divisivo comienza con todas las observaciones en un único clúster y las divide sucesivamente. Este algoritmo permite procesar estructuras de datos como las matrices de distancias.

Este algoritmo, dependiendo si es para el método divisivo o aglomerativo, sigue las siguientes ecuaciones:

- **Divisivo** basada en la distancia máxima, donde para dividir un clúster C depende de la distancia de disimilitud utilizada:
-

$$D(C) = \max_{x,y \in C} d(x,y)$$

donde $d(x,y)$ es la distancia entre los puntos x e y.

- **Aglomerativo** para fusionar dos clústeres C_i y C_j con una distancia mínima, dependiendo de la distancia de similitud utilizada:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_i} d(x,y)$$

donde $d(x,y)$ es la distancia entre los puntos x e y.

2.4. Índices de evaluación

Evaluar modelos no supervisados de agrupamiento no es tan simple como evaluar modelos supervisados, donde podemos contar los errores o medir el "recall" y la precisión (Halkidi et al., 2001)[41]. En el caso de los modelos de agrupamiento, no podemos considerar los valores absolutos de las etiquetas de los clústeres. En cambio, debemos evaluar si el agrupamiento define separaciones de los datos que sean similares a algún conjunto de clases de referencia (si conocemos las etiquetas verdaderas del conjunto de datos) o si satisface alguna suposición, como que las características de los miembros de un mismo grupo sean más similares entre sí que las de los miembros de diferentes grupos. Además, los modelos de agrupamiento están abiertos a la interpretación según las características del estudio objetivo. Esto significa que la elección de la métrica de evaluación y la interpretación de los resultados pueden variar según el contexto y los objetivos específicos del análisis.

2.2.6. Coeficiente de silueta

Como las etiquetas reales no están disponibles evaluamos el rendimiento utilizando el propio modelo. Una puntuación más alta del coeficiente de silueta indica un modelo con grupos más definidos (Rousseeuw, 1987)[42].

El coeficiente de silueta ***s*** se calcula de la siguiente manera:

$$s = \frac{b - a}{max(a, b)}$$

donde:

- **a:** distancia media entre una muestra y todos los demás puntos en la misma clase.
- **b:** distancia media entre una muestra y todos los demás puntos en el siguiente clúster más cercano.

Los valores oscilan entre -1 y 1, mala agrupación y agrupación altamente densa, respectivamente. Valores cercanos a 0 indican clústeres superpuestos.

2.2.7. Índice de Davies-Bouldin

Si las etiquetas reales no están disponibles, podemos usar el índice de Davies-Bouldin (Davies & Bouldin, 1979)[43] para evaluar nuestros modelos. Donde valores bajos de este índice indican un modelo con mejor separación entre los clústeres. Con el índice ***DB*** (Davies-Bouldin) se calcula la similitud promedio entre clústeres, es decir, se compara la distancia entre clústeres con el tamaño de los propios clústeres.

El índice ***DB*** se calcula la similitud promedio entre cada clúster C_i y su clúster más similar C_j. Se define como:

- s_i, distancia promedio entre cada punto del clúster C_i y el centroide de ese clúster (diámetro del clúster).
- d_{ij}, distancia entre los centroides de los clústeres C_i y C_j.

Donde:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Con lo que el índice ***DB*** se define como:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{j \neq i} R_{ij}$$

donde ***n*** es el número de clústeres.

2.2.8. Índice de Calinski-Harabasz

Si no conocemos las etiquetas verdaderas podemos usar el índice de Calinsk-Habarasz (Caliński & Harabasz, 1974)[44] para evaluar nuestros modelos. Valores altos de este índice indican un modelo con clústeres mejor definidos.

El índice Calinsk-Habarasz permite evaluar la calidad de los clústeres considerando la variabilidad entre los clústeres en relación con la variabilidad dentro de los clústeres, favoreciendo una mayor separación y compactación.

Para un conjunto ***X*** de tamaño n_X que ha sido agrupado en k clústeres, el índice Calisnsk-Habarasz (***s***) se define como el ratio entre la dispersión promedio entre clústeres y la dispersión dentro de los clústeres:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_X - k}{k - 1}$$

donde $tr(B_k)$ es la dispersión entre clústeres y $tr(W_k)$ es la dispersión dentro de los clústeres. Se definen como:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_X)(c_q - c_X)^T$$

Donde C_q es el conjunto de puntos en el clúster q, c_q es el centro del clúster, c_X es el centro de X, y n_X el número de putos en el clúster q.

2.5. Página web interactiva

La aplicación web interactiva se desarrolló con Flask, un microframework de Python (Grinberg, 2018)[45]. Se diseñó con el objetivo de introducir nuevas observaciones provenientes de mamografías siguiendo el sistema de descripción de características Bi-Rads, en función de tres descriptores: forma, margen y densidad. Para este propósito, se implementó un formulario donde se pueden introducir estos tres descriptores, los cuales son agrupados en uno de los clústeres del modelo utilizado. La aplicación no reentrena los modelos con los nuevos datos y utiliza los modelos previamente entrenados y guardados en archivos pickle. Dependiendo del clúster, se determina la probabilidad de que esas características sean malignas o benignas. También se muestran los resultados del modelo desplegado en una ventana informativa, proporcionando información visual y estadística sobre las predicciones.

Para facilitar el acceso y uso de la aplicación web se alojó en "render (https://render.com/)" con el nivel de host gratuito, por lo que el tiempo de procesamiento de la aplicación web está restringido debido a las limitaciones de los recursos de los servicios de alojamiento web gratuitos. La aplicación web interactiva está disponible públicamente para su uso en: https://tfm-web.onrender.com/.

Toda la información referente al código, implementación y despliegue de la aplicación web interactiva se guardó en el siguiente repositorio de GitHub: https://github.com/DemetrioMunoz/TFM_Web.

Capítulo 3

3. Resultados

En este apartado presentamos los resultados obtenidos de los modelos de agrupamiento aplicados a los datos de masas mamográficas. Desde el tratamiento de los datos, las transformaciones utilizadas, el análisis para determinar el número óptimo de clústeres, los índices de evaluación y la selección del modelo final para su despliegue en una página web interactiva.

3.1. Modelos de Agrupamiento

3.1.1. K-Means

Las componentes resultantes de la reducción de dimensionalidad UMAP se aplicaron en el modelo de agrupamiento K-Means. Para determinar el número de clústeres optimo se exploró, en un rango de dos a diez clústeres con diferentes métodos de iniciación (“k-means++” y “random”), distintos índices de validación (Figura 5).

Tras examinar los índices de evaluación y analizar la distribución visual de clústeres (Figura 6), se determinó que, aun no siendo los valores óptimos de evaluación comparado a otro valor de clústeres, el número de clústeres que mejor se ajustó a la naturaleza de nuestros datos fue siete (k = 7) con el método de iniciación “k-means++”.

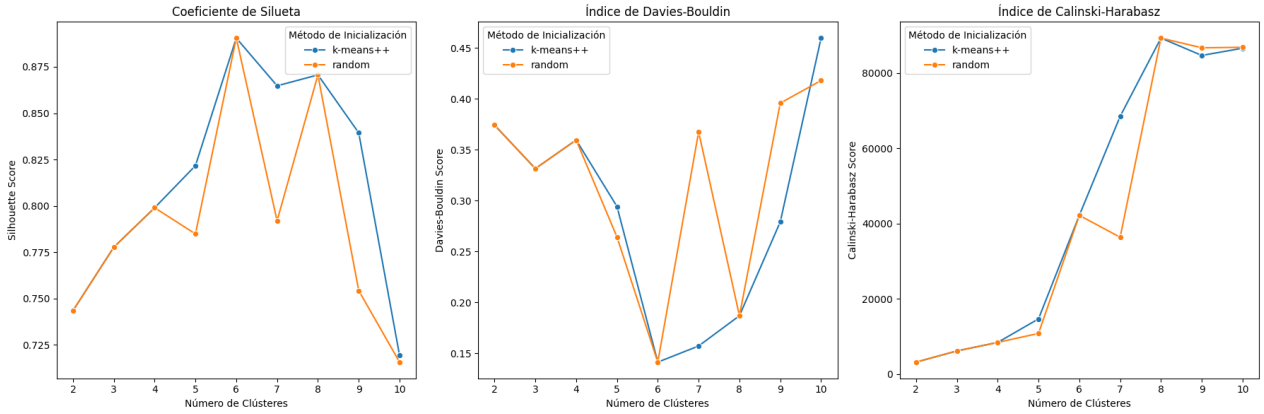


FIGURA 5. VALORES DE EVALUACIÓN PARA CADA VALOR DE K CON DISTINTOS TIPOS DE INICIACIÓN.

Con el número de clústeres seleccionado, los valores de evaluación (Tabla 6) indicaron que los grupos están bien separados, cohesionados y con las características tumorales asignadas a su correspondiente grupo.

TABLA 6. VALORES DE LOS ÍNDICES DE EVALUACIÓN PARA K = 7 CLÚSTERES.

Índices de evaluación K-Means (k = 7)	
Coeficiente de Silueta	0,86
Davies-Bouldin	0,16
Calinski-Harabasz	68486,80

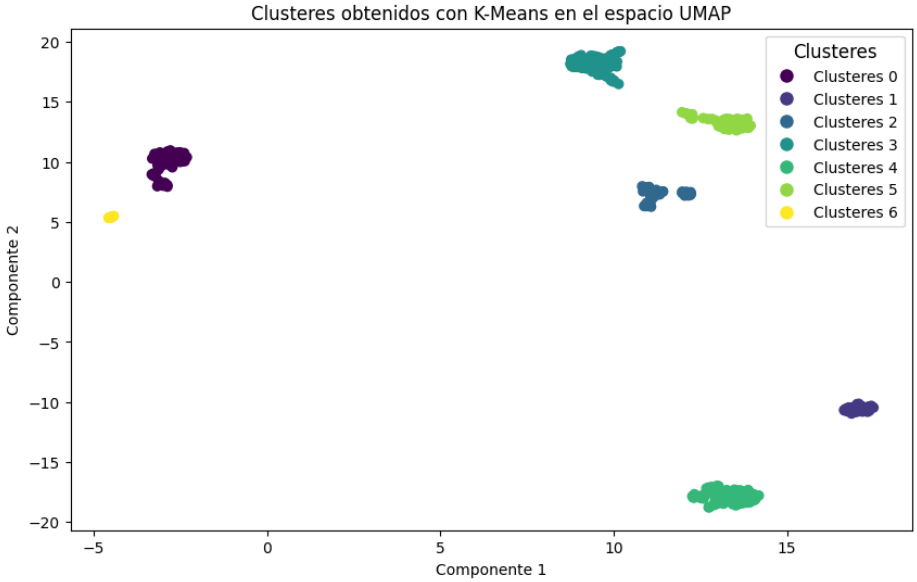


FIGURA 6. DISTRIBUCIÓN DE LOS CLÚSTERES DEL MODELO K-MEANS PARA K = 7 EN EL ESPACIO UMAP.

En la distribución de las características por clúster (Figura 7), se observó que, los grupos compartían características. Para la forma tumoral los grupos se segregaron de forma más definida, los clústeres 0 y 6 comparten la característica ovalada. Los clústeres 1, 3 y 5 se concentran en la forma irregular, con una presencia mínima del clúster 4. El clúster 4 se concentra en la forma redonda, donde también se observan unas pocas observaciones del clúster 3. El clúster 2 solo se observó para la forma lobulada.

Para el margen tumoral, se observó una segregación más irregular, ya que los clústeres comparten más características entre sí. En cuanto al margen circunscrito, solo están presentes los clústeres 0, 2, 4, 5 y 6. En el margen microlobulado, a pesar de tener un número bajo de observaciones, se presentaron todos los clústeres excepto el clúster 1. En el margen oscurecido, están presentes los clústeres 0, 1, 2, 4 y 6. Para el margen mal definido, no están presentes los clústeres 0, 2, 3, 5 y 6. Por último, el margen espiculado presenta los clústeres 0, 2, 4 y 5.

Para la característica densidad, se destacó que los clústeres se concentraron en la característica de densidad baja, con excepción del clúster 6, que está ausente en esta característica y solo está presente en la característica de densidad Iso.

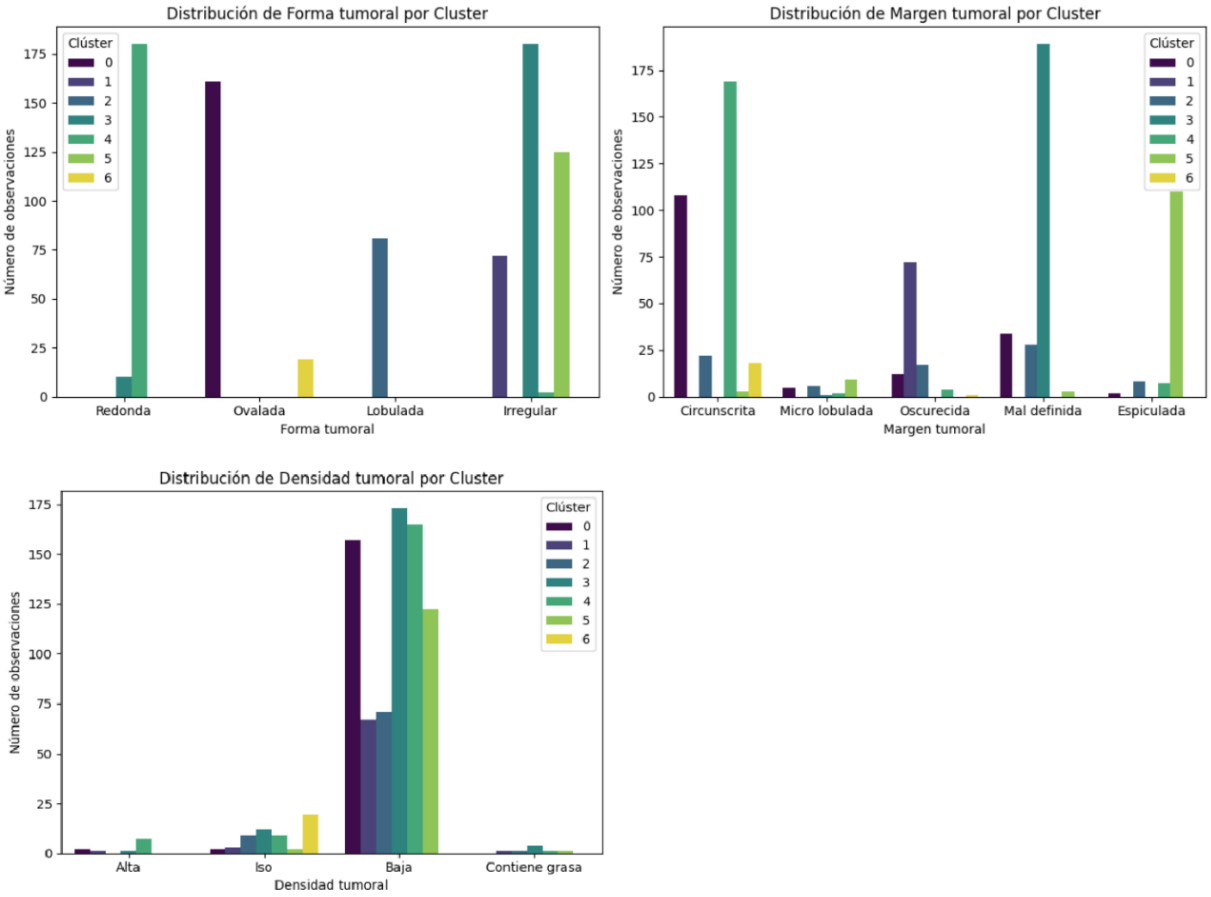


FIGURA 7. DISTRIBUCIÓN DE LAS CARACTERÍSTICAS TUMORALES: FORMA, MARGEN Y DENSIDAD POR CLÚSTER.

Por último, aprovechando que contamos con el diagnóstico de las observaciones (variable "malignant"), se realizó una evaluación de la calidad de nuestros clústeres. Para ello, se llevó a cabo un conteo de la distribución de las observaciones con diagnóstico de carácter maligno y benigno para cada clúster (Figura 8). El porcentaje de casos malignos por grupo fue el siguiente (Tabla 7): el clúster 0 presentó un 20% de casos malignos, el clúster 1 un 77%, el clúster 2 un 52%, el clúster 3 un 77%, el clúster 4 un 14%, el clúster 5 un 80% y el clúster 6 un 0%.

Gracias a esto, se pudo identificar qué grupos tienen características asociadas con diagnósticos de carácter maligno. Los clústeres 1, 3 y 5 presentan más del 75% de diagnósticos malignos cada uno. Estos clústeres comparten características como forma irregular, margen espiculado, margen mal definido y oscurecido, así como una densidad baja. Por otro lado, se identificaron clústeres con un porcentaje de observaciones malignas inferior al 20%, como

los clústeres 0 y 4. Además, se observó un clúster con un 50% de observaciones malignas, el clúster 2, en el cual no se pudo determinar si la naturaleza de sus características es predominantemente maligna o benigna. El clúster 6 destaca por no contener ninguna observación maligna, lo que sugiere que sus características son de índole benignas. Este clúster se corresponde con características como forma ovalada, margen circunscrito y densidad iso.

TABLA 7. NÚMERO DE OBSERVACIÓN SEGÚN EL DIAGNOSTICO, OBSERVACIONES TOTALES Y PORCENTAJE DE MALIGNIDAD PARA CADA CLÚSTER.

Clúster	Observaciones			
	Benigna	Maligna	Totales	Porcentaje (%) obs. Malignas
0	130	31	161	20
1	16	56	72	77
2	39	42	81	52
3	43	147	190	77
4	156	26	182	14
5	24	101	125	80
6	19	0	19	0

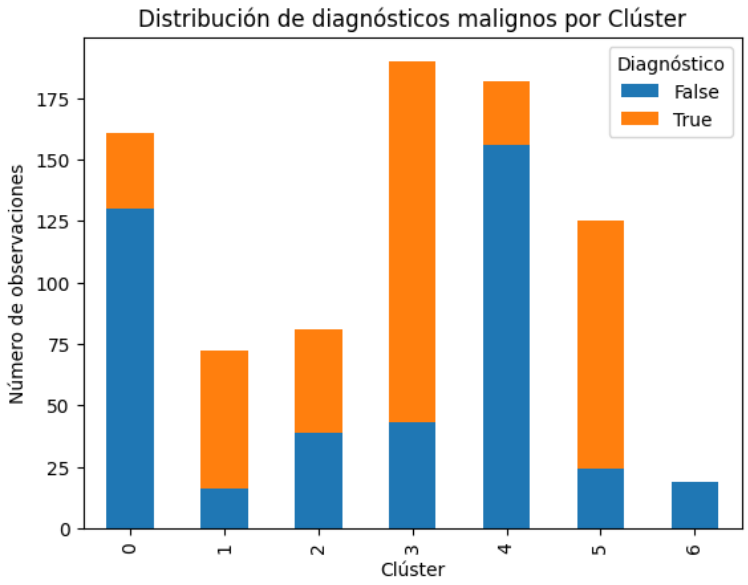


FIGURA 8. DISTRIBUCIÓN DE OBSERVACIONES MALIGNAS Y BENIGNA POR CLÚSTER.

Estos resultados parecen coherentes, ya que las características malignas se relacionan con formas irregulares, márgenes espiculados y mal definidos,

mientras que las características benignas se relacionan con márgenes tumorales bien definidos y circunscritos, así como formas ovaladas y regulares (National Cancer Institute, n.d)[46].

3.1.2. K-Modes

Se utilizaron las componentes categóricas del conjunto de datos para implementar el algoritmo K-Modes. En un primer momento se exploró la posibilidad de usar la variable continua edad (“age”), para usar esta variable se discretizó en intervalos de 10 años y estos intervalos se transformaron a valores numéricos categóricos mediante la codificación de etiquetas. Para determinar el número de clústeres optimo se utilizó un rango de clústeres (de 2 a 10) con distintos tipos de iniciación para el algoritmo (“Huang”, “Cao” y “random”)(Figura 9, Figura 10).

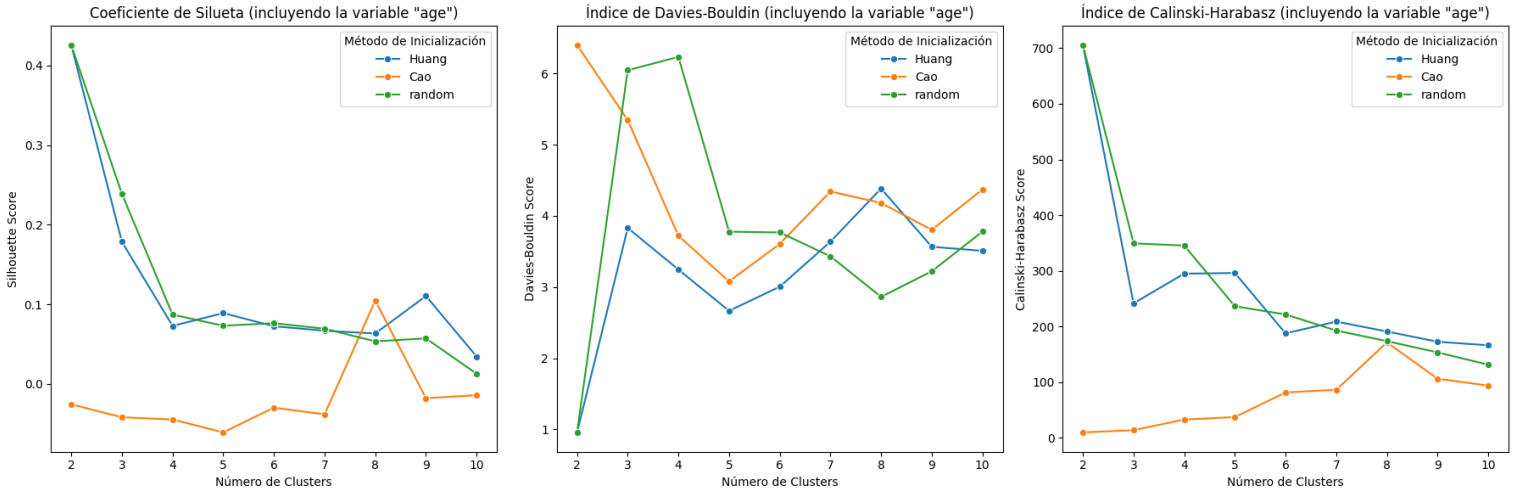


FIGURA 9. VALORES DE EVALUACIÓN PARA CADA VALOR DE K CON DISTINTOS TIPOS DE INICIACIÓN INCLUYENDO LA VARIABLE DISCRETIZADA Y CODIFICADA “AGE”.

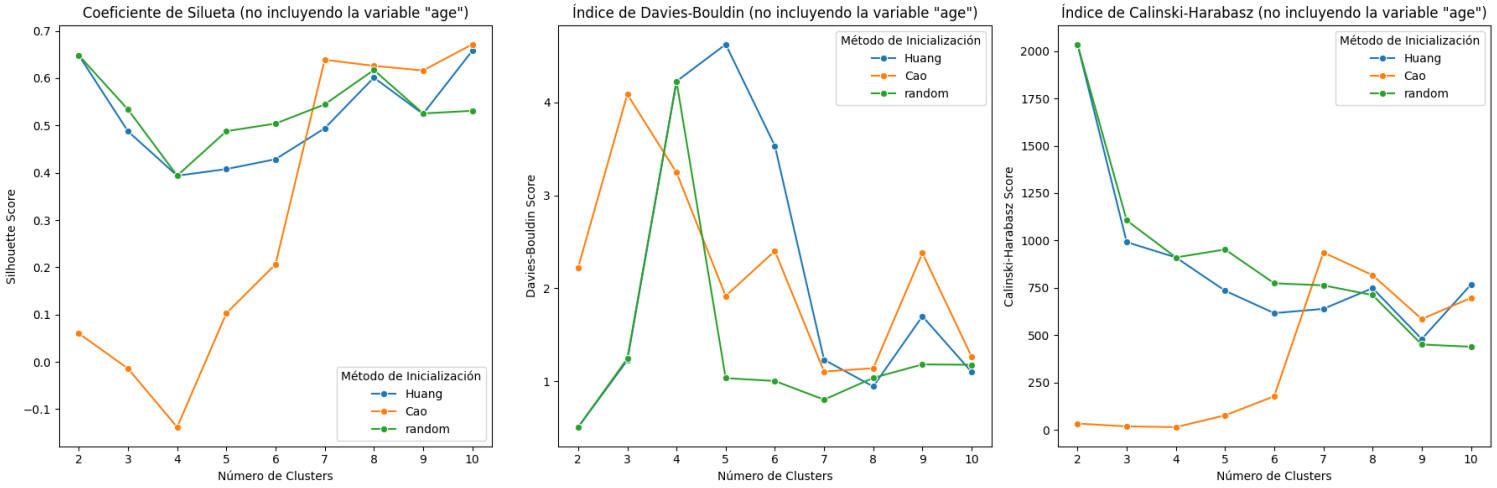


FIGURA 10. VALORES DE EVALUACIÓN PARA CADA VALOR DE K CON DISTINTOS TIPOS DE INICIACIÓN SIN INCLUIR LA VARIABLE "AGE".

En base a los índices obtenidos (Tabla 8) se optó por implementar el modelo sin incluir la variable edad ("age"), con el método de iniciación "random" y estableciendo el número de clústeres a 2 (k =2). Los valores de los índices indican que los clústeres pueden estar cohesionados y relativamente bien separados, aunque los grupos comparten alguna característica.

TABLA 8. VALORES DE LOS ÍNDICES DE EVALUACIÓN PARA K = 2 CLÚSTERES EN EL ALGORITMO K-MODES SIN INCLUIR LA VARIABLE EDAD ("AGE").

Índices de evaluación K-Modes (k = 2)	
Coeficiente de Silueta	0,64
Davies-Bouldin	0,50
Calinski-Harabasz	2035,60

Las características tumorales se distribuyeron en dos clústeres (Figura 11), y se observó que, aunque ambos clústeres compartían características, se agrupaban ciertas características más un clúster que en otro, menos para la densidad tumoral ("density"), que ambos grupos comparten mayoritariamente la característica de densidad baja.

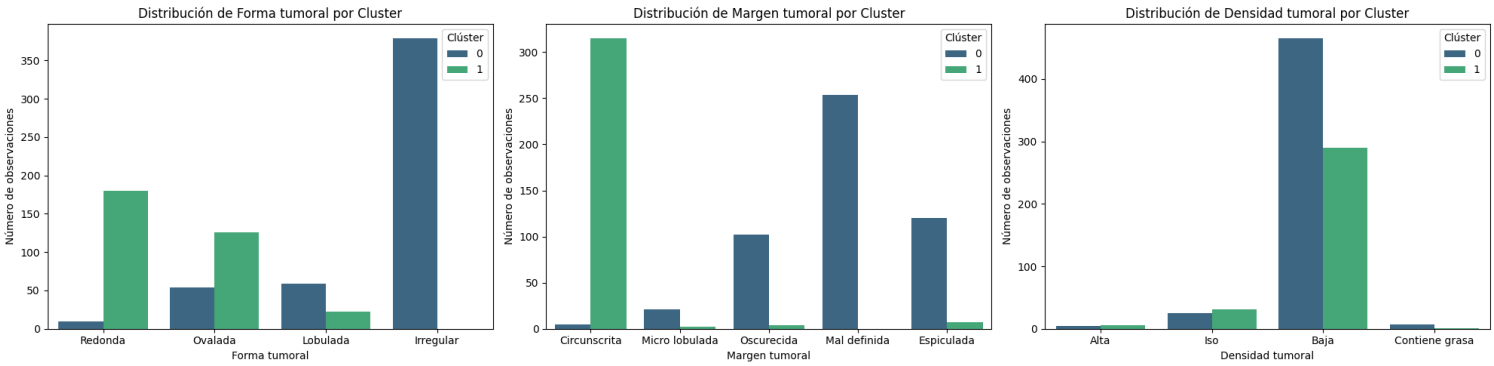


FIGURA 11. DISTRIBUCIÓN DE LAS CARACTERÍSTICAS TUMORALES: FORMA, MARGEN Y DENSIDAD POR CLÚSTER PARA EL ALGORITMO K-MODES.

El clúster 0 presentó características más relacionadas con dolencias malignas, como la formar irregular, donde las observaciones de este clúster se concentran, así como márgenes tumorales poco definidos. Al contrario, el clúster 1, presentó formas más definidas y esféricas (ovaladas y lobuladas), y el margen tumoral que presenta este grupo es sobre todo circunscrito o bien definido, características que suelen presentar los tumores de índole benigno.

Como en otros modelos, aprovechando los diagnósticos de las observaciones, se evaluó la calidad los clústeres (Tabla 9).

TABLA 9. NÚMERO DE OBSERVACIÓN SEGÚN EL DIAGNOSTICO, OBSERVACIONES TOTALES Y PORCENTAJE DE MALIGNIDAD PARA EL ALGORITMO K-MODES PARA K= 2.

Clúster	Observaciones			
	Benigna	Maligna	Totales	Porcentaje (%) obs. Malignas
0	142	360	502	72
1	285	43	328	13

Atendiendo a la distribución de los diagnósticos (Figura 12) se puede concluir que el clúster 0 concentra en gran medida las características malignas, con un 72% de las observaciones con diagnostico maligno. Y el clúster 1 aúna las características benignas, teniendo solo una incidencia de malignidad del 13%.

Aun así, aunque este algoritmo puede segregar características malignas de las benignas, no se consiguió un clúster que solo tuviera observaciones malignas o benignas, es decir, hay características que se solapan en los clústeres.

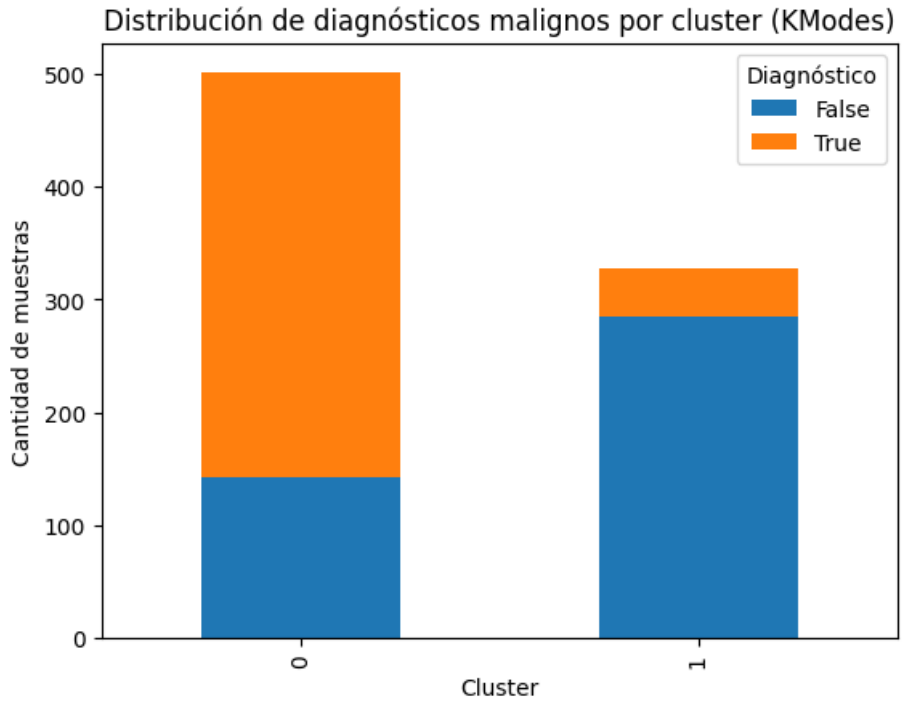


FIGURA 12. DISTRIBUCIÓN DE OBSERVACIONES MALIGNAS Y BENIGNA POR CLÚSTER PARA EL ALGORITMO K-MODES.

3.1.3. K-Prototypes

El algoritmo K-Prototypes nos permitió manejar las variables categóricas y numéricas en un mismo algoritmo sin la necesidad de codificar o transformar las variables. Como en anteriores modelos, se calcularon y se evaluaron los índices de evaluación para un rango de clústeres (de 2 a 10) con diferentes métodos de iniciación ("Huang", "Cao" y "random") (Figura 13).

Atendiendo a los valores de evaluación (Tabla 10) y a una exploración visual se estableció que el número óptimo de clúster fue 5 (k=5).

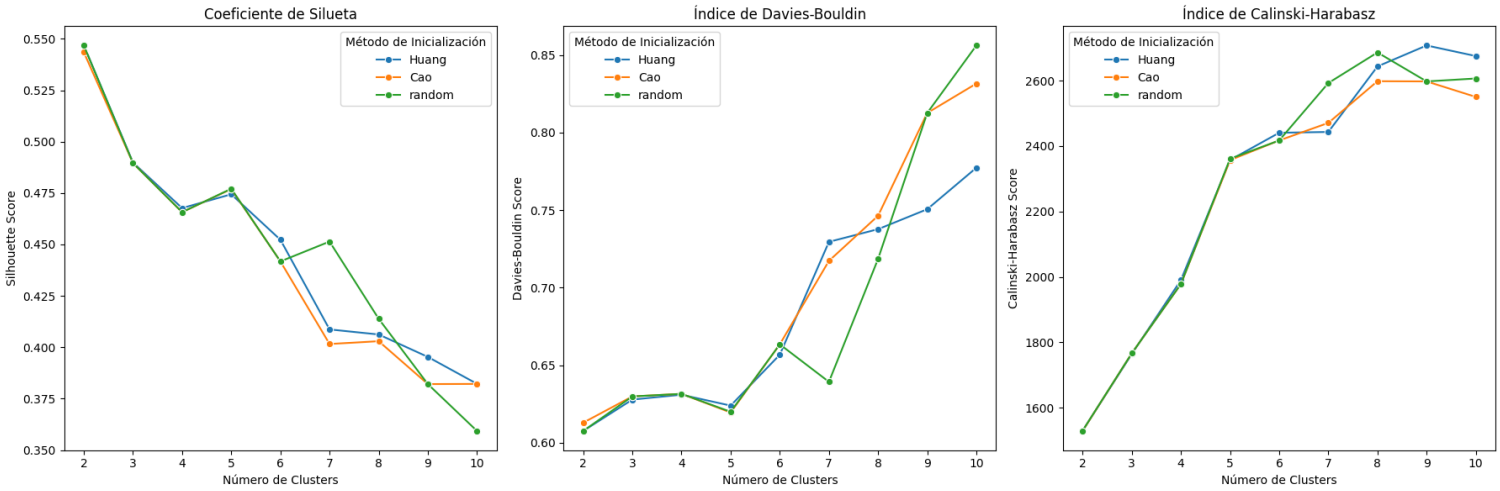


FIGURA 13. VALORES DE EVALUACIÓN PARA CADA VALOR DE K CON DISTINTOS TIPOS DE INICIACIÓN PARA EL ALGORITMO K-PROTOTYPES.

TABLA 10. VALORES DE LOS ÍNDICES DE EVALUACIÓN PARA K = 5 CLÚSTERES EN EL ALGORITMO K-PROTORYPES.

Índices de evaluación K-Prototypes (k = 5)	
Coeficiente de Silueta	0,47
Davies-Bouldin	0,63
Calinski-Harabasz	1990,27

Los valores de los índices no son tan buenos como en otros modelos, esto se observa en la distribución de las características, donde los clústeres no parecen bien segregados y comparten características (Figura 14).

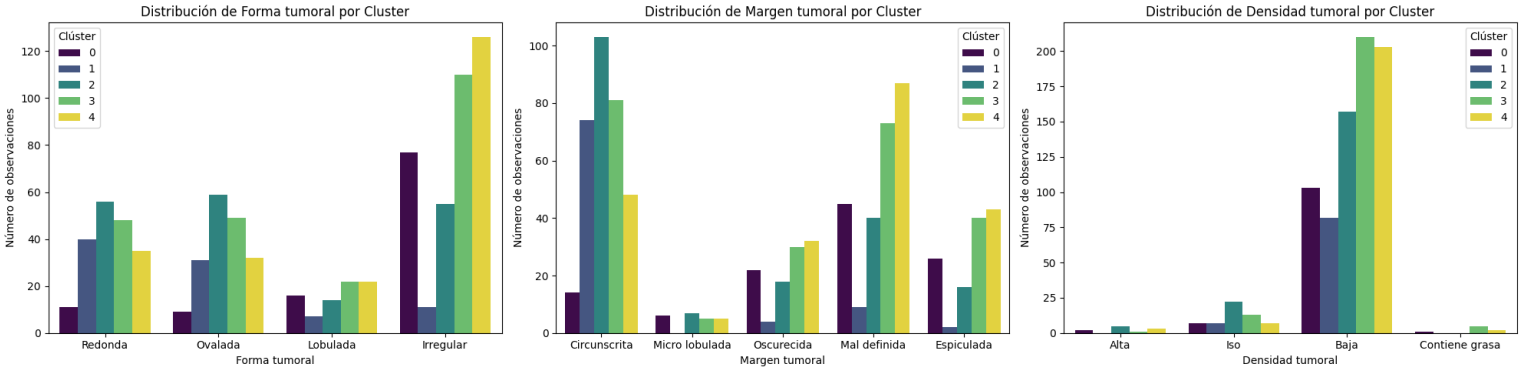


FIGURA 14. DISTRIBUCIÓN DE LAS CARACTERÍSTICAS TUMORALES: FORMA, MARGEN Y DENSIDAD POR CLÚSTER PARA EL ALGORITMO K-PROTOTYPES.

En general, este algoritmo, con nuestros datos no consigue la segregación de las características de manera satisfactoria. Se observó un solapamiento de las características en todos los grupos, e incluso, evaluando los clústeres con el diagnostico de las observaciones, solo se observaron dos clústeres, 0 y 1, en los que los diagnósticos se distribuían mayoritariamente en malignos o benignos, respectivamente (Figura 15). Los demás clústeres muestran una variabilidad alta en la que no se puede determinar significativamente si se agrupan en características benignas o malignas (Tabla 11).

TABLA 11. NÚMERO DE OBSERVACIÓN SEGÚN EL DIAGNOSTICO, OBSERVACIONES TOTALES Y PORCENTAJE DE MALIGNIDAD PARA EL ALGORITMO K-PROTOTYPES PARA K= 5.

Clúster	Observaciones			
	Benigna	Maligna	Totales	Porcentaje (%) obs. Malignas
0	20	93	113	82
1	81	8	89	9
2	130	54	184	29
3	124	105	229	46
4	72	143	215	66

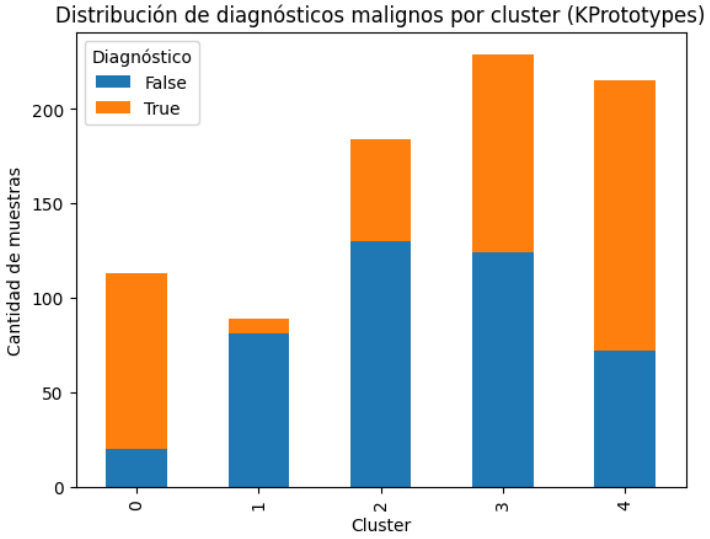


FIGURA 15. DISTRIBUCIÓN DE OBSERVACIONES MALIGNAS Y BENIGNA POR CLÚSTER PARA EL ALGORITMO K-PROTOTYPES.

3.1.4. K-Medoids

El algoritmo K-Medoids nos permitió utilizar matrices de distancia precomputadas, y haciendo uso de la matriz de Gower, que permite trabajar con variables categóricas y continuas a la vez, se implementó este modelo. Igual que en los anteriores modelos se calculó los índices de validación en un rango de clústeres (de 2 a 10) con distintos métodos de iniciación ("pam" y "alternate") (Figura 16).

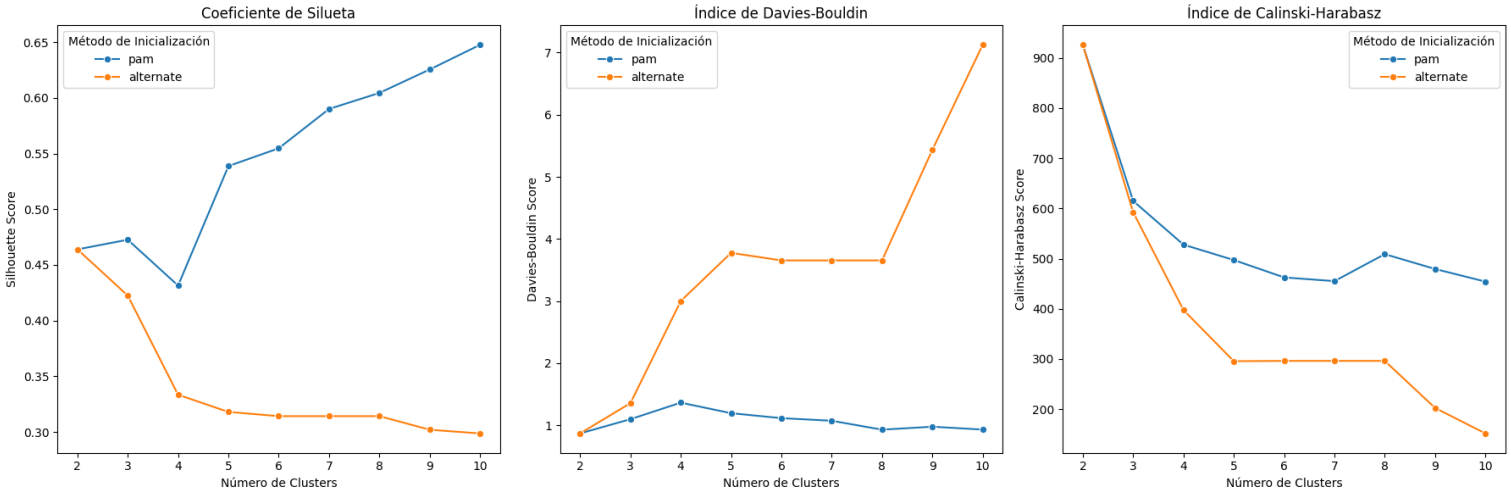


FIGURA 16. VALORES DE EVALUACIÓN PARA CADA VALOR DE K CON DISTINTOS TIPOS DE INICIACIÓN PARA EL ALGORITMO K-MEDOIDS HACIENDO USO DE LA MATRIZ DE GOWER.

Se estableció el número de clúster a 2, donde los valores de los índices mostraban una cohesión y segregación aceptables (Tabla 12), además de por la fácil interpretación visual de las características tumorales. Valores más altos en el número de clústeres diluían demasiado la segregación de las características y empeoraban los índices de evaluación. La distribución de las características en los clústeres (Figura 17), aunque muestran solapamiento en ambos grupos, se denota una tendencia del grupo 0 a agrupar características de formas irregulares y con márgenes micro lobulados, oscurecidos, mal definidos y espiculados, además de un contenido en grasa bajo, que comparte con el grupo 1. Estás características como se ha observado en otros modelos están relacionadas con lesiones tumorales malignas. El clúster 1, por el contrario, concentra características de forma regular y bien definidas o circunscritas comúnmente asociadas a lesiones tumorales benignas. También, en este algoritmo, vemos como se distribuyen los grupos según la edad, se observó que

la incidencia de lesiones relativas al clúster 0 va creciendo desde edades cercanas a 30 años hasta alcanzar el número máximo de observaciones sobre los 67 años (Figura 18). En edades inferiores a 27 años, solo está presente el clúster 1. Esta observación parece coherente ya que la edad juega un papel importante en el riesgo de padecer una afección maligna (Kroenke et al., 2004)[47], y el clúster 0 muestra características relacionas con características de índole maligno. Cabe destacar que el clúster 1 está presente casi todas las edades.

TABLA 12. VALORES DE LOS ÍNDICES DE EVALUACIÓN PARA K = 2 CLÚSTERES EN EL ALGORITMO K-MEDOIDS.

Índices de evaluación K-Medoids (k = 2)	
Coeficiente de Silueta	0,49
Davies-Bouldin	0,87
Calinski-Harabasz	936,71

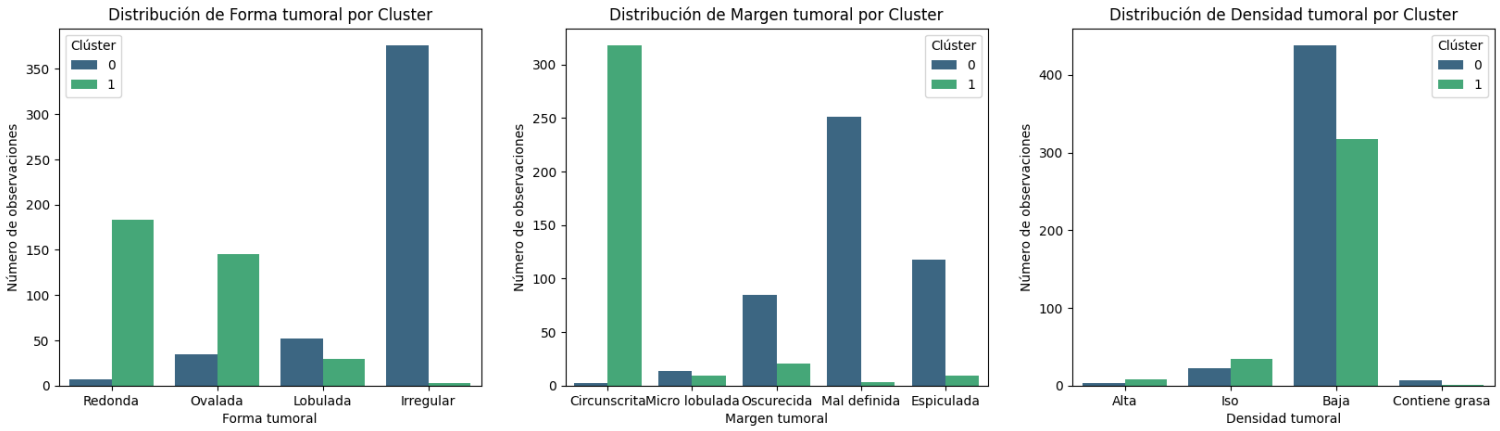


FIGURA 17. DISTRIBUCIÓN DE LAS CARACTERÍSTICAS TUMORALES: FORMA, MARGEN Y DENSIDAD POR CLÚSTER EN EL ALGORITMO K-MEDOIDS UTILIZANDO LA MATRIZ DE DISTANCIA GOWER.

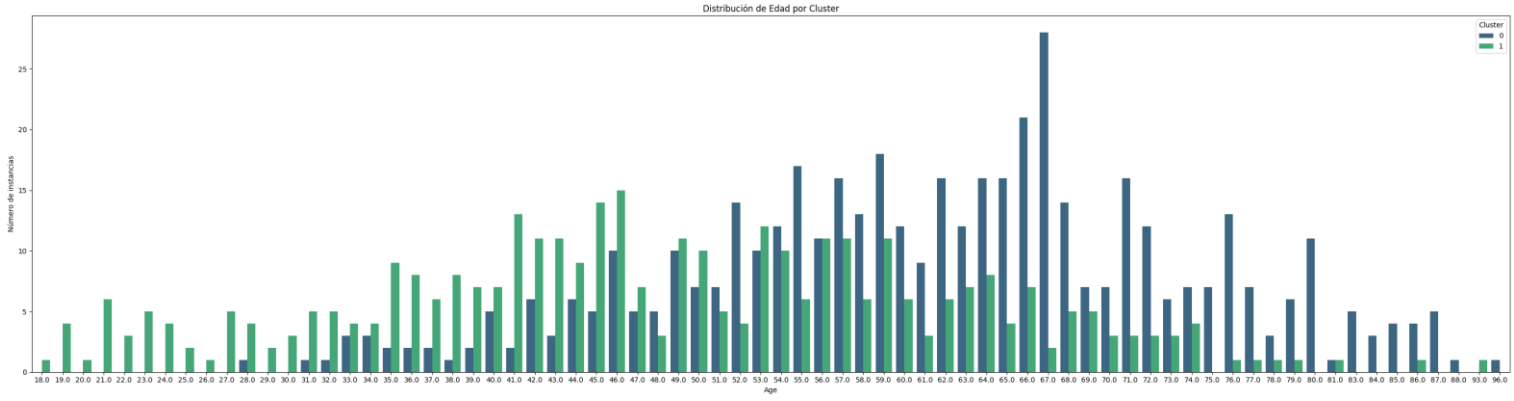


FIGURA 18. DISTRIBUCIÓN DE LA EDAD ("AGE") POR CLÚSTER PARA EL ALGORITMO K-MEDOIDS Y LA MATRIZ DE DISTANCIA GOWER.

Evaluando los clústeres obtenidos con los diagnósticos de las observaciones (Figura 19), obtuvimos que de las observaciones del clúster 0, un 75% fueron de carácter maligno. De las observaciones del clúster 1, el 14% fueron de carácter maligno (Tabla 13).

TABLA 13. NÚMERO DE OBSERVACIÓN SEGÚN EL DIAGNOSTICO, OBSERVACIONES TOTALES Y PORCENTAJE DE MALIGNIDAD PARA EL ALGORITMO K-MEDOIDS PARA K = 2.

Clúster	Observaciones			
	Benigna	Maligna	Totales	Porcentaje (%) obs. Malignas
0	119	351	470	75
1	308	52	360	14

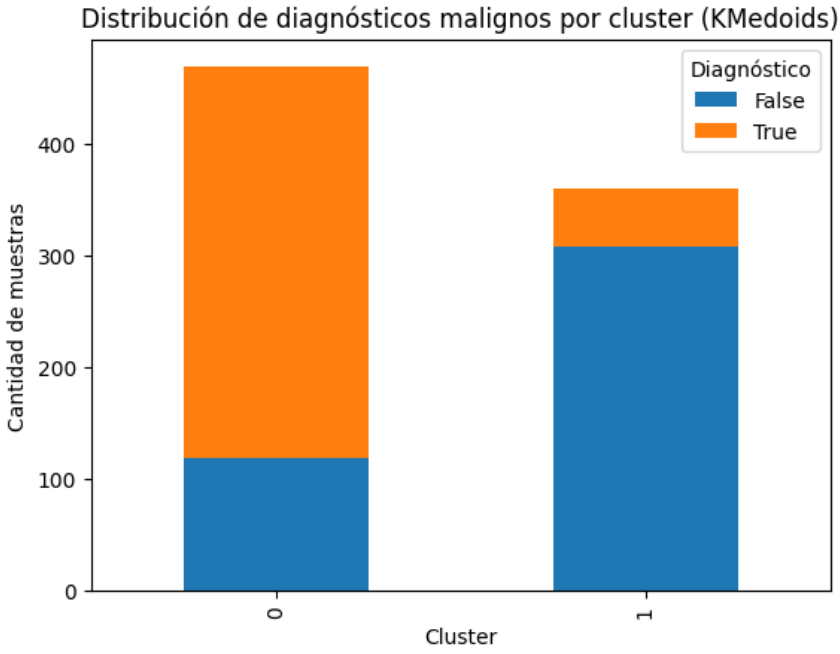


FIGURA 19. DISTRIBUCIÓN DE OBSERVACIONES MALIGNAS Y BENIGNA POR CLÚSTER PARA EL ALGORITMO K-MEDOIDS.

3.1.5. Hierarchical clustering (Agrupación Jerárquica)

El último algoritmo implementado es la agrupación jerárquica aglomerativo haciendo uso de la matriz de distancias Gower. Con la matriz de distancia precomputada se implementó el algoritmo jerárquico con el método "complete" estableciendo la distancia máxima (Figura 20), es decir, la distancia entre dos clústeres se define como la distancia máxima entre cualquier par de puntos, uno de cada clúster.

Se estableció el umbral de corte a una distancia de 9 en el dendrograma, lo que proporciono un número sugerido de 4 clústeres. Estableciendo el número de clúster a un valor inferior o superior empeoraba los incides de evaluación, con lo que establecimos el número óptimo de 4 clústeres para este modelo (Tabla 14).

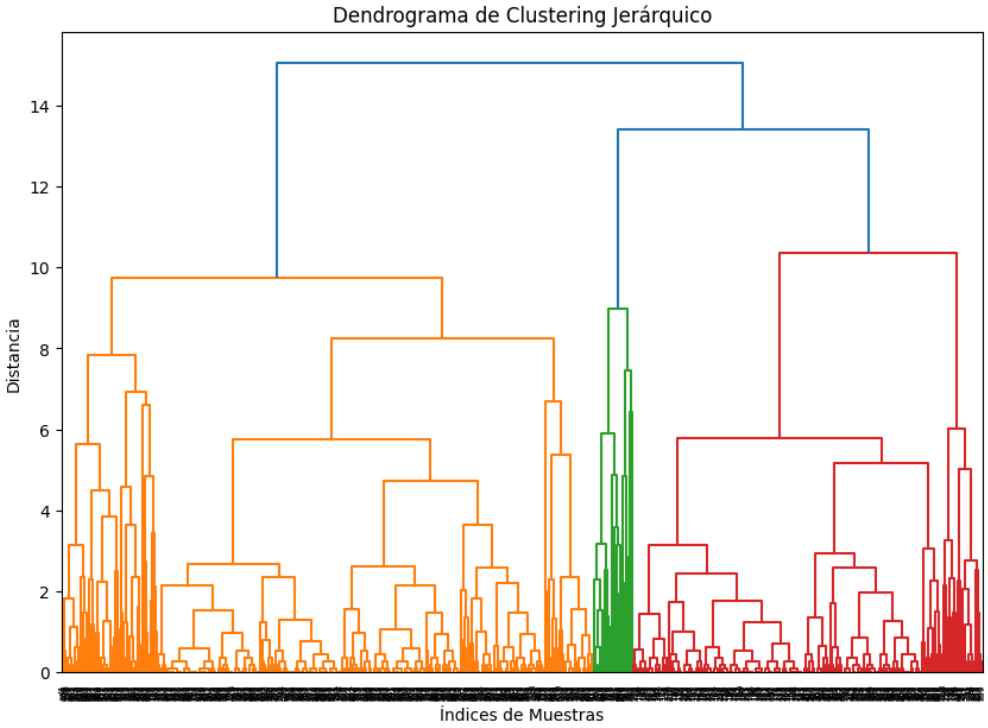


FIGURA 20. DENDROGRAMA DE LA AGRUPACIÓN JERÁRQUICA CON EL MÉTODO "COMPLETE" Y LA MATRIZ DE GOWER.

TABLA 14. VALORES DE LOS ÍNDICES DE EVALUACIÓN PARA 4 CLÚSTERES EN EL ALGORITMO DE AGRUPACIÓN JERÁRQUICO.

Índices de evaluación Agrupamiento Jerárquico (clústeres = 4)	
Coeficiente de Silueta	0,46
Davies-Bouldin	0,91
Calinski-Harabasz	486,35

En la distribución de las características tumorales para estos clústeres (Figura 21) se observó que el clúster 1 esta presenta en menor o mayor medida en todas las observaciones de las características de forma y margen tumoral, aunque están más representado por observaciones con formas irregulares y márgenes mal definidos, de cara a la densidad, se concentra en la característica de densidad baja. El clúster 2 tiene una representación muy baja, es decir, no contiene muchas observaciones (Tabla 15), y se concentra en la forma irregular y márgenes mal definidas, así como con la densidad iso. El clúster 3, esta

caracterizado por contener observaciones con características definidas, formas redondeadas (redondas, ovaladas y lobuladas) y densidades bajas. Por último, el clúster 4, también con pocas observaciones, está representado por características definidas y esféricas, además de densidades altas e iso.

También se observó la distribución de los clústeres por edad (Figura 22), donde los clústeres 1 y 3 son los más representados, mientras que los clústeres 2 y 4 están menos representados y acotadas a edades comprendidas entre los 30 y 80 años. El clúster 1 empieza a tener representación a los 28 años hasta los 98 años, y el clúster 3 se concentra a edades más tempranas.

Estas observaciones casan con los resultados obtenidos en otros modelos, y siguen la misma justificación, la incidencia de tumores malignos se asocia a características en la forma irregular y márgenes poco definidos, además de aumentar con la edad.

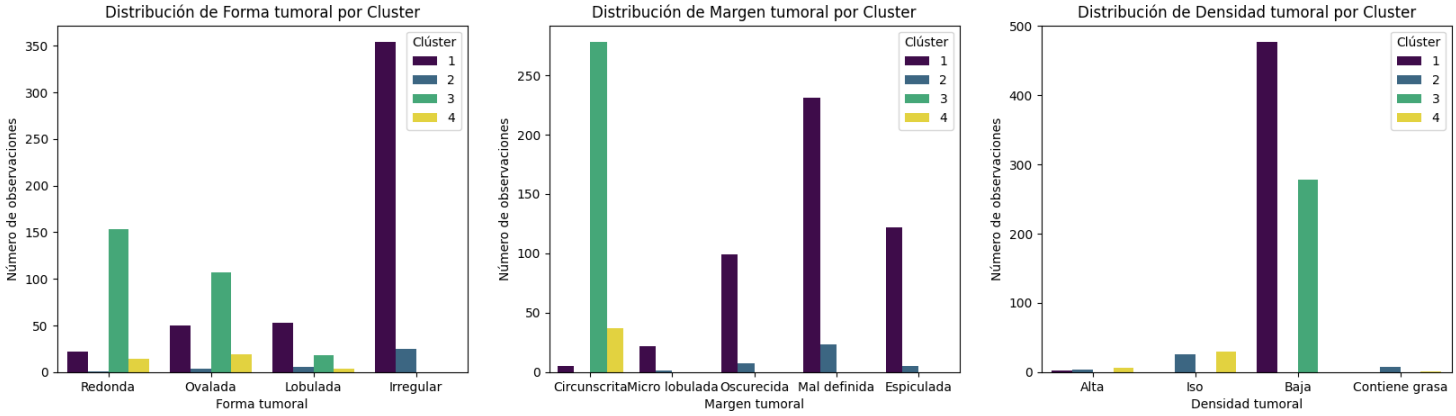


FIGURA 21. DISTRIBUCIÓN DE LAS CARACTERÍSTICAS TUMORALES: FORMA, MARGEN Y DENSIDAD POR CLÚSTER (AGRUPACIÓN JERÁRQUICA).

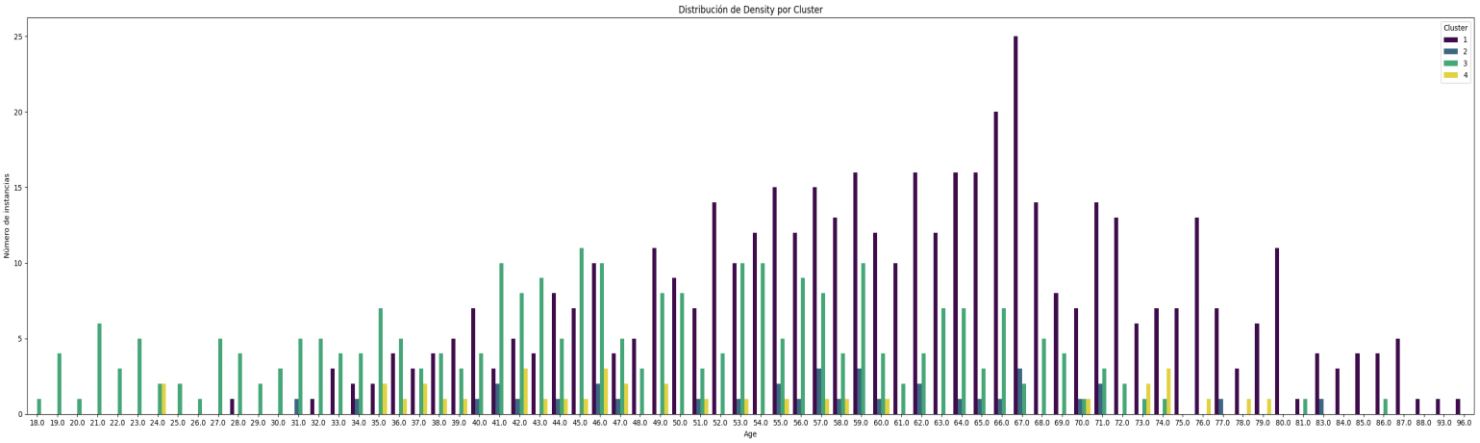


FIGURA 22. DISTRIBUCIÓN DE LA EDAD ("AGE") POR CLÚSTER (AGRUPACIÓN JERÁRQUICA).

La evaluación de la calidad de los clústeres se realizó, nuevamente, con la variable diagnosis de las observaciones (Figura 23).

TABLA 15. NÚMERO DE OBSERVACIÓN SEGÚN EL DIAGNOSTICO, OBSERVACIONES TOTALES Y PORCENTAJE DE MALIGNIDAD.

Clúster	Observaciones			
	Benigna	Maligna	Totales	Porcentaje (%) obs. Malignas
1	135	344		72
2	13	23		63
3	245	33		12
4	34	3		8

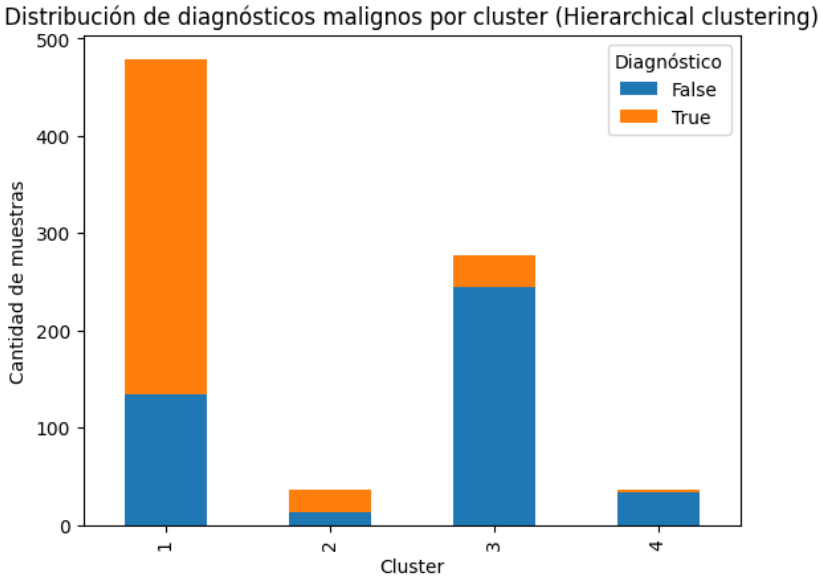


FIGURA 23. DISTRIBUCIÓN DE OBSERVACIONES MALIGNAS Y BENIGNA POR CLÚSTER .

La evaluación mostró dos grupos con una incidencia de casos malignos menor al 15%, clústeres 3 y 4, mientras que los clústeres 1 y 2 mostraron una incidencia mayor al 50%.

3.1.6. Página web interactiva

En este apartado se presenta el uso, las características principales y el diseño de la página web interactiva.

La página web interactiva realiza agrupamientos de características tumorales basadas en la forma, el margen y la densidad tumoral siguiendo la nomenclatura Bi-Rads. Para hacer este agrupamiento se eligió el algoritmo K-Means ya que dio los resultados óptimos e interpretables. Para conseguir esto se creó una "app.py" que contiene la aplicación en Flask, esta aplicación despliega los modelos precargados K-Means, UMAP y One-Hot enconding en formato "pickle", que mediante la introducción de nuevas características a través de un formulario podemos agrupar esta nueva observación a su correspondiente grupo y a su riesgo de malignidad.

De primeras, la web muestra una explicación y varios botones, estos botones redirigen al conjunto de datos utilizado en este proyecto, a los resultados del modelo desplegado en la aplicación y al repositorio del proyecto en GitHub (Figura 24). Mas abajo encontramos una tabla de equivalencias Bi-rads para las características tumorales. Y, por último, encontramos el formulario con el desplegable de características para empezar el proceso de agrupamiento para la nueva observación (Figura 25).

FIGURA 24. VISTA PRINCIPAL DE LA PÁGINA WEB INTERACTIVA.

Estableciendo las características deseadas en el formulario y presionando el botón "Procesar características" empieza el proceso de agrupación, dando resultado en una nueva ventana. Este resultado consta del clúster en el que se asigna la nueva observación y el riesgo de malignidad que tiene este grupo (Figura 26).

FIGURA 25. FORMULARIO DE LA PÁGINA WEB INTERACTIVA PARA INTRODUCIR NUEVAS OBSERVACIONES.

FIGURA 26. RESULTADO DEL PROCESO DE ASIGNACIÓN DE LA PÁGINA WEB INTERACTIVA.

3.1.7. Repositorio público

Este apartado está reservado para los enlaces a los repositorios públicos del proyecto.

- Repositorio del código y memoria del proyecto: https://github.com/DemetrioMunoz/TFM_Code

- Repositorio del código de la página web interactiva: https://github.com/DemetrioMunoz/TFM_Web

Capítulo 4

4. Conclusión

En este proyecto se exploraron varias técnicas de agrupamiento (“clustering”) para segmentar de una manera significativa datos derivados de imágenes mamográficas siguiendo la nomenclatura Bi-rads. Cada modelo se evaluó con una serie de índices y se interpretó de forma visual los resultados, además se aprovechó los diagnósticos de las observaciones para validar la calidad de los clústeres obtenidos.

De entre los modelos, el modelo K-Means fue el seleccionado como método más adecuado para manejar nuestros datos, ya que conseguimos segregar un grupo con solo características benignas. Además, la decisión de elegir este algoritmo se basó en que el rendimiento fue mejor comparado con los demás; es un algoritmo simple de implementar y nos permite introducir nuevas observaciones sin aplicar muchas transformaciones en los datos.

Con estas características se implementó en una web interactiva como base de la asignación de nuevas características. Esta aplicación tiene el objetivo de validar el modelo en un entorno practico proporcionando una interfaz de usuario sencilla de entender.

La implementación del algoritmo K-Means para la segregación de datos de mamografías tiene implicaciones de cara al futuro y si se siguen una serie de futuras direcciones se puede crear un entorno en el que no se precise de la atención de un especialista para interpretar el riesgo de malignidad en el cáncer de mama. Haciendo este proceso automático, se puede conseguir un diagnóstico más temprano y crear terapias personalizadas más precisas.

Para futuros trabajos, se debería considerar la implementación de más técnicas de agrupamiento e incluso explorar la implementación de algoritmos de aprendizaje automático supervisado en combinación con clustering no supervisado.

En resumen, la elección del modelo K-means ha demostrado que puede segregar de forma efectiva y práctica características derivadas de mamografías. El uso de la web interactiva con este modelo tiene el potencial de ser aplicado al mundo real, y se espera contribuir a una mejor detección y análisis de patrones en datos derivados de mamografías.

Capítulo 5

5. Glosario

Algoritmo: Conjunto de instrucciones sistemáticas y previamente definidas que se utilizan para realizar una determinada tarea. Estas instrucciones están ordenadas y acotadas a manera de pasos a seguir para alcanzar un objetivo.

Aprendizaje automático o machine learning (ML): Es el proceso mediante el cual se usan modelos matemáticos de datos para ayudar a un equipo a aprender sin instrucciones directas. Se considera un subconjunto de la inteligencia artificial (IA). El aprendizaje automático usa algoritmos para identificar patrones en los datos, y esos patrones luego se usan para crear un modelo de datos que puede hacer predicciones.

Bi-Rads: Sistema de informes que se usa para describir de manera estandarizada los resultados de las mamografías, ecografías mamarias o imágenes por resonancia magnética de las mamas. El sistema BI-RADS clasifica los resultados de las pruebas según 1 de 7 categorías, que van desde un resultado normal o benigno (no canceroso) hasta altamente sospechoso o maligno (cáncer).

Biopsia: Extracción de células o tejidos para que un patólogo los examine al microscopio o realice otras pruebas.

Clúster (grupo): División llevada a cabo por un algoritmo de agrupamiento.

“Clustering” (Agrupamiento): El “clustering” consiste en agrupar ítems en grupos con características similares que nos permite descubrir patrones y estructuras ocultas en conjuntos de datos.

Inteligencia artificial (IA): Conjunto de tecnologías que permiten que las computadoras realicen una variedad de funciones avanzadas, incluida la capacidad de ver, comprender y traducir lenguaje hablado y escrito, analizar datos, hacer recomendaciones y mucho más.

Mamografía: Tipo de examen médico de imagen que utiliza rayos X de baja dosis para examinar las mamas. Este procedimiento es utilizado principalmente para la detección temprana del cáncer de mama, así como para evaluar otras condiciones anormales en el tejido mamario.

Objetivo de Desarrollo Sostenible (ODS): Meta establecida por las Naciones Unidas (ONU) en el marco de la Agenda 2030 para el Desarrollo Sostenible. Estos objetivos

fueron adoptados en 2015 por todos los Estados Miembros de la ONU como un llamado universal para erradicar la pobreza, proteger el planeta y asegurar que todas las personas disfruten de paz y prosperidad para 2030.

Bibliografía

1. Chhikara, B. S., & Parang, K. (2023). Global Cancer Statistics 2022: the trends projection analysis. *Chemical Biology LETTERS*, 10(1), 451.
2. Wang, L. (2017). Early Diagnosis of Breast Cancer. *Sensors*, 17(7), 1572.
3. Boyd, N. F., Lockwood, G. A., Byng, J. W., Tritchler, D. L., & Yaffe, M. J. (1998). Mammographic densities and breast cancer risk. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 7(12), 1133-1144.
4. Boyd, N. F., Martin, L. J., Bronskill, M., Yaffe, M. J., Duric, N., & Minkin, S. (2010). Breast tissue composition and susceptibility to breast cancer. *JNCI: Journal of the National Cancer Institute*, 102(16), 1224-1237.
5. American College of Radiology. (2003). Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). Reston, VA: American College of Radiology.
6. Aibar, L., Santalla, A., López-Criado, M. S., González–Pérez, I., Calderón, M. A., Gallo, J. L., & Fernández-Parra, J. (2011). Clasificación radiológica y manejo de las lesiones mamarias. Clínica e Investigación en Ginecología y Obstetricia, 38(4), 141-149.
7. Topol, E. (2019). Deep medicine: how artificial intelligence can make healthcare human again. *Hachette UK*.
8. Haq, A., Li, J., Saboor, A., Khan, J., Wali, S., Ahmad, S., . . . Zhou, W. (2021). Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques. *IEEE Access*, 9, 22090-22105.
9. Roy, S., Meena, T., & Lim, S. J. (2022). Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine. *Diagnostics*, 12(10), 2549.
10. Calamuneri, A., Donato, L., Scimone, C., Costa, A., D'Angelo, R., & Sidoti, A. (2017). On machine learning in biomedicine. *Life Safety and Security*, 5(12), 96-99.

11. O'Connor, J. P., Rose, C. J., Waterton, J. C., Carano, R. A., Parker, G. J., & Jackson, A. (2015). Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clinical Cancer Research*, 21(2), 249-257.

12. Naciones Unidas. (14 de Marzo de 2023). Obtenido de Naciones Unidas. (2023, 14 de marzo). Objetivos de Desarrollo Sostenible. Recuperado de https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

13. Elter, M., Schulz-Wendtland, R., & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical physics*, 34(11), 4164-4172.

14. Patel, A. A. (2019). Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data.

15. Jovel, J., & Greiner, R. (2021). An introduction to machine learning approaches for biomedical research. *Frontiers in Medicine*, 8, 771607.

16. Thanoon, M. A., Zulkifley, M. A., Mohd Zainuri, M. A. A., & Abdani, S. R. (2023). A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. *Diagnostics*, 13(16), 2617.

17. Jalloul, R., Chethan, H. K., & Alkhatib, R. (2023). A review of machine learning techniques for the classification and detection of breast cancer from medical images. *Diagnostics*, 13(14), 2460.

18. Nasser, M., & Yusof, U. K. (2023). Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. *Diagnostics*, 13(1), 161.

19. Spak, D. A., Plaxco, J. S., Santiago, L., Dryden, M. J., & Dogan, B. E. (2017). BI-RADS® fifth edition: A summary of changes. *Diagnostic and interventional imaging*, 98(3), 179-190.

20. Boyd, N. F., Guo, H., Martin, L. J., et al. (2007). Mammographic Density and the Risk and Detection of Breast Cancer. *The New England Journal of Medicine*, 356(3), 227-236.

21. Edwards, B. K., Noone, A. M., Mariotto, A. B., et al. (2014). Annual Report to the Nation on the Status of Cancer, 1975–2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State. *Journal of the National Cancer Institute*.

22. Microsoft. (n.d.). Visual Studio Code. Retrieved from https://code.visualstudio.com/
23. Jupyter Project. (n.d.). Jupyter Notebook. Retrieved from https://jupyter.org/
24. The pandas development team. (2020). pandas-dev/pandas: Pandas. *Zenodo*. https://doi.org/10.5281/zenodo.3509134
25. Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021
26. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55
27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html
28. De Vos, P. (n.d.). kmodes: k-modes and k-prototypes clustering algorithms for categorical data. *GitHub*. https://github.com/nicodv/kmodes
29. Lemaître, G., Nogueira, F., & Aridas, C. K. (2020). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18, 1-5. https://github.com/scikit-learn-contrib/scikit-learn-extra
30. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2
31. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426. https://umap-learn.readthedocs.io/en/latest/
32. Kumar, V., Bass, G., Tomlin, C., & Dulny, J. (2018). Quantum annealing for combinatorial clustering. *Quantum Information Processing*, 17, 1-14.
33. Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.
34. Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. *ACM Sigmod record*, 27(2), 73-84.

35. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.

36. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, 8, 80716-80727.

37. Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78, 507-512.

38. Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304.

39. Huang, Z. (1997, February). Clustering large data sets with mixed numeric and categorical values. In Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining, (*PAKDD*) (pp. 21-34).

40. Nielsen, F. (2016). Introduction to HPC with MPI for Data Science. *Springer*.

41. Halkidi, Maria; Batistakis, Yannis; Vazirgiannis, Michalis (2001). "On Clustering Validation Techniques" *Journal of Intelligent Information Systems*, 17(2-3), 107-145.

42. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics* 20: 53-65.

43. Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure" *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-1 (2): 224-227.

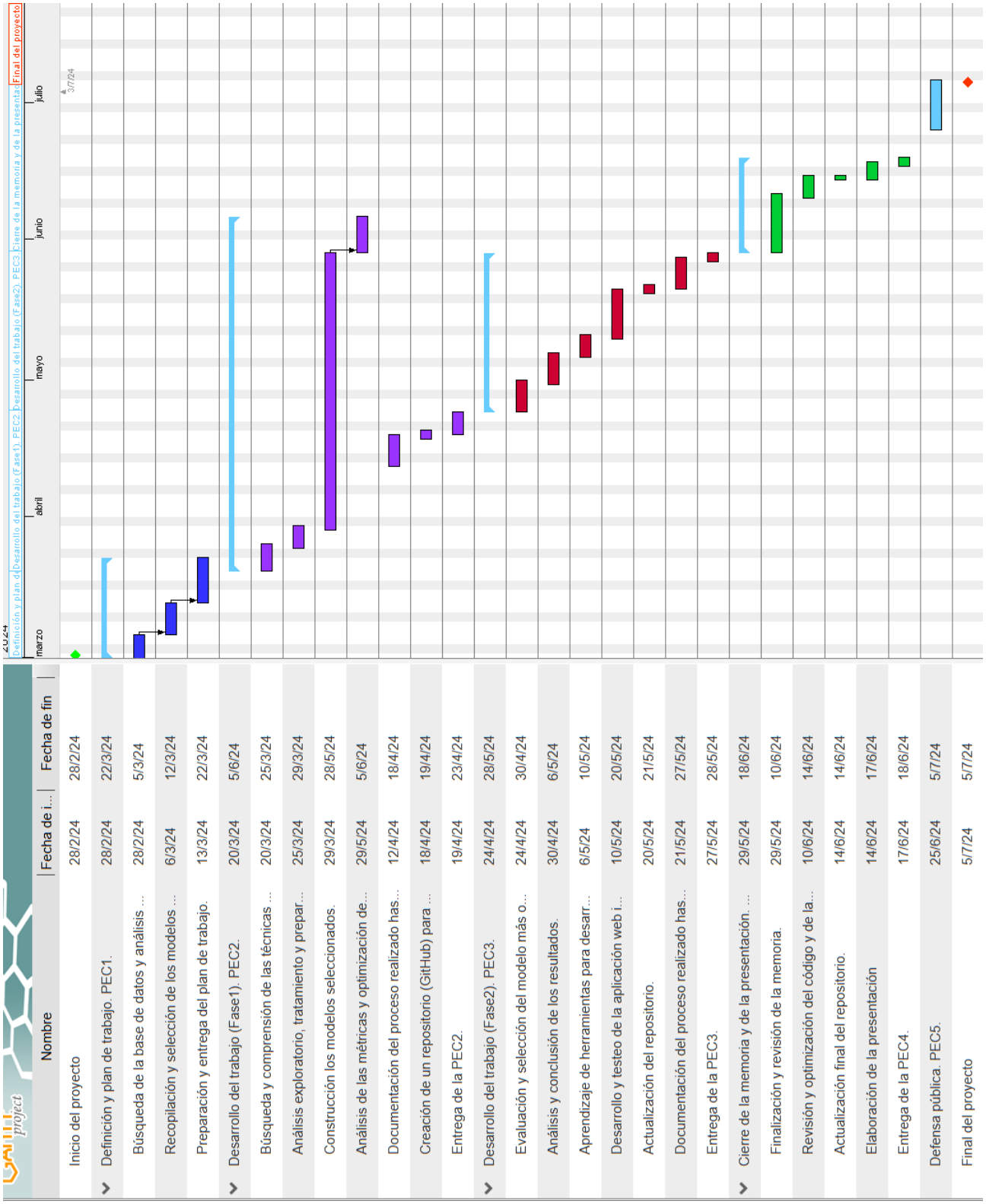
44. Caliński, T., & Harabasz, J. (1974). "A Dendrite Method for Cluster Analysis". *Communications in Statistics-theory and Methods* 3: 1-27.

45. Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python (2nd ed.). *O'Reilly Media*.

46. National Cancer Institute. (s.f.). Breast Cancer. Recuperado de https://www.cancer.gov/

47. Kroenke, C. H., Rosner, B., Chen, W. Y., Kawachi, I., Colditz, G. A., & Holmes, M. D. (2004). Functional impact of breast cancer by age at diagnosis. *Journal of Clinical Oncology*, 22(10), 1849-1856.

Anexo I


Nombre
Fecha de i...
Fecha de fin
Inicio del proyecto 28/2/24 28/2/24
Definición y plan de trabajo. PEC1. 28/2/24 22/3/24
Búsqueda de la base de datos y análisis ... 28/2/24 5/3/24
Recopilación y selección de los modelos ... 6/3/24 12/3/24
Preparación y entrega del plan de trabajo. 13/3/24 22/3/24
Desarrollo del trabajo (Fase1). PEC2. 20/3/24 5/6/24
Búsqueda y comprensión de las técnicas ... 20/3/24 25/3/24
Análisis exploratorio, tratamiento y prepar... 25/3/24 29/3/24
Construcción los modelos seleccionados. 29/3/24 28/5/24
Análisis de las métricas y optimización de... 29/5/24 5/6/24
Documentación del proceso realizado has... 12/4/24 18/4/24
Creación de un repositorio (GitHub) para ... 18/4/24 19/4/24
Entrega de la PEC2. 19/4/24 23/4/24
Desarrollo del trabajo (Fase2). PEC3. 24/4/24 28/5/24
Evaluación y selección del modelo más o... 24/4/24 30/4/24
Análisis y conclusión de los resultados. 30/4/24 6/5/24
Aprendizaje de herramientas para desarr... 6/5/24 10/5/24
Desarrollo y testeo de la aplicación web i... 10/5/24 20/5/24
Actualización del repositorio. 20/5/24 21/5/24
Documentación del proceso realizado has... 21/5/24 27/5/24
Entrega de la PEC3. 27/5/24 28/5/24
Cierre de la memoria y de la presentación. ... 29/5/24 18/6/24
Finalización y revisión de la memoria. 29/5/24 10/6/24
Revisión y optimización del código y de la... 10/6/24 14/6/24
Actualización final del repositorio. 14/6/24 14/6/24
Elaboración de la presentación 14/6/24 17/6/24
Entrega de la PEC4. 17/6/24 18/6/24
Defensa pública. PEC5. 25/6/24 5/7/24
Final del proyecto 5/7/24 5/7/24
2024
marzo
abril
mayo
junio
julio
3/7/24
Definición y plan de
Desarrollo del trabajo (Fase1). PEC2.
Desarrollo del trabajo (Fase2). PEC3.
Cierre de la memoria y de la presenta
Final del proyecto