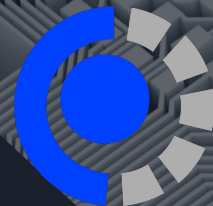


De Datos a Decisiones:



Viabilidad del Aprendizaje de Asociación de Consumos Médicos en el Ámbito de las Obras Sociales.

Lara V. Acuña

Adriano M. Lusso

Técnicas de Depósito y Minería 2024
Sandra Roger



Índice

[Introducción](#)

[Estado del Arte](#)

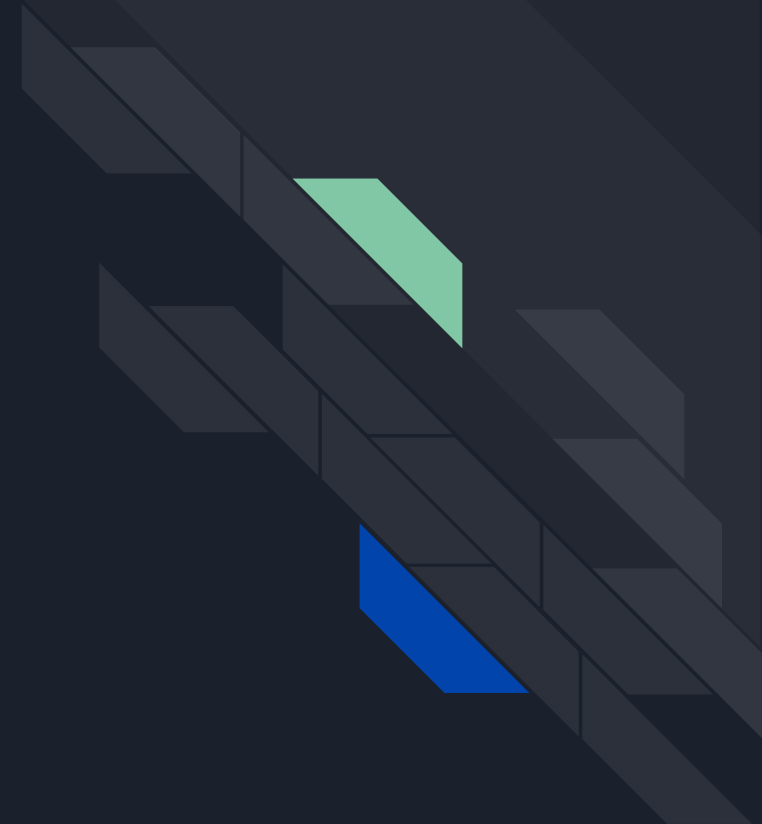
[Trabajos Relacionados](#)

[Datos Crudos](#)

[Pre-Procesamiento de Datos](#)

[Experimentos Realizados](#)

[Análisis y Resultados](#)

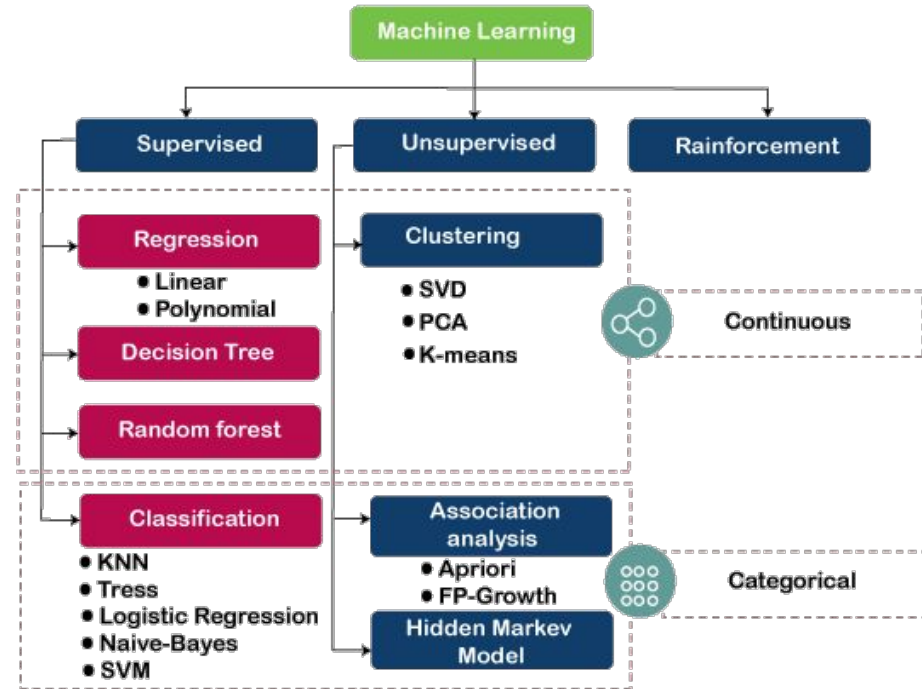


The background is a dark blue gradient. In the top-left corner, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram. In the bottom-left corner, there is a circular inset showing a detailed, grayscale image of a printed circuit board (PCB) with various electronic components. In the top-right corner, there is a grayscale image of a complex, multi-layered circuit board structure.

Introducción

Minería de datos

- Proporciona herramientas para encontrar patrones subyacentes en grandes conjuntos de datos.
- Sirven como capa de abstracción entre los usuarios y las complejas técnicas matemáticas utilizadas.



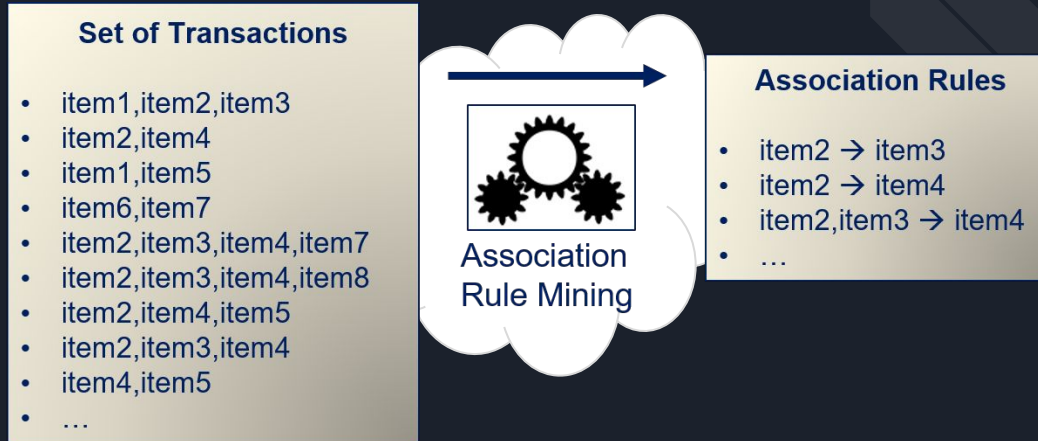


Aplicaciones en la salud

- 01 Enfoques complementarios de Clustering para predicción de la readmisión de pacientes en medicina intensiva.
- 02 Análisis de Discriminante Lineal para predicción de la gravedad en pacientes con Parkinson.
- 03 Redes bayesianas para predicción de riesgo cardiovascular.

Generación de Reglas de Asociación

- Técnica que genera reglas que describe parcialmente relaciones de implicancia o correlación entre eventos.
- Existen diversos algoritmos que implementan esta técnica
 - Apriori
 - FP-Growth
 - Eclat



SOSUNC

- Servicio de Obra Social de la Universidad Nacional del Comahue.
- Base de datos con información de afiliados y su actividad médica.
- El área de Auditoría Médica se encarga de:
 - Realizar el proceso de aprobación de solicitudes de cobertura.
 - Crear y gestionar los planes de cobertura.

Auditoría Médica aún no hace uso de técnicas avanzadas de minería de datos.

Objetivo

Aplicar diferentes algoritmos de generación de reglas de asociación en el dominio de SOSUNC y analizar su rendimiento.

Explorar formas de representación gráfica y cómo puede ayudar a SOSUNC a obtener conocimiento a partir de las reglas.



The background is a dark blue gradient. In the top-left corner, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram. In the bottom-left corner, there is a circular inset showing a detailed, grayscale image of a printed circuit board (PCB) with various electronic components. In the top-right corner, there is a faint, grayscale image of a complex circuit board layout with many traces.

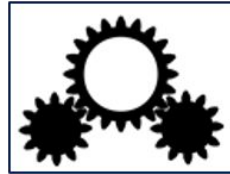
Estado del arte

Generación de Reglas de Asociación

Recapitulando ...

Set of Transactions

- item1,item2,item3
- item2,item4
- item1,item5
- item6,item7
- item2,item3,item4,item7
- item2,item3,item4,item8
- item2,item4,item5
- item2,item3,item4
- item4,item5
- ...



Association
Rule Mining

Association Rules

- item2 \rightarrow item3
- item2 \rightarrow item4
- item2,item3 \rightarrow item4
- ...



Generación de Reglas de Asociación

Soporte

- Frecuencia porcentual de un subconjunto de elementos en el conjunto de transacciones.
- Se utiliza como hyperparámetro para delimitar la frecuencia mínima necesaria para que un subconjunto sea relevante en el análisis.

$$soporte(Tr, X) = \frac{frecuencia(Tr, X)}{|Tr|}$$



Generación de Reglas de Asociación

Confianza

- Define el porcentaje de casos en los que una regla es correcta.
- Se puede utilizar como hyperparámetro para asegurar una calidad mínima en las reglas generadas.

$$confianza(A, B) = \frac{soporte(A, B)}{soporte(A)}$$



Generación de Reglas de Asociación

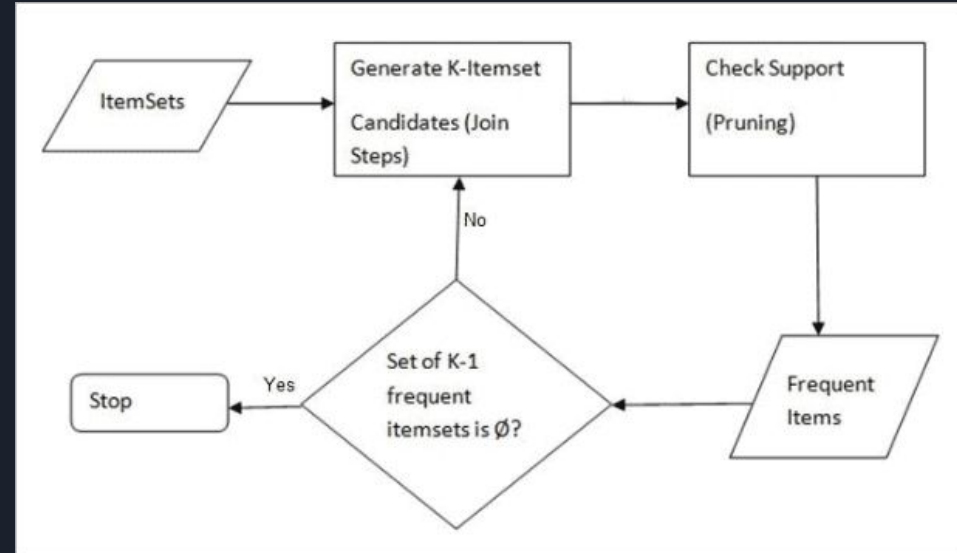
Lift

- Define en qué medida una regla es producto del azar.
- Valores mayores a 1 indican una relación fuerte, que sucede más que lo indicado por el azar.
- Valores menores a 1 indica que la regla aparece con menos frecuencia de lo indicado por el azar (puede tener varias interpretaciones interesantes).
- Un lift igual a 1 indica que regla es azarosa y no tiene validez.

$$lift(A \rightarrow B) = \frac{soporte(A,B)}{soporte(A) \cdot soporte(B)}$$

Algoritmo Apriori

- Búsqueda por niveles Bottom-Up, comenzando por conjuntos de un elemento y aumentando progresivamente su cantidad.
- Consta de dos fases:
 - Generación de candidatos.
 - Poda

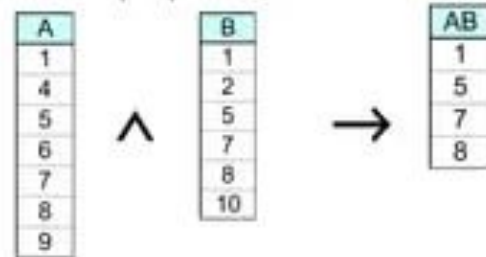


Algoritmo Eclat

- Búsqueda Depth-First en una **base de datos vertical**.
- De cada conjunto de elementos, se almacena que transacciones las contienen.
- A partir de las transacciones identificadas para los K-conjuntos de elementos frecuentes, se generan los de K+1 mediante intersección de transacciones.
- Su principal desventaja es que no utiliza la confianza como hiperparámetro.

ECLAT: Another Method for Frequent Itemset Generation

- Determine support of any k-itemset by intersecting tid-lists of two of its (k-1) subsets.



- 3 traversal approaches:
 - top-down, bottom-up and hybrid
- Advantage: very fast support counting
- Disadvantage: intermediate tid-lists may become too large for memory



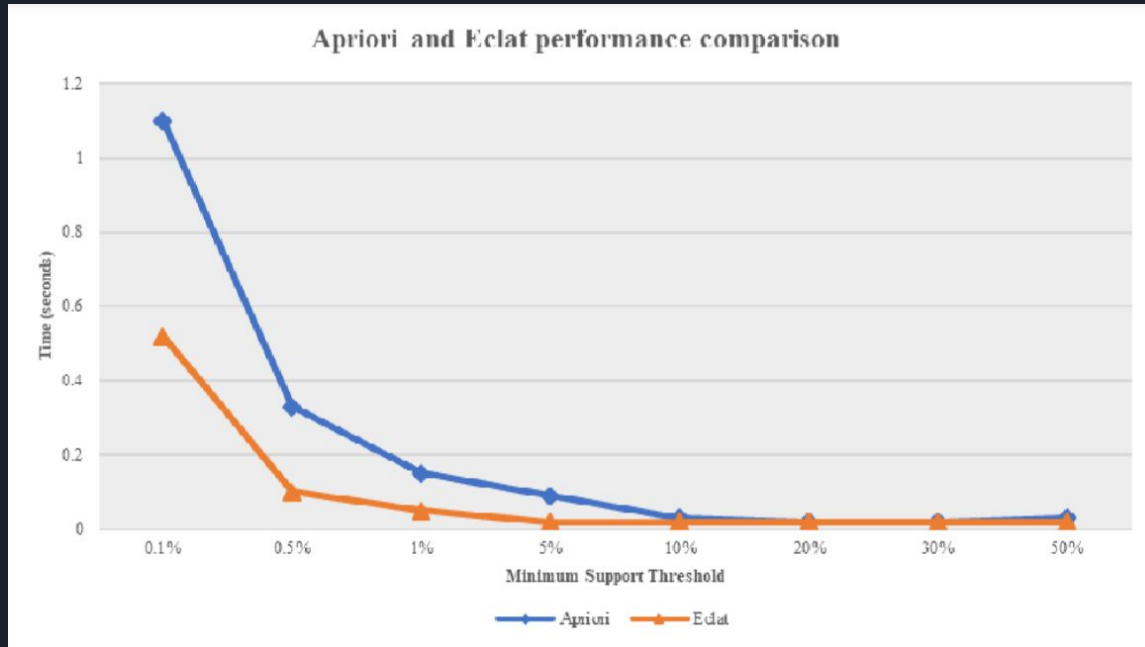
Trabajos relacionados



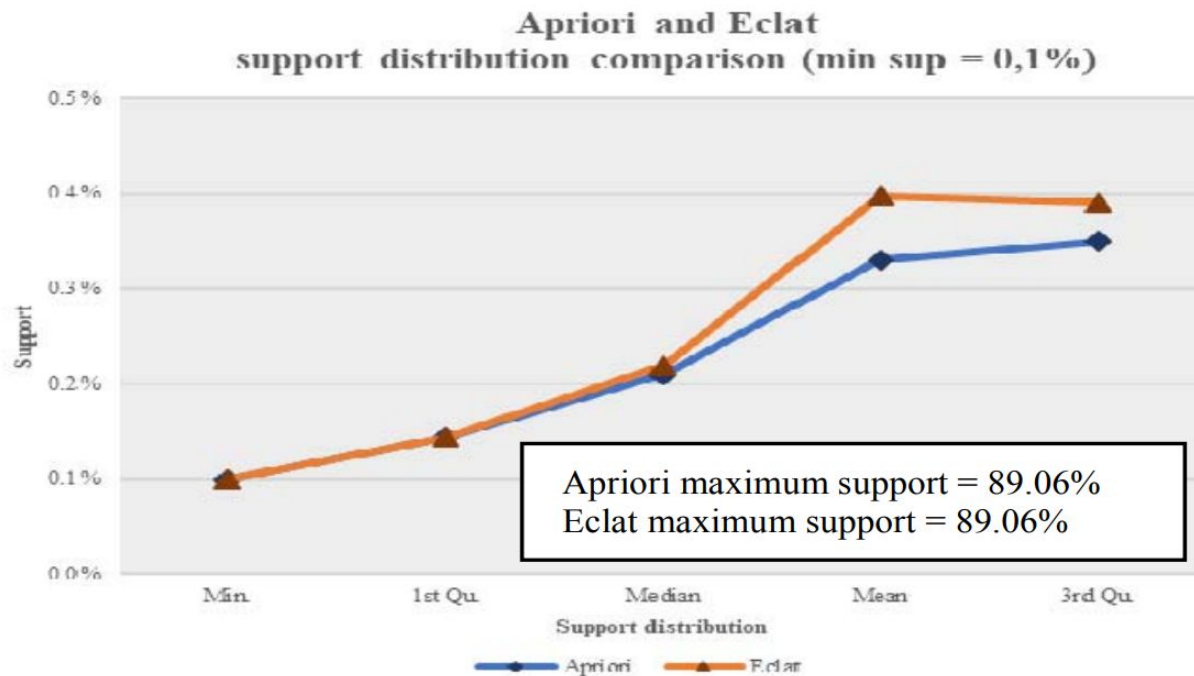
Comparativa del rendimiento de Apriori y Eclat

- Investigación realizada por Vlad Robu y Vitor Duarte dos Santos.
- Trabajan con 52938 transacciones y 26 elementos binarios.
- La base de datos es de elecciones de usuarios en una página web.
- Se comparan resultados de:
 - Tiempos de ejecución
 - Soporte.
 - Cantidad de reglas generadas y cantidad de elementos en estas.

Comparativa del rendimiento de Apriori y Eclat



Comparativa del rendimiento de Apriori y Eclat





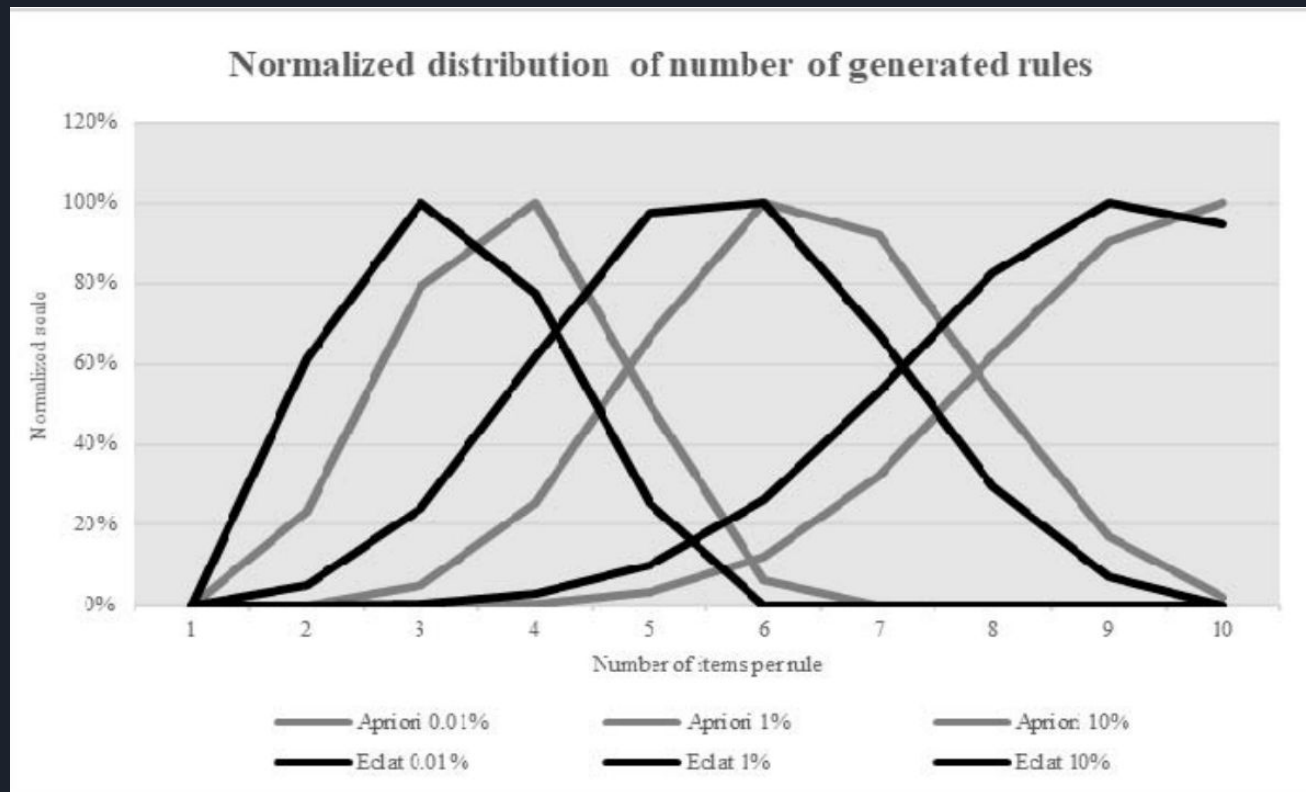
Comparativa del rendimiento de Apriori y Eclat

Comparison of number of rules generated by Apriori and Eclat at different support thresholds		
Min support threshold	Apriori	Eclat
0.01%	7,465,565	1,443,983
0.1%	686,979	154,242
0.5%	90,278	26,735
1%	26,345	9,341
5%	1,211	633
10%	312	187
20%	62	40
30%	34	21
50%	6	4

Comparativa del rendimiento de Apriori y Eclat

Las cantidades se normalizaron utilizando *min-max*:

$$(x - x_{\min}) / (x_{\max} - x_{\min})$$



The background is a dark navy blue. In the top-left corner, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram. In the bottom-left corner, there is a circular inset showing a detailed, grayscale image of a circuit board. In the top-right corner, there is a faint, grayscale image of a circuit board with many small components.

Datos crudos




Consumos médicos

- CSV con los consumos de los afiliados.
- Tiene 286007 consumos registrados desde el 2019 a Septiembre del 2024.
- Estructura:
 - ID aleatoriamente generado y no sensible del afiliado
 - ID de la práctica médica.
 - Nombre de la práctica.
 - Fecha de consumo (completa dd/mm/aaaa y desglosada).

Prácticas de interés

- CSV con 17 con prácticas de interés relacionadas a diabetes.
- Algunas de estas son:
 - Consulta a nutricionista con especialización en diabetes.
 - Controles de diferentes tipos de colesterol.
 - Control de hemoglobina.
 - Control de creatinina.



The background is a dark blue gradient. On the left, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram. Below these, there is a circular inset showing a detailed, grayscale image of a circuit board. In the top right corner, there is a faint, grayscale image of a circuit board with many small components.

Pre-procesamiento de datos



Acotamiento de consumos



Se considera:

- Consumos comprendidos entre el 01/07/2023 al 01/07/2024.
- Consumo de prácticas únicamente relacionadas a diabetes.

Resultado:

- La cantidad de consumos registrados se reduce de 286007 a 9525.

Conversión a formato horizontal

Afiliado	Práctica	Fecha	...
1	X	10/5/2019	...
1	Y	13/5/2019	...
2	X	07/5/2019	...
...

Antes



Afiliado	X	Y	Z	...
1	TRUE	TRUE	FALSE	...
2	TRUE	FALSE	FALSE	...
...

Después

Conversión a formato horizontal: **Resultados**

- De las 17 prácticas quedaron 12: 5 prácticas nunca fueron consumidas en el rango de fechas indicado.
- 9525 registros se reducen a 2227: uno por cada afiliado único
- La cantidad de prácticas consumidas por afiliado varía entre 1 a 8
- La práctica más consumida: GLUCEMIA (67%)



The background is a dark navy blue. In the top-left corner, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram. In the bottom-left corner, there is a circular inset showing a detailed, grayscale image of a printed circuit board (PCB) with various electronic components. In the top-right corner, there is a faint, grayscale image of a complex circuit board layout with many traces.

Experimentos Realizados

Experimento para comparar rendimiento temporal y generación de reglas: Apriori vs Eclat

- Cada algoritmo minará reglas 10000 veces con distintos valores de soporte mínimo: desde 0.01% hasta 50%
- Para cada soporte se promedia el tiempo empleado y se almacenan las reglas obtenidas junto con sus medidas de interés





Consideraciones importantes



- Las reglas generadas únicamente tienen UN ítem en el consecuente.
- La implementación de **Apriori** del paquete utilizado en R posee mejoras respecto al original: utiliza un *prefix tree* para organizar las transacciones.
- **Apriori** usa una confianza mínima del 80%
- **Eclat** no necesita el hyperparámetro de confianza, las reglas se generan aparte.

Demostración de distintos gráficos para analizar reglas

Se exploran las reglas con los siguientes gráficos

- Scatter Plot
- Grouped Matrix
- Graph (grafo)

Se demuestra su utilidad



The background is a dark blue gradient. On the left, there are two overlapping geometric shapes: a blue parallelogram and a light green parallelogram. Below these, a circular inset shows a detailed, grayscale image of a printed circuit board (PCB) with various electronic components. In the top right corner, there is a faint, stylized pattern of white lines resembling a circuit or a topographical map.

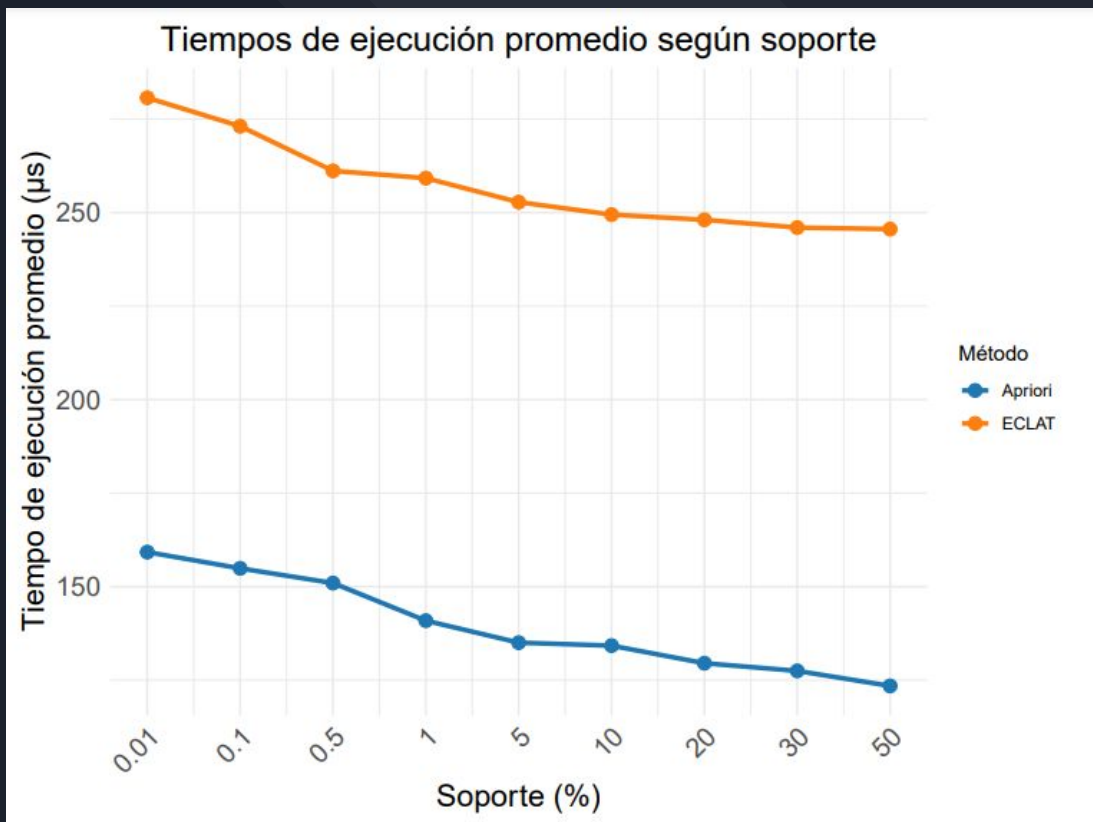
Análisis y Resultados

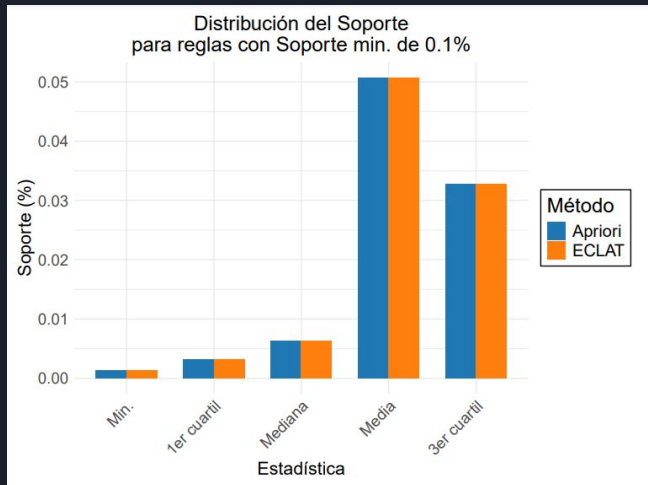
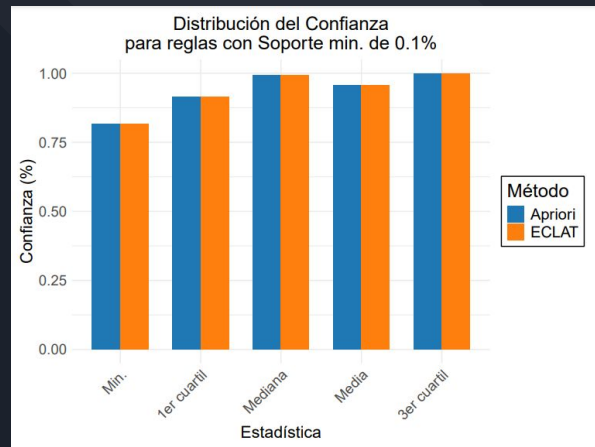
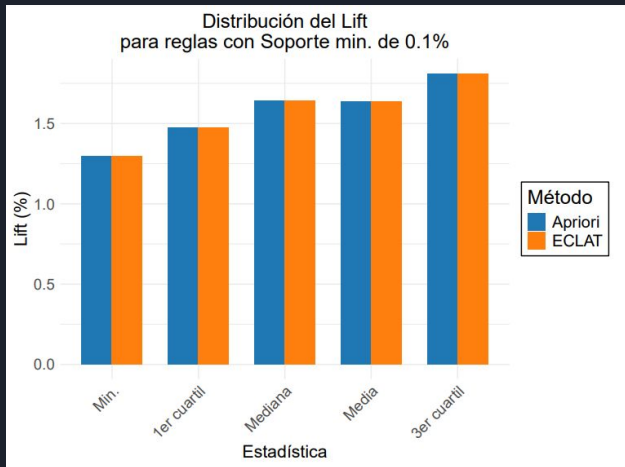


Se analizará

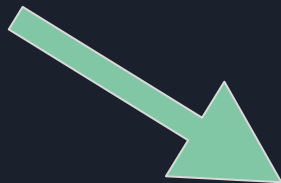
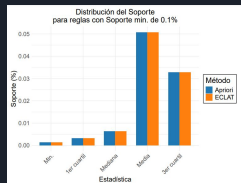
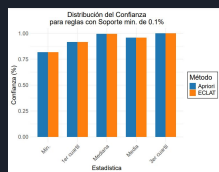
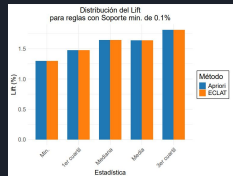
- Tiempos de ejecución
- Estadísticas de Soporte, Confianza y Lift
- Distribuciones normalizadas de cantidad de elementos por regla.
- Cantidad de reglas generadas.
- Descubrimientos sobre las reglas.
- Discusiones.

- Pendientes decreciente a menor Soporte mínimo.
- Apriori tiene menores tiempos de ejecución.





Los valores de estadísticas son idénticos entre los algoritmos.

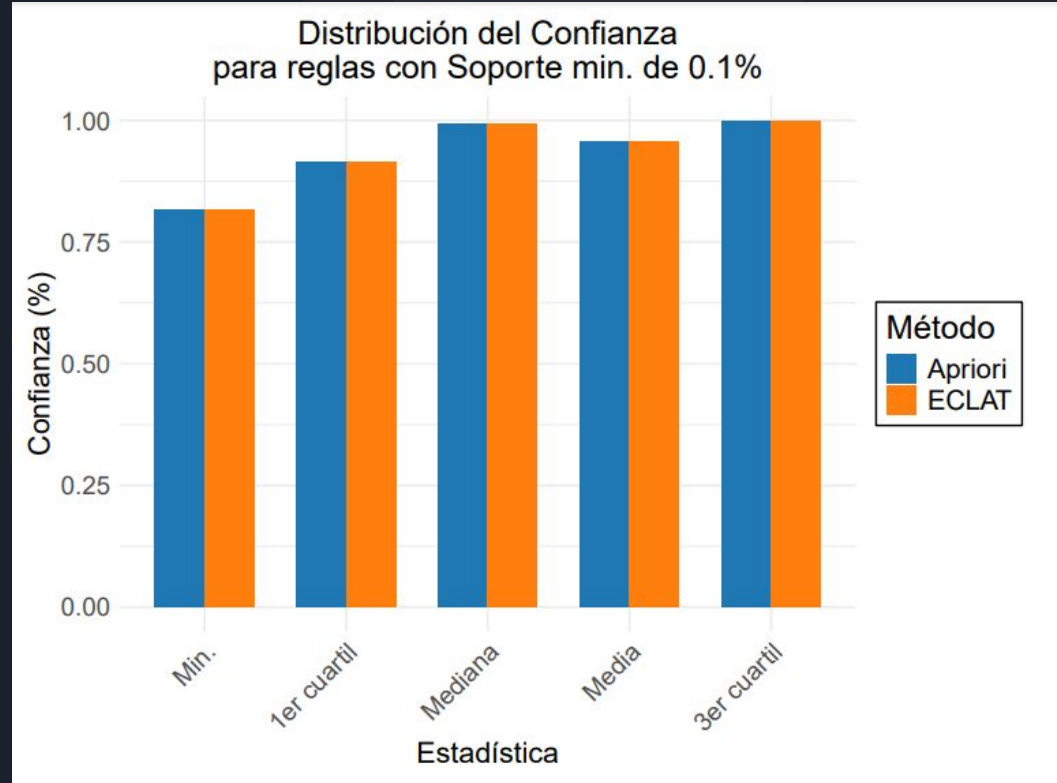


Ambos algoritmos generaron exactamente las mismas reglas

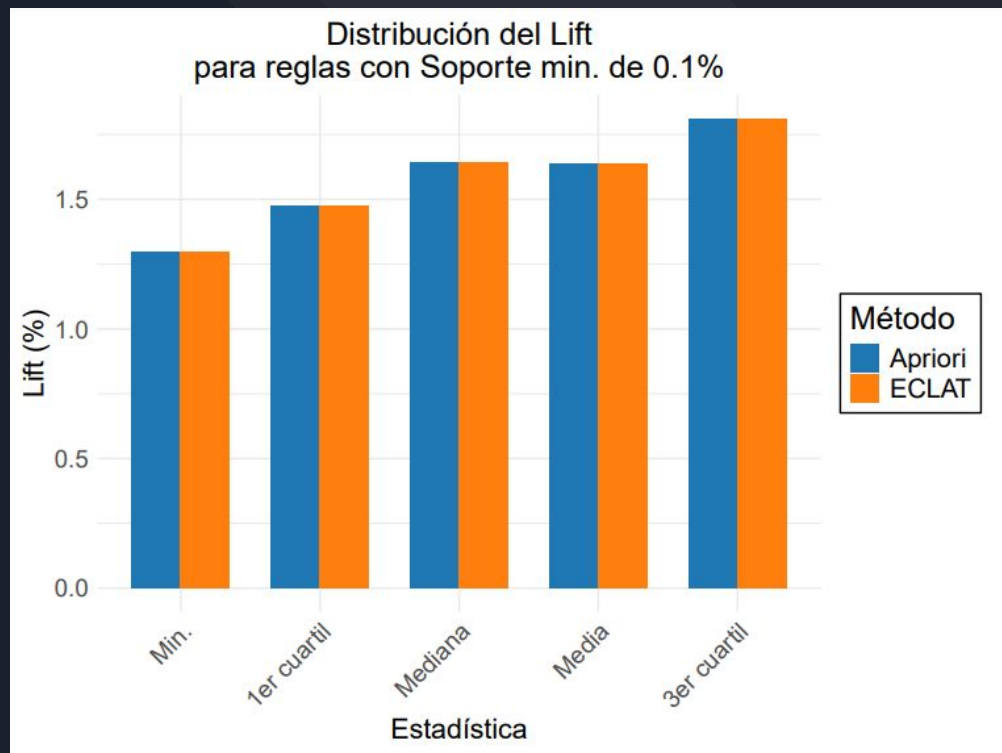
Soporte mínimo	Cantidad de reglas
0,01 %	1144
0,1 %	973
0,5 %	558
1 %	450
5 %	146
10 %	146
20 %	66
30 %	66
50 %	7

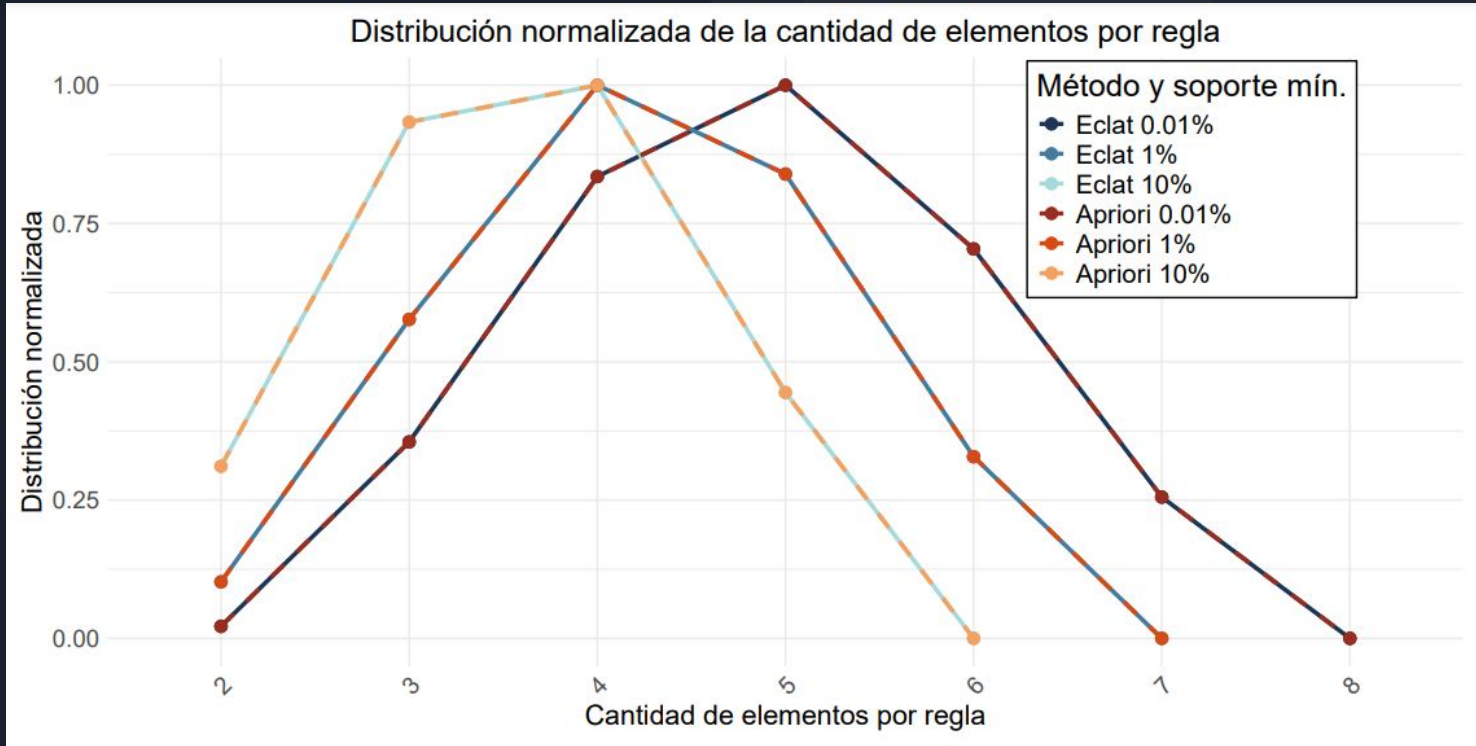
Retomando...

- Valores cercanos a 1 indican reglas con alta validez y confianza.

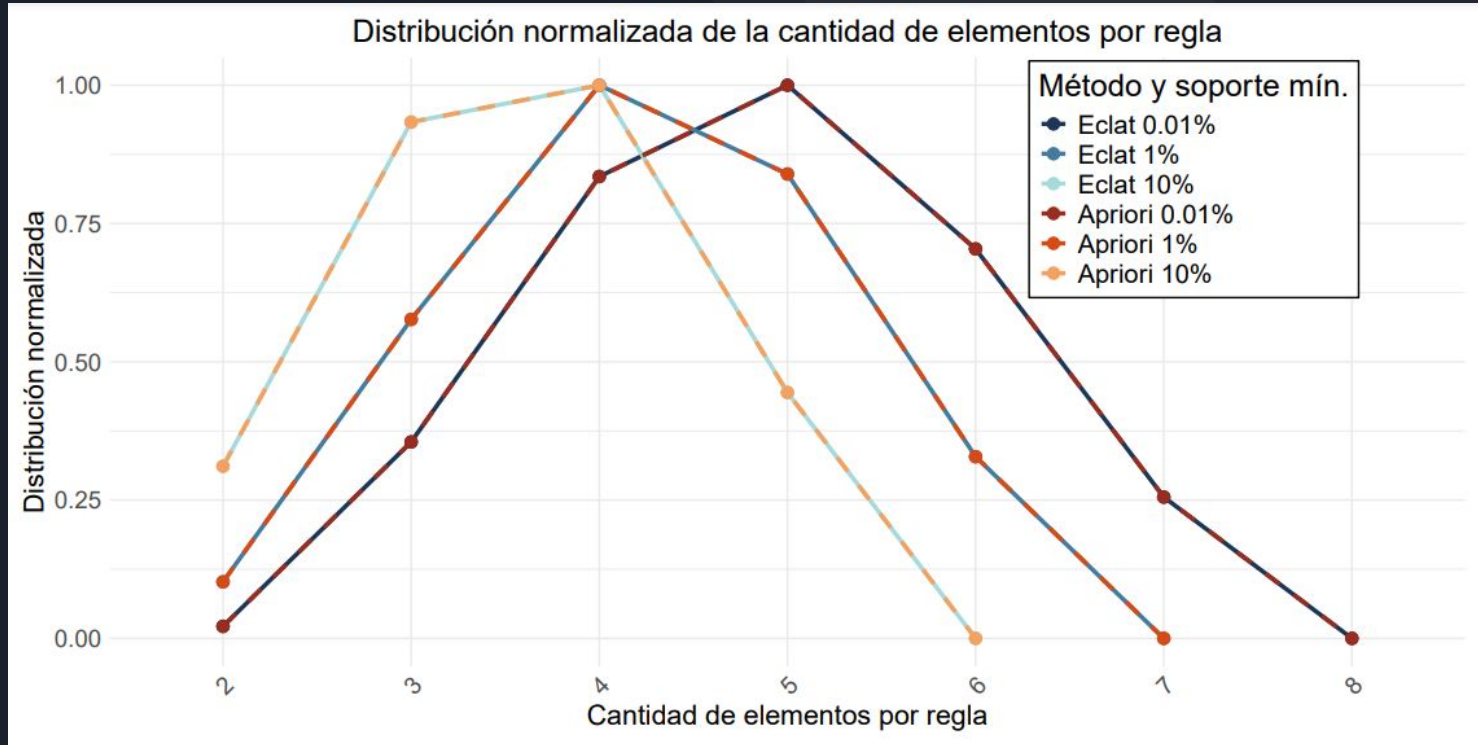


- Valores superiores a 1 que indican fuerza no aleatoria entre antecedente y consecuente.
- Aún así, estos valores promediando a 1.75 no son tan altos como los obtenidos en otros trabajos.

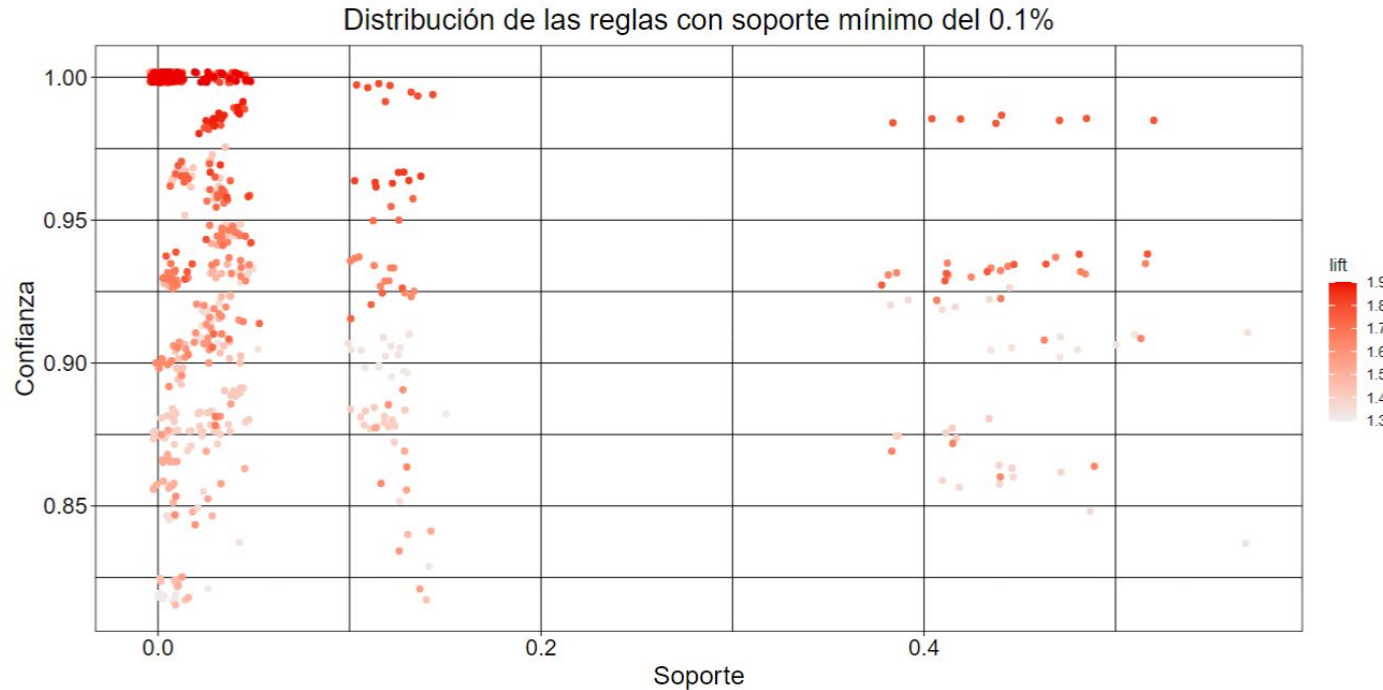




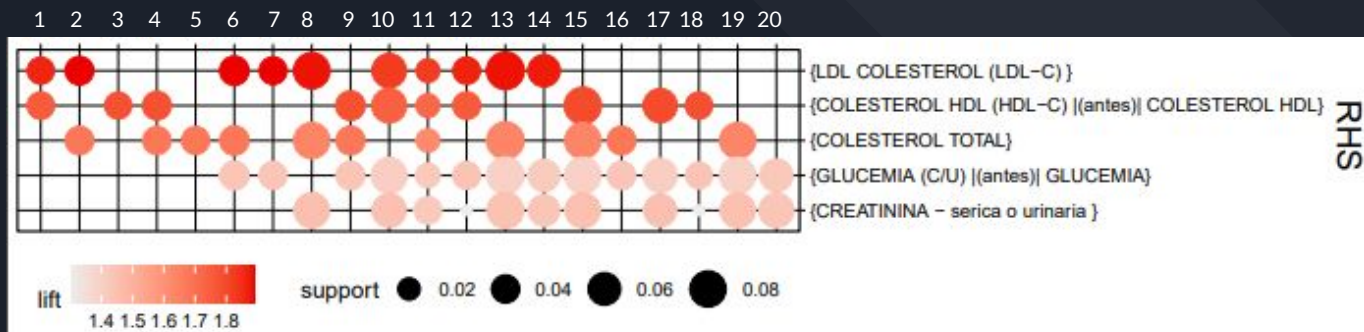
X = frecuencias de cantidad de elementos por reglas
$$\text{min_max}(x) = (x - \min(X)) / (\max(X) - \min(X))$$



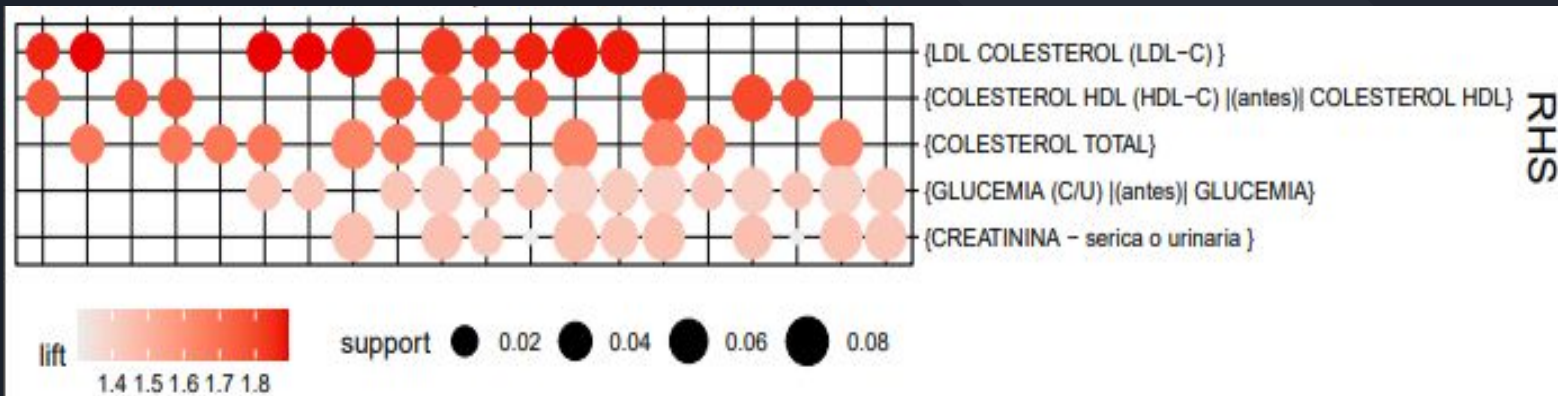
- Distribuciones idénticas para distintos algoritmos y mismo Soporte mínimo.
- A mayor Soporte mínimo, la distribución tiene a menos elementos por regla.



- La tendencia es creciente en el lift en función de la confianza.
- Rango de valores de soporte despoblado.

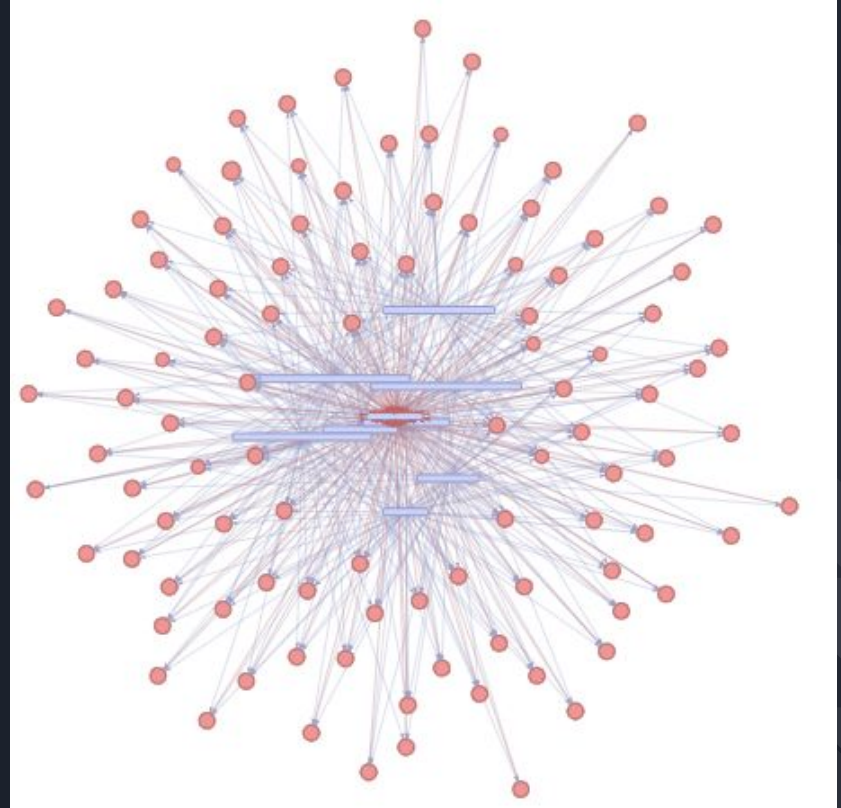


- 1 36 rules: {GLUCEMIA (C/U)} {(antes)} GLUCEMIA, COLESTEROL TOTAL, +5 items}
- 2 32 rules: {GLUCEMIA (C/U)} {(antes)} GLUCEMIA, COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, +5 items}
- 3 18 rules: {GLUCEMIA (C/U)} {(antes)} GLUCEMIA, LDL COLESTEROL (LDL-C), +6 items}
- 4 32 rules: {GLUCEMIA (C/U)} {(antes)} GLUCEMIA, LDL COLESTEROL (LDL-C), +5 items}
- 5 16 rules: {GLUCEMIA (C/U)} {(antes)} GLUCEMIA, LDL COLESTEROL (LDL-C), +6 items}
- 6 48 rules: {COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, CREATININA - serica o urinaria, +4 items}
- 7 51 rules: {COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, COLESTEROL TOTAL, +6 items}
- 8 27 rules: {GLUCEMIA (C/U)} {(antes)} GLUCEMIA, COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, +4 items}
- 9 48 rules: {LDL COLESTEROL (LDL-C), CREATININA - serica o urinaria, +4 items}
- 10 73 rules: {COLESTEROL TOTAL, CONSULTA VESTIDA OFTALMOLOGICA (PEDIATRICA Y DE ADULTO), +4 items}
- 11 147 rules: {CONSULTA ESPECIALISTAS EN DIABETES, GLUCEMIA (C/U)} {(antes)} GLUCEMIA, +4 items}
- 12 58 rules: {COLESTEROL TOTAL, CREATININA - serica o urinaria, +4 items}
- 13 36 rules: {COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, CONSULTA VESTIDA OFTALMOLOGICA (PEDIATRICA Y DE ADULTO), +3 items}
- 14 63 rules: {COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, COLESTEROL TOTAL, +5 items}
- 15 63 rules: {LDL COLESTEROL (LDL-C), CONSULTA VESTIDA OFTALMOLOGICA (PEDIATRICA Y DE ADULTO), +4 items}
- 16 32 rules: {LDL COLESTEROL (LDL-C), COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, +5 items}
- 17 52 rules: {LDL COLESTEROL (LDL-C), COLESTEROL TOTAL, +5 items}
- 18 39 rules: {LDL COLESTEROL (LDL-C), COLESTEROL TOTAL, +5 items}
- 19 45 rules: {LDL COLESTEROL (LDL-C), COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, +5 items}
- 20 57 rules: {LDL COLESTEROL (LDL-C), COLESTEROL HDL (HDL-C)} {(antes)} COLESTEROL HDL, +7 items}

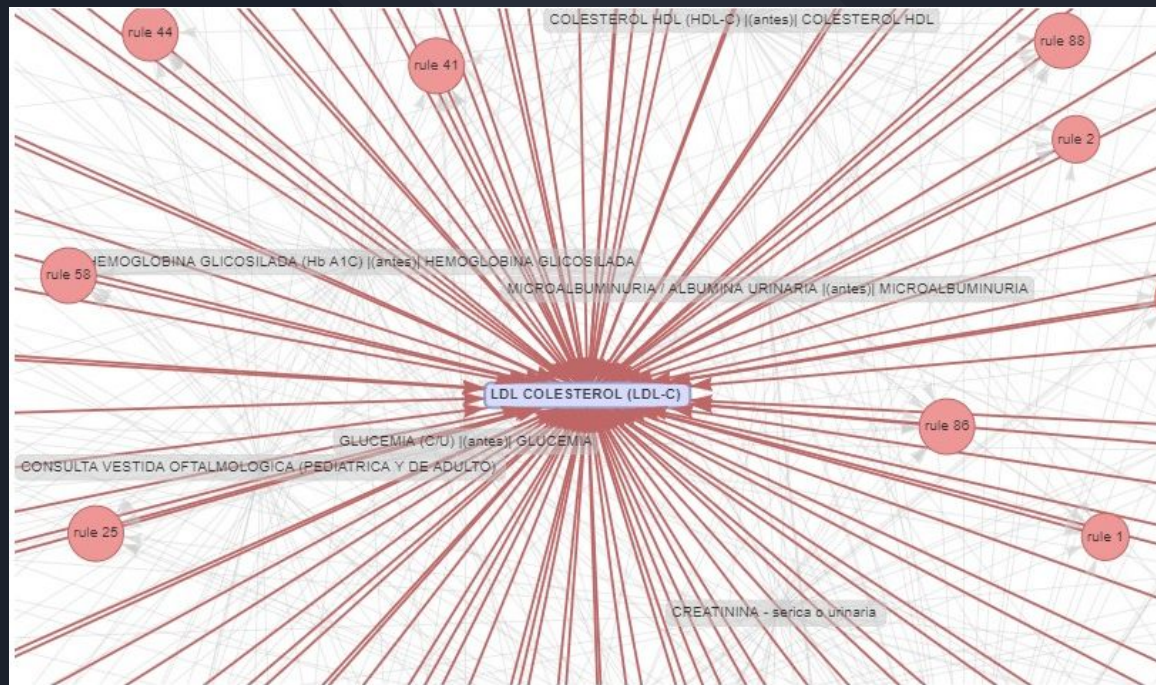


- Las únicas prácticas presentes en los consecuentes son los controles de colesterol LDL, HDL y total, creatinina y glucemia.
- Las reglas con mayor Lift con las de colesterol LDL y HDL.

- Grafo relacional de reglas que muestra las 100 reglas con mayor Lift.



- El control de colesterol LDL es la práctica que aparece en todas las reglas como consecuente.



Discusiones

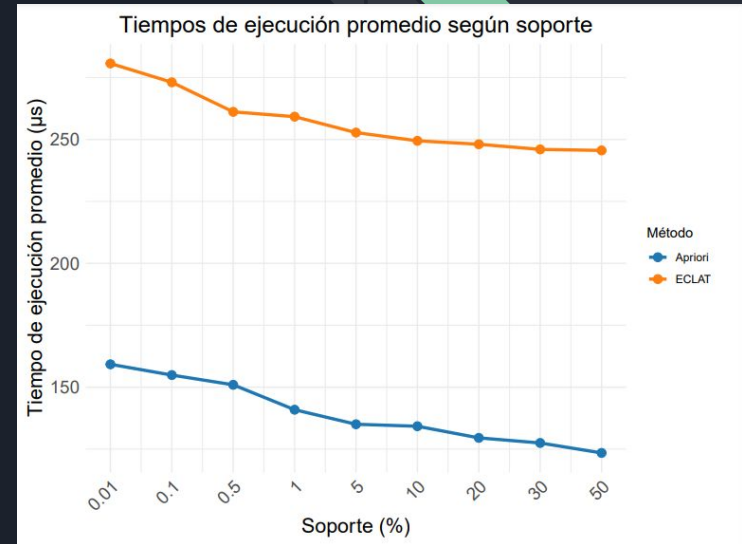
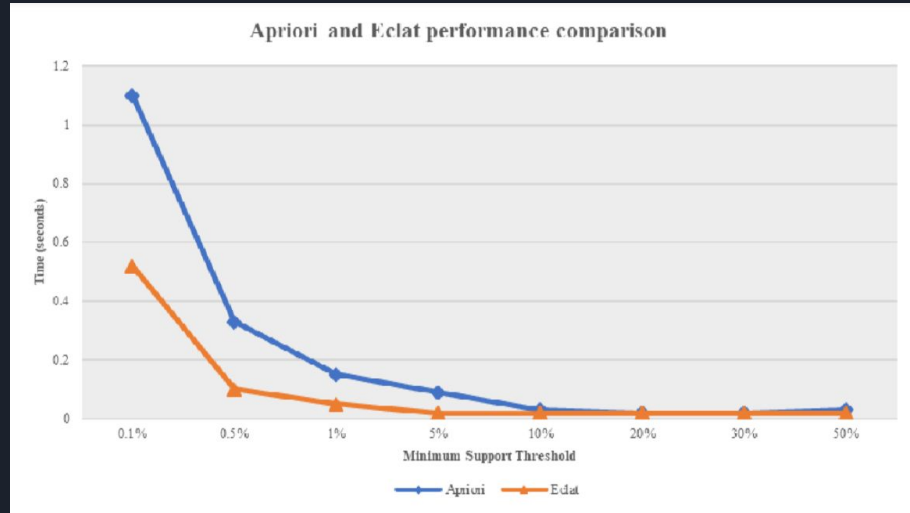
- La diferencia en la cantidad de transacciones (52938 vs 2227) y elementos por transacción (26 vs 12) puede ser causa de las diferencias en resultados respecto al trabajo de referencia.
- Es posible que la búsqueda que hace Eclat en nuestra base de transacciones logre ser exhaustiva, generando así todas las reglas de asociación posibles.

Min support threshold	Apriori	Eclat
0.01%	7,465,565	1,443,983
0.1%	686,979	154,242
0.5%	90,278	26,735
1%	26,345	9,341
5%	1,211	633
10%	312	187
20%	62	40
30%	34	21
50%	6	4

Soporte mínimo	Cantidad de reglas
0,01 %	1144
0,1 %	973
0,5 %	558
1 %	450
5 %	146
10 %	146
20 %	66
30 %	66
50 %	7

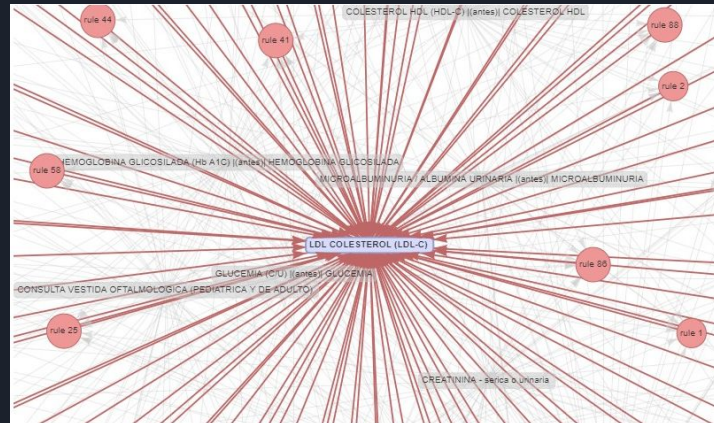
Discusiones

- En la investigación referencia, se plantea que el menor tiempo de ejecución de Eclat puede deberse a que genera menor cantidad de reglas.
- Nuestro trabajo presenta evidencia que sostiene (pero no demuestra) esta afirmación hecha.



Discusiones

- Las prácticas de mayor relevancia son los controles de los tipos de colesterol, glucemia y creatinina.
- Existe un énfasis especial en el control de colesterol LDL.
- Se puede interpretar que una práctica médica común entre todos los afiliados es el control de colesterol LDL, ya que es realizada en combinación al resto de prácticas.





Conclusiones

- Se replicaron exitosamente las pruebas de Robu y Vitor Duarte dos Santos.
- Se encontraron diferencias entre los resultados obtenidos y se justificó una posible causa, siendo esta la diferencia en la cantidad de transacciones y elementos que las conforman.
- Se dió fuerza a la suposición hecha en la investigación referencia, que afirmaba que los tiempos de ejecución de Eclat estaban relacionados a la cantidad de reglas generadas.
- Se evaluó y concluyó sobre la viabilidad en la generación de reglas para SOSUNC.
- Se reforzó la utilidad de completar el estudio con diferentes tipos de gráficos.



Trabajos futuros

- Profundizar sobre otras posibles causas de las diferencias en los tiempos de ejecución.
- Evaluar el punto de quiebre para el cambio en los tiempos de ejecución.
- Utilizar y analizar otras configuraciones de hyperparámetros para Apriori (i.e desactivar el “prefix tree”).
- Realizar un análisis del rendimiento en consumo de memoria.

¡Gracias por
escuchar!

