Polo del Conocimiento

Pol. Con. (Edición núm. 70 Vol. 7, No 8
Agosto 2022, pp. 904-918
ISSN: 2550 - 682X
DOI: 10.23857/pc.v7i8

Analíticas de aprendizaje y análisis estadístico implicativo: comparación de la complejidad temporal de técnicas clúster para bases de datos de tamaño 100000 con variables modales

Learning analytics and statistical implicative analysis: comparison of the temporal complexity of cluster techniques for databases of size 100000 with modal variables

Análise analítica de aprendizagem e análise estatística implicativa: comparação da complexidade temporal das técnicas de agrupamento para bases de dados de tamanho 100.000 com variáveis modais.

Rubén Antonio Pazmiño Maji [1]
rpazmino@espoch.edu.ec
https://orcid,org/0000-0002-6811-7876

Marina Leonor Bonilla Lucero [2]
marina.bonilla@espoch.edu.ec
https://orcid,org/0000-0003-2119-4126

Lourdes Emperatriz Paredes Castelo [3]
Lourdes.paredes@espoch.edu.ec
https://orcid,org/0000-0002-5331-2759

Shirley Estefanía Armas Analuisa [4]
Shirley.armas@espoch.edu.ec

Correspondencia: rpazmino@espoch.edu.ec

[1] PhD en Formación en la Sociedad del Conocimiento, Salamanca, España; Magíster en Informática Educativa y Multimedios, Osorno, Chile; Dr. en Matemática ESPOCH, Grupo de Investigación Ciencia de Datos CIDED, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
[2] Magíster en Educación Sexual. Grupo de Investigación Ciencia de Datos CIDED, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
[3] Magíster en Informática Educativa, ESPOCH. Tecnóloga Química Industrial, ESPOCH. Grupo de Investigación Ciencia de Datos CIDED, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.
[4] Ingeniera en Estadística Informática, Grupo de Investigación Ciencia de Datos CIDED, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador.

Resumen

Al analizar datos provenientes de múltiples procesos económicos, sociales, administrativos, científicos, se tiende a presentar problemas o en ciertos casos llegan a ser procesos irrealizables debido a que no se consideran técnicas óptimas y adecuadas desde el punto de vista de la complejidad temporal (tiempo de ejecución o procesamiento); es por ello por lo que es importante investigar cuáles técnicas clúster son las más rápidas en el procesamiento de información. Las técnicas clúster permiten formar grupos de datos homogéneos con grupos heterogéneos entre sí. El presente trabajo comparó desde el punto de vista de la complejidad temporal las técnicas clúster del Análisis Estadístico Implicativo (ASI) y las de LA (Analíticas de aprendizaje). Para determinar cuál (o cuáles) técnicas son las más rápidas se utilizó una investigación cualitativa pre-experimental del tipo $RGXO_1$, donde RG representa el grupo experimental (aleatorio), X representa el tratamiento que en este caso son las 5 técnicas clúster (3 técnicas de LA y 2 técnicas de ASI) y O es el tiempo de ejecución. Las técnicas Tsim Chic y TcoheChic del ASI se demostraron que son las más rápidas para bases de datos de tamaño 100000 o 1000 observaciones y 100 variables con datos categóricos de hasta 10 categorías, que se sugiere utilizar si se trabaja en hardware no muy actual y se necesitan procesos clúster de rápida ejecución.

Palabras clave: comparación; clúster; complejidad temporal; analíticas de aprendizaje; análisis estadístico implicativo; rchic.

Abstract

By analyzing data from multiple economic, social, administrative, scientific processes, etc. there is a tendency to present problems or in certain cases they become unfeasible processes since optimal and adequate techniques are not considered from the point of view of temporal complexity (execution or processing time); That is why it is important to investigate which cluster techniques are the fastest in information processing. Cluster techniques allow to form homogeneous data groups with heterogeneous groups among themselves. The present work compared, from the point of view of temporal complexity, the cluster techniques of Implicative Statistical Analysis (ASI) and those of LA (Learning Analytics). To determine which (or which) techniques are the fastest, pre-experimental qualitative research of the RGXO1 type was used, where RG represents the

experimental (random) group, X represents the treatment, which in this case are the 5 cluster techniques (3 techniques LA and 2 ASI techniques) and O is the execution time. The ASI TsimChic and TcoheChic techniques were shown to be the fastest for databases of size 100,000 or 1000 observations and 100 variables with categorical data of up to 10 categories, which is suggested to be used if you are working on not very current hardware and need fast-running cluster processes.
Keywords: comparison; cluster; temporal complexity; learning analytics; implicative statistical analysis; rchic.

Resumo

Ao analisar dados de múltiplos processos econômicos, sociais, administrativos e científicos, tende a haver problemas ou em alguns casos tornam-se processos impraticáveis porque não são considerados técnicas óptimas e adequadas do ponto de vista da complexidade temporal (tempo de execução ou processamento); é por isso que é importante investigar quais as técnicas de agrupamento que são as mais rápidas no processamento da informação. As técnicas de agrupamento permitem a formação de grupos de dados homogéneos com grupos heterogéneos. O presente trabalho comparou, do ponto de vista da complexidade temporal, as técnicas de agrupamento da Análise Estatística Implicativa (SIA) e da LA (Análise de Aprendizagem). Para determinar que técnica (ou técnicas) é (são) a mais rápida, foi utilizada uma investigação qualitativa pré-experimental do tipo RGXO1, onde RG representa o grupo experimental (aleatório), X representa o tratamento, que neste caso são as 5 técnicas de cluster (3 técnicas LA e 2 técnicas ASI) e O é o tempo de execução. As técnicas ASI TsimChic e TcoheChic mostraram ser as mais rápidas para bases de dados de tamanho 100000 ou 1000 observações e 100 variáveis com dados categóricos até 10 categorias, que são sugeridas para serem utilizadas se estiver a trabalhar em hardware pouco actual e necessitar de processos de cluster de execução rápida.

Palavras-chave: comparação; conjunto; complexidade temporal; análise de aprendizagem; análise estatística implicativa; chique.

Introducción

En la actualidad, la mayoría de autores de literatura sobre LA (Pazmiño-Maji Rubén et al., 2021), continúan adoptando la siguiente definición de las analíticas de aprendizaje (LA), ofrecida en el 1ª

Conferencia Internacional de Analítica de Aprendizaje (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011), la traducción se muestra a continuación: La Analítica de aprendizaje es la medición, recopilación, análisis y comunicación de datos sobre los estudiantes y sus contextos, a efectos de comprender y optimizar el aprendizaje y los entornos en que se producen. LA trabaja en el entorno R con tres técnicas cluster Thclustvector, Tdiana y TClustOfVar.

El conocimiento se construye con hechos y sus relaciones (Gras y Kuntz, 2009), es decir los hechos son importantes y aportan al conocimiento, el encontrar relaciones entre ellos ayuda a que el conocimiento no se lo vea en forma aislada. El Análisis Estadístico Implicativo (ASI del francés Analyse Statistique Implicative) en forma sencilla y con un fundamento teórico fuerte permite establecer relaciones asimétricas de cuasi-implicación que incrementan el conocimiento basado en hechos ya conocidos. El ASI, está formado por un conjunto de técnicas de análisis que trabajan con diversidad de variables, que en forma general tiene técnicas tales como la similaridad, implicación, cohesión y reducción, que se fortalecen con opciones adicionales como los nodos significativos, la tipicalidad y la contribución y utiliza la entropía para grandes conjuntos de datos. Se utilizó el Rchic, que es una de las formas de automatizar las técnicas del ASI (Coutrier et al., 2015; Couturier y Gras, 2005). El ASI trabaja con dos técnicas cluster TsimChic (callSimilarityTree) y TcoheChic (callHierarchyTree).

Los algoritmos deben ser capaces de resolver problemas amplios y también utilizar un menor tiempo, pero no es tan común que se encuentre algoritmos que sean capaces de cumplir esta característica, por lo que se buscan algoritmos que intentan cumplir con esto. La complejidad algorítmica temporal ayuda a describir el comportamiento de un algoritmo en términos de tiempo de ejecución, es decir, el tiempo que tarda un algoritmo en resolver un problema y por tanto permite determinar la eficiencia de dicho algoritmo, a esto se conoce como complejidad temporal (Dorta et al., 2003).

La medida del tiempo tiene que ser independiente de la máquina, del lenguaje de programación, del compilador y de cualquier otro elemento hardware o software que influya en el análisis. La complejidad temporal se expresa como T(n), en esta investigación analizamos la Tmed(n) (que expresa la complejidad temporal en el caso promedio y es una medida apropiada para la comparación) y no la Tmax(n) (que representa la complejidad temporal en el peor de los casos) ni la Tmin(n) (que trata sobre la complejidad en el mejor de los casos posibles).

La complejidad de un algoritmo se encuentra en función del tamaño del problema. A un conjunto de funciones que comparten un mismo comportamiento se denomina un orden de complejidad. Habitualmente estos conjuntos se denominan O, de esta manera se agrupan todas las complejidades que crecen de igual forma, es decir, que pertenecen al mismo orden que puede ser O(1),O(log n), O(n), etc. (Vásquez, 2004).

A continuación, indicamos algunos estudios comparativos entre las técnicas ASI y otras técnicas de análisis.

Un primer estudio se basa en el artículo de (Michael et al., 2010), donde se desea conocer las características y ventajas del método implicativo del ASI y dos métodos estadísticos de análisis: la agrupación jerárquica de variables y el análisis factorial confirmatorio (CFA). Se utilizaron los resultados en la aplicación de las tres técnicas en la aprehensión operativa de la figura geométrica, se trabajó con datos de 125 alumnos de sexto curso. Mediante el Análisis Factorial Confirmatorio, se desarrolla y verifica un modelo que proporciona información sobre el papel significativo de la modificación mereológica[5], óptica y de la forma del lugar en la aprehensión operativa de la figura geométrica. Utilizando la agrupación jerárquica de las variables, se proporciona evidencia al fenómeno de la segmentación entre las modificaciones en la aprehensión operativa de los estudiantes. En general, se encontró que los resultados de los tres métodos coinciden y pueden ser complementarios para captar las formas en que los estudiantes utilizan los diferentes tipos de modificación de la figura (Michael et al., 2010).

En el artículo (Pazmiño Maji et al., 2017), se analiza la posibilidad de que el árbol jerárquico del ASI pueda cumplir la principal función del clúster jerárquico aglomerativo que es la de agrupar objetos (además se midió el nivel de acuerdo con las agrupaciones realizadas), a las conclusiones se llegaron mediante la observación directa realizada por 35 estudiantes universitarios. Se comprobó que el 69,14% de participantes están fuertemente de acuerdo con las agrupaciones.

Sobre la complejidad algorítmica entre técnicas ASI y otras técnicas clúster, se encontraron los siguientes trabajos:

El artículo científico (Pazmiño-Maji et al., 2017) fija la metodología y características a utilizar para aplicar la comparación de la complejidad entre los árboles jerárquicos del ASI y el clúster jerárquico utilizado en LA. Las principales conclusiones a las que se llegaron con un nivel de error

[5] La mereología es, dentro de la lógica matemática y la filosofía, el estudio de las partes de un conjunto, analizando la relación de las partes entre sí y la de las partes con el todo («Mereología | Qué es, Definición y Concepto.», 2021)

del 5% fueron: las muestras son independientes debido a que son aleatorias, la homogeneidad de varianzas es falsa (con un p-valor <2,2e-16), las muestras no han sido extraídas de una población normal (con un p-valor <2,2e-16). La diferencia en la complejidad temporal entre los algoritmos de cohesión, similaridad, agnes y hclust es altamente significativa (con p-valor <2,2e-16), son necesarias post pruebas 2 a 2 en el futuro. Además, se sugiere continuar investigando con otros sistemas operativos, que se utilicen más de 100000 datos y los diferentes métodos, métricas y opciones como factores.

La elaboración de la tesis titulada Estudio comparativo del ASI y LA en relación con el uso de las técnicas de exploración de datos educativos, fue motivada por el autor de este artículo y elaborada juntamente con el Ing. Mauricio Naranjo. Los objetivos propuestos fueron (1) identificar las técnicas similares entre el ASI y LA, mediante la adaptación del método de estudio de similitud entre modelos y estándares (MSSS), (2) identificar el sistema operativo con mejor manejo de recursos y (3) identificar la técnica óptima en el análisis de datos educativos. Las principales conclusiones a las cuales se llegaron fueron: que existen técnicas similares de agrupación entre LA (dendro_variable, dendro_diana y hclust vector) y ASI (hrarchy y simlrty) y las técnicas similares de reglas de asociación entre LA (apriori, eclat, weclat) y ASI (implicativeGraph). El sistema operativo Ubuntu presenta mejor administración de los recursos, como la asignación de procesos a memoria, existe homogeneidad en el uso de memoria entre las técnicas reglas de asociación similares de LA y ASI, las técnicas de LA y ASI son similares entre ellas (la óptima weclat por ocupar menos memoria), pero esto no implica que sea el que tenga menores tiempo de respuesta. Existe homogeneidad en el tiempo de ejecución entre las técnicas reglas de asociación similares de LA y ASI (la más óptima por tener menor tiempo de respuesta es implicativeGraph) y la menos óptima pero no menos importante met_apriori (Naranjo Serrano y Pazmiño Maji, 2018).

El artículo reciente de (Fotiadis y Anastasiadou, 2019) compara las técnicas ASI (similaridad y cohesión) con el Análisis de Componentes Principales (PCA), con respecto al comportamiento del consumidor. El PCA permite el reconocimiento de patrones, es un método no supervisado, que se basa en el principio de la no existencia de información a priori (los componentes principales no se conocen de antemano), pero se logran como resultado de la aplicación del método PCA. Los componentes principales se calculan jerárquicamente. Los resultados de la aplicación de los métodos han señalado sus diferencias y similitudes, pero también su complementariedad. Se observa que la aplicación de PCA dio como resultado una reducción de datos y mostró que hay

tres componentes principales (variables latentes) que interpretan toda la variabilidad, así como los resultados del ASI en los árboles de similaridad y cohesión.

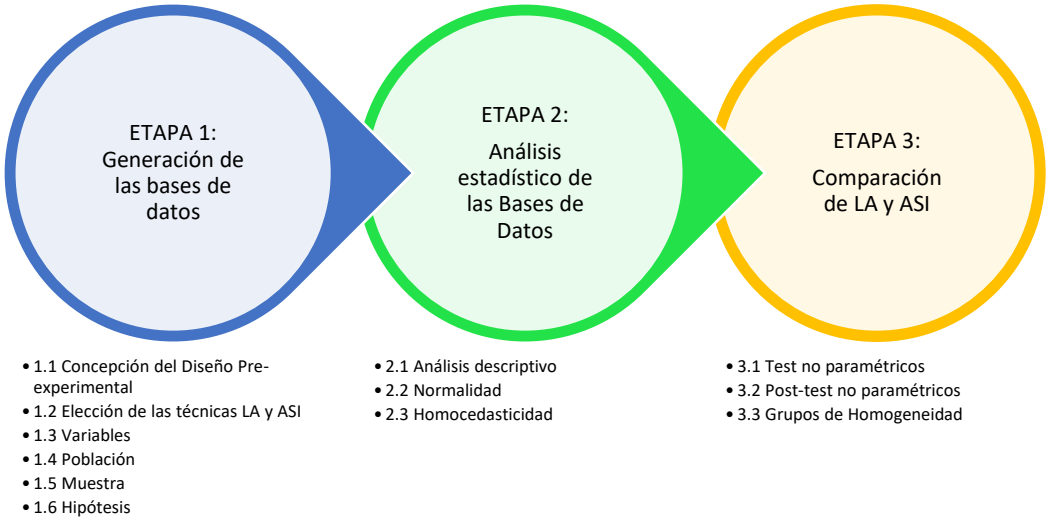
(Fotiadis y Anastasiadou, 2019) demostraron que los dos métodos (PCA y ASI) operan de manera complementaria, cada uno acentuando una dimensión diferente para la interpretación de los datos, cuya interpretación no habría sido determinante sin la participación de los especialistas en marketing.

Metodología

Por el paradigma de investigación es de tipo cuantitativo, por el tipo de diseño utilizado es pre-experimental, por el tiempo de estudio es transversal, Los materiales utilizados fueron: un computador con microprocesador: Intel® Core ™ i7-CPU @ 2,2 Ghz y 8Gb de memoria RAM, se ha instalado el sistema operativo Windows 8-64 bits, Se trabajó con el software estadístico libre R, versión 3,4,1; el entorno de desarrollo integrado RStudio, versión 1,0,143 y el paquete Rchic, versión 0,24, Las bases de datos se generaron aleatoriamente utilizando la función runif() perteneciente al paquete estándar de R, Los datos utilizados fueron categóricos generados por la función runif() y round(),

La Figura 1, muestra el proceso completo seguido en la investigación sobre la complejidad temporal entre técnicas de LA y ASI.

Figura 1: Metodología para la comparación del tiempo de ejecución entre LA y ASI



Fuente: Autores, 2022

ETAPA 1

1.1 Concepción del Diseño Pre-experimental

Para demostrar las hipótesis se planteó un pre-experimento (Connaway, 2015) en la ingeniería de software de tipo $RGXO_1$, Donde RG representa el grupo aleatorio del grupo experimental (tanto-inter como intra-grupos), X representa el tratamiento que en este caso son las 3 técnicas clúster jerárquicos utilizadas en LA y 2 técnicas usadas en ASI (Connaway y Radford, 2016).

1.2 Elección de las técnicas LA y ASI

Las técnicas utilizadas en LA fueron Thclustvector, Tdiana y TClustOfVar y las utilizadas en ASI fueron: TcoheChic y TsimChic.

1.3 Variables

La variable independiente son los métodos clúster tanto de LA (Thclustvector, Tdiana y TClustOfVar) como de ASI (TcoheChic y TsimChic), La variable dependiente fue la variable tiempo que es de tipo numérico.

1.4 Población

La población de estudio lo conforman las 100000 bases de datos aleatorias categóricas formadas por máximo 1000 observaciones y 100 variables, por la amplitud del estudio se seleccionó una muestra de 383 bases de datos aleatorias categóricas.

1.5 Muestra

Por el gran tamaño de la población, se escogió una muestra utilizando el método de muestreo aleatorio simple con parámetro de interés la media, se consideró la fórmula para el cálculo de la muestra $n = \frac{S^2}{\frac{E^2}{Z^2_{\frac{\alpha}{2}}} + \frac{S^2}{N}}$, Para aplicar la fórmula se utilizaron los parámetros desviación estándar=1; α=5%; Z=1,96; E=10,01%; N=100000 y se generó un tamaño de muestra de 383,2 que redondeado es 383.

1.6 Hipótesis

Las hipótesis estadísticas que se demostraron fueron normalidad, test de hipótesis de Kruskal-Wallis y su respectivo post test, Se trabajó con un nivel significancia de α=0,05,

La ETAPA 1, se ejecutó por completo en la Metodología, y las ETAPAS 2 y 3 se ejecutarán en la siguiente sección de resultados, la hipótesis principal por demostrar es:

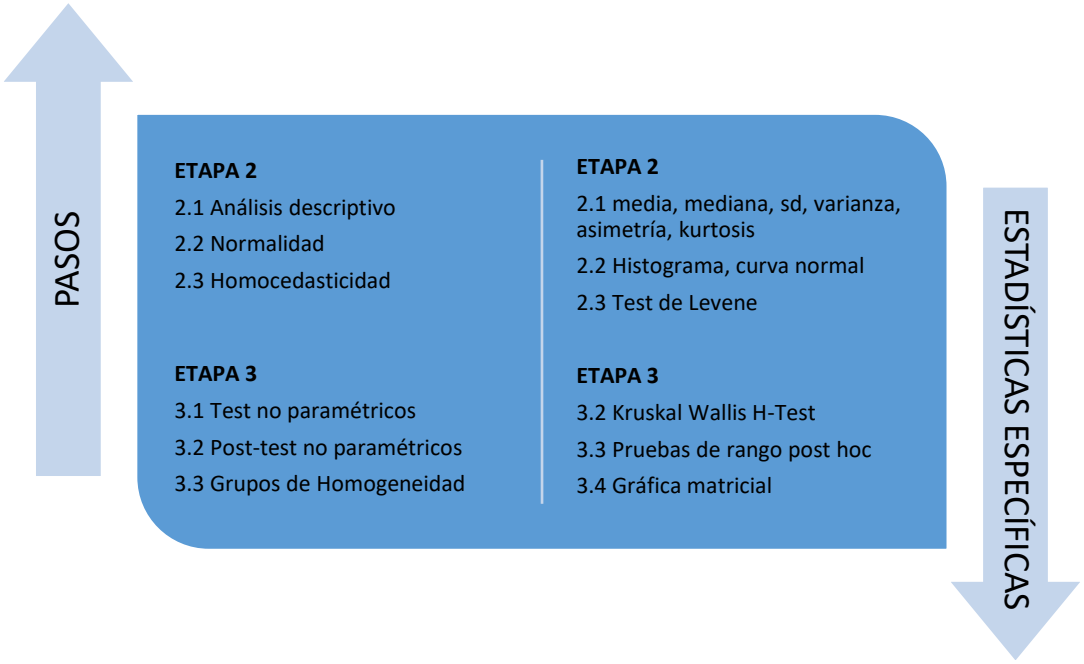
H_o: $\tilde{\mu}_{\text{TsimChic}} = \tilde{\mu}_{\text{Tcohechic}} = \tilde{\mu}_{\text{Thclustvector}} = \tilde{\mu}_{\text{Tdiana}} = \tilde{\mu}_{\text{Thclustvar}}$

$\mathbf{H_1}$: $\exists i, j = \{\text{Tsimchic}, \text{Tcohechic}, \text{Thclustvector}, \text{Tdiana}, \text{Thclustvar}\} / \tilde{\mu}_i \neq \tilde{\mu}_j$

Resultados

A continuación, se desarrollan estadísticamente los 3 pasos de la ETAPA 2 y los 3 pasos de la ETAPA 3, La **Figura 1**, muestra las dos etapas y los 6 pasos asociándose a los comandos estadísticos específicos utilizados.

Figura 2: Comandos estadísticos específicos asociados a los pasos de las ETAPAS 1 y 2



Fuente: Autores, 2022

ETAPA 2

A continuación, se desarrollan estadísticamente los 3 pasos de la ETAPA 2.

2.1 Análisis descriptivo

Se detallan cada una de las técnicas usadas para dar tratamiento a los datos obtenidos para lo cual se realiza un análisis descriptivo de los datos tiempo de ejecución.

Tabla 1, Análisis descriptivo de la variable tiempo de ejecución (medido en segundos)

	TsimChic	TcoheChic	Thclustvector	Tdiana	TClustOfVar
Media	422,41	440,31	34706,38	14236,87	852,48
Mediana	330,67	351,61	9569,41	7236,08	595,62
Sd	447,17	415,43	130598,78	17327,6	1364,17
Varianza	199959,12	172581,41	17056040731,92	300245583,58	1860963,25
Asimetría	3,42	1,58	17	1,53	11,14
Kurtosis	26,19	7,79	316,49	5,99	178,46

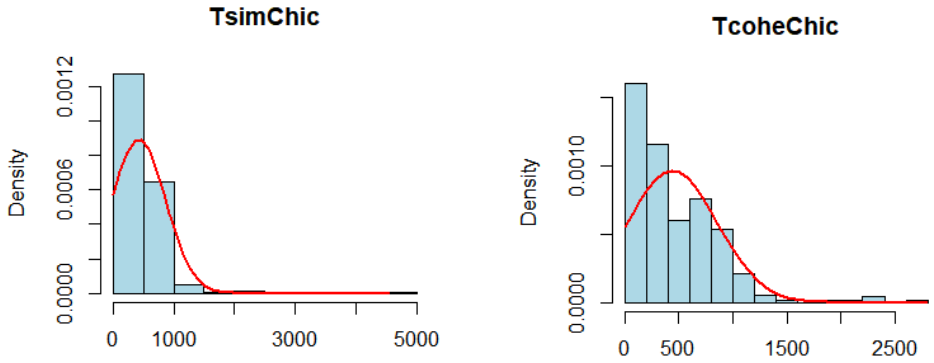
Fuente: Autores, 2022

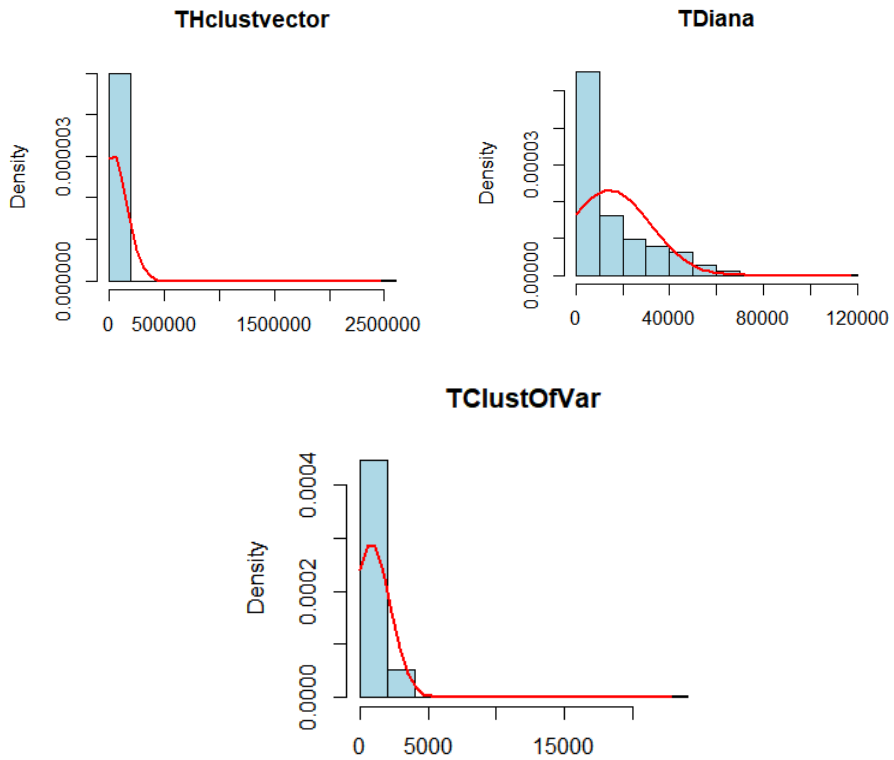
La **Tabla 1**, muestra el análisis de la variable tiempo de ejecución para los métodos clúster de SIA y LA de los cuales podemos decir que en promedio el método que ocupa menor tiempo de ejecución es TcoheChic (función callHierarchyTree) con 422,41 segundos con una mediana que indica que la mitad del tiempo empleado para el método es menor o igual a 351,61 y la otra mitad es mayor o igual a 351,61 , con respecto al método que ocupa mayor tiempo de ejecución Thclustvector (hclustvector) con 34706,38 segundos se tiene que el valor de la mediana indica que la mitad del tiempo empleado para el método es menor o igual a 9569,41 y la otra mitad es mayor o igual a 9569,41, con relación a que tan dispersos se encuentran dichos datos analizados con respecto al valor promedio se obtuvo que presenta menor dispersión TcoheChic (callHierarchyTree) con un valor igual a 415,43 y mayor dispersión Thclustvector con 130598,78. La asimetría de los datos con respecto al tiempo de ejecución permite notar que para todos los métodos en comparación la asimetría es positiva ya que todos los coeficientes son mayores a uno, obteniendo el valor máximo Thclustvector con 17 y Diana siendo el de menor proporción con un valor de 1,53, El coeficiente de Kurtosis refleja que la distribución que sigue cada uno de los métodos es leptocúrtico debido a que el coeficiente obtenido para cada uno es positivo lo cual quiere decir que hay una mayor concentración de los datos en torno a la media.

2.2 Normalidad

A continuación, se muestran los cinco histogramas de los datos de tiempo y la aproximación por la curva normal, para las cinco técnicas estudiadas (Ver **Figura 3**).

Figura 3: Histogramas de los datos de tiempo y la aproximación por la curva normal





Fuente: Autores, 2022

En la sección 2.1 se evaluaron los coeficientes de Asimetría y Kurtosis en donde se determina que presentan una asimetría positiva con Kurtosis platicúrtica, dicha aseveración se confirma también de manera gráfica (Ver **Figura 3**) dando la idea de que posiblemente los datos no siguen una distribución normal.

2.3 Homocedasticidad

Para comprobar la homocedasticidad o prueba de igualdad de varianza se utilizó el Test de Levene. El planteamiento de las hipótesis estadísticas se muestra a continuación.

H_o: $\sigma^2_{\text{TsimChic}} = \sigma^2_{\text{Tcohechic}} = \sigma^2_{\text{Thclustvector}} = \sigma^2_{\text{Tdiana}} = \sigma^2_{\text{Thclustvar}}$

H_1:$\exists i, j = \{\text{Tsimchic}, \text{Tcohechic}, \text{Thclustvector}, \text{Tdiana}, \text{Thclustvar}\}/\sigma_i^2 \neq \sigma_j^2$

En nivel de significancia utilizado fue de $\alpha = 0{,}05$, el resultado del estadístico de prueba es:

```
Levene's Test for Homogeneity of Variance (center = "median")
       Df F value                Pr(>F)
group   4  22.938 < 0.00000000000000022 ***
      1905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Utilizando la siguiente regla, si el p-valor es menor que 0,05 (p-value < 0,05) entonces se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla. La decisión a tomar es que debido a que el p-value obtenido es igual a 0,0000000000000022 el cual es menor al nivel de significancia propuesto, por lo tanto se rechazó la hipótesis nula (Ho) y se concluye que las varianzas de los grupos de tiempo de ejecución no son iguales, los datos sobre el tiempo de ejecución para cada uno de los métodos clúster son heterocedásticos (o no homocedásticos).

ETAPA 3

A continuación, se desarrollan estadísticamente los 2 pasos de la ETAPA 3. Al no cumplir con los supuestos de normalidad (gráficamente) y homocedasticidad (con prueba de hipótesis) se determinó que no se pueden utilizar métodos paramétricos, se comprueba el uso de pruebas no paramétricas.

3.1 Test no paramétricos

Una vez analizados los prerrequisitos se concluyó que no se cumple con los supuestos, por lo que se procedió a realizar una prueba no paramétrica para muestras independientes. Se realizó el planteamiento de hipótesis para tiempo de ejecución a través de la prueba de Kruskal Wallis H-Test:

H_o: $\tilde{\mu}_{\text{TsimChic}} = \tilde{\mu}_{\text{Tcohechic}} = \tilde{\mu}_{\text{Thclustvector}} = \tilde{\mu}_{\text{Tdiana}} = \tilde{\mu}_{\text{Thclustvar}}$

H_1: $\exists i, j = \{\text{Tsimchic}, \text{Tcohechic}, \text{Thclustvector}, \text{Tdiana}, \text{Thclustvar}\} / \tilde{\mu}_i \neq \tilde{\mu}_j$

El nivel de significancia utilizado fue de α = 0,05. Los resultados del estadístico de prueba, se muestra a continuación.

```
        Kruskal-Wallis rank sum test

data:  Datos by Metodo
Kruskal-Wallis chi-squared = 465.34, df = 4, p-value < 0.00000000000000022
```

La regla de decisión, utilizada fue si p-value < 0,05 entonces se rechaza Ho, la decisión realizada es que con un p-valor igual a 0,00000000000000022 el cual es menor a un nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula y se concluye que al menos una de las medianas de los métodos clúster del tiempo de ejecución son diferentes.

3.2 Post-test no paramétricos

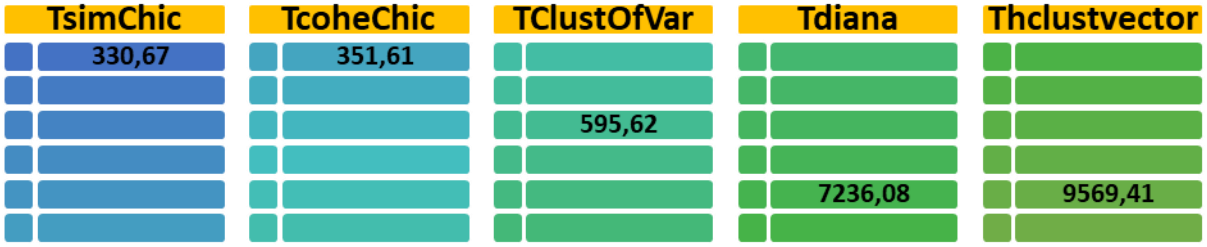
Se realiza las pruebas de rango post-test donde se verificó que existe diferencia entre los métodos de tiempo de ejecución a excepción del método TcoheChic y TsimChic, además de los métodos TDiana y THclustvector.

```
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                              obs.dif critical.dif difference
TClustOfVar-TcoheChic      154.70157     112.0179       TRUE
TClustOfVar-TDiana         401.88743     112.0179       TRUE
TClustOfVar-THclustvector  458.52094     112.0179       TRUE
TClustOfVar-TsimChic       175.96859     112.0179       TRUE
TcoheChic-TDiana           556.58901     112.0179       TRUE
TcoheChic-THclustvector    613.22251     112.0179       TRUE
TcoheChic-TsimChic          21.26702     112.0179      FALSE
TDiana-THclustvector        56.63351     112.0179      FALSE
TDiana-TsimChic            577.85602     112.0179       TRUE
THclustvector-TsimChic     634.48953     112.0179       TRUE
```

3.3 Grupos de homogeneidad

Los grupos de homogeneidad se muestran a continuación.

Figura 4: Gráfico de homogeneidad



Fuente: Autores, 2022

En la **Figura 4,** se observa que las técnicas de análisis que menos tiempo de ejecución tienen son TsimChic y TcoheChic, que son los mejores. Los peores métodos desde el punto de vista del tiempo de ejecución son Tdiana y Thclustvector.

Conclusiones

El diseño pre-experimental utilizado permitió determinar la técnica clúster que menos tiempo de ejecución utiliza. Además, se demostró mediante una prueba de hipótesis que no se cumple la homocedasticidad, gráficamente se puede observar que ninguna de las técnicas cumple normalidad. El test no paramétrico de Kruskal Wallis y las post pruebas demostraron la existencia de diferencias significativas entre los grupos de tiempo de ejecución de las técnicas clúster en ASI y LA en donde se identificó como al método que ocupa menor tiempo a TsimChic (callSimilarityTree) y TcoheChic (callHierarchyTree) de ASI siendo considerados los óptimos mientras que los no óptimos serían Thclustvector y Tdiana de LA debido a que registran mayor consumo de tiempo.

Referencias bibliográficas

1. Connaway, L. S. (2015). Retos de la investigación: El camino hacia el compromiso y el progreso. *BiD: textos universitaris de biblioteconomia i documentació*, *35*.
2. Connaway, L. S., y Radford, M. L. (2016). *Research methods in library and information science*. ABC-CLIO.
3. Coutrier, R., Pazmiño Maji, R. A., Conde González, M. Á., y García-Peñalvo, F. J. (2015). *Statistical implicative analysis for educational data sets: 2 analysis with RCHIC*.
4. Couturier, R., y Gras, R. (2005). *CHIC: traitement de données avec l'analyse implicative*. 679-684.
5. Dorta, I., León, C., Rodríguez, C., Rodríguez, G., y Rojas, A. (2003). Complejidad Algorítmica: De la Teoría a la Práctica. *III Jornadas de Enseñanza Universitaria de Informática*.
6. Fotiadis, T. A., y Anastasiadou, S. (2019). *Contemporary advanced statistical methods for the science of marketing: Implicative Statistical Analysis vs Principal Components Analysis*.
7. Gras, R., y Kuntz, P. (2009). El Análisis Estadístico Implicativo (ASI) en respuesta a problemas que le dieron origen. *Teoría y aplicaciones del Análisis Estadístico Implicativo: primera aproximación en lengua hispana. Castellón: Departamento de Matemática de la Universitat Jaume I*, 3-51.
8. *LAK 2011: 1st International Conference Learning Analytics and Knowledge*. (2011). http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=11606
9. Mereología | Qué es, Definición y Concepto. (2021). *Enciclopedia Online*. https://enciclopediaonline.com/es/mereologia/
10. Michael, P., Elia, I., Gagatsis, A., y Kalogirou, P. (2010). *Examining primary school students' operative apprehension of geometrical figures through a comparison between the hierarchical clustering of variables, implicative statistical analysis and confirmatory factor analysis*. Citeseer.
11. Naranjo Serrano, M. M., y Pazmiño Maji, R. A. (2018). *Estudio comparativo del anàlisis estadìstico implicativoy el Learning Analytics en relaciòn al uso de las tècnicas de exploracoòn de datos educativos*. http://repositorio.pucesa.edu.ec/handle/123456789/2387

12. Pazmiño Maji, R., García Peñalvo, F. J., y Conde González, M. Á. (2017). *Is it possible to apply Statistical Implicative Analysis in hierarchical cluster Analysis? Firsts issues and answers*.
13. Pazmiño-Maji, R., García-Peñalvo, F. J., y Conde-González, M. A. (2017). *Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics*. 1-7.
14. Pazmiño-Maji Rubén, Conde-Gonzales M.A., y Garcia-Penalvo F.J. (2021). *What are Learning Analytics ?: Analysis from its definition*. *1*, 10.
15. Vásquez, A. C. (2004). Teoría de la complejidad computacional y teoría de la computabilidad. *Revista de investigación de Sistemas e Informática*, *1*(1), 102-105.