ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"ANALÍTICAS DE APRENDIZAJE Y ANÁLISIS ESTADÍSTICO IMPLICATIVO: COMPARACIÓN DE TÉCNICAS CLUSTERS PARA VARIABLES MODALES"

Trabajo de Titulación

Tipo: Trabajo Experimental

Presentado para optar al grado académico de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

AUTOR: SHIRLEY ESTEFANIA ARMAS ANALUISA

DIRECTOR: DR. RUBÉN ANTONIO PAZMIÑO MAJI

Riobamba - Ecuador

2022

Yo, Shirley Estefania Armas Analuisa**,** declaro que el presente Trabajo de Titulación es de mi autoría y los resultados de este son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este Trabajo de Titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 10 de marzo de 2022

Shirley Estefania Armas Analuisa
180521723-7

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

El Tribunal del Trabajo de Titulación certifica que: El Trabajo de Titulación: Tipo: Trabajo Experimental, **"ANALÍTICAS DE APRENDIZAJE Y ANÁLISIS ESTADÍSTICO IMPLICATIVO: COMPARACIÓN DE TÉCNICAS CLUSTERS PARA VARIABLES MODALES"**, realizado por la señorita: **SHIRLEY ESTEFANIA ARMAS ANALUISA**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

	FIRMA	**FECHA**
Ing. Natalia Alexandra Pérez Londo **PRESIDENTE DEL TRIBUNAL**		2022-03-10
Dr. Mat. Rubén Antonio Pazmiño Maji **DIRECTOR DEL TRABAJO DE TITULACIÓN**		2022-03-10
Ing. Lourdes Emperatriz Paredes Castelo **ASESORA DEL TRABAJO DE TITULACIÓN**		2022-03-10

DEDICATORIA

A mis padres Gloria Analuisa e Iván Efraín Armas por haberme apoyado en todo este proceso de formación tanto de manera económica como emocional, a ustedes por siempre mi agradecimiento sincero, en especial a mi madre por nunca haberme dejado sola y por siempre haberme estado motivando a no decaer y continuar a pesar de todo hasta el final, por enseñarme a creer en mí misma y por defenderme de todos y, ante todo, este logro en mi vida es para usted con todo mi corazón.

A mis amados hermanos Iván y Sebas por haberme alentado a seguir siempre y por haber confiado en mí y en que lo lograría, por estar orgullosos de mí y yo de ellos, por ser el pilar en mi vida y mi fortaleza, los amo con mi vida.

Al ingeniero Douglas Alejandro y familia, por su empuje y ayuda en los momentos más difíciles, mi mayor gratitud a mami Ruchi y su esposo por siempre habernos apoyado, y aunque ya no están para compartir este momento conmigo; ese sueño del que un día me habló al partir de su hogar; hoy se ha materializado, un abrazo hasta el cielo y mi eterno agradecimiento por siempre.

A mi hermosa familia de cuatro patas: Bella, Sheyco, Sasha, Nana, Luna, Gizmo, Laia, Princesa, Annie, Nucita, Caramelo, Manchas, Negrita, Gatito, Baco y Lia que son mi refugio, alegría y compañía, a mis angelitos que recuerdo siempre Tobby, Doki y Tere, en especial a mi Tobby por cambiar totalmente mi perspectiva de la vida, por haberme amado tanto y enseñado a amar sin esperar nada a cambio, desde aquí hasta el cielo gracias mi bebé.

A don Carlitos Sánchez, por haber sido ese apoyo incondicional en los momentos buenos y malos, por haberme extendido su mano siempre, por cada abrazo, por escucharme cuando lo necesitaba y por nunca haberme dejado sola. Gracias por ser una luz en medio de la obscuridad. ¡Dios lo bendiga siempre!

Finalmente, a mi Don Lopecito que sé que nos cuida desde el cielo, a mamá Olguita; mis amados abuelitos que me dio la vida y a Armando López, un gracias sincero por su enorme cariño, motivación y apoyo.

Shirley

AGRADECIMIENTO

Primeramente, agradezco a Dios por brindarme la salud y no abandonarme nunca, lo cual me ha permitido llegar hasta donde estoy.

A la Escuela Superior Politécnica de Chimborazo por darme la oportunidad de formarme como profesional y a mis maestros por compartir sus conocimientos.

Un agradecimiento muy especial al Doctor Rubén Pazmiño Maji director del trabajo de titulación y a quién tengo un gran aprecio y admiración ya que con su tiempo, experiencia, cooperación y apoyo me ha orientado y ayudado hasta el final de mi formación profesional para culminar con éxito la presente investigación.

A la Ingeniera Lourdes Paredes por su apoyo como miembro del trabajo de titulación y por a lo largo de mi carrera haber compartido sus conocimientos conmigo.

Al grupo de investigación Ciencia de Datos - CIDED por haberme brindando la oportunidad de realizar mis prácticas pre-profesionales, por la guía brindada y finalmente por haberme proporcionado las bases de datos para mi trabajo de titulación.

A mis padres Gloria e Iván que me han cuidado y apoyado siempre, que me han enseñado a seguir en todo aspecto de mi vida.

A mis hermanos Iván y Sebas por ser mi fortaleza y confiar en mí, por sacarme risas y animarme a lograrlo.

Al Ing. Douglas Alejandro y familia por en su momento habernos brindado su ayuda lo cual fue muy importante en mi vida para llegar a cumplir este logro.

A quiénes en lo largo de mi carrera estuvieron conmigo, en los momentos alegres y muchos más en los momentos tristes, a quién no me dejó sola y estuvo gran parte de mi carrera compartiendo su conocimiento y calidad humana conmigo, ¡millón gracias!

Shirley

INDICE DE CONTENIDO

CAPÍTULO II

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

ÍNDICE DE ABREVIATURAS

AA	Analíticas de Aprendizaje
LA	Learning Analytics
AEI	Análisis Estadístico Implicativo
EVA	Entornos Virtuales de Aprendizaje
EDM	Educational Data Mining
CHIC	Clasificación Jerárquica Implicativa y Cohesiva
KDD	Knowledge Discovery in Database
EDA	Exploratory Data Analysis
IDE	Entorno de Desarrollo Integrado
UPS	Participación de Estadísticas Implicativas
DIANA	DIvisive ANAlysis Clustering
CFA	Análisis factorial confirmatorio

RESUMEN

En un mundo cada vez más digitalizado el análisis de datos ha ido ganando relevancia día a día ya que su estudio se ha convertido en un aspecto fundamental e imprescindible al momento de brindar soluciones estratégicas para la toma de decisiones, por tal motivo al no utilizar las técnicas adecuadas para el tratamiento de esta información se vienen a crear procesos repetitivos y en ciertos casos hasta irresolubles, ocasionando lentitud en los cálculos y el no aprovechamiento de recursos como tiempo de ejecución y espacio de memoria. Dada esta problemática esta investigación propuso realizar un análisis comparativo del tiempo de ejecución y espacio de memoria utilizado entre los métodos clúster similares al Análisis Estadístico Implicativo (ASI) y Learning Analytics (LA). Para ello se utilizó un diseño preexperimental con variables modales considerando para ASI los métodos callHierarchyTree y callSimilarityTree y para LA los métodos hclust.vector, hclustvar y diana. El grupo de estudio fue de 100000 bases de datos, el tipo de muestreo fue aleatorio simple con un tamaño de muestra de 382 bases de datos categóricas (de hasta 10 categorías) las cuales fueron generadas aleatoriamente. Se demostró que no existe diferencia significativa en espacio de memoria entre los cinco métodos, con respecto al tiempo de ejecución se determinó que los métodos que ocupan menor tiempo son callHierarchyTree y callSimilarityTree seguido por hclustvar y finalmente los métodos que ocuparon más tiempo fueron hclust.vector y diana. Estos resultados presentados servirán para utilizar un método clúster más eficiente en tiempo de ejecución y espacio de memoria al analizar datos, permitiendo así optar por algoritmos óptimos desde el punto de vista de la complejidad algorítmica.

Palabras clave: <ANÁLISIS ESTADÍSTICO IMPLICATIVO>, <ANALÍTICAS DE APRENDIZAJE>, <VARIABLES MODALES> <ANÁLISIS CLÚSTER>, <SIMILARIDAD>, <COHESION>.

ABSTRACT

In an increasingly digitalized world, data analysis has been gaining relevance day by day due to its study has become a fundamental and essential aspect when providing strategic solutions for decision making, therefore, not using the appropriate techniques for the treatment of this information creates repetitive and, in some cases, even unsolvable processes, causing slowness in the calculations and not taking advantage of resources such as execution time and memory space. Given this problem, this research proposed to perform a comparative analysis of the execution time and memory space used between clustering methods similar to the Implicative Statistical Analysis (ASI) and Learning Analytics (LA). For this purpose, a pre-experimental design with modal variables was used, considering for ASI the callHierarchyTree and callSimilarityTree methods and for LA the hclust. vector, hclustvar and diana methods. The study group was 100000 databases, the type of sampling was simple random with a sample size of 382 categorical databases (up to 10 categories) which were randomly generated. It was demonstrated that there is no significant difference in memory space between the five methods, with respect to execution time it was determined that the methods that take the least time are callHierarchyTree and callSimilarityTree followed by hclustvar and finally the methods that took the most time were hclust, vector and diana. These results presented will serve to use a more efficient clustering method in execution time and memory space when analyzing data, thus allowing to opt for optimal algorithms from the point of view of algorithmic complexity.

Keywords: <STATISTICAL IMPLICATIVE ANALYSIS>, <LEARNING ANALYTICS>, <MODAL VARIABLES>, <CLUSTER ANALYSIS>, <SIMILARITY>, <COHESION>.

Edgar Mesías Jaramillo Moyano
0603497397

INTRODUCCIÓN

Learning Analytics como lo menciona (Lias y Elias, 2011: p. 1-5) es un área de aprendizaje que se ha venido potenciando durante los últimos años con el desarrollo de la tecnología y con la continua interacción entre alumnos y docentes , generando así una gran cantidad de datos que se mueven en línea y que aportan información relevante sobre todo este conjunto de interacciones, partiendo de esta temática surge la necesidad de su análisis con técnicas óptimas con el fin de mejorar la enseñanza y aportar a una mejor toma de decisiones futuras en procesos de aprendizaje efectivos. (Rojas-Castro, 2017a, pp. 106–128) hace mención de que las analíticas de aprendizaje son un campo emergente que hacen uso de herramientas de análisis sofisticados para mejorar el aprendizaje y la educación ya que permiten capturar, informar, procesar y actuar sobre los datos históricos y actuales para aportar en el uso de procesos efectivos para la mejora de la calidad de educación.

El libro Learning Analytics: La narración del aprendizaje a través de los datos (Amo y Santiago,2017: p. 5-20) indica que es fundamental considerar el tiempo en el que es adquirida la información ya que al comprender el pasado se podrá obtener un por qué de los comportamientos y así mismo esto será clave para brindar soluciones a futuro debido a que se podrán otorgar herramientas que permitan entender comportamientos de decesión en los estudios por parte de los alumnos.

Por otra parte el Análisis estadístico Implicativo busca ir a la par con los desafíos que se imponen hoy en día, ya que sus primeras aplicaciones vienen dadas en la matemática en el ámbito educativo concluyéndose así que existen gran cantidad de experiencias en dicha área lo cual permitirá el uso de nuevos enfoques y modelos para obtener un aprendizaje óptimo con todas las herramientas formales necesarias, las cuales serán la puerta para un gran sin número de técnicas de aprendizajes imponderables que considerarán todos los factores que vienen causando problemas de decesión siendo aquellos los que se quieren diagnosticar para poder promocionar el aprendizaje al momento de obtener los datos en dichas áreas (Pazmiño, Mullo, Conde 2019a, pp. 24-39).

Es importante considerar que el AEI va tomando desarrollo en cuestión a problemas encontrados ya que se basa en generar la estructuración de datos en relación a sujetos con sus respectivas variables ya sean estas de tipo binario o como se ha ido añadiendo progresivamente de otros tipos siendo frecuenciales, de intervalo, difusas o como en el presente estudio se centrará en el tratamiento de las de tipo modal considerando que el AEI trabaja bajo los conceptos de cohesión e implicación (Gras et al. 2009, pp. 3-50) .

Al analizar datos provenientes de múltiples procesos educativos ya sean evaluaciones, lecciones, uso de plataformas tecnológicas , etc se tiende a presentar problemas o en ciertos casos llegan a ser procesos irrealizables debido a que no se consideran técnicas óptimas y adecuadas acorde con el tipo de datos en estudio para la optimización de espacio de memoria y tiempo de procesamiento; es por esto que es importante un correcto empleo de técnicas apropiadas según la cantidad y el

tipo de datos para no caer en la temática de redundancia de metodologías lo cual puede generar procesos que duren horas provocando lentitud en los cálculos causando así que las técnicas de análisis que se utilicen lleguen a ser inaplicables para ciertos casos. Habiendo mencionado lo anterior radica la importancia del estudio comparativo realizado para obtener técnicas óptimas para el análisis de datos similares empleadas en el AEI y AA ya que se busca determinar un método clúster óptimo que contribuya en la optimización de tiempo y de memoria de ejecución para el tratamiento de variables modales, esto con el objetivo de brindar alternativas al momento de trabajar con grandes cantidades de datos ya que el método clúster que será obtenido como óptimo contribuirá en la optimización de memoria y de tiempo de ejecución para las variables en cuestión.

Este trabajo está dividido en tres capítulos los cuáles se describen a continuación:

En el **Capítulo I** se desarrolla la base teórica para el presente trabajo experimental, en el cual se presenta información sobre Analíticas de aprendizaje (AA) y sobre el análisis estadístico implicativo (AEI), se describen y conceptualizan todos los métodos y técnicas estadísticas empleadas, así mismo se describe cada función de RStudio utilizada.

En el **Capítulo II** se realiza el planteamiento de la parte metodológica de la investigación, el tipo y diseño de investigación llevada a cabo, población en estudio, materiales y métodos empleados, operacionalización de variables, diseño y experimentación y procedimiento experimental. Se detalla también la elaboración de funciones para cada técnica de análisis clúster planteada para el tiempo y cantidad de memoria en ambas áreas de AA y AEI y la función general para medir los parámetros planteados para cada una de las bases de datos utilizada.

En el **Capítulo III** se exponen los resultados que sustentan el trabajo experimental, mediante tablas se presenta un análisis descriptivo de los datos, seguido del análisis de supuestos como normalidad y homocedasticidad, aproximación a normalidad, y la prueba no paramétrica de Kruskall Wall con su respectiva post prueba para conjuntamente determinar los métodos que optimicen tiempo y espacio de memoria.

Para concluir en la última sección se presentan las conclusiones y recomendaciones elaboradas del trabajo experimental , se presenta la bibliografía que fue empleada para la investigación y en el apartado de Anexos se puede visualizar los códigos realizados para el desarrollo de las funciones que permitieron obtener la base de datos final que comprende el tiempo y cantidad de memoria en 3 iteraciones para cada una de las 382 bases de datos, también se muestra una guía de instalación para el software utilizado.

Antecedentes

Antecedentes metodológicos

Existen diversos estudios sobre la comparación de los tiempos empleados en el procesamiento asimismo como del espacio de memoria entre las técnicas de agrupación clúster más usadas en

AEI y LA. En el siguiente apartado se hace mención algunas investigaciones por su similaridad en las técnicas y por el enfoque de profundizar en el conocimiento de LA y AEI, recalcando la importancia que viene presentando día a día conjuntamente de la mano del desarrollo tecnológico y por ende la gran cantidad de información que se viene manejando.

En el artículo titulado "Learning Analytics: una revisión de la literatura" se identifica que a partir del siglo XXI la tecnología web ha dado un giro impresionante que ha traído consigo la posibilidad de interactuar a los usuarios, permitiendo así obtener nueva información acerca de las actividades que realizan en la web, así mismo este desarrollo permitió la incorporación de entornos virtuales de aprendizaje EVA lo cual generó un aporte significativo de información tomando en cuenta cantidad y tipo de datos. Considerando este aspecto se menciona que surge la profundización de LA para mejorar los entornos de aprendizaje ya que este campo considera distintas técnicas y métodos para comprender y optimizar el conocimiento en las aulas con la finalidad de brindar soluciones óptimas y evitar el uso de malas técnicas ya que esto produce ambigüedad en los métodos e incluso lentitud en los cálculos, se debe tener presente que el LA abarca aspectos de medición, recopilación, análisis y presentación de datos (Rojas-Castro, 2017b, pp. 106–128).

Otro artículo muy similar al anterior titulado "Comparison of Clustering Algorithms for Learning Analytics with Educational Datasets" identifica también a LA como una herramienta clave para el análisis y tratamiento de información ya que se reconoce a esta área como fundamental en el mejoramiento de procesos con técnicas adecuadas para el aprendizaje dado de la mano con el desarrollo digital que se ha venido presentando en los últimos tiempos. El gran manejo de datos causa la necesidad de usar técnicas de análisis y herramientas adecuadas que no provoquen lentitud en los procesos es por eso por lo que se identificaron los algoritmos clúster jerárquicos y de partición más relevantes en LA y se efectuó una comparación entre ellos para determinar cuál posee mejor funcionalidad. Se analizaron siete algoritmos de los cuáles se determinó que para los algoritmos jerárquicos DIANA posee mejor rendimiento mientras que en los algoritmos de partición; K-means y PAM fueron los mejores. Para esto emplearon diferentes softwares de trabajo , en cuanto a manejo de datos se utilizó los archivos de Excel en formato CSV (valores separados por comas) y para análisis estadísticos se usó la plataforma de acceso libre R (Navarro y Ger, 2018, pp. 9–16).

El artículo de investigación titulado "Learning Analytics in Ecuador: An Initial Analysis based in a Mapping Review" menciona la importancia de LA ya que recalca que permite describir, diagnosticar, predecir y prescribir el aprendizaje con grandes cantidades y tipos de datos generados a nivel de tipo educativo especialmente en la educación Superior, se menciona que desde el 2014 al 2019 la línea de interés para LA ha presentado un gran crecimiento lo cual permitió descubrir características específicas de las investigaciones en Learning Analytics

accediendo así a responder a interrogantes sobre como LA optimiza el aprendizaje y entorno en el que se desarrolla. El tipo de metodología que se usó fue una revisión sistemática de mapeo de la literatura con preguntas de investigación claras y con un enfoque bien marcado (Pazmiño-Maji et al. 2019, pp. 304-311).

El artículo titulado "La vida social de Learning Analytics: análisis de conglomerados y el 'rendimiento' de la educación algorítmica" menciona que los métodos usados para el análisis de datos no suelen ser precisos ya que no consideran la cantidad y tipo de datos en estudio, debido a esta problemática surge la necesidad de implementar LA ya que se denota como una tecnología en desarrollo permitiendo el uso de metodologías adecuadas que buscan resolver un problema con análisis detallados a través de mecanismos, técnicas y algoritmos oportunos, se hace énfasis en el uso de análisis de conglomerados ya que al profundizar la investigación en dicha área se pueden encontrar métodos adecuados que guardan relación entre el tipo de datos en estudio, considerando así a este tipo de técnicas como óptimas (Perrotta y Williamson, 2018, pp. 3–16).

En un artículo de la revista Identidad Bolivariana titulado "El análisis estadístico implicativo como estrategia para la promoción del aprendizaje en la educación media: simulaciones para su aprendizaje" (Pazmiño, Mullo y Conde, 2019b, pp. 24–39) se identifica al AEI como una posibilidad de ofrecer soluciones óptimas a través del uso de técnicas adecuadas en el tratamiento de datos mediante la similaridad e implicación, brindando así la posibilidad de emitir diagnósticos y promocionar de una manera adecuada el aprendizaje. Se usó metodología cuantitativa, descriptiva, transversal y no experimental para el tratamiento de los datos generados mediante encuestas y simulaciones en donde se concluyó que al realizar el estudio con métodos adecuados se obtendrán mejores interpretaciones que ayudarán a brindar soluciones favorables y contextualizadas de progreso para motivación en el aprendizaje, se recalca que el uso del software R es importante porque permite el uso del paquete Rchic con menor dificultad (Pazmiño, Mullo y Conde, 2019b, pp. 24–39).

Antecedentes aplicativos

El LA y AEI son una serie de metodologías que permiten el estudio y tratamiento de grandes cantidades de datos, permiten conocer y dar seguimiento a acciones de alumnos y profesores por la información que se llega a obtener, pudiendo dar así solución y técnicas adecuadas a temáticas de estudio. Hoy en día se ve reflejada esta necesidad debido a que por la situación actual de pandemia que se atraviesa se han visto generados ambientes de estudio digitales siendo esto causante de generación de gran cantidad de datos que requieren ser estudiados con más profundidad y con las técnicas adecuadas, por esta razón y por el desarrollo que ha venido presentando el LA y AEI varios investigadores se han visto en la necesidad de profundizar en el tema. A continuación, se detallan algunas de estas investigaciones:

En la tesis titulada “Estudio Comparativo del análisis estadístico implicativo y el Learning Analytics en relación al uso de las técnicas de exploración de datos educativos” se recalca que el uso de técnicas óptimas según el tipo y cantidad de datos es imprescindible ya que de dicha manera se puede optimizar procesos que ocasionan lentitud en los cálculos llegando en ciertos casos a ser inaplicables, es así que para brindar una solución a esta temática se realizó el planteamiento de hipótesis de que existe diferencia significativa entre el espacio de memoria y tiempo de ejecución de los algoritmos de los cuáles se efectuó la comparación en LA y AEI probándose a través de un cuasiexperimento de tipo RGXO dando como resultado que el método que entrega los resultados en menos tiempo es el método hclust_vector de LA con un uso de memoria razonable en comparación a los demás métodos clustering usados los cuáles presentan homogeneidad en el tiempo de ejecución siendo así los siguientes: de Learning Analytics los métodos dendro_diana, dendro_variable y hclust_vector y en AEI los métodos callHierarchyTree y callSimilarityTree que se plantearon como funciones hrarchy y simlrty (Naranjo Serrano 2018, pp. 10-62).

En la tesis “Aportes del análisis Estadístico Implicativo a Learning Analytics” se destaca como se ha ido desarrollando la analítica de aprendizaje en los últimos años la cuál brinda conocimiento de cómo se desarrolla el entorno de los estudiantes y los contextos que los rodean guardando estrecha relación entre el aprendizaje y la educación, dichos aspectos permiten profundizar la comparación de complejidad algorítmica entre las técnicas comunes de LA y ASI todo bajo la normativa de la frecuente necesidad de trabajar con grandes cantidades de datos en diferentes tipos ; con este propósito se obtiene un diseño preexperimental del tipo un solo grupo aleatorio de la forma RGX01 en donde mediante el estudio descriptivo realizado se obtuvo que la técnica clúster que usa en promedio menor cantidad de memoria es hclust_vector con un valor de 233.0 y por ende dicha técnica resulta también en ser la de menor uso del recurso tiempo con un valor de 0.242727, con respecto al análisis de supuestos se tuvo que tanto los datos de tiempo y cantidad de memoria no son normales para ninguna de las pruebas planteadas a niveles de significancia de α=0,01, α=0,05 y α=0,1, sobre igualdad de varianzas se corroboró que existía homocedasticidad en los datos de tiempo pero no en los de memoria y finalmente las técnicas clúster (hclust_vector, dendro_diana, hrarchy, Similrty, dendro_variables) para memoria y tiempo no son independientes (Pazmiño Maji 2021a, pp. 32-50).

En el artículo titulado “Métodos de agrupamiento LA & SIA: Comparación computacional” recalca la importancia de las analíticas de aprendizaje como una tecnología emergente que va en desarrollo día a día, es por dicha temática que surge la necesidad del estudio y nuevas técnicas de análisis considerando el tipo y cantidad de datos, con respecto al análisis estadístico implicativo se menciona que trabaja sobre un conjunto de variables o sujetos de los cuáles permite descubrir reglas de la forma $a \rightarrow b$ pudiendo representar estás reglas de manera gráfica mediante

dendogramas. Cumpliendo el objetivo de comparar la ocupación de memoria se observó que usaron las funciones clúster más utilizadas en Learning Analytics como son hclust.vector, dendro.variables y diana y las funciones callHierarchyTree y callSimilarityTree utilizadas en el Análisis Estadístico Implicativo de las cuales mediante un análisis comparativo bajo un diseño cuasi experimental bifactorial del tipo RGXO con un tamaño de muestra de 383 bases de datos y con arquitecturas de hardware de características como procesador Core I7, Velocidad 2,2 Ghz y Memoria RAM 8Gb y de software Windows 8 con versión de instalación de R en R v3.4.1 y RStudio v1.0.153 verificando el supuesto normalidad en donde se denotó que con un p-valor de 2.2e-16 los datos no provienen de una distribución normal bastando este supuesto para determinar el uso de un análisis estadístico ANOVA no paramétrico con lo que se concluye que los métodos simlrty, dendro_diana y hclust_vector son aquellos que ocupan menos memoria (Naranjo et al. 2018, pp. 74-77).

En el artículo denominado "Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics" se realizó la comparación de árboles jerárquicos en AEI y algunos grupos jerárquicos en AA , además hace mención a que LA se ha venido presentando como una tecnología emergente con un notable desarrollo por lo mismo este campo ha desarrollado un gran avance investigativo durante los últimos años integrando métodos de cálculo y análisis importantes para el tratamiento de datos. Para la comparación se utilizó un diseño cuasiexperimental con datos de tipo binario los cuales fueron generados de manera aleatoria para ejecutar así los algoritmos de clúster: árbol de cohesión (ASI), árbol de similitud (ASI), agnes (paquete de clúster R) y hclust (función base de R) en un computador con microprocesador Intel® Core™ i7-4770 CPU @ 3.40ghZ 3.40 GHz con software libre R de versión 3.4.1, entorno de RStudio 1.0.153 y Rchic de versión 0.24, en donde se evaluó los supuestos a un nivel de significancia del 5% dando como resultado que los datos son independientes, no son homogéneos y tampoco cumplen el supuesto de normalidad, debido a esto se aplicó la prueba estadística no paramétrica de Kruskal-Wallis concluyendo que el tiempo de ejecución de los algoritmos no es el mismo lo cual da impulso para futuras investigaciones en donde se considere incluso como parámetro a estudiar la cantidad de memoria ocupada (Rubén A. Pazmiño-Maji, García-Peñalvo Conde-González, 2017, pp. 2–6).

Planteamiento del problema

Enunciado del problema

Analizar gran cantidad de datos se ha convertido en la temática del día a día debido a la masiva generación de información digital originada por el elevado uso de tecnologías y entornos digitales sobre todo en el campo educativo, es por tal motivo que el uso de técnicas y métodos adecuados acorde al tipo y cantidad de datos beneficiarían la obtención de resultados ya que se evitarían cálculos repetitivos y en ciertos casos cálculos erróneos, permitiendo así un mejor

aprovechamiento de recursos digitales al otorgar una rápida propuesta de soluciones a problemáticas de estudiantes y entornos educativos con el manejo de lineamientos adecuados para mejorar el aprendizaje todo esto conociendo que técnicas ocupan menor tiempo de ejecución y memoria al momento de realizar el análisis de información.

Se contempla que las analíticas de aprendizaje (LA) aportan impulso para la recopilación de todo tipo de información y así mismo para su análisis bajo el fin de mejorar el éxito del aprendizaje a través del uso de técnicas con datos adecuados y evaluaciones formativas que permitirán marcar el comienzo de un proceso de toma de decisiones para la mejora del conocimiento con la detección de anomalías en tiempo real, aspecto que concede al docente una rápida manera de proceder orientando a una mejor adquisición de competencias individuales del estudiante (Maji et al. 2019, pp. 1-5), por otra parte cuando se habla de análisis estadístico implicativo se hace alusión a los problemas encontrados a los cuales se les busca una solución óptima con el objetivo principal que contempla estructuración de datos examinando su relación entre sujetos y variables de estudio (Pérez, Pazmiño, Andaluz 2014, p. 123). Con el avance de las técnicas también se ha dado apertura a indagar nuevas variables , debido a que en el ámbito educativo ya no solo se ponen a expectativa respuestas binarias sino también léxico que se podría clasificar en categorías las cuales pueden tener una jerarquía, la extensión de este campo da lugar al caso de variables modales con números del intervalo 0,1 haciendo descripción a grados de satisfacción o pertenencia (Gras, Kuntz 2009, pp. 3-50).

En la sociedad en la que nos encontramos actualmente el análisis de datos es primordial en los procesos educativos con enfoque en el rendimiento, seguimiento, manejo de tecnologías, entre otros; por lo que es importante conocer las técnicas más optimas y adecuadas que no originen estancamiento en los resultados permitiendo optimizar recursos como tiempo y en ciertos casos hasta económicos. Por ello se consideran los métodos clúster óptimos referente al tiempo de ejecución y espacio de memoria para el tratamiento de variables modales con el fin de brindar soluciones óptimas en el tratamiento masivo de datos.

Formulación del Problema

Pregunta General:

¿El método clúster que será obtenido como óptimo contribuye en la optimización de memoria y de tiempo de ejecución para las variables modales?

Justificación

Justificación Teórica

En la actualidad se ha evidenciado que el incremento de la producción de datos ha tenido un crecimiento elevado debido al uso de tecnologías y ambientes virtuales que mantienen distinto enfoque y entorno, lo cual ha puesto en necesidad el estudio de dicha información a través del uso de técnicas y métodos adecuados según el tipo y cantidad de datos; todo esto con la finalidad de

procesarlos de una manera más rápida y precisa para optimizar recursos relevantes como tiempo y memoria (Naranjo Serrano and Pazmiño Maji, 2018|, pp. 10–62). Conocer estas técnicas que guardan relación en ASI y LA serán de gran utilidad en sectores donde se esté generando gran cantidad de información, sobre todo donde existan actividades de aprendizaje ya que permitirá evaluar su entorno y relación entre alumno-docente, llevando así a brindar soluciones aptas según el ambiente de aprendizaje con la optimización de recursos al momento de analizar dicha información ya que incluso debido al avance de tecnologías se cuenta con softwares aptos para analizar cada enfoque (R y Rchic) permitiendo tener una idea clara de los resultados porque se basan en estadísticas en donde el significado llega a ser muy intuitivo (Couturier, Pazmiño 2016, p. 39). El análisis estadístico implicativo es un método de estadística multivariada el cual permite indagar acerca del conocimiento que presentan los estudiantes en su ámbito educativo a través de la generación de distintos ítems de evaluación con aspectos de interés entre ellos y los sujetos que han sido evaluados. Una de sus principales ventajas a diferencia de los métodos clásicos es que ASI tiene la capacidad de permitir la detección de relaciones de tipo "si a, entonces, casi siempre b" las cuales se denominan reglas o cuasi-implicación, son estas relaciones las cuales a través de un grafo implicativo permiten la mejor interpretación y comprensión de resultados caracterizando aspectos como visualización, rapidez y exequible comprensión (Mendoza, Caputo, Porcel 2019, pp. 53-60), por otra parte las Analíticas de aprendizaje (Learning Analytics) sigue un lineamiento de estudio para el tratamiento de datos que implica la medición, recopilación, análisis e informe de los datos sobre los alumnos y sus entornos de estudio con el objetivo de brindar un aprendizaje optimo a través de la compresión de todos los factores que intervienen en esta área; mucho más en esta última década en donde se puede notar que el campo académico se ha ido relacionando intrínsicamente con la tecnología. Sus principales ventajas son: efectividad, eficiencia y optimización de tiempo al comprender los recursos educativos a usar para resolver temáticas de deserción en ámbitos empresariales y educativos (Elias 2011, pp. 1-22). Por ende al comprender la estrecha relación de los métodos mencionados anteriormente con la combinación de las técnicas clúster más usadas en cada campo se puede establecer una comparación entre ellos, llegando a obtener un mejor proceso de toma de decisiones al probar supuestos y aplicar métodos adecuados para conseguir la interpretación de resultados al dar tratamiento a datos considerando características relevantes como cantidad y tipo (Rubén Antonio Pazmiño Maji, García Peñalvo, Conde González 2017, p. 1).

Teniendo presente la utilidad de las técnicas clúster antes mencionadas en Learning Analytics y el Análisis Estadístico Implicativo, en esta investigación se propone realizar un análisis comparativo en donde se pueda determinar que técnica clúster presenta mayor efectividad y rendimiento en cuanto a menor tiempo de ejecución de las funciones comparadas y su respectivo

uso de memoria con el fin de proporcionar una técnica adecuada para el tratamiento de datos según su tipo.

Justificación Práctica

El análisis de gran cantidad de datos y la aplicación del método correcto para algunos de ellos es un problema frecuente en el tratamiento de esta información, las analíticas de aprendizaje y el análisis estadístico implicativo son temas en auge que han despertado el interés de la comunidad investigadora debido a las facilidades que presentan en esta área. Esta metodología se adapta al uso de cierto tipo de información permitiendo así la optimización de recursos a través del uso del método adecuado para su tratamiento.

El Análisis Estadístico Implicativo (ASI) cuenta con herramientas que permiten el análisis de bases de datos que se estructuran con un conjunto de variables , se conoce que el ASI dispone de una gran utilidad como es el descubrimiento de reglas asimétricas entre las variables y sus respectivas clases , para su utilización ha contribuido en gran escala el desarrollo tecnológico y asimismo la investigación de teorías con lo cual se ha complementado para llevar un continuo avance educativo ; prueba de esto fue el desarrollo de la herramienta informática que automatiza los procesos del ASI, que es llamada CHIC la misma que ha facilitado en gran manera el manejo de gran cantidad de datos y ha permitido entregar resultados favorables optimizando recursos, por otra parte las analíticas de aprendizaje (LA) permiten recopilar y analizar grandes conjuntos de información con el fin de mejorar el aprendizaje a través de la comprensión del entorno de desarrollo identificando causas y consecuencias que facilitan la entrega e interpretación de resultados (Pazmiñoom et al. 2018, pp. 122-139).

El presente trabajo de investigación hará uso de software libre concediendo con esto autonomía tecnológica y ahorro de recursos logrando la innovación e incentivación por optimizar gastos para fortalecer el desarrollo de la sociedad y fomentar la inclusión digital. Adicionalmente el proyecto es completamente factible ya que la matriz de información a emplearse será generada para su respectivo análisis y correcta obtención de resultados bajo la propuesta de realizar un análisis experimental con el objetivo de comparar computacionalmente la complejidad de las técnicas clúster comunes al análisis estadístico implicativo y las analíticas de aprendizaje para variables modales; para esto se considera la estadística ya que posee numerosas ventajas para guiar hacia la solución óptima.

Esta temática será de gran utilidad para docentes y estudiantes debido a que permitirá usar grandes cantidades de datos y conocer que técnica favorece a la hora de optimizar recursos y toma de decisiones, además de que servirá como guía para interesados en esta área que va ganando relevancia en los trabajos de investigación en los últimos años.

Objetivos

Objetivo General

Comparar computacionalmente la complejidad de las técnicas clúster comunes al análisis estadístico implicativo y las analíticas de aprendizaje para variables modales.

Objetivos específicos

- Conceptualizar las Analíticas de Aprendizaje, el Análisis Estadístico Implicativo y las técnicas clúster comunes.
- Comparar el tiempo de ejecución para variables modales entre las técnicas comunes del AEI y de las AA
- Comparar el espacio de memoria para variables modales entre las técnicas comunes del AEI y de las AA

CAPÍTULO I

1. MARCO TEÓRICO

1.1. Definiciones y conceptos

En el siguiente capítulo se presentan los fundamentos teóricos de la investigación, el Análisis Estadístico Implicativo y Learning Analytics.

1.1.1. Análisis Estadístico Implicativo

El análisis estadístico implicativo es un método de análisis de datos creado por Régis Gras hace casi treinta años que tiene un impacto significativo en una variedad de áreas que van desde la investigación pedagógica y psicológica hasta la minería de datos. El análisis estadístico implicativo (ASI) proporciona un marco para evaluar la fuerza de las implicaciones; tales implicaciones se forman a través de técnicas comunes de adquisición de conocimientos en cualquier proceso de aprendizaje, humano o artificial. Este nuevo concepto se ha convertido en una metodología unificadora y ha generado una poderosa convergencia de pensamiento entre matemáticos, estadísticos, psicólogos, especialistas en pedagogía y, por último, pero no menos importante, informáticos especializados en minería de datos (Gras et al. 2008, pp. 12-50).

Otra definición dada por (Vázquez et al. 2019, pp. 157-170) menciona que el análisis estadístico implicativo, conocido por la sigla ASI de Analyse Statistique Implicative del idioma francés donde se originó, es una herramienta de la minería de datos basada en las técnicas estadísticas multivariadas, la teoría de la cuasi-implicación, la inteligencia artificial y el álgebra booleana, para modelar la cuasi-implicación entre los sucesos y variables de un conjunto de datos.

1.1.2. Origen y Desarrollo del ASI

Hoy día, el análisis estadístico implicativo designa un campo teórico centrado en el concepto de implicación estadística o más precisamente en el concepto de quasi-implicación para distinguirlo del de implicación lógica de los dominios de la lógica y de las matemáticas. El estudio de este concepto de cuasi-implicación en tanto que objeto matemático, en los campos de las probabilidades y de la estadística, ha permitido construir herramientas teóricas que instrumentalicen un método de análisis de datos. Es necesario constatar que las raíces epistemológicas de este concepto se han nutrido de cuestiones que han surgido habitualmente en otro campo: el de la didáctica de las matemáticas (Gras, Baquero, Guillet 2009, p. 1).

Más detalladamente en un artículo realizado por (Gras, Ratsimba-Rajohn 1996, pp. 217-232) se menciona que había concebido a priori en 1976 una clasificación que comprendía una serie de objetivos cognitivos, es decir un preorden parcial entre las capacidades que se esperan obtener y desarrollar en el alumno en todo el proceso del aprendizaje y en el funcionamiento operatorio de los conceptos matemáticos. Por ejemplo, "Elección y ordenación de argumentos " precedería a "Critica de argumentación y construcción de contraejemplos " y sería posterior a "Realización de algoritmos simples". A través de diversos test presentados a alumnos de colegio que comprendían una edad de 13 a 15 años, con variantes de ejercicios en este sentido, esperaba la validación de la taxonomía establecida a priori. Bajo la forma de grafo orientado sin ciclos, la organización de las capacidades observadas debería permitir estudiar la adecuación de la taxonomía al preorden representado por el grafo y de forma adicional estudiar las distorsiones ligadas a dos métodos de enseñanza diferentes. Todo enseñante, como todo investigador en didáctica de las matemáticas, sabe por la práctica o por la observación, que surgen contraejemplos en las situaciones observadas en relación con las hipótesis formuladas sobre la competencia. Un instrumento estadístico aparecía entonces como necesario, para evaluar y representar las cuasi-reglas emanadas de la contingencia, sobre la base de los resultados obtenidos. Durante los últimos treinta años, el desarrollo teórico del análisis estadístico implicativo ha presentado gran estimulación y avance basándose en la dialéctica combinada entre la práctica y la teoría, en una tensión entre dos marcos: la estadística aplicada y la estadística matemática. En diversos campos científicos como la didáctica de la matemática, la psicología, la sociología, la bioinformática, etc., los datos construidos han sido sometidos a este método de análisis. Esta forma de actuar ha mostrado la eficiencia del método respecto a su capacidad para hacer emerger las propiedades que otras aproximaciones no permiten, pero también ha permitido mostrar sus límites quienes han suscitado a su vez nuevas problemáticas en torno al concepto-objeto de la quasi-implicación. El razonamiento que fundamenta la interpretación de los resultados del análisis estadístico implicativo es esencialmente de naturaleza estadística y probabilista. Este modo de razonamiento se inscribe en una perspectiva accesible por el desarrollo del pensamiento estadístico, del espíritu estadístico (Claude 2021, pp. 2-5).

1.1.3. Minería de datos y ASI

En los últimos años se ha incrementado nuestra capacidad de generar, transformar, almacenar, analizar y visualizar datos, básicamente por la alta potencia de procesamiento y el bajo costo de las máquinas. Sin embargo, dentro de estos enormes y diferentes tipos de datos hay mucha información desconocida. El descubrimiento de esta información es posible gracias a la Minería de Datos (DM), que entre otras técnicas sofisticadas aplica inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, que son

representaciones abstractas de la realidad. Las tareas comunes en Knowledge Discovery in Databases (KDD) son la inducción de reglas, la clasificación y los problemas de agrupación, el patrón de reconocimiento, el modelado predictivo, la detección de dependencias, etc. Se determinó que ASI Y KDD están fuertemente relacionados en el tratamiento de grandes cantidades de datos con un porcentaje de acercamiento del 66% por lo que dicha aproximación dio paso a la proposición de pasos que constituyen el análisis del proceso que sigue ASI (R. A. Pazmiño-Maji, García-Peñalvo, Conde-González 2017, pp. 1-4).

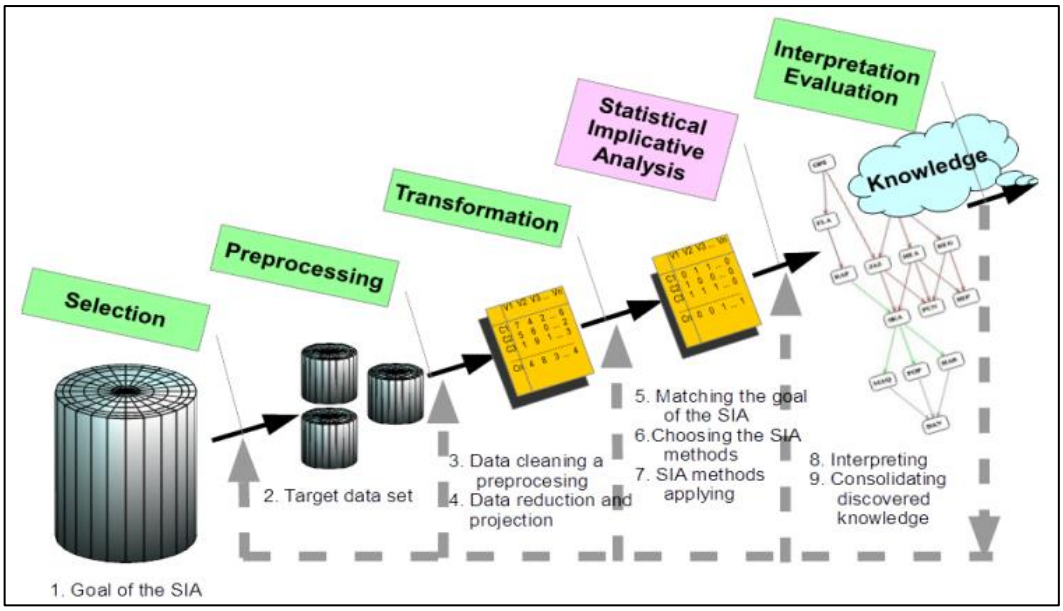


Figura 1-1. Propuesta basada en KDD, de los pasos que constituyen el proceso ASI

Fuente: Pazmiño, García y Conde, 2017.

1.1.4. Principales Categorías del ASI

En el ASI la validez de la regla a ⇒b depende de la probabilidad o fuerza de la cuasi-implicación, que se determina al comparar el número de contraejemplos presentes que invalidan dicha regla con los que aparecerían bajo una ausencia de relación estadística. El ASI consta de tres procedimientos (Sagaró del Campo et al. 2019, pp. 88-103):

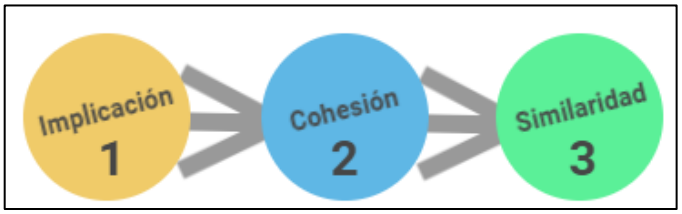


Figura 2-1. Procedimientos de ASI

Fuente: Sagaró y Zamora, 2019.

1.1.4.1. La implicación

En esta área se destacan tres conceptos básicos los cuáles se mencionan a continuación:

- **La intensidad Implicativa:** medida probabilística de la validez de la regla $a \Rightarrow b$. La decisión de aceptar o no la regla está en función del nivel de significación α elegido por el investigador y se dirá que la regla $a \Rightarrow b$ es admisible para un dado α si la cantidad de contraejemplos esperados es menor que los observados (Sagaró del Campo et al. 2019, pp. 88-103).
- **El índice de implicación**: indicador de la no implicación de a sobre b. Este índice es no simétrico y no coincide con el coeficiente de correlación u otros índices simétricos que miden asociación (Sagaró del Campo et al. 2019, pp. 88-103).
- **El índice de implicación-inclusión o de implicación entrópica:** versión entrópica del índice de implicación que supera la poca discriminación de este en muestras grandes. Este índice permite determinar el criterio entrópico al integrar la información a partir de la presencia de un escaso número de contra ejemplos, tanto por la regla $a \Rightarrow b$ como por su negación $\neg a \Rightarrow \neg b$ (Sagaró del Campo et al. 2019, pp. 88-103).

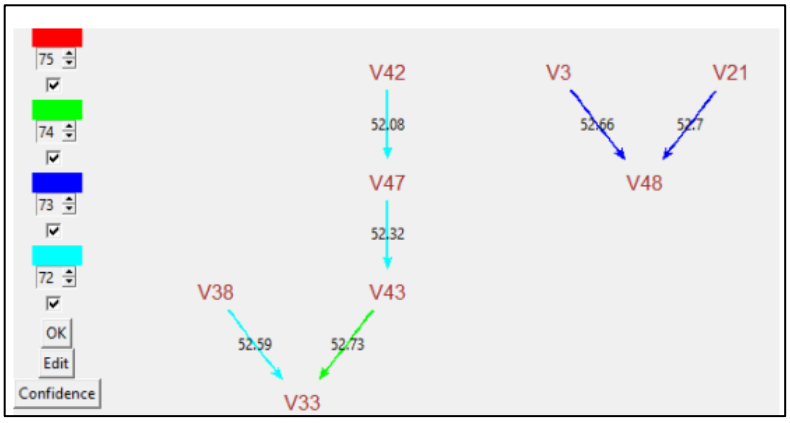


Figura 3-1. Ejemplo de grafo de implicación en Rchic

Fuente: Pazmiño Maji, 2021.

1.1.4.2. Cohesión

Permite estructurar el conocimiento en forma de reglas y meta reglas y superar la simple articulación de las partes de una tipología clásica, a fin de alcanzar un todo significativo al ser de carácter no lineal, asimétrico, jerárquico y dinámico. Las reglas y meta reglas que surgen se pueden presentar en tres esquemas (Sagaró del Campo et al. 2019, pp. 88-103):

- $R \Rightarrow c$, donde R: $a \Rightarrow b$ que se interpreta como que c es consecuencia de la regla R.

- $a \Rightarrow R$, donde $R: b \Rightarrow c$, que se interpreta como que a se dedujo de la regla R o que la regla R es consecuencia de a.
- $R_1 \Rightarrow R_2$, donde $R_1: a \Rightarrow b$ y $R_2: c \Rightarrow d$, que se interpreta como R_2 se dedujo de la regla R_1 o que la regla $\neg a \vee b$ es consecuencia de $a \wedge \neg b$

Intuitivamente, la cohesión mide el desequilibrio de las frecuencias de los eventos $\neg a \vee b$ y $a \wedge \neg b$ a favor del primero (Sagaró del Campo et al. 2019, pp. 88-103).

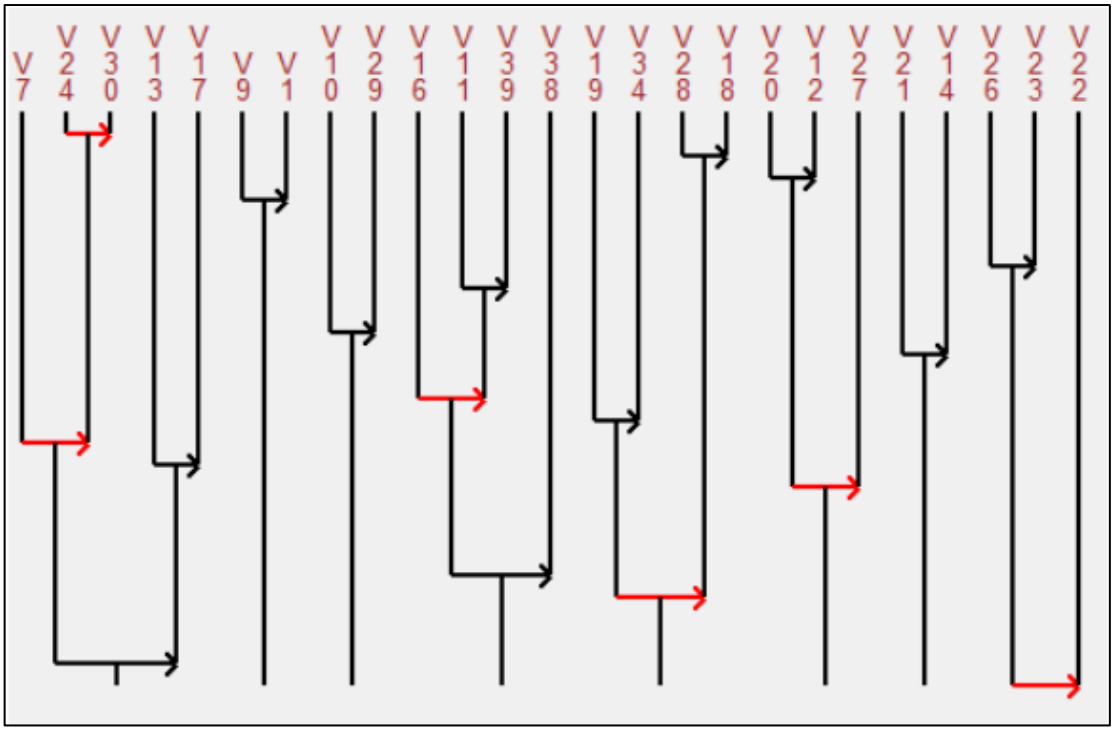


Figura 4-1. Ejemplo de árbol de cohesión realizado en Rchic

Fuente: Pazmiño Maji, 2021.

1.1.4.3. Similaridad

Como medida de correspondencia o semejanza entre los objetos que van a ser agrupados. A diferencia de los métodos de clasificación usualmente empleados, en el ASI se emplea el índice de similaridad de Lerman, que es la probabilidad de que el número observado de copresencias entre dos variables sea mayor o igual que el de las copresencias esperadas por el azar.

Además de los procedimientos antes descritos, el ASI permite cuantificar el aporte de cada individuo en la formación de las estructuras que se obtienen a partir de los índices de cohesión y de similaridad, para lo cual emplea la contribución o la tipicalidad de cada sujeto. La tipicalidad es un índice porcentual que mide cómo se comporta un individuo con relación a la regla o a la clase, llamando sujeto típico a aquél que verifica todas las implicaciones (similaridades) que poseen mayor intensidad de implicación (índice de similaridad) en la formación de las reglas

(clases). La contribución cuantifica el aporte de un determinado individuo en la formación de la regla o de la clase (Campo, Matamoros 2019, pp. 88-103).

1.1.5. Grandes Conjuntos de datos

La versión entrópica considera el contra positivo $\bar{b} => \bar{a}$, que podría robustecer la afirmación de la implicación que se mantiene entre a y b, asimismo podría mejorar la calidad de la discriminación de ϕ cuando la transacción se establece que aumenta: si A y B son pequeños, sus conjuntos complementarios son grandes y viceversa, para el análisis de estos grandes conjuntos de datos se integra el uso del software CHIC a través del uso de las estadísticas. Las comparaciones experimentales han destacado dos características interesantes cuando estas medidas no seleccionan las mismas reglas; en varias bases de datos se encontraron un subconjunto de reglas no sorprendentes con una buena confianza, demostrando la relevancia de algunas de estas reglas sobre datos de la vida real para la toma de decisiones (Gras, Kuntz, Briand 2001, pp. 10-23).

1.1.6. Variables por su contenido

Según el tipo de datos en estudio se consideran las variables a usar en ASI las de tipo binario, modal, frecuencial y de intervalo, asimismo se encuentran en desarrollo el poder usar variables vectoriales y en casos particulares variables difusas (Scimone, Spagnolo 2005, p. 9).

1.1.6.1. Variables de tipo Binario

Las variables de este tipo solo toman dos valores que comúnmente se los debe representar por 0 y 1. Estos números simbolizan dos estados opuestos, como si o no, la existencia y la ausencia, bueno o malo, la verdad y la falsedad, que cumple o no cumple, la posesión y la no posesión, etc. La suma por columnas representa el número de sujetos que poseen o satisfacen la propiedad. La suma por filas representa el número de variables satisfechas por el sujeto (Bernard, Charron 1996, pp. 5-38).

1.1.6.2. Variables de tipo modal

Este tipo de variables guardan asociación a valores que son número en el intervalo [0, 1] y describen grados de satisfacción o de pertenencia , así se podría ejemplificar: "Nada satisfecho", "Poco satisfecho", "Neutral", "Muy satisfecho", "Totalmente satisfecho", utilizadas generalmente en cuestionarios de opinión Likert (Canto de Gante et al. 2020, pp. 38-39), según como se requiera se asumen y se transforman en valores 0, 0,25, 0,50, 0,75, 1 respectivamente.

1.1.6.3. Variables de tipo Frecuencial

Este tipo de variables utilizan porcentajes, que se asocian a fenómenos en la escala 0% a 100%, pero que luego se representarán en el intervalo [0, 1] (Zamora, Gregori, Orús 2009, pp. 65-101). Por ejemplo, la variable porcentaje de estudiantes que les gusta las matemáticas podrían tomar valores de 0%, 5%, 15%, 50%, que en la escala de [0, 1] serían 0, 0,05, 0,15, 0,50.

1.1.6.4. Variables de tipo Intervalo

Son conjuntos finitos C de números reales cualquiera que para aplicar las técnicas del ASI se deben trasladar a la escala [0, 1], donde 0 corresponderá al mínimo valor y 1 al máximo valor del conjunto C, luego para asignar cada valor de x de C al intervalo [0, 1], se realiza una traslación de $x + abs(\min(C))$ y luego una contracción con una regla de tres asociando $\max(C) - \min(C)$ correspondiente a 1, la fórmula final que lleva un x de C a un x' en [0, 1], está dada por $x' = \frac{x+abs(\min(C))}{\max(C)-\min(C)}$. Luego los x' se ingresan como el caso de las variables frecuenciales. Para indicar que una variable nvar es de intervalo se ingresa de la siguiente manera: nvar i (Pazmiño Maji 2021b, pp. 37-40).

1.1.7. Learning Analytics

Analíticas de aprendizaje (Learning Analytics) es un área de rápido crecimiento de la investigación en aprendizaje mejorado por tecnología. Tiene fuertes raíces en una variedad de campos, en particular inteligencia empresarial, analítica web, minería de datos, etc. Sus fuertes conexiones con estos campos han significado que los investigadores ahora deben trabajar juntos para identificar los desafíos y obtener nuevas herramientas tecnológicas que puedan ayudar al estudio de Learning Analytics (Velandia-Vega, Flores-Cabañas 2020, pp. 40-45).

Cuando hablamos de la analítica de aprendizaje podemos encontrar muchas definiciones de este campo las cuáles suelen presentan leves variaciones entre sí, pero a la final engloban el mismo concepto y mensaje sobre esta rama.

Teniendo las más específicas las siguientes:

- Learning Analytics es la medida, recolección, análisis y reporte de datos sobre los alumnos y sus contextos, con el propósito de comprender y optimizar el aprendizaje y el entorno en el que ocurre.

- Learning Analytics es el uso de datos inteligentes, de datos producidos por los alumnos y de modelos de análisis, para descubrir información y conexiones sociales que permitan predecir y asesorar el aprendizaje de las personas.

Al analizar ambas definiciones que han sido planteadas se puede constatar y destacar la importancia y relevancia de tres elementos: los datos, que se consideran como la materia prima de este proceso, el análisis, que permite añadir valor a los datos por medio de algoritmos y procesos, y finalmente la acción a emprender, como respuesta proactiva a los resultados del proceso de analítica. Los datos para sustentar esta metodología de análisis suelen recogerse durante el periodo que dura la formación, y se centran en el propio estudiante, en el entorno de aprendizaje, en las interacciones que tienen lugar durante el proceso y en los resultados académicos (Rodriguez 2019, p. 2).

1.1.8. Enseñanza en entornos digitales

El origen del Learning Analytics está estrechamente relacionado con el proceso paulatino de digitalización –para algunos visionarios excesivamente lento- que lleva a cabo la educación. El uso progresivo de plataformas educativas, como Moodle, implica que el alumno va dejando trazas digitales de su actividad (número de accesos, horarios de conexión, tareas realizadas, participación en *chats* y foros) que pueden ser recopiladas y analizadas. Sin embargo, lo que definitivamente impulsa el desarrollo de este tipo de analítica es el formato MOOC, los cursos masivos online, en donde el gran volumen de alumnos convierte en una tarea casi imposible el seguimiento individual del itinerario formativo de los mismos a través de métodos más tradicionales, como son los aplicados en las clases presenciales de la universidad. Por supuesto, para llegar al desarrollo, avance y florecimiento de la analítica se ha coincidido en el tiempo con todos los grandes y avanzados progresos de la tecnología, brindando así mayor apertura a big data debido a la gran generación de datos en todas las áreas, lo que ha permitido gestionar con mayor facilidad grandes cantidades de información de distinto tipo y tamaño a través del uso de técnicas acordes para una adecuada obtención de resultados (, 2019, p. 2).

1.1.9. Proceso de implementación de analíticas de aprendizaje

Este proceso está basado en la experiencia de desarrollar distintos proyectos de analítica de aprendizaje, y guarda cierta similitud con los presentados en los trabajos relacionados, pero se encuentra más centrado en las etapas de implementación que sigue un proyecto de analítica de aprendizaje, así como los elementos y preguntas que se deben tener en cuenta. El proceso tiene las siguientes cinco etapas en donde cada una comprende un papel fundamental para el respectivo desarrollo y comprensión de las analíticas de aprendizaje (Ruipérez-Valiente 2020, pp. 85-101):

- *Entornos de aprendizaje:* ¿Cuál es el contexto y cuáles son los estudiantes?
- *Recolección de datos en crudo:* ¿Qué datos se deben generar y cómo almacenarlos?
- *Manipulación de datos e ingeniería de características:* ¿Qué características son necesarias y cómo obtenerlas?
- *Análisis y modelos:* ¿Qué análisis y modelos se deben implementar?
- *Aplicación educacional:* ¿Cuál es la aplicación educacional objetivo y el usuario?

Figura 5-1. El proceso de implementación de analíticas de aprendizaje

Fuente: Ruipérez-Valiente, 2020.

1.1.10. Técnicas de Análisis en las analíticas de aprendizaje

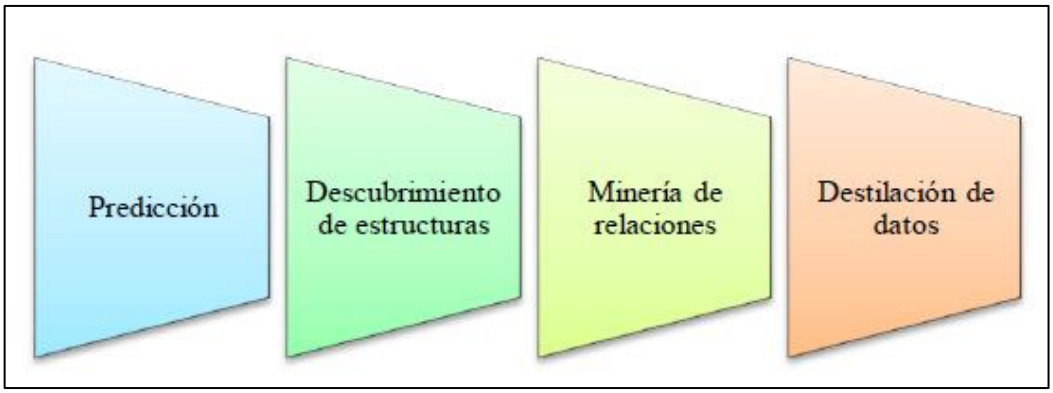


Figura 6-1. Técnicas de Análisis en LA

Fuente: Klašnja-Milićević et al., 2017.

A continuación se describen cada una de las técnicas en LA (Klašnja-Milićević, Ivanović, Budimac 2017, pp. 1066-1078):

1.1.10.1. Predicción:

Consiste sobre los usos futuros del entorno de aprendizaje en función de los datos de aprendizaje disponibles recopilados durante los procesos y actividades de aprendizaje, es posible predecir los usos futuros de las secuencias de aprendizaje, predecir las calificaciones finales de los alumnos o predecir el comportamiento del conocimiento de los alumnos (Klašnja-Milićević, Ivanović, Budimac 2017, pp. 1066-1078).

1.1.10.2. Descubrimiento de estructuras

Se realiza en base a los datos de aprendizaje disponibles, es posible determinar relaciones y patrones significativos entre los niveles de conocimiento de los estudiantes, los tiempos de uso del sistema de aprendizaje electrónico y las calificaciones de los estudiantes (Klašnja-Milićević, Ivanović, Budimac 2017, pp. 1066-1078).

1.1.10.3. Minería de relaciones

Sobre la base de la información recopilada de la interacción entre el usuario y el entorno de aprendizaje, es posible descubrir la relación entre la usabilidad de los materiales del curso y el rendimiento de aprendizaje de los estudiantes.

1.1.10.4. Destilación de datos

Es importante destilar los datos de diferentes formas y para diferentes propósitos para su uso posterior en la gestión humana (Klašnja-Milićević, Ivanović, Budimac 2017, pp. 1066-1078).

1.1.11. Analíticas de aprendizaje y Big data

Distintos autores muestran su perspectiva sobre las analíticas de aprendizaje y big data. El proceso de recopilar, reunir y analizar grandes conjuntos de datos ("big data") útiles para descubrir información útil y alguna forma de patrones se define como análisis de big data. Además, la analítica de big data brinda oportunidades para reconocer mejor la información que podría ser importante para decisiones futuras (Klašnja-Milićević, Ivanović, Budimac 2017, pp. 1066-1078).

(Merceron, Blikstein, Siemens 2015, pp. 4-8) menciona que, dentro de la analítica de aprendizaje, Big Data adquiere una seria de instancias diferentes. Primero, el campo se está diversificando hacia un conjunto más amplio de fuentes y modalidades de datos, lo cual es esencial para realizar una investigación que tenga un gran impacto. Al mismo tiempo, estas nuevas fuentes de datos y plataformas de aprendizaje están generando oportunidades para que la comunidad desarrolle nuevas técnicas analíticas.

Un estudio realizado por (Sin, Muthu 2015, pp. 1035-1037) recalca que el uso de sistemas de gestión del aprendizaje en la educación ha aumentado en los últimos años debido al incesante desarrollo tecnológico, lo cual ha permitido conocer nuevas metodologías y procesos de análisis. Los estudiantes han comenzado a usar teléfonos móviles, principalmente teléfonos inteligentes que se han convertido en parte de su vida diaria, para acceder a contenido en línea. Las actividades en línea de los estudiantes generan una enorme cantidad de datos no utilizados que se desperdician ya que los análisis de aprendizaje tradicionales no son capaces de procesarlos debido a parámetros que suelen poner limitaciones como la cantidad y el tipo originando así que esta información no pueda ser aprovechada en su totalidad. Esto ha resultado en la penetración de tecnologías y herramientas de Big Data en la educación, para procesar la gran cantidad de datos involucrados. Este estudio analiza las aplicaciones recientes de las tecnologías de Big Data en la educación y presenta una revisión de la literatura disponible sobre minería de datos educativos y análisis del aprendizaje considerando todos los enfoques necesarios lo que ha permitió obtener información relevante de estudios en estos campos que aún no suelen ser indagados en su totalidad.

Tabla 1-1: Resultados de búsqueda para cada palabra clave

Keyword	Search Results
Educational Data Mining	5290
Learning Analytics	5890
Educational Data Mining and Learning Analytics	1370

Fuente: Sin y Muthu, 2015.
Realizado por: Armas, Shirley, 2022

La toma de decisiones basada en datos, popularizada en las décadas de 1980 y 1990, está evolucionando hacia un concepto mucho más sofisticado conocido como big data que se basa en enfoques de software generalmente conocidos como análisis. El big data y la analítica para aplicaciones de instrucción están en su infancia y tardarán algunos años en madurar, aunque su presencia ya se siente y no debe ignorarse. Si bien el big data y la analítica no son la panacea para abordar todos los problemas y decisiones que enfrentan los administradores de la educación superior, pueden convertirse en parte de las soluciones integradas en las funciones administrativas e instructivas (Picciano 2012, pp. 9-20).

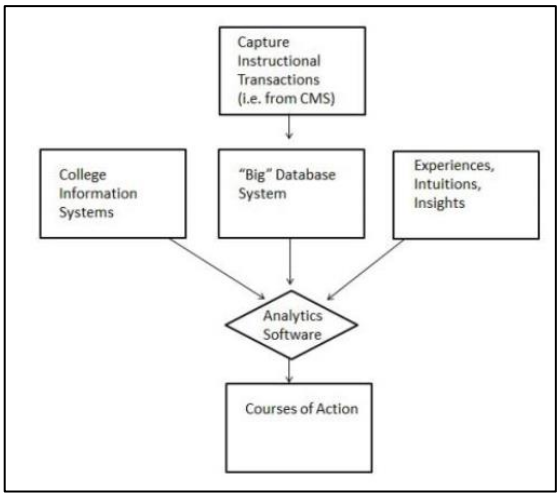


Figura 7-1. Modelo de flujo de análisis de aprendizaje

Fuente: Picciano, 2012.

(Picciano 2014, pp. 35-43) hace énfasis en que para que las aplicaciones de análisis de aprendizaje y big data funcionen bien, los datos deben ser precisos y oportunos. El software de análisis de aprendizaje funciona mejor para los cursos que se imparten de forma completamente electrónica, como los cursos en línea. Los cursos tradicionales presenciales que requieren un tiempo significativo de conversión de datos son problemáticos. Los cursos de aprendizaje combinado (parte presencial y parte online) también presentan problemas de recopilación de datos. Debido a que los cursos de aprendizaje mixto varían mucho en la naturaleza de su entrega, el software de análisis de aprendizaje puede tener importantes lagunas de datos. Las transacciones de instrucción que tienen lugar en el entorno cara a cara se perderán a menos que el miembro de la facultad o el asistente de enseñanza esté dispuesto a ingresarlas manualmente en el sistema de información del estudiante.

1.2. Teoría Estadística

1.2.1. Análisis Exploratorio de Datos

El análisis exploratorio de datos (Exploratory Data Analysis, EDA) o estadística descriptiva es un paso previo e imprescindible a la hora de comprender los datos con los que se va a trabajar y altamente recomendable para una correcta metodología de investigación. El objetivo de este análisis es explorar, describir, resumir y visualizar la naturaleza de los datos recogidos en las variables aleatorias del proyecto o investigación de interés, mediante la aplicación de técnicas simples de resumen de datos y métodos gráficos sin asumir asunciones para su interpretación (Grolemund, Garret, Hadley 2017, p. 2).

Es importante tener presente ciertos parámetros que son imprescindibles tener claro al momento de realizar el análisis siendo los siguientes:

- Una **variable** es una cantidad, cualidad o característica medible, es decir, que se puede medir.
- Un **valor** es el estado de la variable en el momento en que fue medida. El valor de una variable puede cambiar de una medición a otra.
- Una **observación** es un conjunto de mediciones realizadas en condiciones similares (usualmente todas las mediciones de una observación son realizadas al mismo tiempo y sobre el mismo objeto). Una observación contiene muchos valores, cada uno asociado a una variable diferente. En algunas ocasiones nos referiremos a una observación como un punto específico (data *point* en inglés).
- Los *datos* **tabulares** son un conjunto de valores, cada uno asociado a una variable ya una observación. Los datos tabulares están ordenados si cada valor está almacenado en su propia "celda", cada variable cuenta con su propia columna, y cada observación corresponde a una fila.

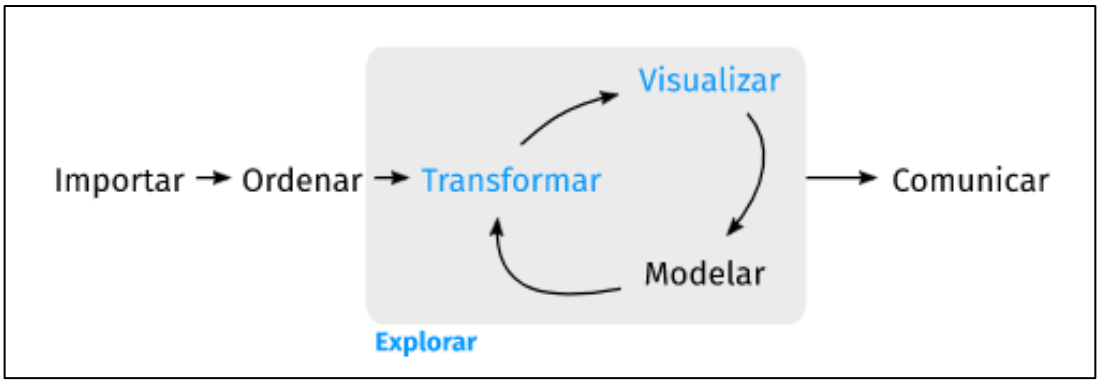


Figura 8-1. Proceso del EDA

Fuente: Grolemund, Garret y Hadley 2017.

1.2.2. Análisis Clúster

El análisis de conglomerados o clúster es una técnica multivariante que busca agrupar elementos o variables tratando de conseguir la máxima homogeneidad en cada grupo y la mayor diferencia entre ellos, mediante una estructura jerarquizada para poder decidir qué nivel jerárquico es el más apropiado para establecer la clasificación. Los métodos de clustering se agrupan dentro de las técnicas de machine Learning y de aprendizaje no supervisado basados en agrupar o identificar

clústeres (subconjuntos similares entre sí) dentro de un conjunto de datos, de acuerdo con una determinada medida de similitud entre las observaciones, pudiendo obtener diferentes clústeres en función de la mediad utilizada. La finalidad pues, es la de particionar los datos en distintos grupos de manera que las observaciones dentro de cada grupo sean bastante similares entre sí, y distintas a otros grupos. El concepto de "similar" dependerá del caso de estudio. El método de clustering se relaciona con el análisis de componentes principales en el sentido de que ambos buscan simplificar los datos, aunque el mecanismo de ambos es distinto: mientras que PCA pretende encontrar una representación de los datos en pocas dimensiones que expliquen gran parte de la varianza, el método de clustering se aplica para encontrar subgrupos homogéneos de observaciones. Siendo un método de data mining bastante popular en muchos campos, existe un gran número de métodos de clustering, siendo dos de los más conocidos (Vilà Baños et al. 2014, pp. 113-127):

- *K-means* clustering: partición de las observaciones en un número predefinido de clústeres.
- *Hierarchical clustering:* no partimos de un número predefinido de clústeres. Representación de datos en un dendograma (representación en forma de árbol).

1.2.2.1. K- means Clustering

El método de K-means clustering es un método no jerárquico para agrupar objetos (no variables) que particiona el set de datos en K clústeres distintos y no solapantes, lo que significa que ninguna observación puede pertenecer a más de un clúster. El número de clústeres o subgrupos requeridos se ha de establecer al inicio.

Siendo $C_1,\ldots, C_K$ el número de sets, la varianza intra-clúster para el clúster C_K es una medida $W(C_k)$ de la cantidad que difieren las observaciones dentro del mismo. Por tanto, se busca minimizar

$$\sum_{k=1}^{K} W(C_k) \tag{1.1}$$

de manera que la varianza total dentro de cada clúster, sumada sobre todos los K clústeres, sea lo más pequeña posible. Una forma común de establecer esta varianza es mediante la distancia euclídea, con lo que obtenemos

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{1.2}$$

Siendo $|C_k|$ el número de observaciones en el k-ésimo clúster. De esta manera la varianza se mide como la suma de todas las distancias euclídeas al cuadrado entre pares de observaciones del clúster k, dividido por el número total de observaciones en ese mismo clúster.
Combinando ambas ecuaciones anteriores obtenemos el problema de optimización que define K-means clustering (Gil Martínez 2018, p. 3):

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k¿1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{1.3}$$

1.2.2.2. Algoritmo

Existen al menos K^n maneras de particionar n observaciones en K clústeres, por lo que este puede ser un número muy alto si K y n no son pequeños. En este caso, el algoritmo iterativo de K-means clustering proporciona un óptimo local:

1. Asignar un clúster inicial (de 1 a K) de manera aleatoria a cada observación.
2. Iterar hasta que la asignación de cada clúster deje de cambiar:

a) Para cada uno de los K clústeres, calcular el centroide del clúster (vector de medias de las variables j para las observaciones del clúster k).
b) Asignar cada observación al clúster cuyo centroide esté más próximo.

El proceso de clustering mejora de manera continua hasta que el resultado deja de cambiar habiéndose alcanzado el óptimo local. El nombre de K-means deriva del hecho de que en el paso 2 (a) los centroides (medias) se calculas como la media de las observaciones asignadas a cada clúster. Debido a que el algoritmo encuentra un óptimo local en lugar del óptimo global, los resultados obtenidos dependerán de la asignación inicial y aleatoria de cada observación en el paso 1 del algoritmo. Por esta razón, es importante aplicar el algoritmo múltiples veces con distintas asignaciones iniciales, seleccionando la mejor solución (Gil Martínez 2018, pp. 4-5).

1.2.2.3. Clustering Jerárquico

Una desventaja de K-means clustering es su requerimiento para seleccionar de manera previa un determinado número de clústeres K. El clustering jerárquico o hierarchical clustering supone un enfoque alternativo que no requiere esta selección inicial. Una ventaja adicional de este método es la posibilidad de obtener representaciones (basadas en árboles) de las observaciones, conocidas como dendogramas. El tipo más común de clustering jerárquico es el aglomerante, y se refiere al

hecho de que el dendograma se crea empezando por las hojas, combinando subgrupos hasta el "tronco" (Gil Martínez 2018, p. 3).

1.2.2.4. Dendograma

Un dendograma es una representación que ilustra la organización jerárquica entre elementos (puede representarse horizontal o verticalmente).

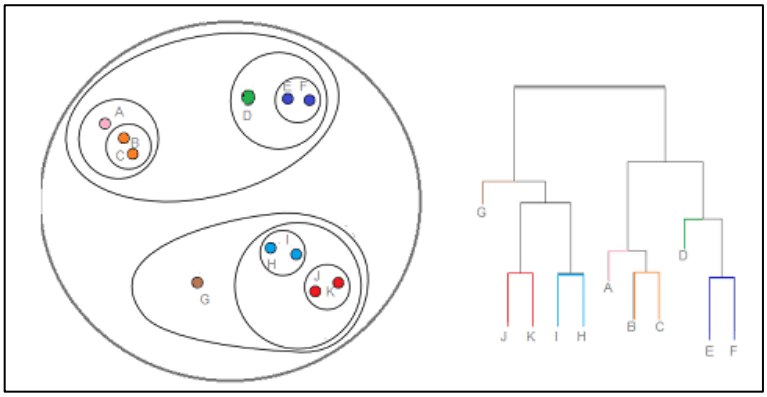


Figura 9-1. Dendograma representando clústeres jerárquicos anidados

Fuente: Gil Martínez, Cristina 2018.

Cada hoja del dendograma representa un elemento u observación. Conforme ascendemos por el árbol, algunas de las hojas se fusionan en ramas. Estas corresponden a observaciones que son similares unas a otras. Si ascendemos más en el árbol, las ramas se fusionan con hojas o con otras ramas. Las uniones más tempranas (más abajo en el árbol) corresponden con grupos de observaciones más similares entre sí. Por el contrario, las observaciones que se unen más arriba del árbol (cerca del final del árbol, más tardías) tienden a ser bastante diferentes.

La clave para interpretar un dendograma es centrarse en la altura a la que dos observaciones se unen. Podemos sacar conclusiones acerca de la similitud de dos observaciones en base a su localización en el eje vertical donde las ramas que contienen esas observaciones se unen por primera vez. Por otro lado, la posición horizontal de cada división da información sobre la distancia (disimilitud) entre dos clústeres (Gil Martínez 2018, pp. 4-5)..

1.2.2.5. Algoritmo

Como primer paso es necesario establecer la medida de disimilitud a utilizar entre cada par de observaciones. Comúnmente se emplea la distancia euclídea, pero existen otras (distancia de Mahalanobis, distancia de Minkowski, etc.). Por otro lado, se encuentra la disimilitud entre pares

de grupos de observaciones, donde aparece el concepto de método de unión o linkage, que mide esta disimilitud. Los cuatro tipos de linkage más comunes son (Gil Martínez 2018, pp. 4-5).:

- *Complete:* Distancia máxima entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la máxima de las distancias.
- *Average:* Distancia media entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la media de las distancias.
- *Single:* Distancia mínima entre clústeres. Se calculan por parejas las disimilitudes entre las observaciones en el clúster A y el B, escogiendo la mínima de estas medidas. Puede dar lugar a dendogramas donde las observaciones se fusionan una a una, obteniendo clústeres muy extendidos. Puede crear grupos muy homogéneos.
- *Centroid:* Distancia entre centros. Medida de disimilitud entre el centroide del clúster A y el centroide del clúster B. Suele utilizarse con frecuencia en genómica, pero puede dar lugar a inversiones indeseables que dificulten la visualización e interpretación (Gil Martínez 2018, pp. 4-5).

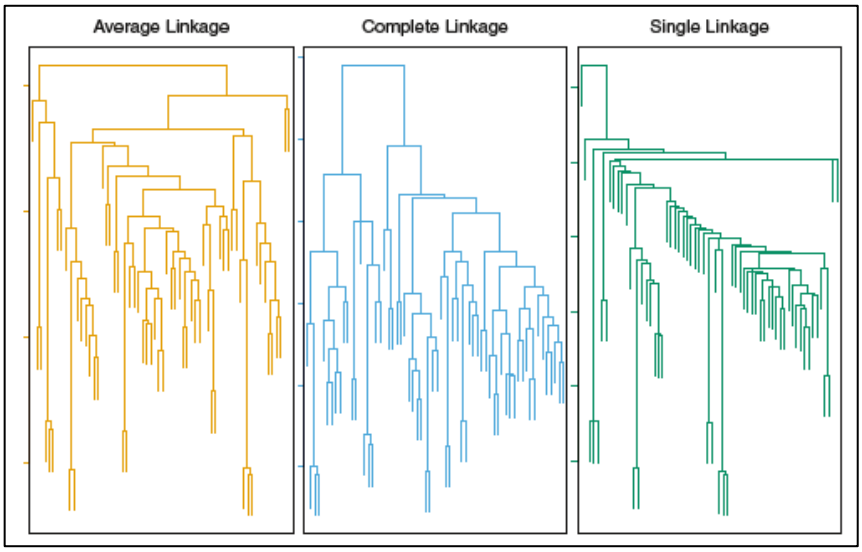


Figura 10-1. Tipos de Dendogramas

Fuente: Gil Martínez, Cristina 2018.

1.2.2.6. Medidas de similitud

La elección de la medida de similitud es muy importante, ya que de ella puede depender el dendograma resultante. Es por ello importante también tener en cuenta del tipo de datos con que se trata y el problema en cuestión. Además de la distancia euclídea como medida de similitud, existen otras que pueden preferirse a esta primera. Por ejemplo, la distancia basada en la correlación considera dos observaciones como similares si sus características asociadas están

altamente correlacionadas, incluso aunque los valores observados estén alejados en términos de distancia euclídea. Puede calcularse, más comúnmente, entre variables en lugar de entre observaciones (Gil Martínez 2018, pp. 4-5).

1.2.3. Verificación de supuestos

1.2.3.1. Normalidad

En Estadística, las pruebas de normalidad se utilizan para determinar si un conjunto de datos está bien modelado por una distribución normal y calcular la probabilidad de que variable aleatoria subyacente al conjunto de datos que se distribuirá normalmente.
Más precisamente, las pruebas son una forma de selección de modelo, y se puede interpretar de varias maneras, dependiendo de la interpretaciones de probabilidad (Filliben 1975, pp. 111-117):

- En estadísticas descriptivas términos, uno mide una bondad de ajuste de un modelo normal a los datos: si el ajuste es deficiente, los datos no están bien modelados a ese respecto mediante una distribución normal, sin emitir un juicio sobre ninguna variable subyacente.
- En estadísticas frecuentistas prueba de hipótesis estadística, los datos se prueban contra la hipótesis nula que se distribuye normalmente.
- En Estadísticas bayesianas, uno no "prueba la normalidad" per se, sino que calcula la probabilidad de que los datos provengan de una distribución normal con parámetros dados μ, σ (para todos μ, σ), y lo compara con la probabilidad de que los datos provengan de otras distribuciones bajo consideración, más simplemente usando un Factor de Bayes (dando la probabilidad relativa de ver los datos dados diferentes modelos), o más finamente tomando una distribución previa sobre posibles modelos y parámetros y calculando un distribución posterior dadas las probabilidades calculadas (Filliben 1975, pp. 111-117).

Se utiliza una prueba de normalidad para determinar si los datos de la muestra se han extraído de una población distribuida normalmente (dentro de cierta tolerancia). Varias pruebas estadísticas, como la prueba t de Student y el ANOVA unidireccional y bidireccional, requieren una población de muestra distribuida normalmente (Rot, Babativa 2010, pp. 127-131).
Se menciona que en la vida real no existe una distribución perfectamente normal, sin embargo, con modelos, que se sabe que son falsos, a menudo se puede derivar resultados que coinciden, con una aproximación útil a los que se encuentran en el mundo real". Entonces, de acuerdo con lo mencionado por Box, es evidente que las pruebas orientadas a probar una perfecta normalidad (como lo hacen los pretest tradicionales) no tienen sentido, a lo mejor tienen algún tipo de utilidad, pero no prueban algo real. En este sentido podríamos mencionar lo que el mismo autor concluyó

de forma más general: "Todos los modelos son erróneos pero algunos son útiles" (Flores, Ocaña, Sánchez 2018, pp. 5-22).

Las hipótesis de una prueba de normalidad vienen dadas de la siguiente manera:

Ho: La muestra proviene de una población con distribución normal
Ha: La muestra no proviene de una población con distribución normal

El contraste de la hipótesis de normalidad de los datos puede efectuarse utilizando dos tipos de pruebas, a saber, representaciones gráficas y test de hipótesis, tales como Anderson-Darling, Ryan-Joiner, Shapiro-Wilk y Kolmogórov-Smirnov, siendo de gran utilidad realizar una prueba de normalidad y producir una gráfica de probabilidad normal en el mismo análisis, por cuanto la prueba de normalidad y la gráfica de probabilidad suelen ser las mejores herramientas para evaluar la normalidad (Tapia, Cevallos 2021, pp. 83-106). La Tabla 2-1 muestra los estadísticos para las pruebas de normalidad.

Tabla 2-1: Estadísticos para las pruebas de Normalidad

Prueba	Estadístico de Prueba	Función en R
Shapiro-Wilk	$W = \frac{(\sum a_i y(i))^2}{\sum (y_i - \bar{y})^2}$	shapiro.test
Anderson-Darling	$A = -n - \frac{1}{n}\sum (2_i - 1)\left[\log (P_{(i)} + \log (1 - P_{(n-i+1)})\right]$	ad.test
Jarque-Bera	$\gamma = \frac{n}{6}\left\{\beta_1 + \frac{(\beta_2 - 3)^2}{4}\right\}$	jarque.bera.test
Pearson	$P = \sum \frac{(C_i - E_i)^2}{E_i}$	pearson.test
Cramér-Von Mises	$W = \frac{1}{12n} + \sum (P_{(i)} - \frac{2_i - 1}{2n})$	cvm.test
Shapiro-Francia	$W' = \frac{(\sum a_i y(i))^2}{\sum (y_i - \bar{y})^2}$	sf.test
Kolmogorov-Smirnov	$D^+ = max_{i=1,\dots,n}\left\{\frac{i}{p_{(i)}}\right\} D^- = max_{i=1,\dots,n}\left\{p_{(i)} - \frac{i-1}{n}\right\}$	lillie.test
D' Agostino	$DA = \frac{\sum \left(1 - \frac{n+1}{2}\right) X_i^*}{n^2 \sigma_n}$	agostino.test

Fuente: Tapia y Cevallos, 2021.
Realizado por: Armas, Shirley, 2022

1.2.3.2. Variabilidad

El supuesto de homogeneidad de varianzas, también conocido como supuesto de homocedasticidad, considera que la varianza es constante (no varía) en los diferentes niveles de un factor, es decir, entre diferentes grupos.

A la hora de realizar contrastes de hipótesis o intervalos de confianza, cuando los tamaños de cada grupo son muy distintos ocurre que (Amat 2016, p. 1) :

- Si los grupos con tamaños muestrales pequeños son los que tienen mayor varianza, la probabilidad real de cometer un error de tipo I en los contrastes de hipótesis será menor de lo que se obtiene al hacer la prueba. En los intervalos, los límites superior e inferior reales son menores que los que se obtienen. La inferencia será por lo general más conservadora.
- Si, por el contrario, son los grupos con tamaños muestrales grandes los que tienen mayor varianza, entonces se tendrá el efecto contrario y las pruebas serán más liberales. Es decir, la probabilidad real de cometer un error de tipo I es mayor que la devuelta por la prueba y los intervalos de confianza verdaderos serán más amplios que los calculados.

Existen diferentes pruebas que permiten evaluar la distribución de la varianza. Todos ellos consideran como hipótesis nula que la varianza es igual entre los grupos y como hipótesis alternativa que no lo es. La diferencia entre ellos es el estadístico de centralidad que utilizan:

- Las pruebas que trabajan con la media de la varianza son los más potentes cuando las poblaciones que se comparan se distribuyen de forma normal.
- Utilizar la media truncada mejora la prueba cuando los datos siguen una distribución de Cauchy (colas grandes).
- La mediana consigue mejorarlo cuando los datos siguen una distribución asimétrica.

Por lo general, si no se puede alcanzar cierta seguridad de que las poblaciones que se comparan son de tipo normal, es recomendable recurrir a test que comparen la mediana de la varianza (Amat 2016, p. 1).

1.2.3.3. Independencia

La prueba de independencia de ji cuadrado es una prueba estadística de hipótesis que se usa para determinar si dos variables categóricas o nominales pueden estar o no relacionadas (Cerda L, Villarroel Del P 2007, p. 414).

1.2.4. Estadística no Paramétrica

Las técnicas estadísticas de estimación de parámetros, intervalos de confianza y prueba de hipótesis son, en conjunto, denominadas estadística paramétrica y son aplicadas básicamente a variables continuas. Estas se basan en especificar una forma de distribución de la variable aleatoria y de los estadísticos derivados de los datos. En estadística paramétrica se asume que la

población de la cual la muestra es extraída es NORMAL o tienen distribución normal. Esta propiedad es necesaria para que la prueba de hipótesis sea válida (Quispe et al. 2019, pp. 12-17).

Las pruebas no paramétricas reúnen las siguientes características:

1) son más fáciles de aplicar;

2) son aplicables a los datos jerarquizados;

3) se pueden usar cuando dos series de observaciones provienen de distintas poblaciones;

4) son la única alternativa cuando el tamaño de muestra es pequeño

5) son útiles a un nivel de significancia previamente especificado.

Variable dependiente	Una muestra (bondad de ajuste)	Muestras relacionadas		Muestras independientes	
		2 muestras	>2 muestras	2 muestras	>2 muestras
Nominal	Binomial Chi-Cuadrado Rachas	McNemar	Cochran	-	-
Ordinal/ Intervalo	Kolmogorov-Smirnov	Signos Wilcoxon	Friedman Kendall	Rachas de Wald-Wolfowitz U de Mann-Whitney Moses Kolmogorov-Smirnov	Mediana Kruskal-Wallis Jonckheere-Terpstra

Figura 11-1. Resumen de las principales pruebas estadísticas no paramétricas

Fuente: Quispe,2019.

1.2.4.1. Ventajas de las pruebas estadísticas no paramétricas

Una serie de ventajas son mencionadas por (Ardila 1966, pp. 89-102), a continuación se da a conocer cada una de ellas:

- Las probabilidades que se obtiene con la mayoría de ellas son exactas, sin tener en cuenta la forma de la distribución de la población de la cual se recaudó la muestra. Algunas pueden presuponer identidad de forma de dos o más distribuciones y otras presuponen distribuciones simétricas de las poblaciones. En ciertos casos presupone que la distribución es continua.
- Si los tamaños de la muestra son tan pequeños como N=6, no hay alternativa posible, debe usarse una prueba estadística no paramétrica a menos que se conozca exactamente la naturaleza de la distribución de la población.

- Existen pruebas estadísticas no paramétricas para tratar muestras tomadas de observaciones de diferentes poblaciones. Ninguna prueba paramétrica puede tratar tales datos sin hacer presupuestos irreales.
- Hay pruebas estadísticas no paramétricas para tratar datos que se dan en rangos y datos cuyos puntajes numéricos tienen solo la fortaleza de rangos; o sea datos en los cuales no es posible realizar verdadera cuantificación.
- Los métodos no paramétricos pueden tratar datos que son simplemente de clasificación, que se miden solo en la escala nominal. Los métodos paramétricos no pueden hacerlo.
- Las pruebas estadísticas no paramétricas son mucho más fáciles de aprender y de aplicar que las paramétricas (Ardila 1966, pp. 89-102).

1.2.4.2. Desventajas de las pruebas estadísticas no paramétricas

Asimismo (Ardila 1966, pp. 89-102) también presenta las desventajas que puede tener el uso de las pruebas estadísticas no paramétricas.

- Si cumplen todos los presupuestos del modelo paramétrico, y si la medida requiere cierta fortaleza, entonces las pruebas estadísticas no paramétricas constituyen un desperdicio de datos. El grado de desperdicio se mide por la relación poder-eficacia.
- No existen métodos no paramétricos para probar interacciones en el modelo del análisis de varianza a menos que se hagan ciertos presupuestos sobre aditividad.

CAPÍTULO II

2. MARCO METODOLÓGICO

2.1. Tipo y Diseño de Investigación

La presente investigación es:

- Según en el método de investigación es de tipo cuantitativa
- Según el objetivo es teórica, puesto que la investigación busca identificar los métodos clúster similares en ASI y LA para efectuar la respectiva comparación entre ellos.
- Según el nivel de profundización en el objeto de estudio es explicativa, ya que la comparación de las técnicas clúster en SIA y LA permitirá determinar que método maneja mejores recursos de espacio de memoria y tiempo.
- Según la manipulación de variables es experimental puesto que la matriz de datos para esta investigación proviene de una fuente de información primaria (simulaciones de datos).
- Según el tipo de inferencia es hipotético-deductiva ya que busca determinar cuáles son los métodos clúster que optimizan tiempo de ejecución y memoria ocupada con las técnicas en auge en SIA y LA
- Según el periodo temporal es de tipo transversal esto se debe a que el objetivo principal de la investigación tiene un enfoque en un único período de tiempo

2.2. Localización del estudio

El objetivo es analizar las diferentes alternativas de ubicación espacial, temporal, temática, etc. del proyecto. La localización espacial tiene por objetivo analizar los diferentes lugares donde es posible ubicar el proyecto.

2.3. Población en estudio

La población está formada por 100000 bases de datos (y la combinación de todos sus datos en el intervalo [0;1]) formadas hasta por un máximo de 1000 observaciones y 100 variables.

2.4. Tamaño de la muestra

Por el gran tamaño de la población, se procedió a escoger una muestra utilizando el método de muestreo aleatorio simple con parámetro de interés la media. Para este proceso se consideró la fórmula (2.1) para el cálculo de la muestra:

$$n = \frac{s^2}{\frac{E^2}{Z^2_{\frac{\alpha}{2}}} + \frac{s^2}{N}} \quad (2.1)$$

Cada uno de los parámetros a usarse se describen a continuación:

n: tamaño de la muestra a obtenerse

Z: nivel de confianza

E: precisión o error admitido

s: desviación estándar

N: tamaño de la población o universo

Para la aplicación de la fórmula basándose en los datos obtenidos para el presente trabajo se utilizaron los siguientes parámetros:

desviación estándar = 1

$\alpha = 5\%$

Z = 1,96

E=10%

N = 100000

Con la aplicación de dicha información proporcionada anteriormente se generó un tamaño de la muestra igual a 382,6758 bases de datos a estudiar.

2.5. Método de muestreo

El método de muestreo es aleatorio simple por estratos debido a que se trabajó con información generada mediante simulaciones para efectuar la comparación entre técnicas, garantizando de esta manera la selección aleatoria de los datos para considerar una muestra en estudio que sea representativa con respecto a la población.

2.6. Recolección de información

El presente trabajo de investigación se realizó y sustentó mediante tesis y papers sobre el tema en estudio, de dichos trabajos e investigaciones se indagó a profundidad con el objetivo de extraer toda la información necesaria y pertinente que avalará el trabajo permitiendo cumplir los objetivos

planteados. Los datos que se usarán para el análisis experimental se generarán por cuenta del investigador a través de funciones en el software R el mismo que es gratuito, permitiendo de esta manera que el tema sea completamente accesible y además asegurando que serán variables modales que son las que se requieren y proponen para el análisis comparativo.

2.7. Variables en estudio

Considerando los parámetros que se pretenden evaluar de los métodos clúster en ASI y LA se estableció como variables dependientes al espacio de memoria y al tiempo de ejecución las cuales son de tipo numérico.

Para establecer las variables independientes se consideró a los métodos clúster en ASI y LA, las mismas que son de tipo cualitativo.

Las variables en estudio se describen en la **Tabla 1-2**.

2.7.1. Operacionalización de variables

Tabla 1-2: Descripción de variables en estudio

VARIABLE	CONCEPTO	INDICADOR	INSTRUMENTO
Variable independiente: Métodos de análisis clúster (ASI y LA)	El análisis clúster es un conjunto de técnicas multivariantes utilizadas para clasificar a un conjunto de individuos en grupos homogéneos	Función de los métodos de análisis seleccionados	Ficha de observación y chequeo de las características de las técnicas de análisis.
Variable dependiente: Espacio de memoria y Tiempo de ejecución	El tiempo requerido por un algoritmo expresado como una función del tamaño de la entrada del problema se denomina complejidad en tiempo del algoritmo y se denota T(n). El comportamiento límite de la complejidad a medida que crece el tamaño del problema se denomina complejidad en tiempo asintótica.	Tiempo en segundos Memoria en Mb	Variables (variables, vectores, matrices, listas, data frames) de captura de datos de memoria y tiempo

Fuente: Elaboración Propia.

Realizado por: Armas, Shirley, 2022

2.8. Materiales y métodos

Los materiales que se usaron para realizar el estudio comparativo planteado con respecto a las técnicas clúster en Análisis Estadístico Implicativo y Learning Analytics se detallan en la **Tabla 2-2**.

Tabla 2-2: Material informático (Software y hardware)

Requisitos	Computadora
Procesador	Core I5
Velocidad	2.50GHz
Memoria RAM	8,00 GB
Sistema Operativo	Windows 10 Pro 64 bits
Programas	• Software R (versión 4.1.2) • RStudio (versión 1.1.456) • Rchic (versión 0.27)

Fuente: Elaboración propia.
Realizado por: Armas, Shirley, 2022

2.8.1. Software R

R es un software de lenguaje estadístico gratuito que está conformado por una gran variedad de herramientas que posee la gran facilidad de implementar funciones a través de paquetes que ya hayan sido desarrollados. Este software es considerado como una herramienta muy poderosa para la manipulación de datos y obtención de resultados ya que facilita la comprensión y además consume pocos recursos considerando también que se encuentra al alcance de todos ya que como se mencionó anteriormente es un software libre siendo uno de los más usados en la investigación por la comunidad estadística (Ver Anexo A).

2.8.2. RStudio

RStudio es un entorno de desarrollo integrado (IDE) para R, dicho entorno consta de una consola la cual ayuda a la ejecución del código planteado; brindando herramientas útiles para el desarrollo del trabajo en el software lo cual permite realizar una mejor gestión de cálculos, memoria y espacio de trabajo (Ver Anexo B). Este entorno se desarrolla bajo modalidad de código abierto admitiendo con esto la combinación de diversos componentes de R facilitando así el aprendizaje para usuarios con herramientas de alto rendimiento con una accesible comprensión de resultados (Allaire 2012, pp. 165-171).

Dentro de este entorno se da la posibilidad de instalar los paquetes que se encuentren disponibles en R, uno de los cuales es de gran relevancia en el desarrollo del diseño cuasi experimental que es Rchic.

2.8.3. Rchic

Es un programa que permite realizar análisis estadísticos bajo los conceptos de similitud y cohesión entre las variables de interés, este paquete trabaja bajo el entorno de RStudio (Ver Anexo C).

CHIC facilita la comprensión de AEI ya que permite el uso de la mayoría de los métodos definidos en esta área. Tiene como objetivo la búsqueda de implicaciones con mayor relevancia entre las variables de un conjunto de datos en estudio. Se basa en la temática de organización de las implicaciones como una jerarquía cohesiva o gráfico implicativo, también proporciona una jerarquía de similitudes de las variables de interés (Couturier, Almouloud 2009, pp. 3-5).

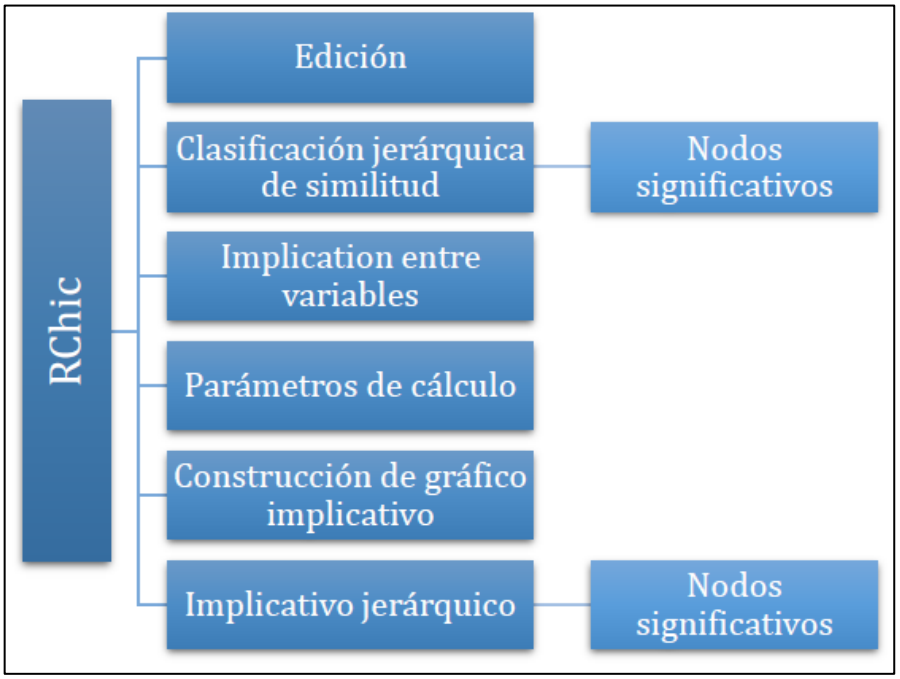


Figura 1-2. Diagrama de Flujo del Software Rchic

Fuente: Couturier y Almouloud,2009.

Anteriormente se había mencionado que CHIC permite construir dos tipos de árboles y un gráfico. De los árboles más conocidos se tiene al de similitud el cual realiza la construcción de una jerarquía ascendente o también hay la posibilidad de generar un gráfico original al cuál se le denomina gráfico implicativo debido que brinda la posibilidad de que el usuario seleccione las reglas de asociación y variables deseadas para su aparición (Couturier, Ag Almouloud 2007, pp. 41-49).

2.8.4. Otras herramientas

Con el pasar del tiempo y con las nuevas necesidades que se han generado al realizar investigaciones con distintos tipos de datos, R ha presentado un gran desarrollo en los paquetes a integrarse; teniendo en cuenta que deben ser instalados posterior a la instalación de R y seguidamente de RStudio, en el siguiente apartado se detallan cada uno de estos paquetes que se usaron para el desarrollo del diseño cuasiexperimental.

2.8.4.1. Clust Of Var

Este paquete se centra en el análisis de conglomerados sobre un conjunto de variables. Las variables en estudio pueden ser de tipo cuantitativo, cualitativo o mixtas. Se debe considerar que en el caso de tener valores faltantes en la información se deben reemplazar por medias para variables de tipo cuantitativo y por ceros en la matriz de indicadores para variables cualitativas (Chavent et al. 2017, pp. 1-3).

- **Hclustvar:**

Realiza la agrupación jerárquica de variables ya sean de tipo, cualitativo, cuantitativo o de tipo mixto. Se detallan a continuación las particularidades de dicha librería (Chavent et al. 2017, pp. 1-3).

Uso*:*

hclustvar(X.quanti = NULL, X.quali = NULL, init = NULL)

Argumentos:

X.quanti: una matriz numérica de datos, o un objeto que puede ser forzado a tal matriz (como un vector numérico o un marco de datos con todas las columnas numéricas).

X.quali: una matriz categórica de datos, o un objeto que puede ser forzado a tal matriz (como un vector de caracteres, un factor o un marco de datos con todas las columnas de factores).

Init: una partición inicial (un vector de enteros que indica el grupo al que se asigna cada variable).

Código implementado:

Para el tratamiento de los datos en estudio, con el fin de evaluar el tiempo y memoria se implementó la siguiente función.

```
HclustVar <- function(x){
```

```
  hc <- read.csv(x,sep = ";")
  hc <- hc[,-1]
  v <- hclustvar(X.quanti = hc, init = NULL)
  plot(v)
  return(v)
}
```

2.8.4.2. Función gc

Para realizar la recolección de los elementos que no están siendo utilizados en memoria se lo puede realizar con gc ya que dicha función realiza la acción mencionada anteriormente con el fin de liberar espacio en la memoria y así brindar la posibilidad de optimizar procesos (Schork 2021, p. 1).

Uso:

gc(verbose = getOption("verbose"), reset = FALSE, full = TRUE)

Argumentos:

Verbose: lógico; si es VERDADERO, la recolección de basura imprime estadísticas sobre las celdas contras y el espacio asignado para los vectores.

Reset: lógico; si es VERDADERO, los valores para el espacio máximo utilizado se restablecen a los valores actuales.

Full: lógico; si es VERDADERO, se realiza una recopilación completa; de lo contrario, solo se pueden recopilar objetos asignados más recientemente.

2.8.4.3. Microbenchmark

Este paquete proporciona funciones que permiten al usuario medir el tiempo de ejecución de las expresiones planteadas, su finalidad comprende la evaluación y comparación de distintos métodos para implementar funciones con aspectos destacados en rendimiento. En contraste con la combinación de system.time y replicate se nota una gran ventaja con microbenchmark ya que este paquete intenta utilizar el método más apropiado acorde a los datos en evalúo y al sistema operativo usado; proporcionando funciones útiles para comparar de una manera rápida los resultados de tiempo para distintas expresiones (Mersmann, Krey 2011, p. 142).

Uso:

microbenchmark(...,list = NULL, times = 100L,unit = NULL, check = NULL,control = list(),setup = NULL)

Argumentos:

…: Expresiones para comparar
List: Lista de expresiones no evaluadas para comparar
Times: Número de veces para evaluar cada expresión
Unit: Unidad predeterminada utilizada en summary y print

Check: Una función para verificar si las expresiones son iguales. Por defecto NULL el que omite el check. Además de una función, se puede proporcionar una cadena. La cadena 'igual' comparará todos los valores usando all.equal, 'equivalente' comparará todos los valores usando all.equaly check.attributes = FALSE, e 'idéntico' comparará todos los valores usando identical.

Control: Lista de argumentos de control

Setup: Una expresión no evaluada que se ejecutará (sin tiempo) antes de cada expresión de referencia|

2.8.4.4. Cluster

Este paquete constituye una serie de métodos para el análisis de conglomerados o agrupamiento de datos, consiste en una tarea de aprendizaje automático no supervisada. (Maechler 2021, p. 1) menciona que implica descubrir automáticamente la agrupación natural en los datos. A diferencia del aprendizaje supervisado (como el modelado predictivo), los algoritmos de agrupación solo interpretan los datos de entrada y encuentran grupos naturales o agrupaciones en el espacio de características.
Dentro de este paquete se encuentran varias funciones las cuáles permiten el análisis clúster de un conjunto de datos, es así como se destaca la siguiente:

Diana

Calcula una agrupación jerárquica divisiva del conjunto de datos que devuelve un objeto de clase diana (Patnaik, Bhuyan, Krishna Rao 2016, pp. 407-418).

- **Uso**

diana(x, diss = inherits(x, "dist"), metric = "euclidean", stand = FALSE, stop.at.k = FALSE, keep.diss = n < 100, keep.data = !diss, trace.lev = 0)

- **Argumentos**

x: matriz de datos o marco de datos, o matriz de disimilitud u objeto, según el valor del argumento diss

diss: indicador lógico: si es VERDADERO (predeterminado para objetos de disimilitud o disimilitud), entonces x se considerará como una matriz de disimilitud. Si es FALSO, entonces x se considerará como una matriz de observaciones por variables.

metric: cadena de caracteres que especifica la métrica que se utilizará para calcular las diferencias entre las observaciones.

Las opciones disponibles actualmente son "euclides" y "manhattan". Las distancias euclidianas son la raíz de la suma de los cuadrados de las diferencias, y las distancias de Manhattan son la suma de las diferencias absolutas. Si x ya es una matriz de disimilitud, entonces este argumento será ignorado.

Stand: lógico; si es cierto, las medidas en x se estandarizan antes de calcular las diferencias. Las mediciones se estandarizan para cada variable (columna), restando el valor medio de la variable y dividiendo por la desviación absoluta media de la variable. Si x ya es una matriz de disimilitud, entonces este argumento será ignorado.

- **Código implementado:**

```
Diana <- function(x){
  dian <- read.csv(x, sep = ";")
  d <- diana(x = dian, metric = "euclidean", stand = TRUE)
  dend <- as.dendrogram(d)
  fviz_dend(dend)
  return(d)
}
```

2.8.4.5. FastCluster

Este es un paquete dos en uno que proporciona interfaces tanto para R como para 'Python'. Implementa rápidas rutinas de agrupamiento jerárquico y aglomerante. Parte de la funcionalidad está diseñada como reemplazo directo de las rutinas existentes: linkage() en el paquete 'SciPy' 'scipy.cluster.hierarchy', hclust() en el paquete 'stats' de R y el paquete 'flashClust'. Proporciona la misma funcionalidad con el beneficio de una implementación mucho más rápida. Además,

existen rutinas de ahorro de memoria para la agrupación de datos vectoriales, que van más allá de lo que proporcionan los paquetes existentes (Müllner, Inc 2021, p. 2).

Hclust.vector

Esta función realiza una rápida agrupación jerárquica y aglomerativa de datos vectoriales a través del uso de algoritmos de ahorro de memoria.

- **Uso**

hclust.vector(X, method="single", members=NULL, metric='euclidean', p=NULL)

- **Argumentos**

X: una matriz (N×D) de valores 'dobles': N observaciones en D variables

Method: el método de aglomeración a utilizar. Debe ser (una abreviatura inequívoca de) uno de "único", "barrio", "centroide" o "mediana"

Members: NULL o un vector con longitud el número de observaciones.

Metric: la medida de distancia a utilizar. Este debe ser uno de "euclidean", "maximum", "manhattan", "canberra", "binary" o "minkowski". Se puede dar cualquier subcadena inequívoca.

P: parámetro para la métrica de Minkowski.

- **Código implementado:**

```
hclust_vector <- function(x){
  hv <- read.csv(x, sep = ";")
  dists <- dist(hv, method = "euclidean")
  hc <- hclust.vector(X = dists, members=NULL, metric='euclidean', p=NULL)
  dend <- as.dendrogram(hc)
  fviz_dend(dend)
  #plot(hc)
  return(dend)
```

}

2.8.4.6. callSimilarityTree

Dicha función viene dada al momento de la instalación de rchic la cual consiste en el cálculo de un árbol de similitud con el previo conocimiento de que dicha medida es simétrica, el árbol obtenido consta de la agrupación de las variables según su medida de similitud.

- **Uso**

callSimilarityTree(fileName, contribution.supp = FALSE, typicality.supp = FALSE, verbose = FALSE)

- **Argumentos**

fileName: nombre del archivo que contiene los datos

contribution.supp: booleano para calcular la contribución de las variables suplementarias

typicality.supp: booleano para calcular la tipicidad de las variables suplementarias

verbose: booleano para dar muchos detalles

- **Código implementado:**

```
Similarity_Tree<- function(x){
  Similarity <-  callSimilarityTree(x,contribution.supp=FALSE,typicality.supp=
                                    FALSE,verbose=FALSE)
  return(Similarity )
}
```

2.8.4.7. callHierarchyTree

La función permite realiza el cálculo de un árbol de jerarquía con el índice de cohesión que no es simétrico, la agrupación de variables en el árbol viene dada bajo dicha medida.

- **Uso**

callHierarchyTree(fileName, contribution.supp = FALSE, typicality.supp = FALSE, computing.mode = 1, verbose = FALSE)

- **Argumentos**

fileName: nombre del archivo que contiene los datos

contribution.supp: booleano para calcular la contribución de las variables suplementarias

Typicality.supp: booleano para calcular la tipicidad de las variables suplementarias

Computing.mode: controla el modo de cálculo: 1=implicación clásica, 2=implicación clásica+confianza, 3=implicación

Verbose: Booleano para dar muchos detalles

- **Código implementado:**

```
hierarchy_Tree <- function(x){

 cohesion <- callHierarchyTree(x, contribution.supp = FALSE,typicality.supp

                    = FALSE,computing.mode = 3, verbose = FALSE)

 return(cohesion)

}
```

2.9. Diseño y experimentación

Para demostrar las hipótesis se planteó un cuasi experimento en la ingeniería de software de tipo RGXO. 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 (hclust.vector, hclustvar y diana) y 2 técnicas usadas en ASI (callHierarchyTree y callSimilarityTree). Se trabajará con un nivel de significancia del 95%, considerando que en el campo de la investigación el método explicativo permite determinar la relación de causa efecto entre los algoritmos más comunes de LA y AEI para lo cual se ejecutarán los siguientes pasos (Campbell, Stanley 1996, pp. 13-25).

Proceso del diseño Cuasiexperimental

1. Generación de la base de datos informática (con variables v1, v2, v3; las misma que almacenan números categóricos)

2. Determinación de variables dependientes, factores, variables intervinientes,

3. Definición del diseño cuasiexperimental a utilizar.

4. Análisis del tipo de datos.

5. Selección de la prueba estadística a utilizar.

6. Comprobación de supuestos.

7. Ejecución del experimento.

8. Conclusiones sobre las hipótesis.

2.10. Procedimiento Experimental

Se busca determinar que técnica clúster posee mejor rendimiento en recursos de memoria y tiempo de ejecución, desarrollándose primeramente una función la cual permite calcular dichos parámetros en cada una de las 382 bases de datos en estudio. Estos datos están comprendidos por variables modales que son aquellas que están en el intervalo de 0 a 1.
Para efectuar la comparación se tiene como variable independiente a los métodos de análisis clúster (ASI y LA) y como variable dependiente al espacio de memoria y tiempo de ejecución. Se realizará este experimento comprendidas dos áreas las cuales se especifican a continuación

1. Comparación de tiempos de cada una de las funciones clúster en estudio, este tiempo viene medido en segundos.
2. Comparación de cantidad de memoria para cada uno de los métodos clúster, dicha cantidad vienen medida Mb.

Para la obtención de resultados se seguirá una serie de pasos fundamentales los cuáles consideran el diseño cuasi experimental con el uso del software estadístico R, RStudio y Rchic, y el tipo de datos en estudio de los que se tiene previo conocimiento que fueron generados de manera aleatoria.

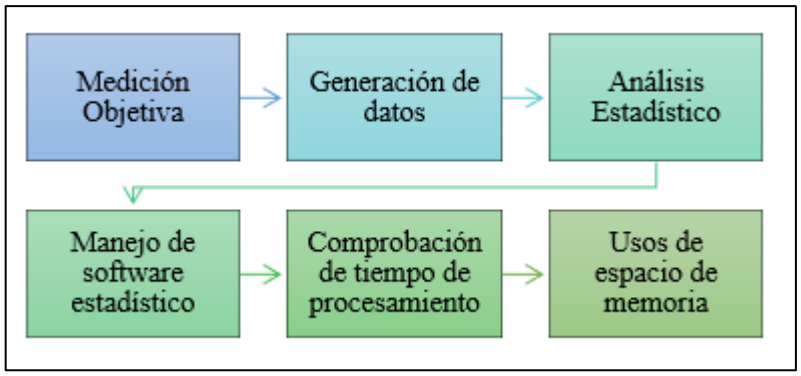


Figura 2-2. Proceso para la obtención de resultados

Fuente: Naranjo y Pazmiño, 2018.

CAPÍTULO III

3. MARCO DE ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

3.1. Trabajos previos relacionados

A continuación, se muestran estudios comparativos entre las técnicas ASI y otras técnicas de análisis.

3.1.1. ASI, CFA y Agrupación Jerárquica.

Un primer estudio se basa en el artículo de, 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, 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 (**Tabla 1-3**) para captar las formas en que los estudiantes utilizan los diferentes tipos de modificación de la figura (Michael et al., 2010, pp. 227-230).

Tabla 1-3: Comparación entre CFA, agrupación jerárquica y el método implicativo

CFA	Agrupación jerárquica	Método implicativo
Estructura factorial de la comprensión de figuras geométricas. Desarrollo de un modelo que incluye dos factores latentes para los efectos de tres tipos de modificación de la figura y un factor de segundo orden que representa la aprehensión operativa de la figura geométrica. Diferencia en la fuerza de las relaciones de los tres factores de primer orden con el factor de segundo orden.	Clasificación jerárquica y consistencia de las respuestas a las modificaciones de figuras geométricas. Agrupaciones de similitud entre las medidas observadas en las respuestas a las tres formas de modificar una figura geométrica. Agrupación separada de las variables de la modificación mereológica y la óptica. Similitud relativamente débil de las modificaciones.	Relaciones entre las respuestas de los alumnos a las modificaciones de las figuras geométricas. Implicaciones entre las variables observadas en las respuestas de los alumnos a los tres tipos de modificaciones de las figuras geométricas. Las tareas de modificación de la forma fueron más complejas que las tareas de modificación de la forma mereológica u óptica. Las tareas de modificación óptica fueron las más fáciles.

Fuente: Michael et al. 2010.

Realizado por: Michael,2010

3.1.2. Árbol jerárquico del ASI y clúster

En el artículo, 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 (Pazmiño Maji, Rubén, García Peñalvo y Conde González, 2017,pp 1-7).

3.2. Identificación de métodos clúster para ASI y LA

Mediante la investigación de trabajos similares se pudieron identificar los métodos clúster más usados en ASI y LA. Con respecto a una de las funciones propuestas como fue dendro.variables perteneciente al paquete CluMix, se tuvo la limitación de que ya no trabaja para versiones actuales de R y RStudio debido a que dependía de una función C de otro paquete que ya no se encuentra disponible en CRAN por lo que se optó por buscar otra función similar para proceder con la comparación. Esta aseveración se corroboró con la autora Manuela Hummel creadora del paquete CluMix (**Figura 1-3**).

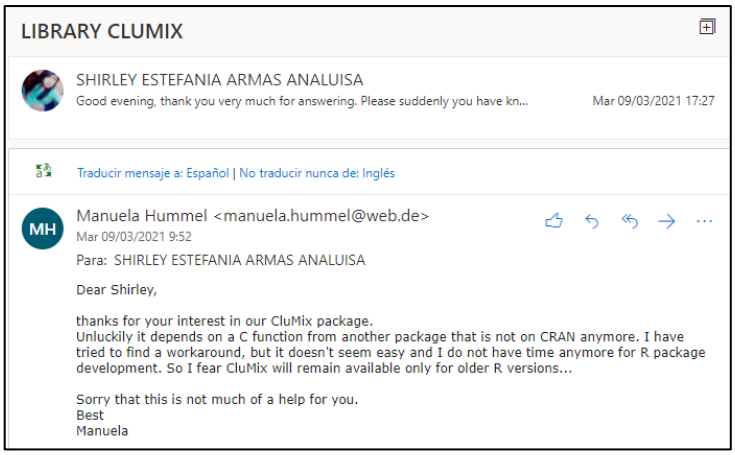


Figura 1-3. Verificación de funcionalidad de CluMix

Realizado por: Armas Shirley, 2022.

A continuación, en la **Tabla 2-3** se detallan los métodos a emplearse para efectuar las comparaciones

Tabla 1-3: Métodos clúster en ASI y LA

Análisis Estadístico Implicativo		Learning Analytics	
Rchic	callSimilarityTree	Clúster	Diana
	callHierarchyTree	**Fastcluster**	hclust.vector
		ClusOfVar	hclustvar

Fuente: Elaboración Propia.

Realizado por: Armas, Shirley, 2022

3.3. Construcción de la base de datos

Para la generación de la base de datos final se tuvo como punto de partida las 382 bases de datos, mismas que fueron procesadas a través de una función generada en RStudio para así obtener los tiempos de ejecución y espacio de memoria de cada una de ellas para los métodos clúster en estudio.

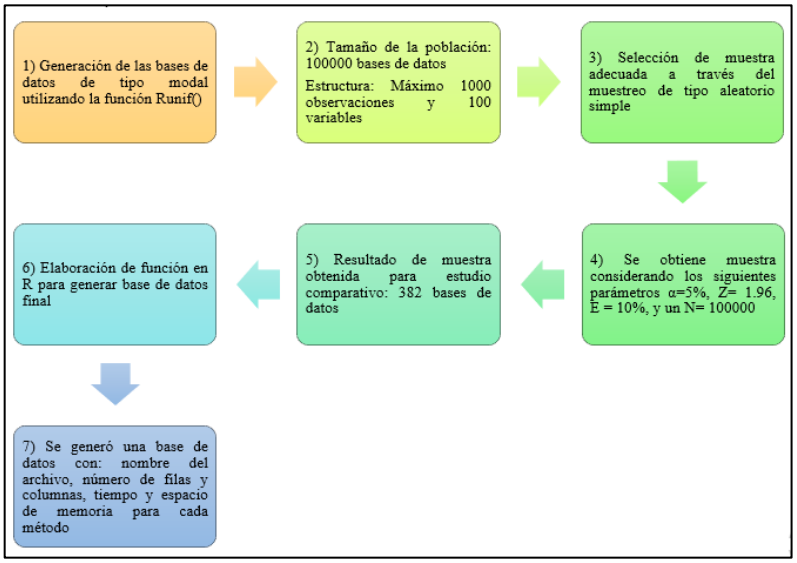


Figura 2-3. Proceso de elaboración de la base de datos final

Realizado por: Armas Shirley, 2022.

La base de datos final se obtuvo en un tiempo de 50 horas totalmente trabajadas, se generaron en 4 grupos de 60 (240 bases de datos) y un grupo final de 42, el código implementado para obtener dichos resultados se muestra en el Anexo C.

3.4. Técnicas de análisis

Se detallan cada una de las técnicas usadas para dar tratamiento a los datos obtenidos, se realizó primeramente un análisis descriptivo de los datos tiempo de ejecución (medido en segundos) y espacio de memoria (medido en megabytes).

3.4.1. Análisis Descriptivo: Variable Tiempo de ejecución

Tabla 3-2: Análisis Descriptivo de tiempo de ejecución

	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: Elaboración Propia.
Realizado por: Armas, Shirley, 2022

En la **Tabla 3-3** se analiza 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 (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 leptocúrtica 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.

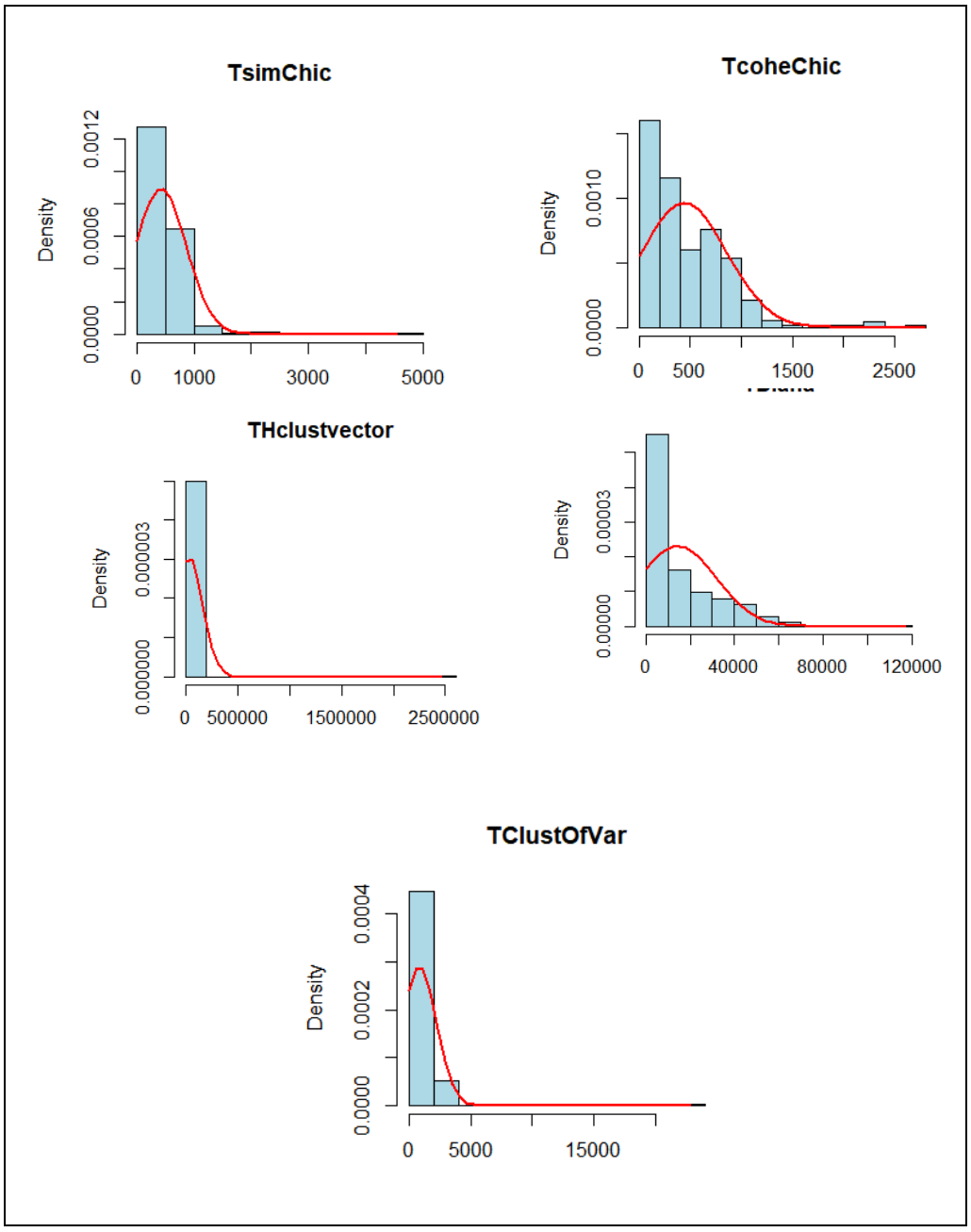


Gráfico 1-3. Histogramas Tiempo de ejecución

Realizado por: Armas Shirley, 2022.

En el apartado anterior se evaluaron los coeficientes de Asimetría y Kurtosis en donde se determinó que presentaban una asimetría positiva con Kurtosis platicúrtica, dicha aseveración se confirma también de manera gráfica (**Ver Gráfico 1-3**) dando a una idea de que posiblemente los datos no parecen seguir una distribución normal.

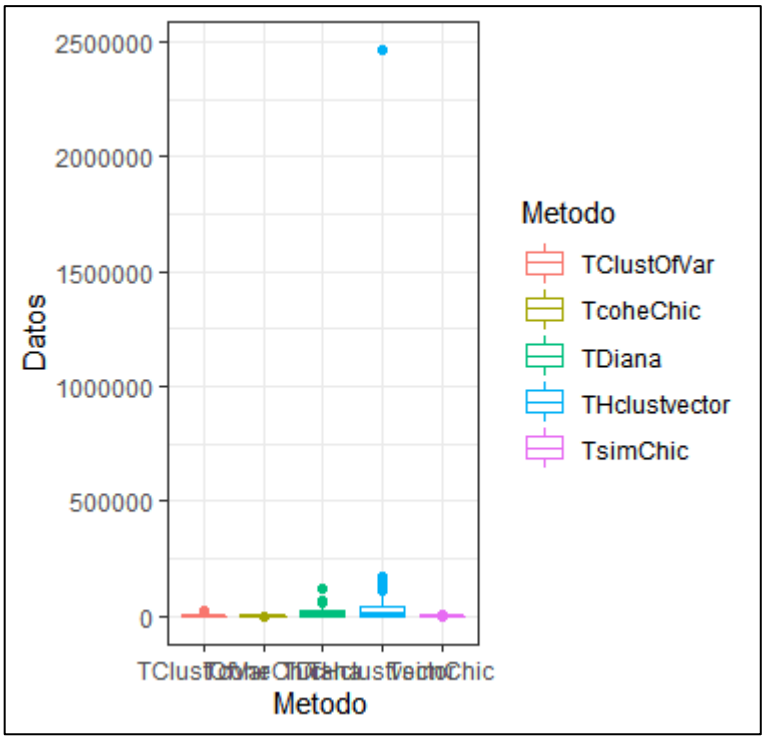


Gráfico 2-3. Diagrama de cajas correspondiente al tiempo

Realizado por: Armas Shirley, 2022.

Se evidenció que los datos promedio correspondientes al tiempo de ejecución son similares para cada método clúster de ASI y LA, siendo el método Thclustvector aquel que presentó mayor variabilidad, además se pudo notar la presencia de datos atípicos en los métodos TDiana y con mayor notoriedad en Thclustvector con una observación muy distante lo cual ocasionaría distorsión en los resultados de los análisis (**Gráfico 2-3**).

3.4.2. Comprobación de supuestos: Variable tiempo de ejecución

Se procedió a la verificación de los supuestos que deben cumplir los datos lo que permitió determinar si se requerirá para el análisis pruebas paramétricas o no paramétricas. Los supuestos fundamentales a verificarse son normalidad y homogeneidad de varianza

3.4.2.1. Supuesto de Normalidad

A continuación, se muestra la gráfica de cuartiles (**Gráfico 3-3**) que nos provee una idea gráfica de la normalidad de los datos sobre el índice de rangos (cuartiles) en los diferentes métodos.

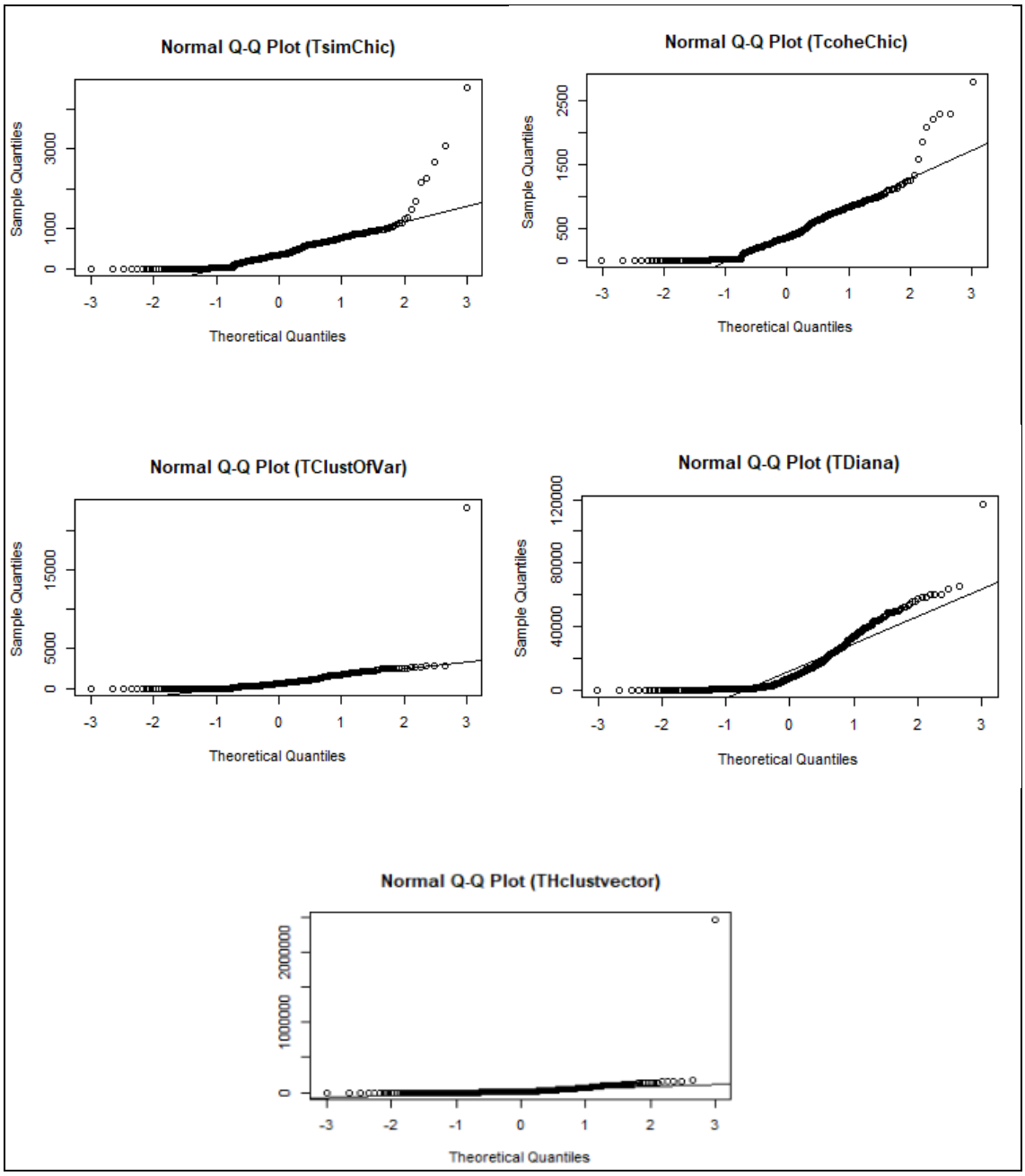


Gráfico 3-3. Gráficos de cuartiles para los métodos de tiempo

Realizado por: Armas Shirley, 2022.

Al analizar la cercanía de la recta a la curva presentada, sobre el tiempo de ejecución usado por cada uno de los métodos clúster se observa que cierta cantidad de puntos están situados en la línea recta mientras que otros no, esto se evidencia claramente ya que sobresalen de la recta por lo cual para verificar dicha aseveración se realizaron los respectivos tests de normalidad a un nivel de significancia de 0.05.

- *Planteamiento de Hipótesis*

H_0: Tiempo de ejecución $\sim \mathcal{N}(\mu, \sigma^2)$

H_1: Tiempo de ejecución $\nsim \mathcal{N}(\mu, \sigma^2)$

- *Nivel de significancia*

$\alpha = 0.05$

- *Estadístico y valor p*

```
Data$Metodo: TClustOfVar

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.43336, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: TcoheChic

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.43336, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: TDiana

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.43336, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: THclustvector

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.43336, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: TsimChic

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.43336, p-value < 0.00000000000000022
```

Figura 3-3. Test de Normalidad - Kolmogorov Smirnov

Realizado por: Armas Shirley, 2022.

- *Regla de decisión*

Si el p valor es menor que 0.05 (p-value < 0.05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla

- *Decisión*

Existe suficiente evidencia para rechazar la hipótesis nula, es decir que ninguno de los 5 métodos correspondientes al tiempo de ejecución sigue una distribución normal a un nivel de confianza del 95%, el valor p obtenido para cada uno de los métodos es igual a 0.0000000000000022.

3.4.2.2. Transformación a normalidad

En busca de la transformación de los datos de la variable tiempo de ejecución a una distribución normal se usó la transformación de Box Cox con el objetivo de determinar la potencia óptima a la que se debe elevar la variable de interés y así buscar obtener normalidad en los datos.

- *Planteamiento de Hipótesis*

H_o: Los datos siguen una distribución normal
H_1: Los datos no siguen una distribución normal

- *Nivel de significancia*

$\alpha = 0.05$

- *Estadístico y valor p*

	TsimChic	TcoheChic	THclustvector	TDiana	TClustOfVar
Log	1.042516e-70	6.509209e-70	1.539209e-19	1.012224e-29	2.020290e-53
sqrt	1.341992e-19	1.758028e-22	4.904310e-31	9.353180e-13	1.425020e-16
1/x	2.037147e-224	1.738456e-225	9.800122e-163	2.938127e-154	1.856225e-214
Box Cox	7.402359e-23	3.118272e-24	1.033750e-10	6.578169e-16	2.897955e-24

Figura 4-3. Transformaciones para normalidad – Tiempo

Realizado por: Armas Shirley, 2022.

- *Regla de decisión*

Si el p valor es menor que 0.05 (p-value < 0.05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla

- *Decisión*

Existe suficiente evidencia para rechazar la hipótesis nula debido a que los valores p obtenidos para cada una de las variables son muy pequeños siendo menores al nivel de significancia de 0.05, es decir ninguno de los 5 métodos correspondientes al tiempo de ejecución siguen una distribución normal a pesar de usar distintas transformaciones (Ver **Figura 4-3**).

3.4.2.3. Supuesto de Homocedasticidad

Test de Levene

- *Paso 1: Planteamiento de Hipótesis*

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

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

- *Paso 2: Nivel de significancia*

$\alpha = 0.05$

- *Paso 3: Estadístico de Prueba*

```
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
```

Figura 5-3. Resultados Test de Levene

Realizado por: Armas Shirley, 2022.

- *Paso 4: Regla de Decisión*

Si el p valor es menor que 0.05 (p-value < 0.05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla

Paso 5: Decisión

El p-value obtenido es igual a 0.0000000000000022 el cual es menor al nivel de significancia propuesto (0.05) por lo que 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.

Al no cumplir con los supuestos de normalidad y homocedasticidad se determinó que no se pueden utilizar métodos paramétricos para el estudio de los datos, se constata el uso de pruebas no paramétricas para la variable tiempo de ejecución.

3.4.3. Análisis Descriptivo: Variable Espacio de memoria

Tabla 3-3: Análisis Descriptivo de Espacio de memoria

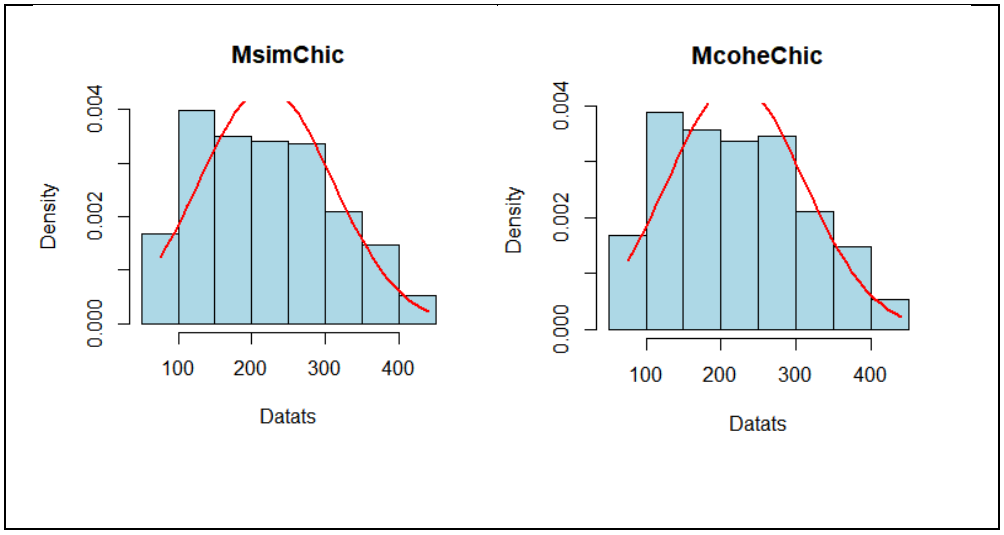
	MsimChic	McoheChic	Mhclustvector	Mdiana	MClustOfVar
Media	219.55	219.81	219.22	218.94	218.95
Mediana	213.8	214.27	213.2	213.08	213.1
Sd	90.77	90.69	90.67	90.67	90.67
Varianza	8238.82	8225.44	8221.05	8220.27	8220.34
Asimetría	0.35	0.35	0.35	0.35	0.35
Kurtosis	2.21	2.21	2.21	2.21	2.21

Fuente: Elaboración Propia.

Realizado por: Armas, Shirley, 2022

En la **Tabla 4-3** se realiza el análisis descriptivo de la variable espacio de memoria para los métodos clúster de SIA y LA de los cuales podemos decir que: en promedio el método que ocupa menor cantidad de memoria es Mhclustvector (callHierarchyTree) con 219.22 megabytes con una mediana que indica que la mitad del tiempo empleado para el método es menor o igual a 213.2 y la otra mitad es mayor o igual a 213.2 , con respecto al método que ocupa en promedio mayor cantidad de memoria es hclustvar (ClustOfVar) con 218.95 megabytes del cual se tiene que el valor de la mediana indica que la mitad del tiempo empleado para el método es menor o igual a 213.1 y la otra mitad es mayor o igual a 213.1, con relación a que tan dispersos se encuentran dichos datos analizados con respecto al valor promedio se obtuvo que tres métodos presentan menor dispersión, hclustvector, diana y hclustvar con un valor igual a 90.67 y mayor dispersión callSimilarityTree con 90.77.

La asimetría de los datos con respecto al espacio de memoria usado permite notar que para todos los métodos en comparación la asimetría es positiva ya que todos los coeficientes determinados son positivos, obteniendo el mismo valor de 0.35 para todos. El coeficiente de Kurtosis refleja que la distribución que sigue cada uno de los métodos es leptocúrtica 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.



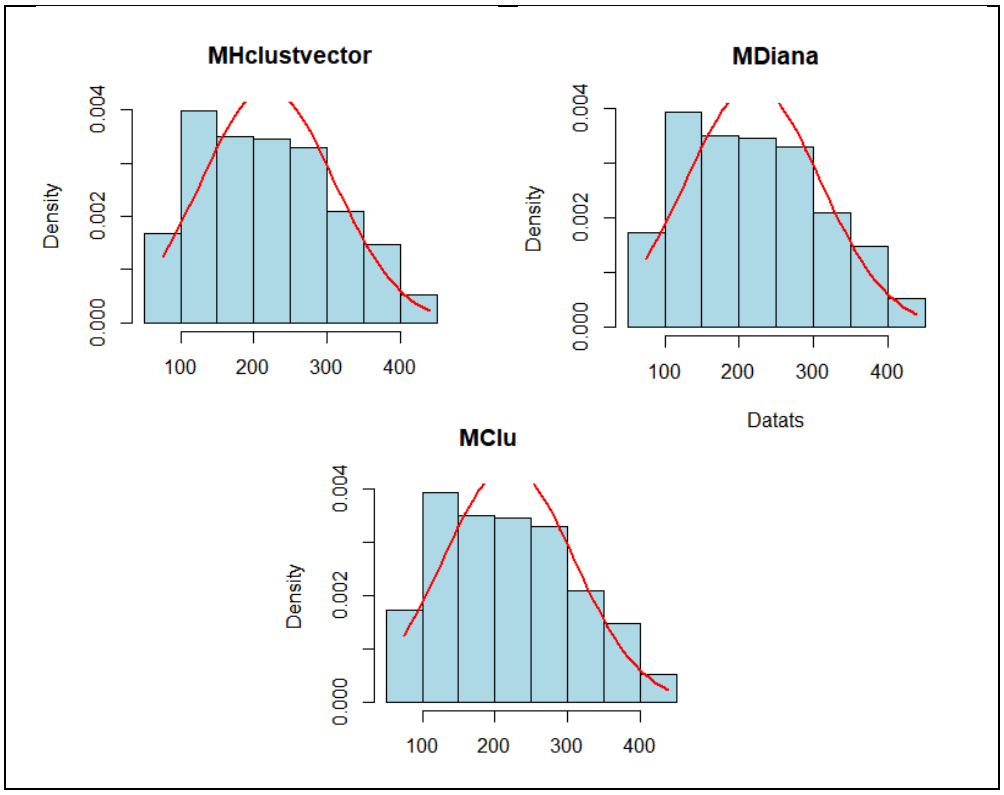


Gráfico 4-3. Histogramas Espacio de memoria

Realizado por: Armas Shirley, 2022.

Anteriormente se evaluaron los coeficientes de Asimetría y Kurtosis en donde se determinó que presentaban una asimetría positiva con Kurtosis platicúrtica, dicha afirmación se reconfirma también de manera gráfica (**Ver Gráfico 3-4**) dando a una idea de que posiblemente los datos no parecen seguir una distribución normal.

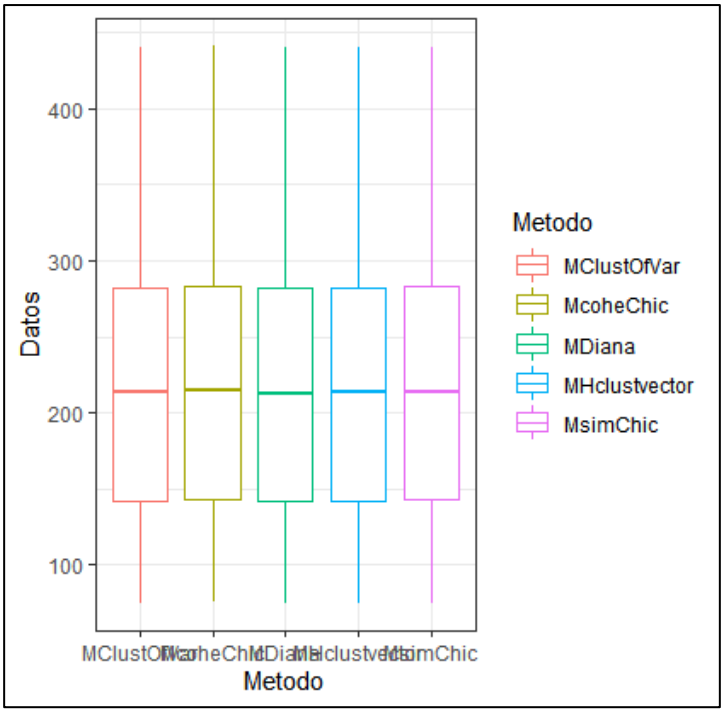


Gráfico 5-3. Diagrama de cajas de espacio de memoria

Realizado por: Armas Shirley, 2022.

El diagrama de cajas evidenció que los datos no presentan observaciones atípicas aparentemente, se reflejó que el consumo de memoria parte desde los 100 megabytes hasta 400 megabytes como se visualiza en la gráfica (Ver gráfico 5-3).

3.4.4. Comprobación de supuestos: Variable espacio de memoria

Anteriormente se verificó que los datos de la variable tiempo de ejecución para cada uno de los métodos no posee una distribución normal, dicho proceso se procedió a realizar para espacio de memoria en donde se obtuvieron los siguientes resultados del análisis.

3.4.4.1. Supuesto de Normalidad

A continuación, se muestra la gráfica de cuartiles (Grafico 3-6) que nos provee una idea gráfica de la normalidad de los datos sobre el índice de rangos (cuartiles) en los diferentes métodos.

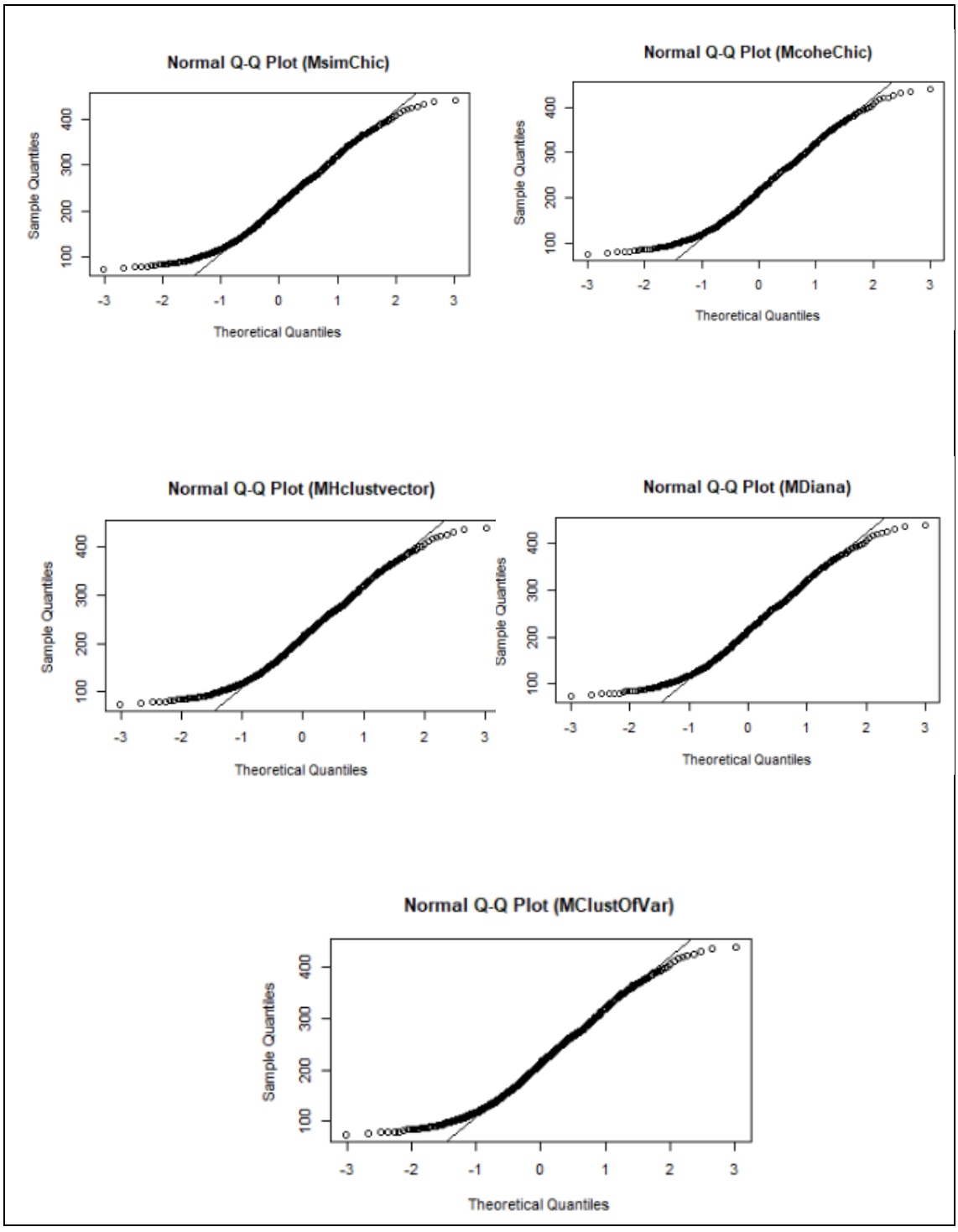


Gráfico 5-3. Gráficos de cuartiles de espacio de memoria

Realizado por: Armas Shirley, 2022.

Al analizar la cercanía de la recta a la curva presentada, sobre el espacio de memoria usado por cada uno de los métodos clúster se observa que cierta cantidad de puntos están situados en la línea recta mientras que otros no, esto se evidencia claramente ya que sobresalen de la recta por lo cual

para verificar dicha aseveración se realizó la respectiva prueba de normalidad a un nivel de significancia del 5%.

- *Planteamiento de Hipótesis*

H_o: Espacio de memoria ~ $\mathcal{N}(\mu, \sigma^2)$
H_1: Espacio de memoria $\nsim \mathcal{N}(\mu, \sigma^2)$

- *Nivel de significancia*

$\alpha = 0.05$

- *Estadístico y valor p*

```
Data$Metodo: MClustOfVar

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.066502, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: McoheChic

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.066502, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: MDiana

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.066502, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: MHclustvector

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.066502, p-value < 0.00000000000000022

------------------------------------------------------------
Data$Metodo: MsimChic

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Data$Datos
D = 0.066502, p-value < 0.00000000000000022
```

Figura 6-3. Test de Normalidad - Espacio de memoria

Realizado por: Armas Shirley, 2022.

- *Regla de decisión*

Si el p valor es menor que 0.05 (p-value < 0.05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla

- *Decisión*

Existe suficiente evidencia para rechazar la hipótesis nula, es decir que ninguno de los 5 métodos correspondientes al espacio de memoria sigue una distribución normal a un nivel de confianza del 95%, el valor p obtenido para cada uno de los métodos es igual a 0.0000000000000022 el cuál es menor para el nivel de significancia de 0.05.

3.4.4.2. Transformación a normalidad

En busca de la transformación de los datos de la variable espacio de memoria a una distribución normal se usó la transformación de Box Cox con el objetivo de determinar la potencia óptima a la que se debe elevar la variable de interés y así buscar obtener normalidad en los datos.

- *Planteamiento de Hipótesis*

H_0: Los datos espacio de memoria siguen una distribución normal
H_1: Los datos espacio de memoria no siguen una distribución normal

- *Nivel de significancia*

$\alpha = 0.05$

- *Estadístico y valor p*

	MsimChic	McoheChic	MHclustvector	MDiana	MClustOfVar
Log	1.422924e-04	1.422924e-04	1.422924e-04	1.422924e-04	1.422924e-04
sqrt	1.341992e-19	1.341992e-19	1.341992e-19	1.341992e-19	1.341992e-19
1/x	4.914508e-18	4.914508e-18	4.914508e-18	4.914508e-18	4.914508e-18
Box Cox	1.160870e-02	1.160870e-02	1.160870e-02	1.160870e-02	1.160870e-02

Figura 7-3. Transformaciones para aproximar a normalidad – memoria

Realizado por: Armas Shirley, 2022.

- *Regla de decisión*

Si el p valor es menor que 0.05 (p-value < 0.05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla

- *Decisión*

Existe suficiente evidencia para rechazar la hipótesis nula debido a que los valores p obtenidos para cada una de las variables son muy pequeños siendo menores al nivel de significancia de 0.05, es decir ninguno de los 5 métodos correspondientes al espacio de memoria siguen una distribución normal a pesar de usar distintas transformaciones los valores p que presenta son muy bajos (**Ver Figura 7-3**).

3.4.4.3. Supuesto de Homocedasticidad

Test de Levene

- *Paso 1: Planteamiento de Hipótesis*

$\mathbf{H_o}$: $\sigma^2_{\text{MsimChic}} = \sigma^2_{\text{Mcohechic}} = \sigma^2_{\text{Mhclustvector}} = \sigma^2_{\mathbf{Mdiana}} = \sigma^2_{\mathbf{Mhclustvar}}$

$\mathbf{H_1}$:$\exists i, j = \{\text{Msimchic}, \text{Mcohechic}, \text{Mhclustvector}, \text{Mdiana}, \text{Mhclustvar}\}/\sigma_i^2 \neq \sigma_j^2$

- *Paso 2: Nivel de significancia*

$\alpha = 0.05$

- *Paso 3: Estadístico de Prueba*

```
Levene's Test for Homogeneity of Variance (center = "median")
        Df F value Pr(>F)
group    4  0.0002      1
      1905
```

Figura 8-3. Resultados Test de Levene

Realizado por: Armas Shirley, 2022.

- *Paso 4: Regla de Decisión*

Si el p valor es menor que 0.05 (p-value < 0.05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla

Paso 5: Decisión

El p-value obtenido es igual a 0.0002 el cual es menor al nivel de significancia propuesto (0.05) por lo que se rechazó la hipótesis nula (Ho) y se concluye que las varianzas de los grupos de espacio de memoria no son iguales, los datos para cada uno de los métodos clúster en estudio son heterocedásticos.

Al no cumplir con los supuestos de normalidad y homocedasticidad se determinó que no se pueden utilizar métodos paramétricos para el estudio de los datos, se constata el uso de pruebas no paramétricas para espacio de memoria.

3.5. Comprobación de hipótesis

Una vez analizados los prerrequisitos se concluyó que no se cumple con los supuestos, ni tampoco se puede lograr la misma normalidad mediante las transformaciones realizadas, por lo que se efectuó a realizar una prueba no paramétrica para muestras independientes. Se realiza el planteamiento de hipótesis para tiempo de ejecución y espacio de memoria a través de la prueba de Kruskal Wallis:

3.5.1. Kruskal Wallis H-Test – Tiempo de ejecución

Paso 1: Planteamiento de Hipótesis

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

H_1: $\tilde{\mu}_i \neq \tilde{\mu}_j$ al menos para un par (i, j)

Paso 2: Nivel de significancia

$\alpha = 0.05$

Paso 3: Estadístico de Prueba

```
        Kruskal-Wallis rank sum test

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

Figura 9-3. Test No paramétrico para tiempo de ejecución

Realizado por: Armas Shirley, 2022

Paso 4: Regla de Decisión

p value < 0.05 → *Rechazo Ho*

Paso 5: Decisión

Se obtuvo un valor p 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 (Ver Figura 3-9).

Se realiza las pruebas de rango post hoc en donde se verificó Existe diferencia entre los métodos de tiempo de ejecución a excepción del método Tcohechic y tsimChic, además de los métodos Tdiana –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
```

Figura 10-3. Pruebas de rango post hoc para tiempo de ejecución

Realizado por: Armas Shirley, 2022.

3.5.1.1. Grupos homogéneos de tiempo de ejecución

La **Tabla 5-3** permite identificar los métodos que usaron mayor tiempo de ejecución siendo así para Thclustvector y Tdiana, los métodos que ocupan el primer lugar y por ende serían los más recomendables son TsimChic y Tcohechic (callSimilarityTree y callHierarchyTree) ya que ocupan el menor tiempo en ejecución.

Tabla 4-3: Grupos homogéneos - Tiempo de ejecución

Nivel	TsimChic	Tcohechic	Thclustvector	Tdiana	TClustOfVar
1	X (330.67)	X (351.61)			
2					Z (595.62)
3			Y (9569.41)	Y (7236.08)	

Fuente: Elaboración Propia.

Realizado por: Armas, Shirley, 2022

3.5.2. Kruskal Wallis H-Test – Espacio de almacenamiento

Paso 1: Planteamiento de Hipótesis

H$_o$: $\tilde{\mu}_{\text{MsimChic}} = \tilde{\mu}_{\text{Mcohechic}} = \tilde{\mu}_{\text{Mhclustvector}} = \tilde{\mu}_{\textbf{Mdiana}} = \tilde{\mu}_{\textbf{Mhclustvar}}$

H$_1$: $\tilde{\mu}_i \neq \tilde{\mu}_j$ al menos para un par (i,j)

Paso 2: Nivel de significancia

$\alpha = 0.05$

Paso 3: Estadístico de Prueba

```
        Kruskal-Wallis rank sum test

data:  Datos by Metodo
Kruskal-Wallis chi-squared = 0.055759, df = 4, p-value = 0.9996
```

Figura 11-3. Test No paramétrico para espacio de memoria

Realizado por: Armas Shirley, 2022.

Paso 4: Regla de Decisión

p value < 0.05 → *Rechazo Ho*

Paso 5: Decisión

Se obtuvo un valor p igual a 0.9996 el cual es mayor a un nivel de significancia de 0.05 por lo que se acepta la hipótesis nula y se concluye que al menos una de las medianas de los métodos clúster de espacio de memoria son estadísticamente iguales, es decir que no existe diferencia significativa entre los grupos (Ver Figura 11-3). Esta aseveración se verificó mediante la ejecución de las pruebas de rango post hoc en donde se pudo constatar que los métodos presentan un parecido espacio de memoria (Ver Figura 12-3).

```
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                           obs.dif critical.dif difference
MClustOfVar-McoheChic    7.4541885     112.0179      FALSE
MClustOfVar-MDiana       0.1767016     112.0179      FALSE
MClustOfVar-MHclustvector 3.0143979    112.0179      FALSE
MClustOfVar-MsimChic     5.3560209     112.0179      FALSE
McoheChic-MDiana         7.6308901     112.0179      FALSE
McoheChic-MHclustvector  4.4397906     112.0179      FALSE
McoheChic-MsimChic       2.0981675     112.0179      FALSE
MDiana-MHclustvector     3.1910995     112.0179      FALSE
MDiana-MsimChic          5.5327225     112.0179      FALSE
MHclustvector-MsimChic   2.3416230     112.0179      FALSE
```

Figura 12-3. Pruebas de Rango post hoc para memoria

Realizado por: Armas Shirley, 2022.

3.5.2.1. Grupos homogéneos de espacio de almacenamiento

La **Tabla 6-3** permite identificar los métodos que usaron mayor espacio de almacenamiento al momento de realizar la medición, evidenciándose así que la diferencia entre los grupos no es significativa por lo que cualquier método en cuánto a memoria es apto, sin embargo, con el respectivo análisis se recomienda al método Mdiana ya que es aquel que ocupa el primer lugar en

menos ocupación de espacio de memoria, seguido por MClustOfVar, Mhclustvector, MsimChic y finalmente el que ocupa el último lugar como menos recomendable es Mcohechic.

Tabla 5-3*:* Grupos homogéneos – Espacio de almacenamiento

Nivel	MsimChic	Mcohechic	Mhclustvector	Mdiana	MClustOfVar
1				D (213.08)	
2					E (213.1)
3			C (213.2)		
4	A (213.8)				
5		B (214.27)			

Fuente: Elaboración Propia.

Realizado por: Armas, Shirley, 2022

CONCLUSIONES

- Existen técnicas similares de análisis clúster entre LA y ASI las cuales son: callSimilarityTree y callHierarchyTree para el análisis estadístico implicativo (ASI) y diana, hclutsvector y hclustvar (ClustOfVar) para Learning Analytics (LA).
- No existe diferencia significativa entre los grupos de espacio de memoria de las técnicas clustering similares de ASI y LA ya que todos los métodos callSimilarityTree, callHierarchyTree, diana, hclutsvector y hclustvar registran un consumo parecido entre ellas, se definiría como el método óptimo a hclustvector de LA y como óptimo a callHierarchyTree de ASI.
- Existen diferencias significativas entre los grupos de tiempo de ejecución de las técnicas clúster similares en AEI y LA en donde se identificó como al método que ocupa menor tiempo a callSimilarityTree y callHierarchyTree de ASI siendo considerados los óptimos mientras que los óptimos serían Thclustvector y Tdiana de LA debido a que registran mayor consumo de tiempo.
- Al analizar los resultados obtenidos en cuanto a tiempo de ejecución y espacio de memoria para cada uno de los métodos clúster se evidenció que los métodos más recomendables serían callSimilarityTree y callHierarchyTree de ASI ya que con respecto al espacio de memoria no se registró mayor variación entre las técnicas como pasó con el tiempo.

RECOMENDACIONES

- Se recomienda que se identifique y trate de manera adecuada los datos atípicos sobre el tiempo de ejecución presentados en el método TDiana y Thclustvector para obtener resultados óptimos.
- Fortalecer la investigación con respecto a los factores que puede influir internamente en la variación del tiempo de ejecución al momento de trabajar con variables de tipo modal en Learning Analytics y el análisis estadístico Implicativo.
- Profundizar de manera más exhaustiva el uso de técnicas de agrupación clúster en Learning Analytics para conocer a más detalle que limitaciones han venido presentado dichos métodos a través del tiempo sobre todo al momento de usarlos en el software.
- Debido a la gran cantidad de información que se genera en la actualidad en todas las áreas, especialmente en la educación lo cual hace improbable el análisis con cálculos manuales y en ciertos casos hasta un bajo rendimiento de un software tradicional se recomienda para futuros estudios seguir aprovechando el potencial de softwares especializados como R, RStudio y Rchic.

BIBLIOGRAFÍA

ALLAIRE, J., 2012. RStudio: integrated development environment for R. *Boston, MA*. Vol. 770, número 394, pp. 165-171.

AMAT, Joaquin, 2016. Análisis de la homogeneidad de varianza (homocedasticidad) con R. [en línea]. 2016. Recuperado a partir de : https://www.cienciadedatos.net/documentos/9_homogeneidad_de_varianza_homocedasticidad.html [accedido 13 febrero 2022].

ARDILA, Rubén Ardila, 1966. Técnicas estadísticas no paramétricas. *Revista Colombiana de Psicología*. Vol. 11, número 1-2, pp. 89-102.

BERNARD, Jean-Marc y CHARRON, Camilo, 1996. L'analyse implicative bayésienne, une méthode pour l'étude des dépendances orientées. I: données binaires. *Mathématiques et Sciences humaines*. Vol. 134, pp. 5-38.

CAMPBELL, Donald T y STANLEY, JG, 1996. Experimental and Quasi-experimental. *Design for Research. Ran McNally, Chicago, III*.

CAMPO, Nelsa María Sagaró del y MATAMOROS, Larisa Zamora, 2019. ¿Por qué emplear el análisis estadístico implicativo en los estudios de causalidad en salud? *Revista Cubana de Informática Médica*. Vol. 11, número 1, pp. 88-103.

CANTO DE GANTE, Ángela Guadalupe et al., 2020. Escala de Likert: Una alternativa para elaborar e interpretar un instrumento de percepción social. *Revista de la alta tecnología y sociedad*. Vol. 12, número 1.

CERDA L, Jaime y VILLARROEL DEL P, Luis, 2007. Interpretación del test de Chi-cuadrado (X^2) en investigación pediátrica. *Revista chilena de pediatría*. Vol. 78, número 4, pp. 414-417. DOI 10.4067/S0370-41062007000400010.

CHAVENT, Marie et al., 2017. *ClustOfVar: Clustering of Variables* [logiciel]. Version 1.1. 12 agosto 2017. [accedido 14 febrero 2022]. Recuperado a partir de : https://CRAN.R-project.org/package=ClustOfVar [accedido 14 febrero 2022].

CLAUDE, Jean, 2021. A.S.I. 5 - 5° Coloquio Internacional de Análisis Estadístico Implicativo. [en línea]. 2021. Recuperado a partir de : https://sites.univ-lyon2.fr/asi/5/?page=0&lang=es [accedido 13 febrero 2022].

COUTURIER, Raphaël y AG ALMOULOUD, S., 2007. CHIC: utilisation et fonctionnalités. *GRAS, R.; ORÚS, P.; PINAUD, B.* pp. 41-49.

COUTURIER, Raphaël y ALMOULOUD, Saddo Ag, 2009. *Historique et fonctionnalités de CHIC.*

COUTURIER, Raphaël y PAZMIÑO, Rubén, 2016. Use of Statistical Implicative Analysis in Complement of Item Analysis. *International Journal of Information and Education Technology*. Vol. 6, número 1, p. 39.

ELIAS, Tanya, 2011. Learning analytics. *Learning*. pp. 1-22.

FILLIBEN, James J., 1975. The Probability Plot Correlation Coefficient Test for Normality. *Technometrics*. Vol. 17, número 1, pp. 111-117. DOI 10.1080/00401706.1975.10489279.

FLORES, Pablo, OCAÑA, Jordi y SÁNCHEZ, Tania, 2018. Verificación de supuestos en las pruebas de comparación de medias. Una revisión. *Ciencia Digital*. Vol. 2, número 4.1., pp. 5-22.

GIL MARTÍNEZ, Cristina, 2018. RPubs - Métodos de clustering. [en línea]. 2018. Recuperado a partir de : https://rpubs.com/Cristina_Gil/Clustering [accedido 13 febrero 2022].

GIL MARTÍNEZ, CRISTINA, 2018. RPubs - Métodos de clustering. [en línea]. 2018. Recuperado a partir de : https://rpubs.com/Cristina_Gil/Clustering [accedido 7 octubre 2020].

GRAS, R, BAQUERO, J y GUILLET, F, 2009. A.S.I. 7 - 7° Coloquio Internacional de Análisis Estadístico Implicativo. [en línea]. 2009. Recuperado a partir de : https://sites.univ-lyon2.fr/asi/7/?page=0&lang=es [accedido 13 febrero 2022].

GRAS, R. y RATSIMBA-RAJOHN, H., 1996. Analyse non symétrique de données par l'implication statistique. *RAIRO - Operations Research - Recherche Opérationnelle*. Vol. 30, número 3, pp. 217-232.

GRAS, Régis et al., 2008. *Statistical Implicative Analysis: Theory and Applications*. Springer. ISBN 978-3-540-78983-3. Google-Books-ID: _AltCQAAQBAJ

GRAS, Régis et al., 2009. El Análisis Estadístico Implicativo (ASI) en respuesta a problemas que le dieron origen. En : .

GRAS, Régis y KUNTZ, Pascale, 2009. El Analisis Estadistico Implicativo (ASI) en respuesta a problemas que le dieron origen. En : . ISBN 978-84-692-3925-4.

GRAS, Régis, KUNTZ, Pascale y BRIAND, Henri, 2001. Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et sciences humaines. Mathematics and social sciences*. Número 154. DOI 10.4000/msh.2849.

GROLEMUND, GARRET y HADLEY, Wickham, 2017. *7 Análisis exploratorio de datos (EDA) | _main* [en línea]. Recuperado a partir de : https://es.r4ds.hadley.nz/an%C3%A1lisis-exploratorio-de-datos-eda.html [accedido 13 febrero 2022].

KLAŠNJA-MILIĆEVIĆ, Aleksandra, IVANOVIĆ, Mirjana y BUDIMAC, Zoran, 2017. Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education*. Vol. 25, número 6, pp. 1066-1078.

LIAS, Tanya E. y ELIAS, Tanya, 2011. *Learning Analytics: The Definitions, the Processes, and the Potential*. .

MAECHLER, Martin, 2021. *cluster: «Búsqueda de grupos en los datos»: análisis de conglomerados ampliado Rousseeuw et al.* [logiciel] [en línea]. 2021. [accedido 15 febrero 2022]. Recuperado a partir de : https://CRAN.R-project.org/package=cluster [accedido 15 febrero 2022].

MAJI, Rubén Antonio Pazmiño et al., 2019. LA INVESTIGACIÓN DE PREGRADO EN LA ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO: MAPEO SISTEMÁTICO Y ANALÍTICAS. *Revista Científica ECOCIENCIA*. Vol. 6, número 1, p. 25.

MENDOZA, María E., CAPUTO, Liliana N. y PORCEL, Eduardo A., 2019. Análisis estadístico implicativo de los conocimientos previos sobre propiedades de operaciones con

números reales de ingresantes a carreras de ingeniería de FACENA. *Extensionismo, Innovación y Transferencia Tecnológica*. Vol. 5, número 0, pp. 53-60. DOI 10.30972/eitt.503736.

MERCERON, Agathe, BLIKSTEIN, Paulo y SIEMENS, George, 2015. Learning analytics: From big data to meaningful data. *Journal of Learning Analytics*. Vol. 2, número 3, pp. 4-8.

MERSMANN, Olaf y KREY, Sebastian, 2011. microbenchmark: A package to accurately benchmark R expressions. En : *The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*. . 2011.

MICHAEL, P. et al., 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.

MÜLLNER, Daniel y INC, Google, 2021. *fastcluster: Fast Hierarchical Clustering Routines for R and «Python»* [logiciel]. Version 1.2.3. 24 mayo 2021. [accedido 15 febrero 2022]. Recuperado a partir de : https://CRAN.R-project.org/package=fastcluster [accedido 15 febrero 2022].

NARANJO, Mauricio et al., 2018. Métodos de agrupamiento LA & SIA: Comparación computacional. *Congreso de Ciencia y Tecnología ESPE*. Vol. 13, número 1. DOI 10.24133/cctespe.v13i1.817.

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.(2018). .

NARANJO SERRANO, Mauricio Medardo, 2018. Estudio comparativo del análisis estadístico implicativo y el learning analytics en relación al uso de las técnicas de exploración de datos educativos. [en línea]. Recuperado a partir de : https://repositorio.pucesa.edu.ec/handle/123456789/2387 [accedido 8 febrero 2022]. Accepted: 2018-04-27T17:03:34Z

NAVARRO, Álvaro Alexander Martínez y GER, Pablo Moreno, 2018. Comparison of Clustering Algorithms for Learning Analytics with Educational Datasets. *IJIMAI*. Vol. 5, número 2, pp. 9-16.

PATNAIK, Ashish Kumar, BHUYAN, Prasanta Kumar y KRISHNA RAO, K. V., 2016. Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets. *Alexandria Engineering Journal*. Vol. 55, número 1, pp. 407-418. DOI 10.1016/j.aej.2015.11.003.

PAZMIÑO MAJI, R. A., 2021a. *Aporte del Análisis Estadístico Implicativo a Learning Analytics* [en línea]. Thesis . Programa de Doctorado Formación en la Sociedad del Conocimiento. Recuperado a partir de : https://repositorio.grial.eu/handle/grial/2487 [accedido 9 febrero 2022]. Accepted: 2021-12-17T16:40:09Z

PAZMIÑO MAJI, R. A., 2021b. *Aporte del Análisis Estadístico Implicativo a Learning Analytics*. . Programa de Doctorado Formación en la Sociedad del Conocimiento.

PAZMIÑO MAJI, Rubén Antonio, GARCÍA PEÑALVO, Francisco José y CONDE GONZÁLEZ, Miguel Ángel, 2017. Is it possible to apply Statistical Implicative Analysis in hierarchical cluster Analysis? Firsts issues and answers. .

PAZMIÑO MAJI, Rubén, GARCÍA PEÑALVO, Francisco José y CONDE GONZÁLEZ, Miguel Ángel, 2017. Is it possible to apply Statistical Implicative Analysis in hierarchical cluster Analysis? Firsts issues and answers. .

PAZMIÑO, Rubén, MULLO, Jose y CONDE, Miguel, 2019a. EL ANÁLISIS ESTADÍSTICO IMPLICATIVO COMO ESTRATEGIA PARA LA PROMOCIÓN DEL APRENDIZAJE EN LA EDUCACIÓN MEDIA: SIMULACIONES PARA SU APRENDIZAJE. *Identidad Bolivariana*. pp. 24-39. DOI 10.37611/IB0ol024.

PAZMIÑO, Rubén, MULLO, Jose y CONDE, Miguel, 2019b. EL ANÁLISIS ESTADÍSTICO IMPLICATIVO COMO ESTRATEGIA PARA LA PROMOCIÓN DEL APRENDIZAJE EN LA EDUCACIÓN MEDIA: SIMULACIONES PARA SU APRENDIZAJE. *Identidad Bolivariana*. pp. 24-39. DOI 10.37611/IB0ol024.

PAZMIÑO-MAJI, R. A., GARCÍA-PEÑALVO, F. J. y CONDE-GONZÁLEZ, M. Á, 2017. Statistical Implicative Analysis Approximation to KDD and Data Mining: A Systematic and Mapping Review in Knowledge Discovery Database Framework. En : [en línea]. IARIA XPS Press. ISBN 978-1-61208-558-6. Recuperado a partir de : https://repositorio.grial.eu/handle/grial/851 [accedido 13 febrero 2022]. Accepted: 2017-05-23T23:03:37Z

PAZMIÑO-MAJI, Rubén et al., 2019. Learning Analytics in Ecuador: An Initial Analysis based in a Mapping Review. En : *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, pp. 304-311. New York, NY, USA : Association for Computing Machinery. 16 octubre 2019. TEEM'19. ISBN 978-1-4503-7191-9. DOI 10.1145/3362789.3362913.

PAZMIÑO-MAJI, Rubén A., GARCÍA-PEÑALVO, Francisco J. y CONDE-GONZÁLEZ, Miguel A., 2017. Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics. En : *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, pp. 1-7. Cádiz Spain : ACM. 18 octubre 2017. ISBN 978-1-4503-5386-1. DOI 10.1145/3144826.3145399.

PAZMIÑOOM, Rubén et al., 2018. Software Estadístico CHIC: descubriendo sus potencialidades mediante el análisis de percepción sexual universitaria. *Ciencia Digital*. Vol. 2, número 4.1., pp. 122-139. DOI 10.33262/cienciadigital.v2i4.1..194.

PÉREZ, María Gabriela, PAZMIÑO, R. y ANDALUZ, Victor, 2014. Cuasi-implicación estadística y determinación automática de clases de equivalencia en imágenes de resonancia magnética de cerebro. *Revista Politécnica*. Vol. 34, número 1, pp. 123-123.

PERROTTA, Carlo y WILLIAMSON, Ben, 2018. The social life of Learning Analytics: cluster analysis and the 'performance' of algorithmic education. *Learning, Media and Technology*. Vol. 43, número 1, pp. 3-16. DOI 10.1080/17439884.2016.1182927.

PICCIANO, Anthony G., 2012. The evolution of big data and learning analytics in American higher education. *Journal of asynchronous learning networks*. Vol. 16, número 3, pp. 9-20.

PICCIANO, Anthony G., 2014. Big data and learning analytics in blended learning environments: Benefits and concerns. *IJIMAI*. Vol. 2, número 7, pp. 35-43.

QUISPE, A et al., 2019. Estadística no paramétrica aplicada a la investigación científica con software SPSS, MINITAB Y EXCEL. *Enfoque práctico. Colombia. Editorial EIDEC*.

RODRIGUEZ, Pablo, 2019. Learning Analytics: el poder del big data en la educación. *Telos Fundación Teléfonica* [en línea]. 2019. Recuperado a partir de : https://telos.fundaciontelefonica.com/la-cofa/learning-analytics-el-poder-del-big-data-en-la-educacion/ [accedido 13 febrero 2022].

ROJAS-CASTRO, Pablo y ROJAS-CASTRO, Pablo, 2017a. Learning Analytics: una revisión de la literatura. *Educación y Educadores*. Vol. 20, número 1, pp. 106-128. DOI 10.5294/edu.2017.20.1.6.

ROJAS-CASTRO, Pablo y ROJAS-CASTRO, Pablo, 2017b. Learning Analytics: una revisión de la literatura. *Educación y Educadores*. Vol. 20, número 1, pp. 106-128. DOI 10.5294/edu.2017.20.1.6.

ROT, Myriam Vergara y BABATIVA, Giovany, 2010. El supuesto de normalidad:¿ mito o realidad? *Equidad y desarrollo*. Número 13, pp. 127-131.

RUIPÉREZ-VALIENTE, José A., 2020. El Proceso de Implementación de Analíticas de Aprendizaje. *RIED. Revista Iberoamericana de Educación a Distancia*. Vol. 23, número 2, pp. 85-101. DOI 10.5944/ried.23.2.26283.

SAGARÓ DEL CAMPO, Nelsa María et al., 2019. ¿Por qué emplear el análisis estadístico implicativo en los estudios de causalidad en salud? *Revista Cubana de Informática Médica*. Vol. 11, número 1, pp. 88-103.

SCHORK, Joachim, 2021. Clean Up Memory in R (Example) | Garbage Collection with gc() Function. *Statistics Globe* [en línea]. 2021. Recuperado a partir de : https://statisticsglobe.com/clean-up-memory-gc-function-r/ [accedido 14 febrero 2022].

SCIMONE, Aldo y SPAGNOLO, Filip**po**, 2005. The importance of supplementary variables in a case of an educational research. . p. 9.

SIN, Katrina y MUTHU, Loganathan, 2015. APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW. *ICTACT journal on soft computing*. Vol. 5, número 4.

TAPIA, Carlos Ernesto Flores y CEVALLOS, Karla Lissette Flores, 2021. PRUEBAS PARA COMPROBAR LA NORMALIDAD DE DATOS EN PROCESOS PRODUCTIVOS: ANDERSON-DARLING, RYAN-JOINER, SHAPIRO-WILK Y KOLMOGÓROV-SMIRNOV. *Societas. Revista de Ciencias Sociales y Humanísticas*. Vol. 23, número 2, pp. 83-106.

VÁZQUEZ, Karell Galano et al., 2019. Análisis estadístico implicativo en la identificación de factores pronósticos de mortalidad del cáncer renal. *Revista Información Científica*. Vol. 98, número 2, pp. 157-170.

VELANDIA-VEGA, John Alexánder y FLORES-CABAÑAS, Juan Pablo, 2020. Desafíos que debe resolver learning analytics para mejorar los procesos de aprendizaje en los estudiantes. .

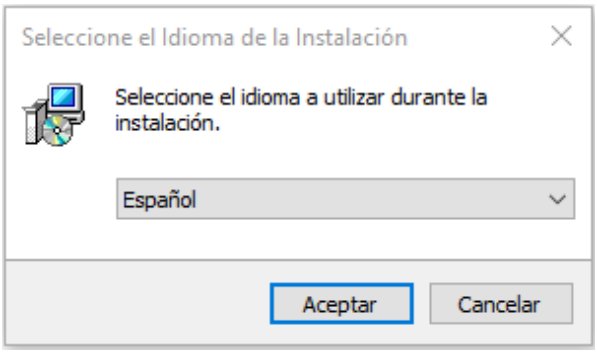
VILÀ BAÑOS, Ruth et al., 2014. Cómo aplicar un cluster jerárquico en SPSS. *REIRE. Revista d'Innovació i Recerca en Educació, 2014, vol. 7, num. 1, p. 113-127*.

ZAMORA, L., GREGORI, P. y ORÚS, P., 2009. Conceptos fundamentales del Análisis Estadístico Implicativo (ASI) y su soporte computacional CHIC. *Contribuciones al ASI*. Vol. 4, pp. 65-101.

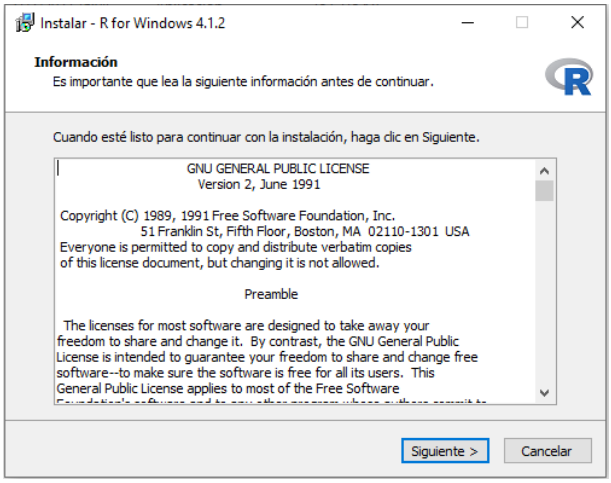
ANEXOS

ANEXO A: INSTALACIÓN SOFTWARE R

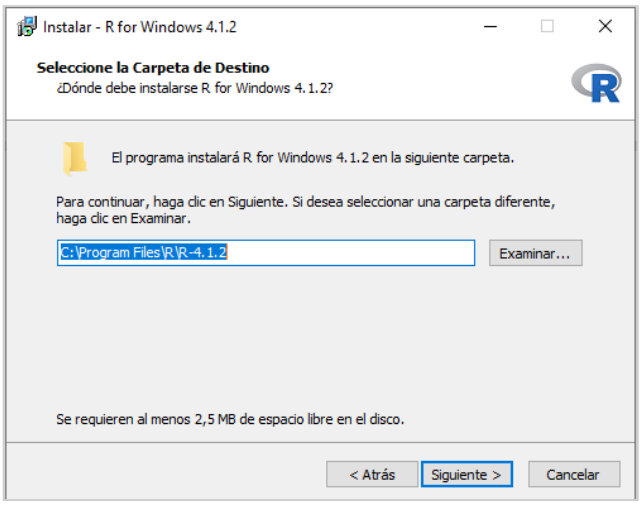
1. Descargar el software R de la siguiente dirección: https://cran.r-project.org/bin/windows/base/
2. Seleccionar el idioma



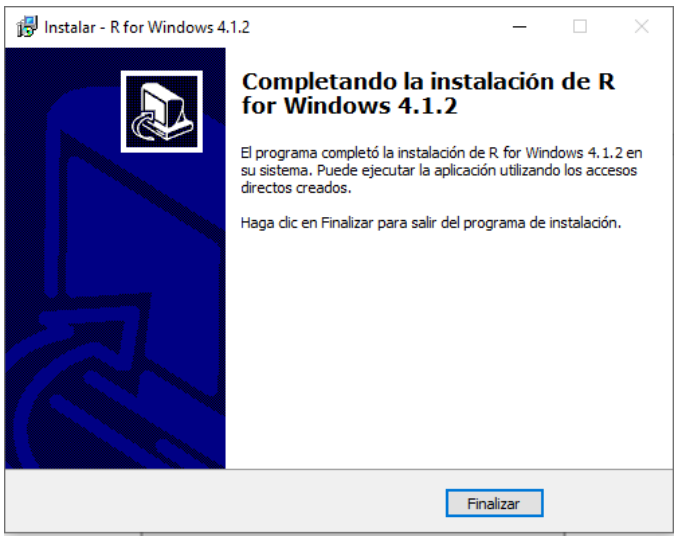
3. Pulsar aceptar y continuar seleccionando siguiente para continuar con la instalación, previamente aparecerá la licencia de la versión de R a instalarse



4. A continuación, seleccionar el directorio de instalación y dar clic en siguiente (se aceptan todas las opciones marcadas por defecto en todas las ventanas que aparecen)

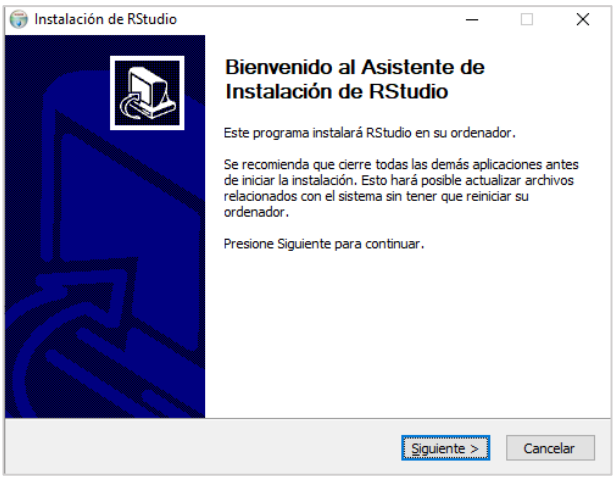


5. Esperar a que se complete el proceso de instalación del software y finalmente dar clic en finalizar

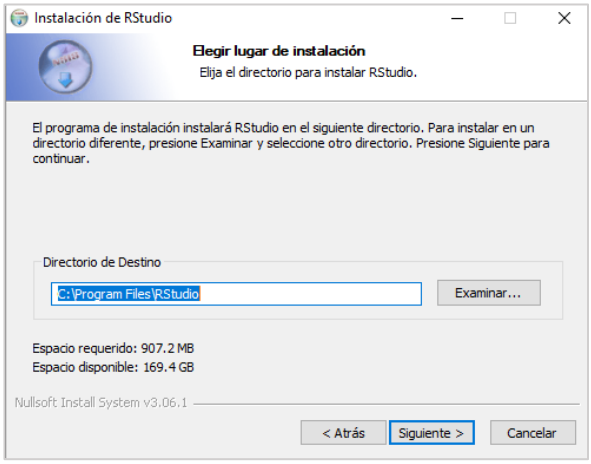


ANEXO B: INSTALACIÓN SOFTWARE R-STUDIO

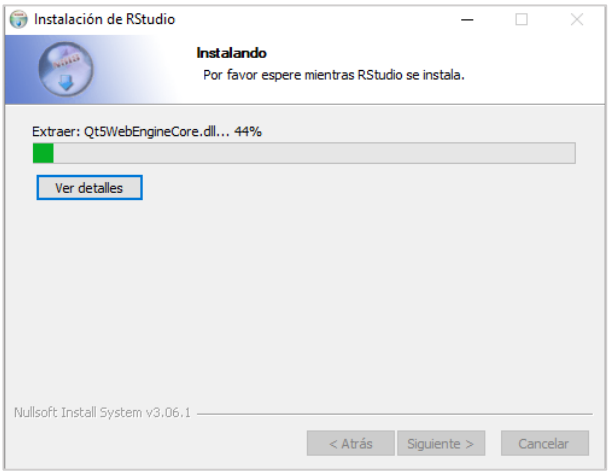
1. Descargar R-Studio desde el siguiente enlace: https://www.rstudio.com/products/rstudio/download/
2. Ejecutar el archivo como administrador y permitir que realice cambios en nuestro equipo
3. Leer las recomendaciones y seguidamente dar clic en siguiente para empezar con la instalación



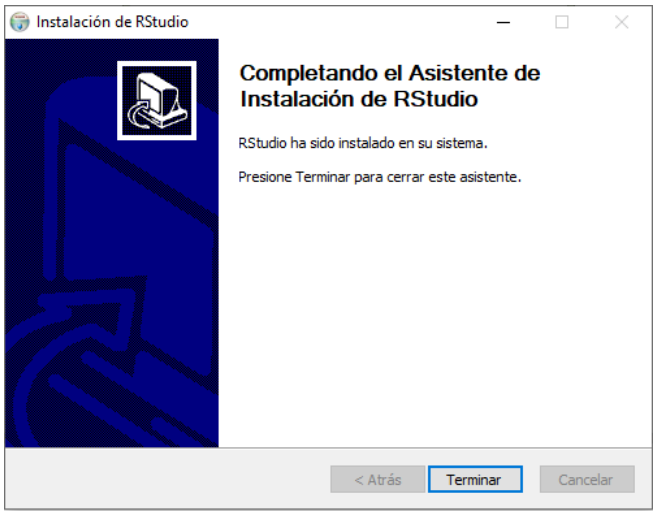
4. Seleccionar el directorio de instalación y dar clic en siguiente para continuar con la instalación



5. Esperar a que termine el proceso de instalación del software



6. Finalmente dar clic en terminar



ANEXO C: INSTALACIÓN PAQUETES R

1. Para descargar los paquetes para la instalación de Rchic lo podemos realizar en el siguiente enlace: https://members.femto-st.fr/raphael-couturier/en/rchic
2. Es necesario tener presente que actualmente Rchic está actualizado para trabajar con la última versión de R y RStudio
3. En la página oficial (que se indica en el enlace del paso número 1) se puede visualizar que ya viene dado el código directo a ejecutarse en RStudio para su instalación

```
# ONLY THE FIRST TIME YOU INSTALL
install.packages(c("stringr", "tcltk2", "Rcpp"))
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager") BiocManager::install("Rgraphviz")
install.packages(" https://members.femto-st.fr/raphael-couturier/sites/femto-st.fr.raphael-couturier/files/content/rchic/rchic_0.27.zip ", repos = NULL, type = "win.binary")
```

ANEXO C: CÓDIGO PARA CONSTRUCCIÓN DE BASE DE DATOS FINAL

```
library("ClustOfVar")
library(factoextra)
library(FactoMineR)
library("openxlsx")
library("WriteXLS")
library("plyr")
library("Rcpp")
library("microbenchmark")
library("rchic")
library("fastcluster")
library("cluster")
rchic()
#============================FUNCIONES=====================
=========
```

```
#Cohesion
hierarchy_Tree <- function(x){
  cohesion <- callHierarchyTree(x, contribution.supp = FALSE,typicality.supp
                          = FALSE,computing.mode = 3, verbose = FALSE)
  return(cohesion)
}
#Similaridad
Similarity_Tree<- function(x){
  Similarity <-  callSimilarityTree(x,contribution.supp=FALSE,typicality.supp=
                              FALSE,verbose=FALSE)
  return(Similarity )
}
#Hclust vector
hclust_vector <- function(x){
  hv <- read.csv(x,sep = ";")
  dists <- dist(hv,method = "euclidean")
  hc <- hclust.vector(X = dists, members=NULL, metric='euclidean', p=NULL)
  dend <- as.dendrogram(hc)
  fviz_dend(dend)
  #plot(hc)
  return(dend)
}
#Dendro diana
Diana <- function(x){
  dian <- read.csv(x,sep = ";")
  d <- diana(x = dian,metric = "euclidean", stand = TRUE)
  dend <- as.dendrogram(d)
  fviz_dend(dend)
  return(d)
}
# Hclust var
HclustVar<- function(x){
  hc <- read.csv(x,sep = ";")
  hc <- hc[,-1]
  v <- hclustvar(X.quanti = hc, init = NULL)
  plot(v)
  return(v)
}
getwd()
Resumen <- function(){
  n = 1
  a = 2
  b = 2
  c = 2
  d = 2
  e = 2
  f = 2
  g = 2
  h = 2
```

```
  k = 2
  l = 2
  m = 2
  o = 2
  p = 2
  q = 2
  #==========Creacion del documento de Excel y nombres de columnas==========
  info <- createWorkbook(creator = "xlsx")
  addWorksheet(info, "Sheet 1")
  writeData(info, sheet = 1, x = "Nombre_Archivo", startCol = 1, startRow = 1)
  writeData(info, sheet = 1, x = "Numero_Filas",startCol = 2,startRow = 1)
  writeData(info, sheet = 1, x = "Numero_columnas",startCol = 3,startRow = 1)
  writeData(info, sheet = 1, x = "Repeticion",startCol = 4,startRow = 1)
  writeData(info, sheet = 1, x = "TsimChic",startCol = 5,startRow = 1)
  writeData(info, sheet = 1, x = "TcoheChic",startCol = 6,startRow = 1)
  writeData(info, sheet = 1, x = "THclustvector",startCol = 7,startRow = 1)
  writeData(info, sheet = 1, x = "TDiana",startCol = 8,startRow = 1)
  writeData(info, sheet = 1, x = "TClustOfVar",startCol = 9,startRow = 1)
  writeData(info, sheet = 1, x = "MsimChic",startCol = 10,startRow = 1)
  writeData(info, sheet = 1, x = "McoheChic",startCol = 11,startRow = 1)
  writeData(info, sheet = 1, x = "MHclustvector",startCol = 12,startRow = 1)
  writeData(info, sheet = 1, x = "MDiana",startCol = 13,startRow = 1)
  writeData(info, sheet = 1, x = "MClustOfVar",startCol = 14,startRow = 1)

#==========FUNCIONGENERAL==========
  total <- 52
  for(j in 1:total){
    s = 1
    bases <- list.files()
    dataS <- read.csv(bases[n],sep = ";", dec = ".",stringsAsFactors = FALSE)
    x <- bases[n]
    #========== REPETICIONES
    rep<- 3
    for(i in 1:rep) {
      #========== Nombre base de datos
      writeData(info, sheet = 1,x = bases[n],startCol = 1,startRow = a)
      #========== Numero de Filas ==========
      writeData(info, sheet = 1, x = nrow(dataS),startCol = 2,startRow = b)
      #========== Numero de columnas =========
      writeData(info,sheet = 1, x = ncol(dataS[,-1]),startCol = 3,startRow = c)
      #========== Repeticiones ==========
      writeData(info,sheet = 1, x = s,startCol = 4,startRow = d)
      s = s+1
      # ========== Tiempo ==========
      tiempo <- microbenchmark(hierarchy_Tree(x),Similarity_Tree(x),
                               hclust_vector(x),Diana(x),HclustVar(x),times = 1)
      t <- summary(tiempo)$median
```

```
      writeData(info, sheet = 1,x = t[1],startCol = 5,startRow = e)
      writeData(info, sheet = 1,x = t[2],startCol = 6,startRow = f)
      writeData(info, sheet = 1,x = t[3],startCol = 7,startRow = g)
      writeData(info, sheet = 1,x = t[4],startCol = 8,startRow = h)
      writeData(info, sheet = 1,x = t[5],startCol = 9,startRow = k)

      # ====================== Memoria=========================
      MHT <- gc(verbose = hierarchy_Tree(x),reset = TRUE)
      MST <- gc(verbose = Similarity_Tree(x) ,reset = TRUE)
      MHV <- gc(hclust_vector(x),reset = TRUE)
      MD <-gc(Diana(x),reset = TRUE)
      MV <- gc(HclustVar(x),reset = TRUE)
      writeData(info,sheet = 1,x = MHT[1,2],startCol = 10,startRow = l)
      writeData(info, sheet = 1,x = MST[1,2],startCol = 11,startRow = m)
      writeData(info, sheet = 1,x = MHV[1,2],startCol = 12,startRow = o)
      writeData(info, sheet = 1,x = MD[1,2],startCol = 13,startRow = p)
      writeData(info, sheet = 1,x = MV[1,2],startCol = 14,startRow = q)
      #Contadores
      a = a+1
      b = b+1
      c = c+1
      d = d+1
      e = e+1
      f = f+1
      g = g+1
      h = h+1
      k = k+1
      l = l+1
      m = m+1
      o = o+1
      p = p+1
      q = q+1
    }
    file.remove("hierarchy.ps")
    n = n+1
    print(n)
  }
  saveWorkbook(info,"DatosRepeticion.xlsx",overwrite = TRUE)
}
Resumen()
```

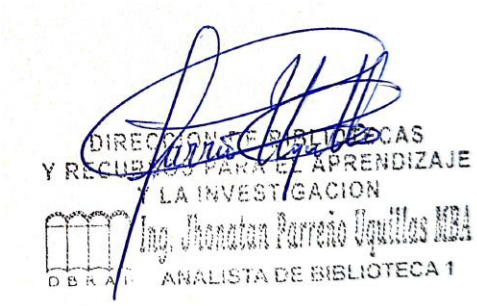
espoch | Dirección de Bibliotecas y Recursos del Aprendizaje

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 07/12/2023

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: SHIRLEY ESTEFANIA ARMAS ANALUISA
INFORMACIÓN INSTITUCIONAL
Facultad: CIENCIAS
Carrera: INGENIERÍA EN ESTADÍSTICA INFORMÁTICA
Título a optar: INGENIERA EN ESTADÍSTICA INFORMÁTICA
f. Analista de Biblioteca responsable: Ing. CPA. Jhonatan Rodrigo Parreño Uquillas**. MBA.**



1831-DBRA-UTP-2023