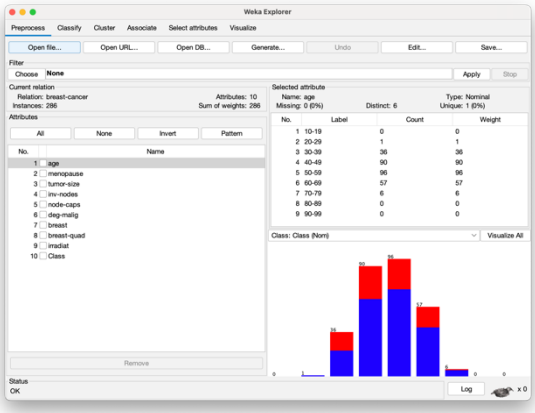
APRENDIZAJE NO SUPERVIDADO

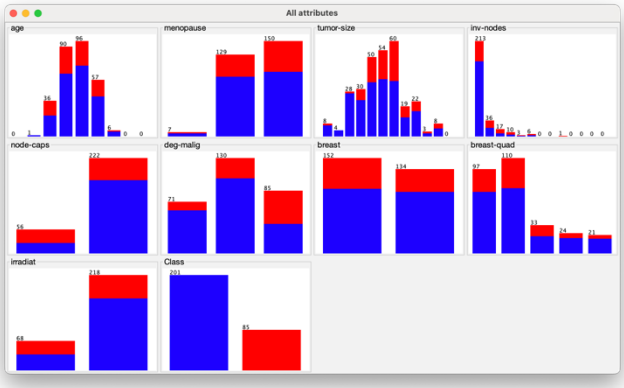
MINERIA DE DATOS 2024

MARTA DE CASTRO LEIRA

1.SELECCIÓN DEL DATASET

- **Dataset**: Breast Cancer (provisto por defecto en Weka)
- **Número de instancias**: 286
- **Número de atributos:** 10
- **Descripción:** Este conjunto de datos contiene información clínica y biológica de tumores de mama, con el objetivo de clasificar los tumores como malignos o benignos
- **Objetivo del análisis:** Predecir la clase (recurrence-events vs no-recurrence-events) aplicando técnicas de preprocesamiento y algoritmos de clasificación



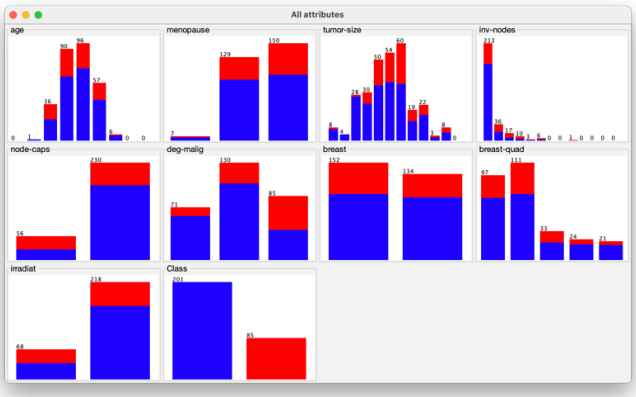


2. TÉCNICAS DE PRPROCESAMIENTO

Dado que la calidad de los datos impacta al rendimiento del modelo, se implementaron las siguientes técnicas de preprocesamiento:

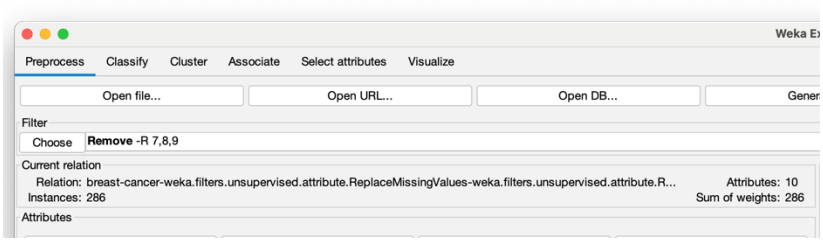
2.1 Manejo de Valores Faltantes

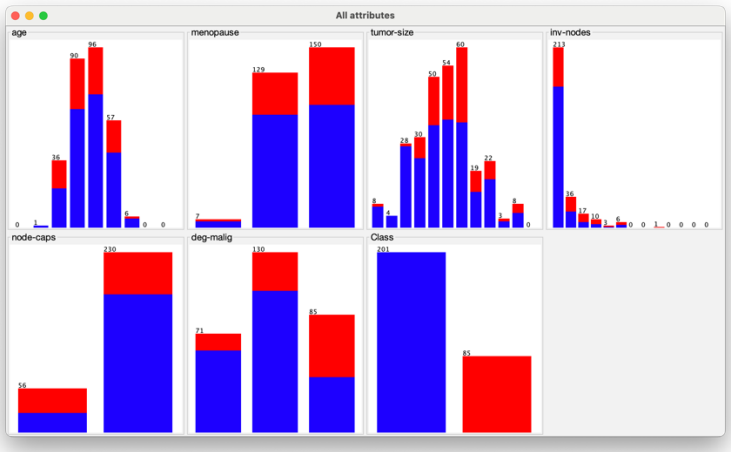
Para abordar los valores faltantes en el conjunto de datos, se utilizó el filtro ***ReplaceMissingValues***, que imputa los datos según el tipo de atributo. En los atributos numéricos, como edad y tamaño del tumor, se reemplazaron los valores faltantes con la media aritmética, lo que permite mantener la distribución general sin introducir sesgos significativos. En los atributos categóricos, como menopausia y node-caps, se utilizó la moda (la categoría más frecuente), asegurando que los datos imputados sean representativos de las características predominantes.



2.2 Eliminación de Atributos

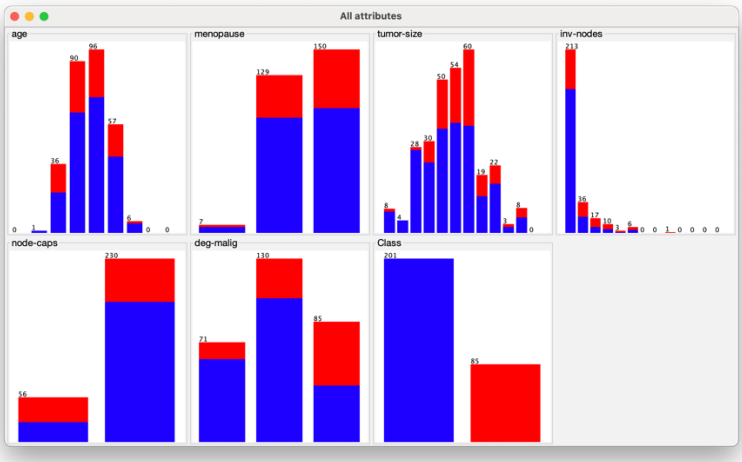
Se aplicó el filtro ***Remove*** para eliminar atributos considerados irrelevantes o redundantes en el contexto del análisis. Entre ellos, el atributo Breast fue descartado, ya que no tiene impacto en la recurrencia del tumor. Asimismo, Breast-quad se excluyó debido a su información redundante, y Irradiat fue eliminado para enfocar el análisis exclusivamente en las características clínicas y biológicas más relevantes.
Esta eliminación busca reducir el ruido en los datos, mejorando la claridad del conjunto y evitando que atributos irrelevantes interfieran en la eficacia del modelo. Además, al trabajar con un menor número de atributos, se acelera el tiempo de entrenamiento y se optimiza el rendimiento general de los algoritmos de clasificación.





2.3 Normalización de Datos

Dado que las variables numéricas (tumor-size, inv-nodes, etc.) presentan escalas diferentes, se aplicó una **normalización min-max** para llevar todas las características a un rango común (0-1). Esto garantiza que ninguna variable domine el proceso de clustering debido a su magnitud.
La normalización es crucial en algoritmos como **K-means**, donde la distancia euclidiana entre puntos es fundamental. Este preprocesamiento asegura que cada atributo contribuya de manera equitativa, mejorando la precisión del modelo en la identificación de clústeres.

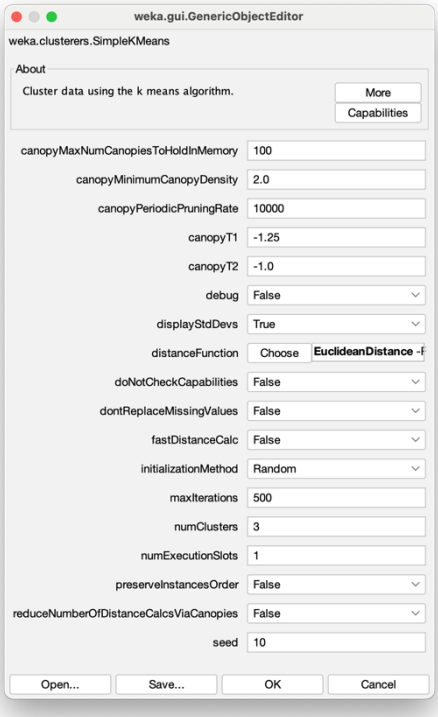


2.3 Discretización de Datos

Para facilitar la extracción de reglas de asociación, se aplicó un preprocesamiento de discretización a las variables. Este proceso convirtió los valores continuos en intervalos categóricos, permitiendo que el algoritmo Apriori pueda identificar asociaciones significativas entre categorías discretas.

La discretización resulta esencial para las reglas de asociación, ya que estos algoritmos trabajan más eficientemente con variables categóricas. Sin embargo, este paso se utilizó exclusivamente para las tareas de minería de reglas de asociación y no fue aplicado en la etapa de clustering. En el análisis de clustering, se mantuvieron las variables numéricas en su forma original para conservar la precisión en el cálculo de distancias y evitar introducir sesgos debido a la transformación de datos.

3. ALGORITMO DE CLUSTERING: SIMPLEKMEANS



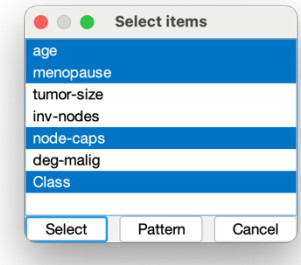
Se implementó el algoritmo **K-means** para realizar el agrupamiento del conjunto de datos, dividiendo las instancias en grupos según sus similitudes. Este método minimiza la variabilidad dentro de cada clúster y maximiza la separación entre ellos. Para este análisis, se configuraron los siguientes parámetros:

- **Número de clusters (numClusters)**: Se probaron configuraciones con **2, 3, 4 y 5 clusters**. Este parámetro determina cuántos grupos formará el algoritmo. Probar distintas configuraciones permitió analizar cuál era el número óptimo, buscando el equilibrio entre simplicidad y una segmentación más detallada.
- **Método de inicialización (initializationMethod)**: Se utilizó el

método **Random**, que selecciona de manera aleatoria los centroides iniciales dentro del rango de los datos.

- **Función de distancia (distanceFunction)**: Se empleó la **distancia euclidiana** como métrica para calcular la similitud entre las instancias y los centroides. Es adecuada para datos normalizados.
- **Visualización de desviaciones estándar (displayStdDevs)**: Se activó esta opción, lo que permitió observar la **dispersión** de los datos dentro de cada clúster.

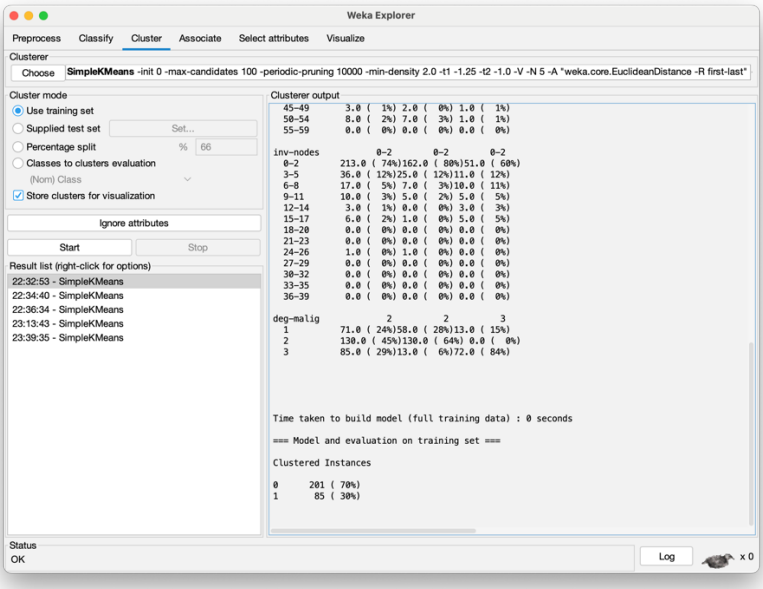
En conjunto, estos parámetros configuraron el modelo de clustering, permitiendo segmentar el dataset en grupos significativos y proporcionando insights útiles para el análisis. Además se ignoraron los atributos **age**, **menopause**, **node-caps** y **class** porque se consideraron menos relevantes para el clustering, ya sea por su baja variabilidad o por aportar información redundante que podría introducir ruido.



Comparación de Resultados

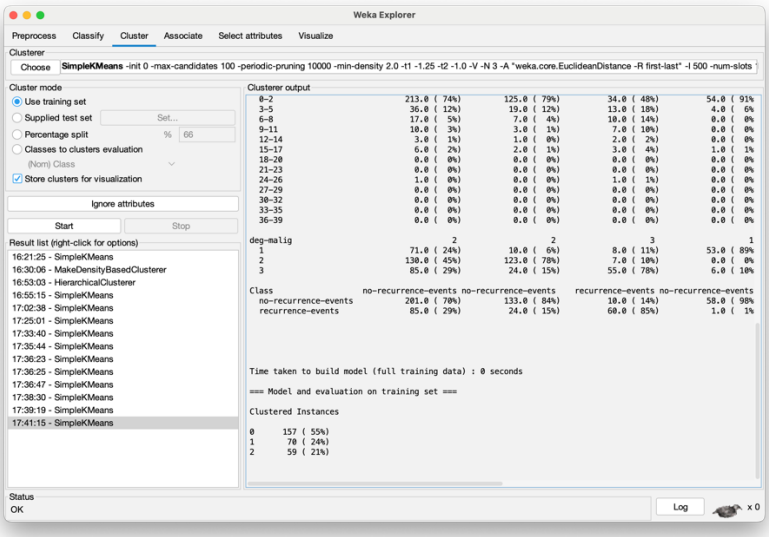
Se evaluaron los resultados del clustering en función de los diferentes números de clusters generados:

El clustering con 2 clústeres separa claramente los tumores benignos y malignos, lo que refleja una clasificación binaria sencilla. Cluster 0 (70%) agrupa tumores menos malignos, mientras que Cluster 1 (30%) contiene los tumores más agresivos. Los centroides muestran diferencias en el tamaño de los tumores y el grado de malignidad. En general, esta segmentación es adecuada para un análisis básico, pero carece de precisión en patrones más complejos.

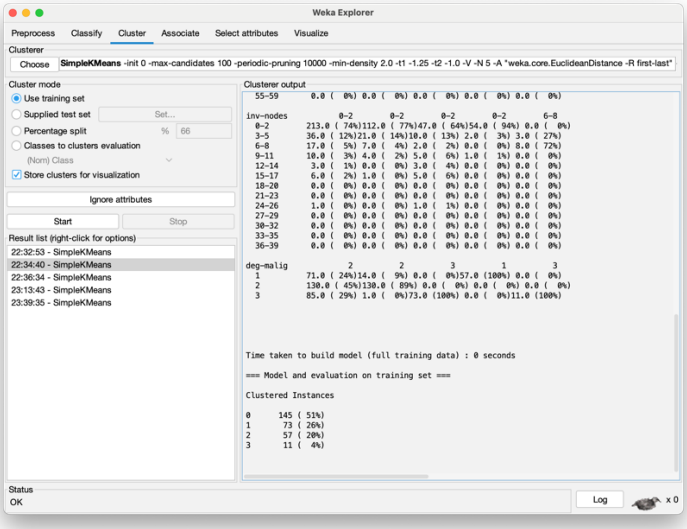


El clustering con 3 clústeres introduce una diferenciación más detallada dentro de los tumores benignos. El **Cluster 0** (55%) incluye tumores benignos con tamaños pequeños y pocos nodos afectados, con una alta proporción de casos sin recurrencias. El **Cluster 1** (24%) agrupa tumores malignos con mayor agresividad, especialmente con recurrencias, y el **Cluster 2** (21%) destaca tumores de tamaño medio, principalmente benignos y sin recurrencias. Este enfoque aporta un buen

balance entre simplicidad y detalle, pero no mejora significativamente la predicción de recurrencias en comparación con dos clústeres.

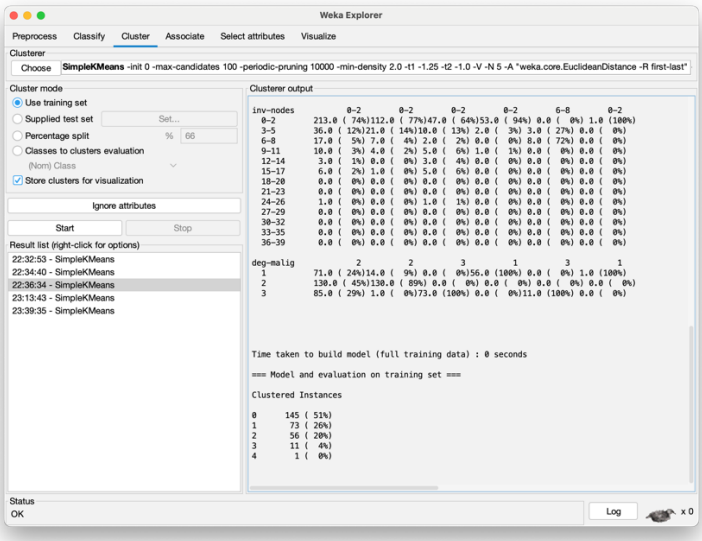


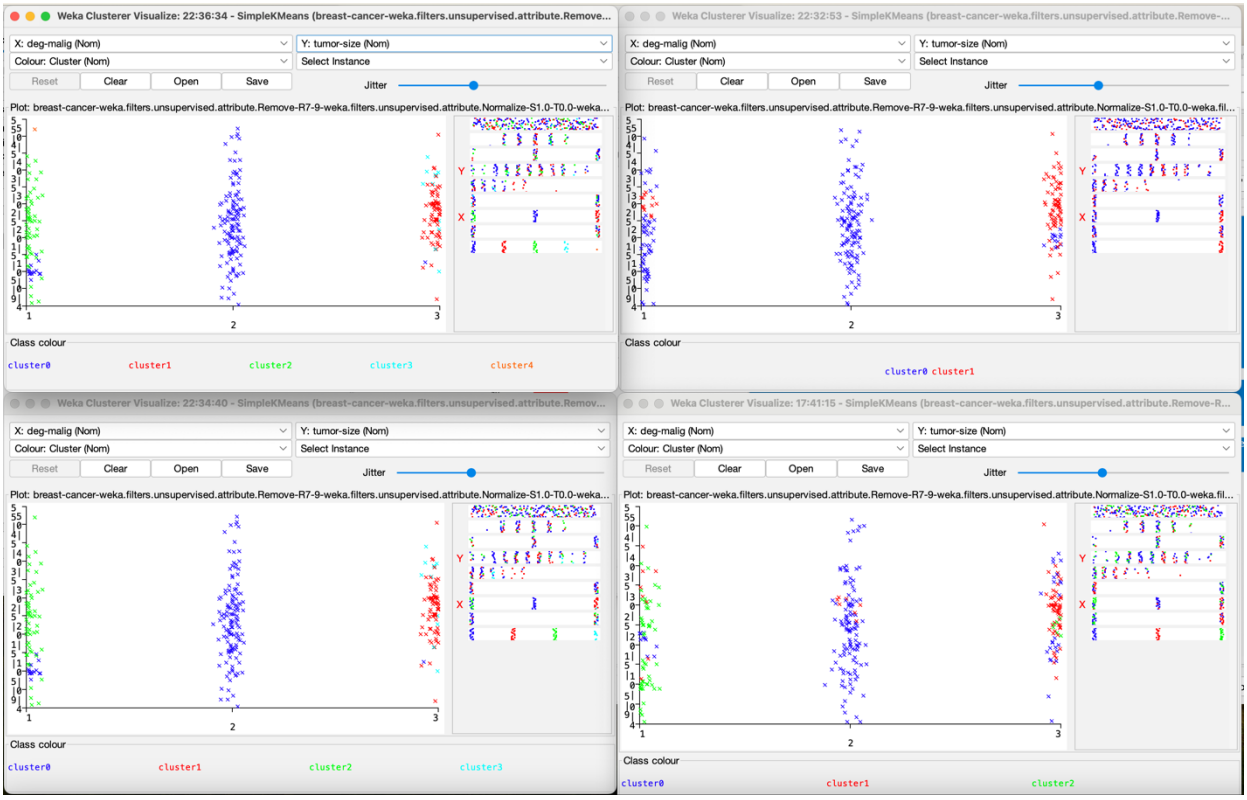
El modelo de 4 clústeres proporciona una segmentación más detallada, diferenciando subgrupos dentro de los tumores malignos, lo que permite una clasificación más específica. Los tumores malignos se agrupan en diferentes clústeres según el tamaño del tumor, los nodos afectados y el grado de malignidad, con un grupo adicional para los casos más graves. Esta segmentación añade valor clínico al permitir una análisis más preciso de las características de cada subgrupo. Sin embargo, la interpretación de estos grupos es más compleja debido a la mayor granularidad, lo que podría dificultar la identificación de patrones claros. En general, aporta más detalles pero no necesariamente mejora la predicción de recurrencias.



Finalmente con 5 clústeres, se observa una mayor segmentación de los datos, pero sin mejoras significativas respecto a los 4 clústeres. El modelo genera clústeres adicionales, pero algunos son poco relevantes, como el clúster 4, que contiene solo 1 instancia lo que sugiere un riesgo de sobresegmentación. La distribución de las instancias muestra que la mayoría se agrupan en los primeros clústeres (0, 1 y 2),

mientras que los clústeres 3 y 4 tienen pocos elementos. En general, no hay una mejora clara en la calidad de la segmentación.





Comparación General de los Clústeres:

Número de Clústeres	Simplicidad	Separación	Interpretación	Error Intra-clúster (WSS)	Distribución de Instancias
2 Clústeres	Alta (básica)	Baja	Captura grandes diferencias	354.0	Clúster 0: 70% Clúster 1: 30%

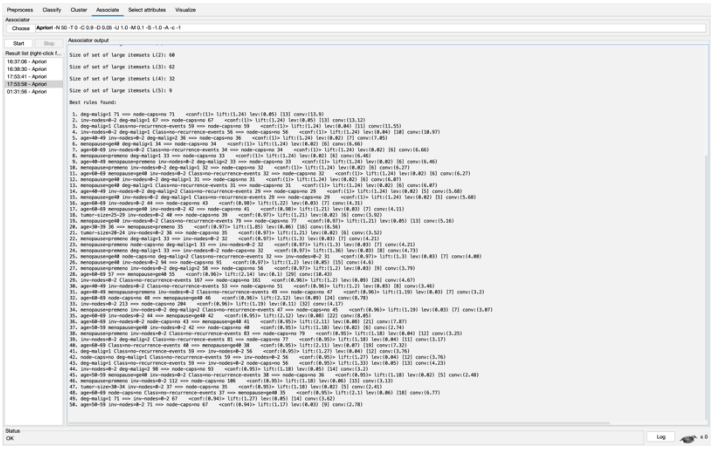
Número de Clústeres	Simplicidad	Separación	Interpretación	Error Intra-clúster (WSS)	Distribución de Instancias
3 Clústeres	Media	Moderada	Buen equilibrio entre simplicidad y detalle	352.0	Clúster 0: 55% Clúster 1: 24% Clúster 2: 21%
4 Clústeres	Moderada-Alta	Alta	Segmentación más granular y detallada	295.0	Clúster 0: 51% Clúster 1: 26% Clúster 2: 20% Clúster 3: 4%
5 Clústeres	Compleja	Muy Alta	Riesgo de sobresegmentación	294.0	Clúster 0: 51% Clúster 1: 26% Clúster 2: 20% Clúster 3: 4% Clúster 4: 0%

4. ALGORITMO DE ASOCIACIÓN

En este análisis de asociación, se aplicó un preprocesamiento de discretización sobre los atributos numéricos del dataset para ajustar las características a una escala común. Este preprocesamiento mejora la interpretación de las reglas generadas por el algoritmo Apriori, facilitando la identificación de patrones relevantes.

Las reglas generadas por el algoritmo revelan una fuerte relación entre la malignidad del tumor (atributo *deg-malig*) y la presencia o ausencia de cápsulas nodales (*node-caps*). En particular, las instancias con *deg-malig=1* (baja malignidad) están asociadas frecuentemente con la ausencia de cápsulas nodales. Además, la clase *no-recurrence-events* muestra una estrecha asociación con la ausencia de cápsulas nodales, lo que podría sugerir un mejor pronóstico en estos casos. Estas relaciones destacan aspectos clínicos clave entre las características de los pacientes con cáncer de mama y la probabilidad de recurrencia de eventos, proporcionando información valiosa para entender y predecir el comportamiento de la enfermedad.

Se realizaron dos ejecuciones del algoritmo Apriori utilizando las mismas configuraciones generales, pero variando la opción del parámetro **car** (Classification Association Rules). Este parámetro determina si las reglas generadas deben enfocarse exclusivamente en predecir la clase del conjunto de datos. En la ejecución con **"car true"**, el algoritmo generó solo 6 reglas, mientras que en **"car false"** se alcanzaron las 50 reglas.



En la ejecución con **"car true"** (imagen izquierda), las reglas más destacadas incluyen asociaciones entre variables como menopausia, nodos afectados y grado de malignidad. Entre estas, la regla más confiable es *"menopause=ge40, inv-nodes=0-2, deg-malig=1 → Class=no-recurrence-events"*, con una confianza del 94%. Así pues se generaron menos reglas debido a un enfoque más restrictivo, y las asociaciones descubiertas se limitan a un número menor de itemsets grandes.

Por otro lado, en la ejecución con **"car false"** (imagen derecha), se generaron reglas con confianzas notablemente más altas, algunas alcanzando el 100%. Un ejemplo de ello es la regla *"deg-malig=1 → node-caps=no"*, con una confianza perfecta. Esto resultó en un mayor número de itemsets grandes generados y una cantidad mucho más amplia de reglas, alcanzando las 50.

En resumen, el cambio en el parámetro **car** tuvo un impacto significativo en el comportamiento del algoritmo. Con **"car true"**, las reglas generadas se centraron exclusivamente en predecir la clase, pero a costa de limitar el número total de asociaciones. En contraste, con **"car false"**, se encontraron más patrones relevantes al relajar esta restricción, logrando una mayor cantidad de reglas de alta confianza y diversidad.