Universidad del Valle de México

Administración de Bases de Datos

Actividad Resumen

ESTUDIANTES:

CARRASQUEDO GUERRERO JUAN PABLO

ÁLVAREZ NAVA CARLOS ALEXIS

PEREZ SILVA MARIEL

CARRERA: ING. EN SISTEMAS COMPUTACIONALES

FECHA DE ENTREGA: 14 DE MARZO DEL 2025

Contenido

Planteamiento

El cáncer de mama es una de las principales causas de muerte en mujeres en todo el mundo. Detectarlo a tiempo puede salvar muchas vidas, y en México se usan las mastografías como herramienta básica para encontrar esta enfermedad de forma temprana. Sin embargo, cuando hay tantos datos (como los más de 855,000 registros que tenemos de bases oficiales), organizar la información y descubrir patrones sin que haya una clasificación previa resulta difícil.

Para enfrentar este reto, podemos aplicar métodos de aprendizaje no supervisado, como el algoritmo K-Means++, que nos ayuda a formar grupos (clusters) con datos que tengan características similares. Con esto, queremos descubrir si hay ciertos conjuntos de pacientes que tienen más probabilidades de presentar un resultado anormal en su mastografía y ver si estos grupos se relacionan con factores de riesgo conocidos (por ejemplo, la edad o la zona geográfica).

En este trabajo, usaremos K-Means++ en el lenguaje de programación Julia, ya que Julia es muy eficiente cuando se maneja un volumen grande de datos y se realizan muchas operaciones matemáticas. Después de formar los grupos, compararemos los resultados con la información disponible en artículos médicos sobre factores de riesgo para el cáncer de mama. Así podremos saber si lo que encontramos en nuestra base de datos coincide o no con lo que dice la literatura científica.

Marco teórico

Generalidades del cáncer de mama

- El cáncer de mama es la neoplasia más frecuente en mujeres. Según algunos estudios en México, cada año se detectan alrededor de 38.4 casos por cada 100,000 mujeres, y el costo promedio de atención es de aproximadamente $110,459.00 pesos por paciente. Esto representa un gran desafío para el sistema de salud mexicano (Joaquín et al., s.f.).
- Factores de riesgo como la edad, la herencia familiar, la obesidad y ciertos cambios hormonales pueden aumentar la probabilidad de que una mujer desarrolle cáncer de mama. Además, la falta de diagnósticos tempranos empeora la situación (OPS/OMS, 2025).

Importancia de la detección temprana

- La principal razón para promover la detección temprana es que entre más pronto se detecta el cáncer de mama, mejores son los resultados del

tratamiento. La mamografía es la prueba más común y funciona como un filtro inicial para encontrar anomalías.

- Sin embargo, cuando tenemos bases de datos enormes (como la que estamos usando, con más de 855,000 registros), es muy complicado revisar caso por caso. Aquí es donde la tecnología de análisis de datos nos puede ayudar a identificar patrones que ayuden a los especialistas a enfocarse en los casos de mayor riesgo.

Algoritmo Kmeans++ en Julia

- K-Means++ es una versión mejorada del algoritmo K-Means, que empieza el proceso de agrupamiento eligiendo de forma más inteligente los "centroides" (puntos representativos de cada grupo). Esto ayuda a que el algoritmo sea más estable y a veces más rápido en encontrar buenas agrupaciones.
- Julia, por otro lado, es un lenguaje de programación que está ganando popularidad en la comunidad científica porque permite realizar operaciones numéricas de forma muy eficiente, incluso con grandes cantidades de datos.

Factores externos

- Después de agrupar los datos con K-Means++, queremos ver si los grupos con mayor probabilidad de tener mastografías anormales coinciden con factores de riesgo reportados en estudios previos, como la edad, el lugar donde viven o incluso el tiempo que tardan en interpretar la mastografía.
- Por ejemplo, si descubrimos que existe un grupo con pacientes que suelen tener más de 50 años y viven en ciertas zonas, podríamos compararlo con la literatura para ver si, efectivamente, la edad y la localización geográfica son factores importantes en el desarrollo del cáncer de mama.

Resumen de estudios

- Diversos artículos y revisiones (como "Cáncer de mama: una visión general", citado por Joaquín et al.) describen la etiología, los principales síntomas y los métodos de detección actuales. Todos coinciden en que el cáncer de mama, aunque tiene múltiples causas, puede ser detectado a tiempo con buenos métodos de tamizaje y con personal médico capacitado.
- De igual forma, se menciona que la incidencia de cáncer de mama va en aumento a nivel mundial, pero la mortalidad disminuye debido a que las personas se están haciendo sus revisiones más seguido y hay tratamientos más efectivos.

Justificación

Este proyecto busca contribuir a la detección temprana y al estudio del cáncer de mama en México. Analizar una base de datos tan grande con técnicas tradicionales puede llevar mucho tiempo, mientras que los métodos de aprendizaje automático (como K-Means++) permiten agrupar información y encontrar patrones de forma más automatizada. Si descubrimos que hay ciertas características (por ejemplo, edad o municipio de residencia) asociadas con diagnósticos anormales, esto podría ayudar a las autoridades y al sector salud a tomar decisiones de prevención y detección más eficientes, reforzando la importancia de la mastografía como herramienta principal.

Preguntas de investigación

1) ¿Qué grupos principales se forman cuando aplicamos el algoritmo K-Means++ a los datos de mastografías de más de 855,000 registros en México?
2) ¿Existen patrones relacionados con la edad, la ubicación geográfica u otros factores que aumenten la probabilidad de un resultado anormal en la mastografía?
3) ¿Coinciden los patrones encontrados en la base de datos con los factores de riesgo descritos en artículos científicos sobre cáncer de mama?
4) ¿Es K-Means++ el método más adecuado para este tipo de agrupación o se requerirían otras técnicas de análisis no supervisado para mejores resultados?

Formulación del problema

En México, el cáncer de mama sigue siendo una de las principales causas de muerte en mujeres. Aunque existen registros masivos de mastografías, no siempre se aprovecha al máximo toda esa información para identificar patrones de riesgo. El problema, entonces, radica en cómo usar estos datos de manera eficiente para hallar grupos de pacientes que puedan requerir mayor atención. Por ello, se propone utilizar K-Means++ con Julia para crear clusters y analizar si dichos grupos reflejan factores de riesgo ya conocidos o si revelan información nueva que pueda servir para mejorar la detección temprana del cáncer de mama.

Conclusión

En resumen, usar el algoritmo K-Means++ para analizar la gran cantidad de datos de mastografías en México puede ayudarnos a encontrar patrones relevantes que podrían indicar riesgos mayores de presentar cáncer de mama. Gracias a la eficiencia de Julia, este estudio no solo podría revelar información valiosa sobre grupos de pacientes con

características en común, sino también facilitar la comparación de estos hallazgos con factores de riesgo descritos en la literatura médica, como la edad y la zona geográfica. Aunque el objetivo no es predecir diagnósticos clínicos, los hallazgos obtenidos podrían guiar estrategias de prevención y detección temprana, con un impacto positivo en la salud pública. Siguiendo estos pasos y documentando cuidadosamente cada etapa, se reforzará la importancia de la mastografía como método efectivo de detección temprana y se abrirán nuevas líneas de investigación basadas en el uso de técnicas de aprendizaje no supervisado.

Referencias

- Joaquín, P. P., Jareth, L. R., Aylin, J. a. L., Alonso, M. N. C., Joaquín, P. P., Jareth, L. R., Aylin, J. a. L., & Alonso, M. N. C. (s.f.). Cáncer de mama: una visión general. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-72032021000300354

- Cáncer de mama. (2025, 4 de febrero). OPS/OMS | Organización Panamericana de la Salud. https://www.paho.org/es/temas/cancer-mama