UNIDAD ACADÉMICA:

OFICINA DE POSTGRADOS

TEMA:

ESTUDIO COMPARATIVO DEL ANÁLISIS ESTADÍSTICO IMPLICATIVO Y EL *LEARNING ANALYTICS* EN RELACIÓN AL USO DE LAS TÉCNICAS DE EXPLORACIÓN DE DATOS EDUCATIVOS

Proyecto de Investigación y Desarrollo previo a la obtención del título de

Magister en Ciencias de la Educación.

Línea de Investigación, Innovación y Desarrollo principal:

Sistemas de información y/o Nuevas Tecnologías de la Información y Comunicación y sus aplicaciones

Caracterización técnica del trabajo: Investigación

Autor:

Mauricio Medardo Naranjo Serrano

Director:

Rubén Antonio Pazmiño Maji, Mg.

Ambato – Ecuador

Abril 2018

Estudio comparativo del Análisis Estadístico Implicativo y el *Learning Analytics* en relación al uso de las técnicas de exploración de datos educativos

Informe de Trabajo de Titulación
presentado ante la
Pontificia Universidad Católica del Ecuador
Sede Ambato

Por

Mauricio Medardo Naranjo Serrano

En cumplimiento parcial de
los requisitos para el Grado de
Magister en Ciencias de la
Educación

Oficina de Postgrados
Abril 2018

Estudio comparativo del Análisis Estadístico Implicativo y el *Learning Analytics* en relación al uso de las técnicas de exploración de datos educativos

Aprobado por:

María Fernanda San Lucas, Mg
Presidente del Comité Calificador
Coordinadora de Postgrados

Fernando Alfredo Flor Tapia, Mg
Miembro Calificador

Rubén Antonio Pazmiño Maji PhD
Miembro Calificador
Director de Proyecto

Dr. Hugo Rogello Altamirano Villarroel
Secretario General

Ricardo Patricio Medina Chicaiza, Mg
Miembro Calificador

Fecha de aprobación:
Abril 2018

Ficha Técnica

Programa: Magister en Ciencias de la Educación

Tema: Estudio comparativo del Análisis Estadístico Implicativo y el Learning Analytics en relación al uso de las técnicas de exploración de datos educativos

Tipo de trabajo: Proyecto de Investigación y Desarrollo (Tesis)

Clasificación técnica del trabajo: Investigación

Autor: Mauricio Medardo Naranjo Serrano

Director: Rubén Antonio Pazmiño Maji, Mg.

Líneas de Investigación, Innovación y Desarrollo

Principal: Sistemas de Información y/o Nuevas Tecnologías de la Información y Comunicación y sus aplicaciones

Secundaria: Pedagogía, Andragogía, Didáctica y/o currículo

Resumen Ejecutivo

El análisis de datos es fundamental dentro de la educación y al no utilizar las técnicas de análisis óptimas al tipo y cantidad de datos, se crea en muchos procesos (repitencia, deserción, bajo rendimiento, alto rendimiento, entre otros), el problema de la obstrucción o lentitud del cálculo, lo que hace que las técnicas de análisis de datos sean inaplicables, por lo tanto, los problemas educativos sean irresolubles. Por tal razón, en esta investigación se plantea realizar un análisis comparativo de los tiempos de procesamiento y espacio de memoria entre las técnicas de agrupación y las técnicas de búsqueda de reglas de asociación similares al Análisis Estadístico Implicativo (AEI) y Learning Analytics (LA). En este trabajo se realizará: (1) identificar las técnicas similares entre el AEI y LA, mediante la adaptación del método de estudio de similitud entre modelos y estándares (MSSS), (2) identificar el sistema operativo con mejor manejo de recursos y (3) identificar la técnica más óptima en el análisis de datos educativos; estos dos últimos con la elaboración de un diseño cuasi-experimental (Campbell & Stanley, 1966). La hipótesis por demostrar es que existe diferencia significativa (función espacio y/o función tiempo) entre los algoritmos similares a LA y AEI, se demostrará mediante ANOVA no paramétrico (Kruskal-Wallis, Kruskal-Canover), de esta forma obtener la técnica óptima que ayude al docente resolver más rápidamente los problemas educativos que utilizan datos masivos.

Declaración y autorización

Yo: **MAURICIO MEDARDO NARANJO SERRANO**, con **CC. 060300594-3**, autor del trabajo de graduación intitulado: "Estudio comparativo del análisis estadístico implicativo y el learning analytics en relación al uso de las técnicas de exploración de datos educativos", previa a la obtención del título profesional de Magister En Ciencias de la Educación, en la escuela de Postgrado.

1.- Declaro tener pleno conocimiento de la obligación que tiene la Pontificia Universidad Católica del Ecuador, de conformidad con el artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de graduación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.
2.- Autorizo a la Pontificia Universidad Católica del Ecuador a difundir a través de sitio web de la Biblioteca de la PUCE Ambato, el referido trabajo de graduación, respetando las políticas de propiedad intelectual de Universidad

Ambato, abril 2018

MAURICIO MEDARDO NARANJO SERRANO
CC. 060300594-3

Dedicatoria

Con todo mi respeto, cariño y amor para las personas que me dieron la vida y porque gracias a ellos hoy pueda cumplir un objetivo más, por motivarme, apoyarme y darme su mano en todo momento lo necesite, a ustedes por siempre mi corazón y mi agradecimiento Padres. A mis hijos Samantha y Sebastian que fueron la fortaleza para poder culminar una etapa más de estudios.

Mauricio Naranjo

Reconocimientos

La realización de este proyecto de investigación fue posible, en primer lugar, a la cooperación y apoyo brindado por Rubén Antonio Pazmiño Maji Mg., quien en calidad de Director del Proyecto de Investigación y Desarrollo, ayudo en la corrección de las versiones preliminares realizadas, en la parte investigativa y experimental. De igual forma a la ayuda brinda por Raphael Couturier PhD., quien con sus conocimientos en *software* R, pudo despejar mis dudas, durante el uso de esta herramienta.

Resumen

El análisis de datos es fundamental dentro de la educación y al no utilizar las técnicas de análisis óptimas al tipo y cantidad de datos, se crea en muchos procesos (repitencia, deserción, bajo rendimiento, alto rendimiento), problemas de obstrucción o lentitud del cálculo, siendo inaplicables e irresolubles. Por tal razón, esta investigación plantea realizar un análisis comparativo de los tiempos de procesamiento y espacio de memoria entre las técnicas de agrupación y las técnicas de búsqueda de reglas de asociación similares al Análisis Estadístico Implicativo (AEI) y Learning Analytics (LA). Las fases desarrolladas son: (1) identificar las técnicas similares entre el AEI y LA, mediante la adaptación del método de estudio de similitud entre modelos y estándares (MSSS)[1], (2) identificar el sistema operativo con mejor manejo de recursos y (3) identificar la técnica más optima en el análisis de datos educativos; estos dos últimos con la elaboración de un diseño cuasi-experimental[2]. La hipótesis por demostrar es que existe diferencia significativa (función espacio y/o función tiempo) entre los algoritmos similares a LA y AEI, la cual será probada a través de un diseño cuasi-experimento de tipo RGXO, (notación de Campbell y Stanley). La demostración de la hipótesis permitirá al docente poder resolver rápidamente los problemas educativos que utilizan datos masivos y de diferente tipo (binarios, modales, numéricos), al seleccionar los algoritmos más óptimos.

Palabras claves: learning analytics, análisis estadístico implicativo, métodos cluster, minería de reglas de asociación, RCHIC

[1] Calvo-Manzano, J., Cuevas, G., Muñoz, M., & San Feliu, T. (2008). Process similarity study: Case study on project planning practices based on CMMI-DEV v1. 2. EuroSPI 2008 Industrial Proceedings.

[2] Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research.

Abstract

Data analysis is essential in education, and when ideal analysis techniques are not used for the type and quantity of data, many processes are created (repetition, dropout, low performance, high performance), problems of obstruction or slow calculation, which are inapplicable and without solution. For this reason, this study proposes to carry out a comparative analysis of processing times and memory space between the techniques of grouping and the search for association rules that are similar to implicative statistical analysis (ISA) and learning analytics (LA). The developed stages are (1) identify similar techniques between ISA and LA through the adaptation of the study method of similarity among models and standards (MSSS) [3], (2) identify an operating system with better resource management, and (3) identify the most ideal technique in educational data analysis, these last two stages with the elaboration of a quasi-experimental design[4].

The hypothesis to be demonstrated is whether or not there is a significant difference (space function and/or time function) between the algorithms that are similar to LA and ISA which will be tested through a quasi-experimental design that is RGXO (notation of Campbell and Stanley). The demonstration of the hypothesis will make it possible for the teacher to quickly solve educational problems that use massive data of different types (binary, modal, numerical) when the most ideal algorithms are selected.

Key words: learning analytics, implicative statistical analysis, cluster methods, association rule mining, RCHIC

[3] Calvo-Manzano, J., Cuevas, G., Muñoz, M., & San Feliu, T. (2008). Process similarity study: Case study on project planning practices based on CMMI-DEV v1. 2. EuroSPI 2008 Industrial Proceedings.

[4] Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Lista de Fórmulas

Capítulo 1

Introducción

Learning Analytics como lo indica (Ferguson, 2016) es un área importante de aprendizaje potenciado por la tecnología que ha surgido durante la última década. Se ha impulsado el desarrollo de análisis de entornos educativos a través de un examen de los factores tecnológicos, educativos y políticos. El aumento de las perspectivas de aprendizaje e influencia de las preocupaciones económicas nacionales han permitido el surgimiento del LA, para implementar estudio de información basados en datos.

(Bogarín Vega, Romero Morales, & Cerezo Menéndez, 2015) mencionan que el LA es la medición, recopilación, análisis y datos sobre los alumnos y sus contextos, con el propósito de comprender y optimizar el aprendizaje y los entornos en los que se produce para cubrir la mayoría de la investigación educativa, pero típicamente se combina con dos suposiciones: que el aprendizaje analítico hace uso de datos preexistentes, legibles por la máquina, y que sus técnicas pueden ser usadas para manejar grandes datos, grandes conjuntos de datos que no sean factibles tratar manualmente.

El reporte horizon realizado por (Johnson et al., 2016) indica que el LA pretende utilizar el análisis de datos para generar información que permita tomar las mejores decisiones en ámbito educativo, para elaborar mejores pedagogías, entender a los estudiantes el porqué de su abandono de los estudios e incrementar la retención, está información ha sido eficaz y deben mantenerse; los resultados obtenidos son importantes para los directivos, los encargados de crear normativas y demás autoridades que son parte del sistema educativo. Para los docentes, el LA es crucial a la hora de buscar cómo interactúan los educandos con los textos y materiales disponibles por Internet. Los educandos también se benefician de los resultados de LA, mediante las diferentes aplicaciones desarrolladas para dispositivos móviles y plataformas por Internet que utilizan datos específicos de cada estudiante para crear sistemas de apoyo que se ajusten a las necesidades de aprendizaje.

Por otra parte, el análisis estadístico implicativo (AEI) según la investigación realizada por (Pazmiño, 2014), indica que su primera aplicación es el ámbito educativo, históricamente en el área de la matemática. De los artículos relacionados a la Educación (71 artículos) duplican aquellos de

desarrollo teórico (27 artículos), por lo cual el investigador concluye que existe muchas experiencias en la aplicación en el área educativa, con lo cual el autor motiva a los educadores a utilizar esta nueva técnica estadística multivariada.

El análisis de los datos provenientes de ciertos procesos educativos (evaluaciones, rendimiento, uso de tecnologías, entre otros) son irrealizables por no utilizar técnicas óptimas que minimicen el espacio de memoria y el tiempo de procesamiento (complejidad algorítmica). El análisis de datos es fundamental en los procesos educativos y al no utilizar las técnicas optimas al tipo y cantidad de datos, se visualiza obstrucción o lentitud en los cálculos, lo que hace que las técnicas de análisis de datos sean inaplicables. He aquí, la importancia de un estudio comparativo, para obtener las técnicas óptimas de análisis de datos similares utilizadas en el ASI y LA. A los educadores, desde las autoridades distritales, autoridades de las instituciones educativas y docentes, encuentran barreras al enfrentarse al análisis de datos, tratan de visualizar la evaluación de temas referentes a la educación y su formación. Entonces los datos en un futuro dentro de la educación van a tener un papel fundamental, en la toma de decisiones dentro del sector educativo.

Las técnicas a ser analizadas son clustering (LA: hclust.vector, dendro.variables, diana; AEI: callHierarchyTree, callSimilarityTree) y minería de reglas de asociación (LA: apriori, weclat, eclat; ASI: implicativeGraph).

Las técnicas conceptuales de clustering según (Fisher & Langley, 1985), son métodos desarrollados por científicos sociales y naturales para crear esquemas de clasificación sobre conjuntos de objetos. Alternativamente, los clustering conceptualmente pueden verse como una forma de aprendizaje por observación o formación de conceptos, en oposición a los métodos de aprendizaje a partir de ejemplos o la identificación de conceptos. (Kaufman & Rousseeuw, 2009) indica que el objetivo de clustering es agrupar datos, en varios conjuntos de datos. Las agrupaciones se obtienen a partir de las variables a ser analizadas (similitudes y diferencias entre estudiantes).

Por otra parte, la segunda técnica, minería de reglas de asociación el objetivo es encontrar reglas (if-then). Las reglas descubiertas revelan co-ocurrencias comunes, las que podrían ser difíciles de descubrir manualmente.

Por lo cual, y en base a la aproximación del ASI y LA (R. A. Pazmiño-Maji, F. J. García-Peñalvo, & M. A. Conde-González, 2016), la implementación de un modelo de similitud entre las técnicas en base al método de estudio de similitud entre modelos y estándares (MSSS) (Calvo-Manzano, Cuevas, Muñoz, & San Feliu, 2008) y la elaboración de un diseño cuasi-experimental de la ingeniería de *software* propuesto por Donald Campbell y Julian Stanley (Campbell & Stanley, 1966), el objetivo

principal es determinar y comparar el tiempo de procesamiento en velocidad y uso de memoria de las técnicas clustering (LA:hclust.vector, dendro.variables, diana; ASI: callHierarchyTree, callSimilarityTree) y minería de reglas de asociación (LA: apriori, weclat, eclat; ASI: implicativeGraph).

Al obtener la mejor técnica similar entre el AEI y LA, en tiempo y uso de memoria para el análisis de datos masivos, los docentes independientemente del nivel podrán aplicar, y conseguir los resultados del análisis en tiempos óptimos. En el caso, de las instituciones educativas que cuenten con sistemas académicos, estas aplicaciones arrojan varios reportes (generalmente archivo en Excel .cvs) de acuerdo con las necesidades de la institución, para luego pre-procesar los datos, a través del software R y RStudio, y aplicar la técnica más óptima para la obtención de los resultados. Los resultados derivados del análisis permitirán al docente tomar las mejores decisiones dentro del campo educativo en análisis. (R. Pazmiño-Maji, F. García-Peñalvo, & M. Conde-González, 2017), plantean una secuencia de procesos para el análisis de datos, la cual constan de: selección, pre-proceso, transformación, aplicación de la técnica a utilizar y evaluación de la interpretación.

1.1. Presentación del trabajo

El análisis de datos es fundamental dentro de la educación y al no utilizar las técnicas de análisis óptimas al tipo y cantidad de datos, se crea en muchos procesos (repitencia, deserción, bajo rendimiento, alto rendimiento, docentes con mayor número de estudiantes aprobados, entre otros) el problema de la obstrucción o lentitud del cálculo, lo que hace que las técnicas de análisis de datos sean inaplicables, por lo tanto, los problemas educativos sean irresolubles. Por tal razón, en esta investigación se plantea: identificar las técnicas similares entre el Análisis Estadístico Implicativo (AEI) y Learning Analytics (LA), establecer técnicas óptimas (referente al tiempo de ejecución y espacio de memoria) de análisis de datos (agrupaciones y búsqueda de reglas de asociación) utilizadas en el AEI y LA, con la elaboración de un diseño cuasi-experimental mediante la ingeniería de *software*.

En este trabajo se realizará un análisis comparativo de los tiempos de procesamiento y espacio de memoria entre las técnicas de agrupación y las técnicas de búsqueda de reglas de asociación similares al Análisis Estadístico Implicativo y Learnign Analytics. Se consideran como factores el número y tipo de variables, número de casos, algoritmos y sistema operativo (variables independientes) y como variables dependientes las funciones espacio y tiempo. La hipótesis por demostrar es que existe diferencia significativa (función espacio y/o función tiempo) entre los algoritmos similares a LA y AEI. Para el estudio se utilizan dos computadores con el mismo microprocesador y con los sistemas operativos Windows 10 y Ubuntu 16.04 y MacOS Sierra 10.12. Todos los computadores y sistemas

operativos trabajan con el *software* estadístico libre R v 3.4.1 y el entorno de desarrollo integrado libre (IDE) RStudio v 1.0.143. Para demostrar las hipótesis se plantea un experimento de tipo RGXO, (notación de Campbell y Stanley) donde RG representa el grupo experimental conformado por datos generados aleatoriamente, X son los tratamientos y O es la observación luego del tratamiento, no se tiene grupo de control ni tampoco una pre-prueba. Luego de la comprobación de los supuestos de normalidad, homocedasticidad e independencia se empleará un test paramétrico ANOVA (probar hipótesis a través de analizar la variabilidad) con cinco factores, se trabajará con un nivel de significancia del 95%.

La demostración de la hipótesis sobre la diferencia entre las funciones espacio y tiempo de los algoritmos de agrupación y de búsqueda de reglas de asociación utilizados en AEI y LA, permitirá que el docente pueda resolver más rápidamente los problemas educativos que utilizan datos masivos y de diferente tipo (binarios, modales, numéricos), al seleccionar los algoritmos más óptimos. Esto implica que los docentes obtendrán los resultados con mayor rapidez, al ser de gran beneficio para poder actuar y tomar las medidas necesarias dentro del ámbito educativo. Los datos para los cuales se realizarán las pruebas serán aquellos que son utilizados en la mayoría de casos en el ámbito educativo, es decir de tipo binario (si/no, verdadero/falso, admite/no admite, con propiedad/sin propiedad, aprueba/desaprueba, admitido/no admitido, promovido/no promovido, motivado/desmotivado), modal (bajo/medio/alto, malo/regular/bueno/muy bueno/sobresaliente, nada de acuerdo/poco de acuerdo / ni en acuerdo ni en desacuerdo / muy de acuerdo / completamente de acuerdo) y numérico (notas, rendimiento, asistencia, número de faltas).

El Learning Analytics (LA) y el Análisis Estadístico Implicativo (AEI) permiten comprender mejor la enseñanza, el aprendizaje, el contenido inteligente, la personalización y la adaptación. Aún en las primeras etapas de la investigación y la implementación, varias organizaciones (Society for Learning Analytics Research y la International Educational Data Mining Society) se han formado para fomentar una comunidad de investigación en torno al papel de la analítica de datos en la educación. LA y AEI aportan tecnologías y metodologías para el desarrollo de análisis analíticos, modelos analíticos, la importancia de aumentar las capacidades analíticas en las organizaciones educativas y los modelos para implementar análisis en entornos educativos. A partir de este estudio se podrá aplicar el mejor método en dependencia al número, tipo, variables y otras características de los datos, dentro del proceso enseñanza – aprendizaje.

1.2. Descripción del documento

El desarrollo del proyecto, está repartido en seis capítulos de la siguiente forma:

El capítulo I, menciona los objetivos del desarrollo del proyecto de investigación y su implicación en el ámbito educativo.

El capítulo II, plantea la propuesta de investigación, enfocándose en el análisis de estudios preliminares,

El capítulo III, vislumbra el marco teórico, el mismo que permite fundamentar aspectos científicos enfocados al tema de investigación, así como las pautas para desarrollar la comparación del Análisis Estadístico Implicativo y el Learning Analytics

El capítulo IV, abarca la metodología de investigación, la cual refleja los procesos de investigación, diseño y experimentación ineludibles en la recolección, procesamiento, pruebas e interpretación de información.

El capítulo V, muestra el estudio de las técnicas similares entre el AEI y LA, los resultados obtenidos del diseño cuasi-experimental, la comprobación de la hipótesis, además de la identificación de las técnicas óptimas.

El capítulo VI, se encuentra las conclusiones y recomendaciones, las mismas que ayudarán a los docentes aplicar la mejor técnica para el análisis de datos masivos, dentro de los procesos educativos.

Capítulo 2

Planteamiento de la Propuesta de Trabajo

2.1. Información técnica básica

Tema: Estudio comparativo del Análisis Estadístico Implicativo y el Learning Analytics en relación al uso de las técnicas de exploración de datos educativos

Tipo de trabajo: Proyecto de investigación y desarrollo.

Clasificación técnica del trabajo: Investigación

Líneas de Investigación, Innovación y Desarrollo

Principal: Sistemas de Información y/o Nuevas Tecnologías de la Información y Comunicación y sus aplicaciones

Secundaria: Pedagogía, Andragogía, Didáctica y/o currículo.

2.2. Descripción del problema

El análisis de datos es fundamental dentro de la educación y al no utilizar las técnicas óptimas al tipo y cantidad de datos (evaluaciones, rendimiento, uso de tecnología, entre otros), se crea en muchos procesos (repitencia, deserción, bajo rendimiento, alto rendimiento, docentes con mayor número de estudiantes aprobados, entre otros) el problema de la obstrucción o lentitud de cálculos estadísticos y representaciones gráficas, lo que hace que las técnicas sean inaplicables, y por tanto los análisis educativos irresolubles. Por tal razón, en esta investigación se plantea establecer técnicas óptimas (referente al tiempo de ejecución y espacio de memoria) de análisis de datos (agrupaciones y búsqueda de reglas de asociación) utilizadas en el Análisis Estadístico Implicativo (AEI) y Learning Analytics (LA).

En relación al LA, los constantes avance tecnológico, permiten obtener nuevas herramientas *software* o mejorar de acuerdo a las necesidades. (Ferguson, 2014) afirma "El LA es un área importante del aprendizaje mejorado por tecnología (TEL) que surgió durante la última década. Esta revisión del campo comienza con un examen de los factores tecnológicos, educativos y políticos que han impulsado el desarrollo de análisis en entornos educativos. Continúa para trazar el surgimiento de LA, incluidos

sus orígenes en el siglo XX, el desarrollo de análisis basados en datos, el aumento de las perspectivas centradas en el aprendizaje y la influencia de las preocupaciones económicas nacionales" (p.138). Al encontrarnos en una sociedad de conocimiento y algunos grandes avances, nuevos inventos y descubrimientos progresarán exponencialmente, el hombre está en la búsqueda constante de nuevos métodos, que simplifiquen las actividades en cualquiera de la áreas o campos, es así que el problema del LA no es tecnológico, pues su constante desarrollo permite al LA tener las herramientas necesarias, con el fin de busca nuevos diseños de algoritmos y más óptimos, en la resolución de problemas.

LA tiene varios desafíos por resolver. (Avella, Kebritchi, Nunn, & Kanai, 2016) afirma "Los desafíos incluyen cuestiones relacionadas con el seguimiento, la recopilación, la evaluación y el análisis de datos; falta de conexión con las ciencias del aprendizaje; optimizar los entornos de aprendizaje y los problemas éticos y de privacidad" (p.13).

Otros desafíos del LA, (Ferguson, 2014) afirma, "debe abordar ahora: integrar la experiencia de la ciencia del aprendizaje, trabajar con una gama más amplia de conjuntos de datos, relacionarse con las perspectivas del alumno y desarrollar un conjunto de pautas éticas" (p.145). En la parte pertinente a nuestro estudio, LA debe mejorar en el manejo de gran cantidad de datos, y al manejar gran cantidad de datos, se debe desarrollar algoritmos o técnicas que entreguen con mayor rapidez los resultados.

Por otra parte, los grupos de investigación LITI (laboratorio de Innovación en Tecnología de la Información) y GIDTIC (Grupo de Investigación e Innovación en Docencia con Tecnología de la Información y la Comunicación), trabaja en esta línea, en la aplicación del Learning Analytics para solucionar problemas derivados del alto coste y esfuerzo que supone aplicar paradigmas de aprendizaje. Uno de sus integrantes (Fidalgo, 2012) afirma "el problema del Learning Analytics no es tanto la tecnología, sino cómo determinar qué datos son los relevantes (y capturarlos LMS Learning Management Systems) para la mejora de los procesos de aprendizaje, la modelización de los comportamientos del alumnado y del profesorado, el establecimiento de diagnósticos y los recursos más adecuados para los distintos modelos de comportamiento". Además, (Chatti et al., 2014) indica que es un desafío clave, los grandes análisis de aprendizaje, cómo agregar e integrar datos sin procesar fuentes múltiples y heterogéneas, a menudo disponibles en diferentes formatos distribuidos en el espacio, los medios y el tiempo; al tener Big Data el tiempo es un desafío técnico porque se deben implementar métodos y herramientas de análisis eficientes para entregar resultados significativos sin demasiada demora, de modo que las partes interesadas tengan la oportunidad de actuar sobre la información recién obtenida a tiempo.

Esto se corrobora, con el análisis realizado por (Avella et al., 2016) quienes afirman "El campo del análisis de aprendizaje ha sido y continuará expandiéndose en gran medida debido en parte a la capacidad de almacenar cantidades crecientes de datos y una gran diversidad de líneas de investigación. Como resultado de una mayor disponibilidad y acceso a los datos, el LA proporcionará una mayor comprensión de los patrones de comportamiento, redes e interacciones del alumno" (p.25). LA tiene la necesidad de proporcionar una mayor comprensión de cómo analizar los datos para optimizar los resultados y utilizar la información para mejorar el proceso educativo en todos los niveles. Es decir, que sea visible el proceso de análisis de datos, para una mejor interpretación de los procesos para la obtención de resultados.

Uno de los unos desafíos del LA, que es importante dentro de nuestra investigación es "trabajar con una gama más amplia de conjuntos de datos" (Ferguson, 2014). De acuerdo a esto, el AEI puede darnos mejores réditos al permitir trabajar, según (Zamora, Gregori, & Orús, 2009), con variables de distintos tipos como: binaria (0-1), modal (1 a 5), frecuencial (0 a 7), intervalo (rangos numéricos), vectorial (grupos) y fuzzy (datos difusos).

Por estas razones es importante el manejo de otras técnicas que pueden competir con las técnicas de LA, he aquí la importancias de las técnicas del AEI las cuales se automatizan mediante el *software* CHIC, que fue desarrollado y es constantemente actualizado por el profesor Raphaël Couturier (Couturier, 2015). Actualmente existe una versión de CHIC en el ambiente estadístico R, esta versión será libre y permite que un mayor número de personas se beneficien de él, está aplicación lleva de nombre RCHIC (Couturier, 2018).

2.3. Preguntas básicas

¿Cómo aparece el problema que se pretende solucionar?

Utilización de diferentes tendencias tecnológicas en la comunicación, las mismas que no se aprovechan en la parte educativa de forma eficiente. Se generan grandes volúmenes de datos y en formato diferente.

¿Por qué se origina?

Desconocimiento de las herramientas de análisis de datos apropiados para este nuevo contexto educativo.

¿Qué lo origina?

La no existencia de técnicas optimas que minimicen el espacio de memoria y el tiempo de procesamiento.

¿Cuándo se origina?

Inicia al utilizar nuevas herramientas de comunicación dentro de la educación.

¿Dónde se origina?

En todas las instituciones educativas en las que hay utilización masiva de herramientas tecnológicas y los docentes desean utilizar esa información para mejorar el aprendizaje.

¿Dónde se detecta?

En la utilización de nuevas herramientas pedagógicas y tecnologías, al no analizar el proceso desarrollado por los estudiantes de forma eficaz.

2.4. Formulación de hipótesis y/o pregunta de investigación

Existe diferencia significativa en el espacio de memoria ocupado entre las técnicas de agrupación (jerárquicas y no jerárquicas) similares al análisis estadístico implicativo y las analíticas de aprendizaje, en base a procesos estadísticos.

Existe diferencia significativa en el espacio de memoria ocupado entre las técnicas de búsqueda de reglas de asociación similares al Análisis Estadístico Implicativo y *Learning Analytics,* en base a procesos estadísticos.

Existe diferencia significativa en el tiempo de ejecución entre las técnicas de agrupación (jerárquicas y no jerárquicas) similares al Análisis Estadístico Implicativo y *Learning Analytics*, en base a procesos estadísticos.

Existe diferencia significativa en el tiempo de ejecución entre las técnicas de búsqueda de reglas de asociación similares al Análisis Estadístico Implicativo y las *Learning Analytics*, en base a procesos estadísticos.

2.5. Variable(s)

Variables independientes: Número y tipo de variables, número de casos, sistema operativo y las técnicas similares de exploración de datos utilizadas tanto en el Análisis Estadístico Implicativo (AEI) como en el *Learning Analytics* (LA).

Variables dependientes: Espacio de memoria ocupado (en megabytes) y el tiempo de ejecución (en minutos).

2.6. Objetivos

Objetivo general. - Comparar las técnicas similares del Análisis Estadístico Implicativo (AEI) y el *Learning Analytics* (LA), mediante un diseño cuasi-experimental (de investigación sobre la enseñanza) en la ingeniería de *software*, para determinar las técnicas más óptimas.

Objetivos específicos. -

1. Identificar las técnicas similares entre el Análisis Estadístico Implicativo (AEI) y el Learning Analytics (LA).
2. Elaborar el diseño cuasi-experimental a utilizar mediante la ingeniería de *software*.
3. Ejecutar el diseño cuasi-experimental.
4. Elaborar las conclusiones estadísticas luego de la ejecución del diseño cuasi-experimental.
5. Identificar las técnicas más óptimas, similares entre el AEI y LA, de los resultados obtenidos del diseño cuasi-experimental.

2.7. Delimitación funcional

¿Qué será capaz de hacer el producto final del trabajo de titulación?

- ✓ El diseño cuasi-experimental en la ingeniería de *software*, permitirá generar automáticamente datos aleatorios por el *software* estadístico libre R y el entorno de desarrollo integrado libre (IDE) RStudio, los cuales servirán para simular la información del sistema educativo y aplicar las técnicas similares entre el AEI y LA, para seleccionar las técnicas de análisis de datos optimas en espacio y tiempo en las teorías del Análisis Estadístico Implicativo y Learning Analytics.
- ✓ A través del diseño cuasi-experimental se simulará los datos provenientes de ciertos procesos educativos (evaluaciones, rendimiento, uso de tecnologías, entre otros), generados aleatoriamente por el *software* estadístico libre R.
- ✓ Adopción de la mejor técnica de análisis de datos, para su aplicabilidad dentro del ámbito educativo.
- ✓ Permitir a los educadores realizar el análisis de datos (evaluaciones, rendimiento, uso de tecnología), de los estudiantes, dentro y fuera del aula, al aplicar la mejor técnica similar entre del AEI y LA, obtenida al ejecutar el diseño cuasi-experimental en la ingeniería de *software*.
- ✓ La técnica más óptima podrá ser aplicada en los sistemas operativos Windows, Ubuntu y Mac (32 y 64 bits), mediante la instalación de las aplicaciones R y RStudio, en sus últimas versiones.

¿Qué no será capaz de hacer el producto final del trabajo de titulación?

- ✓ Extrapolar los resultados obtenidos en condiciones diferentes, como aplicar una información conocida a otro dominio para extraer consecuencia.
- ✓ El diseño cuasi-experimental en la ingeniería de *software*, al trabajar con el *software* estadístico R, no arrojara información errónea e irrelevante, por ser un sistema informático, diseñado para el análisis de datos estadísticos.

Capítulo 3

Marco Teórico

3.1. Definiciones y conceptos

En el siguiente capítulo presento los fundamentos teóricos de la investigación, el Learning Analytics y Análisis Estadístico Implicatvo.

3.1.1. Learning Analytics

La Society for Learning Analytics Research asumió la definición emitida en la Primera Conferencia sobre Learning Analytics and Knowledge (LAK 2011): "Learning Analytics es el control, compilación, análisis e información de datos sobre los estudiantes y sus contextos, con la finalidad de comprender y optimizar el aprendizaje de acuerdo al medio donde se desarrollan". (Rubén A. Pazmiño-Maji et al., 2016)

Según (Elias, 2011) el Learning Analytics es un ámbito emergente en el que se utilizan experimentadas herramientas analíticas con la finalidad de mejorar el aprendizaje y la educación. Este ámbito se basa en una serie de campos de estudio, como la inteligencia empresarial, análisis *web*, análisis académico, minería de datos educativos y análisis de acciones, los mismos que están estrechamente vinculados entre sí.

La NMC Horizon Report en el 2016 determinó a LA como una de las preferencias más importantes en el aprendizaje y la enseñanza mejorados en tecnología (Johnson et al., 2016).

Los autores (Siemens & Long, 2011) afirman que los ensayos de imaginar frecuentemente el futuro de la educación se basa en el ahínco de las nuevas tecnologías - dispositivos de computación universales, diseños flexibles en las aulas y en las innovadoras presentaciones visuales. El aspecto más patético que establece el futuro de la educación superior es que las personas no pueden tocar ni ver: "grandes datos y análisis". Es así que la analítica de aprendizaje aún se encuentra en períodos de implementación y experimentación. Hay muchas preguntas sobre cómo se relaciona el análisis con los sistemas organizativos existentes. Sin duda, que la analítica y la gran cantidad de datos tendrán un papel importante que redimir en el futuro de la educación superior. Los sectores gubernamental y empresarial confirman la tendencia del papel creciente de las técnicas y tecnologías de análisis. En el área educativa, el valor de la analítica y los grandes datos se pueden encontrar en: (1) en las

orientaciones de las actividades de reforma en la educación superior; Y (2) contribuye a los educadores para mejorar la enseñanza y el aprendizaje. La analítica del aprendizaje es importante para despejar inconvenientes que se presentan en gran parte de la educación superior. Los educadores, los estudiantes y los administradores requieren una base sobre la cual esparcir el cambio. La disponibilidad de información en tiempo real sobre el desempeño de los estudiantes, incluidos los estudiantes que están en riesgo, puede ser una ayuda significativa en la planificación de las actividades de enseñanza para los educadores. Los estudiantes, reciben información de su desempeño de manera motivadora y alentadora relacionándolo con la de sus compañeros o sobre su progreso en relación con sus metas personales. Finalmente, en la actualidad los administradores y los responsables de la toma de decisiones afrontan una enorme fluctuación ante los recortes presupuestarios y la competencia global en la enseñanza superior. El Aprendizaje Analítico puede ingresar en los conflictos en torno a cómo asignar los recursos, desarrollar ventajas competitivas, y lo más importante, mejorar la calidad y el valor de la experiencia de aprendizaje.

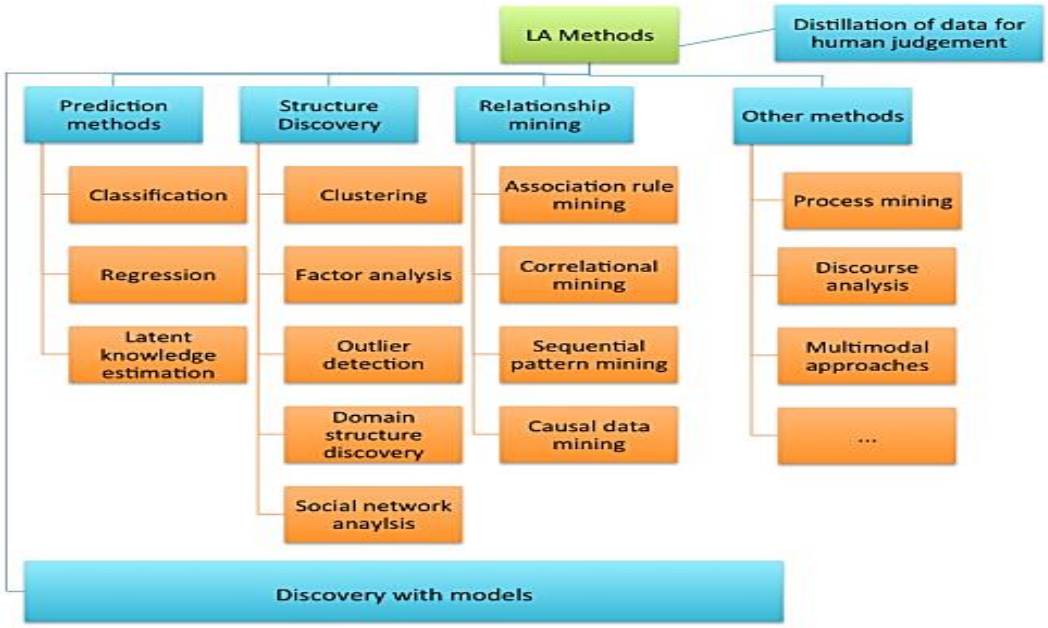
Deseo subrayar que, el comprender y optimizar el aprendizaje de los estudiantes dentro de sus contextos, es el objetivo de los docentes en cualquier nivel educativo, muchos de ellos no conocen o no saben que técnica aplicar para obtener resultados que les permitan tomar medidas dentro del proceso enseñanza aprendizaje. El Learning Analytics mediante un proceso de recopilación, análisis y presentación de información sobre los estudiantes dentro de sus contextos, permite obtener resultados los cuales ayudarán a los docentes a comprender y optimizar el aprendizaje dentro de su entorno.

3.1.1.1. Técnicas utilizadas en el Learning Analytics.

Los autores (Van Harmelen & Workman, 2012), definen al LA como la herramienta que suministra resultados que informan a los maestros, estudiantes o administradores y facilitan una base para la acción e intervención adecuada a nivel de curso o institucional. El tipo de metodologías analíticas empleadas en un proyecto determinado de LA depende de las partes interesadas, sus objetivos y del tipo de datos recopilados.

Este apartado brinda una visión general de las metodologías comunes, es decir, los métodos para analizar los datos educativos y las herramientas que implementan estas técnicas.

Figura 1. Descripción general de los métodos de LA.



Fuente: (Steiner et al., 2014)

Dicho brevemente a partir de datos educativos, el LA utiliza diferentes métodos para extraer patrones significativos. Las técnicas ciertamente utilizadas en un escenario de aplicación determinado dependerán de los objetivos de las tareas de análisis, pero asimismo del tipo de datos recopilados. Baker y Siemens creen que los métodos de la minería y análisis de datos en general, así como la psicometría y la medición educativa son las principales fuentes de inspiración para los métodos y herramientas de LA, las mismas que proporcionan un enfoque sistemático de los métodos clave actualmente aplicados en LA. Se puede mencionar cinco clases principales: métodos de predicción, descubrimiento de estructuras, minería de relaciones, descubrimiento con modelos y destilación de datos para el juicio humano (Baker & Inventado, 2014). LA se basa en (Baker & Inventado, 2014) y (Siemens & Baker, 2012).

Es importante, indicar que varios métodos no han sido explotados, todos estos métodos utilizan diferentes técnicas (ver figura 1), dentro de esta investigación posteriormente se realiza un análisis de las técnicas que son más utilizadas, y que son similares a las técnicas del AEI, las mismas que servirán de base en la implementación del diseño-cuasi experimental, así como también un análisis detallado del funcionamiento de cada una de las técnicas.

3.1.2. Learning Analytics y Educación

Avanzando en nuestro razonamiento, al indica que mayoritariamente la educación está inmersa en el mundo de la información. Cada día se utiliza una variedad de información, la cual es difícil de ponderar, es por eso que se habla de terabytes, gigabytes; al dar paso a una nueva tecnología llamada Big data, puesto que continuará generándose diariamente grandes cantidades de información. El desarrollo de nuevos paradigmas en el análisis de la información se lo ha conseguido por el acceso a internet, tecnologías móviles, redes sociales y las pericias que le brinda al usuario para generar y publicar contenidos. Big Data se encuentra incorporada a un conjunto de tecnologías, con las cuales se puede predecir y analizar el comportamiento de los individuos.

Big Data también es una herramienta de apoyo en el ámbito educativo, dentro del cual consta "*Learning Analytics*" o analítica del aprendizaje para referirse al uso inteligente de los datos que produce el estudiante, permite generar modelos analíticos con la finalidad de personalizar el aprendizaje, comprender y predecir los procesos implicados, así como optimizar los entornos en los que dicho aprendizaje se produce. Al hablar de análisis de datos en general, o al referirse a la analítica del aprendizaje en particular, se tiene que recordar lo siguiente: De manera individual, los datos no indican nada, por lo tanto, es necesario convertir esos datos en información útil. ¿Cómo se consigue esto? El proceso consiste en: generar los datos, almacenarlos, gestionarlos y analizarlos. Porque… ¿para qué se va almacenar tantos datos si no se va a utilizarlos de manera rentable y eficiente?, recuerda: Big Data y por extensión LA permite construir patrones de aprendizaje inteligentes a partir de la información, al lograr un valor de retorno cuantificable y, en definitiva, una experiencia de aprendizaje única y personal, (Kons, 2016).

El LA se relaciona con otros métodos como *Machine learning* por los métodos cluster, que son denominados métodos de clasificación, ejemplos de estos métodos son: máquinas de soporte, redes neuronales, árbol de decisión, bosque aleatorio o potenciación. (Wolfgang & Hendrik, 2012) indica que la minería de datos, machine learning (aprendizaje automático), filtrado colaborativo o análisis semántico latente en Technology-Enhanced Learning (TEL) (Tecnología-Aprendizaje mejorado) se revela a través de un número creciente de conferencias científicas, talleres y proyectos combinados bajo el nuevo término de investigación Learning Analytics. De igual forma se relaciona con AEI, con algunas diferencias como: el AEI no necesita una tabla de aprendizaje, pero tampoco tiene una tabla de Testing.

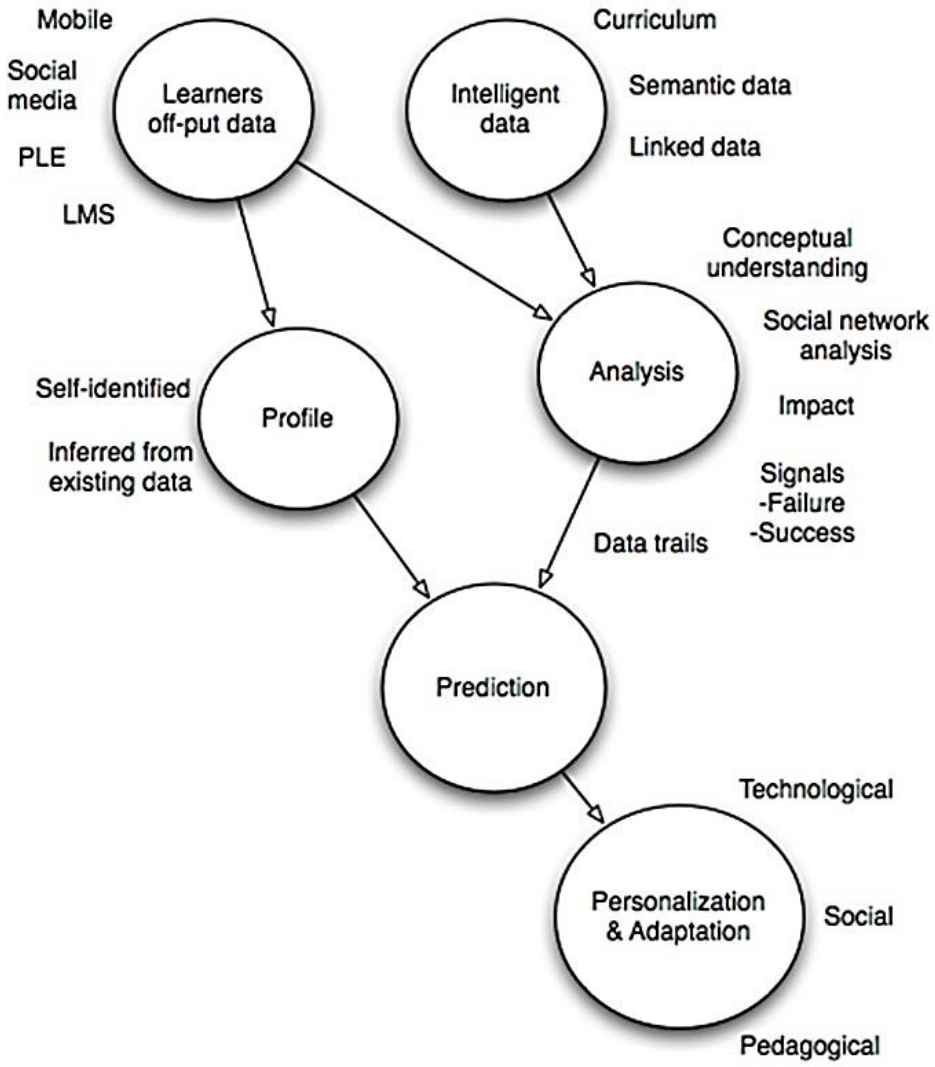
Dicho de otra forma, el LA está integrado por minería de datos, análisis semántico latente, machine learning, entre otros, lo permite visualizar el crecimiento del análisis de datos masivos.

El LA es un tema que incide en el profesor y al alumno. Al docente el análisis del aprendizaje le brinda una ventana para ver el compromiso de los estudiantes en el proceso de aprendizaje, y dónde pueden tener dificultades o problemas. Por medio de las visualizaciones de datos de aprendizaje se pueden identificar a los estudiantes que no alcanzan los resultados, permite que los profesores investiguen con los respectivos alumnos, e intervengan con medidas correctivas apropiadas. (Alcalde, 2015)

Se debe aclarar que el LA no es un instrumento de control o supervisión directa de la tarea tutorial, sino más bien es un proceso de medida-análisis cuantitativo y estudio cualitativo que derivan en la adopción de medidas que contribuyen a modificar, mejorar y diferenciar proyectos educativos (Ver figura 2). Además, permiten estudiar el componente informal que complementa lo establecido en el itinerario formativo. (Siemens, 2012)

Si bien el proyecto a desarrollar, mediante el diseño cuasi-experimental, permitirá visualizar cualidades pero no se podrá dar lectura, debido a que se utilizará datos simulados, generados aleatoriamente, divididos en columnas, las cuales forman una base de datos, no maneja atributos (rendimiento, acceso al internet), si no por variables generadas (v1, v2, v3,..). Pero si se aplica cualquier técnica similar entre el AEI y LA, con datos reales, estás arrojarían información estadística y gráfica de acuerdo al estudio a ejecutar, lo que permitirá realizar un análisis cualitativo y cuantitativo.

Figura 2. Proceso de Learning Analytics



Fuente: (Siemens, 2012)

De acuerdo a lo mencionado, se puede ver que, en la actualidad, el análisis de grandes datos, el análisis sistemático del comportamiento de los usuarios de la *web* y desarrollos específicos de la estadística aplicada al uso de la tecnología facilita y amplía las posibilidades del aprendizaje analítico. En este sentido, una de las áreas de la educación más influidas por el desarrollo de las técnicas analíticas es la evaluación.

3.1.3. Análisis Estadístico Implicativo

Los autores (Couturier & Pazmiño, 2016), mencionan que en muchas situaciones, los estudiantes o, más generalmente, los individuos llenan formularios o encuestas. Tales formas podrían ser usadas para

evaluar el conocimiento de los estudiantes después de una lección en un aula o podrían formar una evaluación global de todos los estudiantes en un país. De manera más general, una encuesta tiene como objetivo reunir la opinión de las personas sobre un tema en particular. En tal caso, el análisis de ítems ofrece información interesante sobre cómo se han respondido los ítems. El Análisis Implicativo Estadístico (AEI) que produce reglas orientadas. Se estudia una encuesta sobre el futuro de los estudiantes en las diferentes escuelas de la ESPOCH para resaltar el interés de utilizar el AEI para poder analizar el comportamiento general de la población.

El Análisis Estadístico Implicativo (AEI) es una poderosa herramienta de investigación propuesta por Regis Gras a la Educación Matemática, ya que permite determinar cuasi-implicaciones (implicaciones con excepciones) entre variables y clases de variables. El método (AEI) se desarrolla en correspondencia con los problemas referente a la didáctica matemática, su objetivo contempla la estructuración de datos, interrelacionados a sujetos y variables, y la extracción de reglas inductivas entre las variables. Además, permite conocer posibles relaciones de similitud, implicación y cohesión entre el rendimiento académico relacionado al ámbito educativo. La visualización de los resultados y su interpretación, se facilita con el *software* C.H.I.C. (Clasificación Jerárquica Implicativa y Cohesiva). (Iurato, 2012), (Pazmiño, 2014)

De manera puntual me refiero a AEI, como una herramienta estadística, la cual fue desarrollada hace varios, pero inicia a sobresalir en los últimos 10 año, y cada vez existe más investigaciones desarrolladas, la cual se fundamenta en el software CHIC.

Según (Pazmiño-Mají, García-Peñalvo, & Conde-González, 2017), manifiestan que desde el comienzo del AEI, con Regis Gras, hace más de treinta años, se creó el Análisis Implicativo Estadístico (AEI) acompañado de un conjunto de herramientas de datos que permiten analizar el conocimiento a partir de la información contenida en la base de datos (individuos y variables). El enfoque del AEI, se realiza a partir de la generación de reglas asimétricas entre variables y clases de variables, representadas por tablas (clusters no jerárquicos), gráficos (reglas de asociación) y dendrogramas (clusters jerárquicos, clusters orientados jerárquicamente). La teoría estadística y la aplicación de AEI están en permanente esparcimiento y desarrollo. La herramienta del *software* AEI se llama CHIC; en la versión de Windows 7.0, la CHIC versión multiplataforma libre se llama RCHIC. Se conoce que AEI cuenta con un grupo internacional de investigadores desde 2000. La CHIC contiene funciones comunes como: Árbol de Similitud, Gráfico Implicativo, Árbol de Cohesión y Reducción. Además, las opciones complementarias implementadas en CHIC son: la entropía que es útil para analizar una muestra de datos grande. Las variables suplementarias son: variables cualitativas como género, nivel educativo o

categoría económica. Este aporte es utilizado para conocer cuáles son los sujetos o clases de sujetos más responsables de las implicaciones calculadas, en tanto que la tipicidad muestra los sujetos típicos de la población para las implicaciones calculadas. (Pazmiño, García-Peñalvo, Coutrier, & Conde-González, 2015)

El autor de este trabajo coincide con los ensayistas antes citados, en que las matemáticas actualmente constituyen una herramienta que permite la comprensión y la modelización de gran número de fenómenos naturales, técnicos y sociales. Lo que consiente explorar, clasificar, analizar, generalizar, estimar, inferir, abstraer, argumentar o tomar decisiones. Esto posibilitará al docente dentro el ámbito educativo, analizar los resultados obtenidos para tomar las medidas y correctivos necesarios tendientes a fortalecer el nivel educativo.

3.1.3.1. Técnicas utilizadas en el Análisis Estadístico Implicativo

El análisis de las técnicas utilizadas por AEI, realizadas por (Couturier & Pazmiño, 2016), son:

- ✓ **Cohesión e implicación según Regis Gras:**

  Los autores (Gras et al., 2008), iniciaron el análisis estadístico implícito. El objetivo principal de este método fue precisar una forma de responder a la interrogación: "Si un objeto tiene una propiedad A, ¿tiene también una propiedad B". La contestación a esta pregunta raramente es positiva. No obstante, es dable notar que hay tendencias generales. AEI asume como objetivo revelar tales tendencias en un conjunto de propiedades.

- ✓ **Similaridad según Israel Lerman.**

  La propuesta por (Lerman, 1981), en base de la intensidad de implicación y la intensidad de similitud, un árbol de similitud, un árbol de jerarquía y un gráfico implicativo, se puede construir en base a CHIC. El árbol de similitud es el más conocido, también conocido como dendrograma. Se fundamenta en el índice de similitud. De manera similar, para construir un árbol de jerarquía se puede utilizar la intensidad de implicación. Se debe indicar que el índice de cohesión se define con la implicación en el árbol jerárquico.

  La intensidad de implicación además se consigue utilizar para definir un gráfico de implicación, que admite al usuario seleccionar las reglas de asociación y las variables que quiere. AEI establece las siguientes propiedades entre las variables que maneja, en oposición a los otros métodos de análisis de datos:

  - Las relaciones entre las variables son disimétricas.
  - Las medidas de asociación no son lineales y se basan en probabilidades.

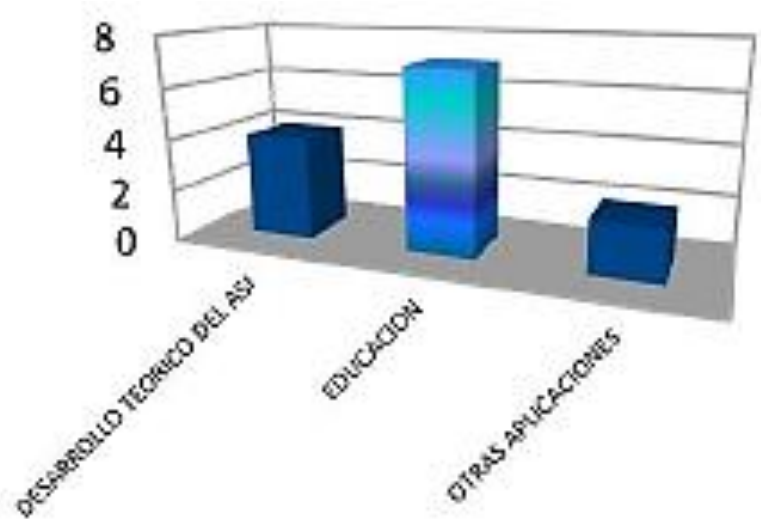
- El usuario puede utilizar representaciones gráficas que siguen la semántica de la relación.

Indiscutiblemente, AEI al trabajar con quasi-implicaciones, mantiene estrictamente relación con las matemáticas, probabilidades y estadística, pero no es un impedimento para el docente, al existir algoritmos desarrollados, que aplican estás técnicas, las cuales se utilizan en el diseño cuasi-experimental.

3.1.4. Análisis Estadístico Implicativo y Educación

El autor (Pazmiño, 2014) realiza un análisis profundo del AEI desde el punto de vista de sus aplicaciones en el campo educativo. Muestra un nuevo método estadístico multivariado aplicado a la didáctica en general, así como a la didáctica de la matemática en particular. En su estudio muestra datos referentes al número de artículos por área, es el caso de Caen-Francia donde se puede visualizar claramente que el AEI es más usado en el área educativa (ver figura 3).

Figura 3. AEI : Caen- Francia, número de artículos por área



Fuente: (Pazmiño, 2014)

De igual forma en la siguiente tabla muestra los resultados de los diferentes congresos internacionales AEI, por áreas.

Tabla 1. Total de publicaciones por congreso y por área

ÁREAS					
	Desarrollo Computacional del AEI	**Desarrollo Teórico del AEI**	**Educación**	**Otras Aplicaciones**	**TOTAL**
AEI1	0	4	7	2	13
AEI2	0	1	5	0	6
AEI3	3	6	17	1	27
AEI4	1	9	19	1	30
AEI5	1	3	13	10	27
AEI6	0	4	10	5	19
TOTAL	**5**	**27**	**71**	**19**	**122**

Fuente: (Pazmiño, 2014)

(Pazmiño, 2014) indica que históricamente el AEI tuvo su primera aplicación en el ámbito educativo, dentro del área de matemática. Se puede observar (ver tabla 1) que existen 71 artículos relacionados a Educación y duplica al segundo que es de desarrollo teórico con 27 artículos, esto indica que en área educativa se tiene mucha experiencia.

Consideremos ahora que el AEI, es aplicado en mayor proporción a la educación, lo cual afirma que las técnicas pueden ser aplicadas en el campo educativo, con resultados altamente confiable. A esto sumar que las técnicas manejan varios tipos y gran cantidad de datos, lo que facilita el uso de estas técnicas en beneficio del entorno educativo.

El análisis que efectúan (Montes & Ursini, 2014), permite conocer la extraordinaria utilidad de AEI en la educación, permite encontrar y resaltar algunas tendencias en las propiedades del objeto de estudio, para tomar decisiones. Actualmente se puede generalizar que los programas de estudios de toda la educación básica (preescolar, primaria y secundaria) se enfocan al desarrollo de competencias. Al concluir la educación básica se espera que el perfil de los estudiantes, se señala que: El perfil de egreso de los estudiantes plantea rasgos que deberán mostrar al término de la Educación Básica, como: garantía de desenvolverse satisfactoriamente en cualquier ámbito en el que decidan continuar su desarrollo. Al reconocer las actitudes de los estudiantes, se producen varios factores, los cuales se puede identificar a través de sus opiniones, creencias y emociones. Al realizar un análisis implicativo de un conjunto de variables en una población, se pueden obtener diferentes tipos de reglas o clases, lo que permite crear un árbol jerárquico (árbol donde se presentan las implicaciones y sus intensidades) y finalmente la gráfica implicativa, permite seleccionar las reglas y las variables que interesan. El

conjunto de cálculos y gráficas que se pueden obtener a través de CHIC, el cual es un programa informático que permite realizar un análisis implicativo.

Un ejemplo es el cual (S. Anastasiadou & Gagatsis, 2005), utilizan el AEI como método para revelar las actitudes de los estudiantes hacia las estadísticas, todo esto bajo el control psicológico, y utilizan el *software* CHIC para el análisis de datos por el método implicativo.

A partir de los referentes teóricos analizados sobre AEI, el autor de la presente investigación considera que el AEI permite revelar posibles relaciones de similaridad, implicación y cohesión entre variables, esto dentro del ámbito educativo es un gran aporte, al poder comparar variables que se den o se necesite analizar en el campo educativo. Con la finalidad obtener posibles soluciones o alternativas de mejora, tanto para el docente, estudiante, padres de familia e institución. Al ejemplificar se puede analizar las relaciones de similaridad, implicación y cohesión entre dos asignaturas dadas, actitudes y desempeño de hombres y mujeres de un nivel establecido, entre otros ejemplos. Al obtener los resultados se podrá realizar los correctivos necesarios para fortalecer o mejorar el desempeño en el ámbito educativo.

3.1.5. Cinco pasos del Análisis Académico según Campbell

El análisis académico es fundamental para tomar decisiones o guiar el comportamiento, el mismo que se logra al considerar los siguientes pasos: capturar, reportar, predecir, actuar y refinar. (Campbell John P. & G., 2007).

3.1.5.1. Capturar

Los datos obtenidos son la base fundamental del análisis. Es así, que el análisis académico se apoya en múltiples fuentes como un SIS, un CMS o sistemas financieros y en varios formatos como hojas de cálculo, informes del sistema financiero de la empresa o registros impresos. Así mismo, estos datos pueden obtenerse dentro o fuera de la institución. La importancia de las decisiones tomadas en datos depende de la calidad e integridad de los datos, al ser un reto la tarea de estas y otras variables en la colección, clasificación y racionalización de los datos.

3.1.5.2. Reportar

Al disponer de datos, se almacenan en un lugar o espacio común que, luego ayudados de herramientas de consulta, informes y análisis, realizar reuniones, examinar la información e identificar tendencias, patrones y excepciones en los datos. Por lo general se efectúan estadísticas descriptivas (media, desviación estándar), que se diferencian de los proyectos analíticos, los informes tradicionales (tablas de datos).

3.1.5.3. Predecir

Los datos recolectados y almacenados, se analizan al emplear procesos estadísticos. Las normas que administran los modelos se establecen en numerosos puntos de datos y algoritmos estadísticos con el objetivo de generar predicciones. Por ejemplo, un estudiante puede estar en riesgo de perder el curso si los datos indican que tiene una preparación académica baja o limitada en matemáticas y no ha asistido por varios días a clases.

3.1.5.4. Actuar

El objetivo de todo proyecto analítico, es consentir que una institución funcione y tome decisiones en base a predicciones y probabilidades. Las gestiones pueden ser tomadas en base a la información y al análisis efectuado. Por ejemplo, un proyecto analítico puede suministrar información a los estudiantes sobre el progreso educativo, comparándolas con sus compañeros, y probablemente sugerir aspectos para mejorar.

3.1.5.5. Refinar / clarificar

Los proyectos de análisis deben incluir en su propuesta un proceso de mejora. El monitoreo del impacto del proyecto se lo debe efectuar de manera continua, con la aplicación de modelos estadísticos actualizados de forma periódica. Las distinciones pueden incluir nuevos datos, mejoras del proceso o acciones diferentes. También, se pueden agregar datos adicionales de los obtenidos como otro componente del banco de datos, permite a las instituciones actualizar sus modelos y evaluar cómo afecta su presencia en el desempeño.

Al conocer los pasos del análisis académico (capturar, reportar, predecir, actuar y refinar/clarificar), el docente tendrá el punto de partida, para el análisis de datos masivos que se genera diariamente en las instituciones educativas. Además, es importante tener en cuenta que el paso más importante es la captura de los datos, por ser fundamentales en el análisis, para obtener resultados efectivos, los cuales me permitan tomar las medidas necesaria en referencia al estudio realizado.

3.1.6. Aproximación del AEI a métodos LA – métodos de análisis

El Learning Analytics ha sido y es una tecnología emergente, cada día aumenta la cantidad de investigación sobre el LA. La integración de nuevas herramientas, métodos y teorías es necesaria. Se parte del análisis de (Pazmiño-Maji, García-Peralvo, & Conde-González, 2016) , donde indican que AEI podría contribuir en los alumnos, la recopilación, el análisis, la comprensión y el aprendizaje de los estudiantes sobre la definición de LA y la fuente de datos; AEI podría contribuir en la captura e informe

de las etapas de LA; AEI podría contribuir mucho en la minería de reglas de asociación y la agrupación de métodos. Por otra parte, el AEI no contribuye en el contexto del alumno, la optimización, la elaboración de informes, el análisis, el proceso escolar de la definición de LA y la fuente de datos; AEI no contribuye en la demografía, las percepciones y el proceso escolar de las etapas de LA; AEI no contribuye a la minería de datos causales de los métodos LA. Desde esta perspectiva los autores realizan la aproximación del AEI con respecto al LA; según Baker e Inventado y según Papamitsiou y Economides.

Es importante analizar los diferentes métodos del AEI que se aproximan al LA, por ser la base de nuestro análisis, al implementar un diseño cuasi-experimental entre las técnicas similares del AEI y el LA.

3.1.6.1. Según Baker e Inventado

En la minería de datos educativos, Educational Data Mining (EDM), existe una amplia gama de métodos. A continuación se detallan los cuatro métodos planteados por (Ryan Shaun Baker & Inventado, 2014), que son: (a) Métodos de Predicción, (b) Descubrimiento de Estructuras, (c) Minería de Relaciones "Relationship Mining", y (d) Descubrimiento con Modelos.

3.1.6.1.1. Métodos de Predicción/ Pronostico

El Data Mining (DM) reúne una amplia gama de técnicas y algoritmos que permiten la extracción de conocimientos desde la base de datos para tomar decisiones de manera oportuna. DM se ha aplicado en diversos campos de estudio, al ser la educación un aspecto de investigación importante. La aplicación de DM en la educación se conoce como datamining educativo (EDM). El objetivo primordial de la EDM es investigar datos de instituciones educativas que apliquen diferentes técnicas como: predicción, agrupación, análisis de series de tiempo, clasificación, entre otros. (Moscoso-Zea & Luján-Mora, 2016)

En la predicción, la finalidad es desarrollar un modelo para inferir un solo aspecto de los datos, a lo que se conoce como variable predicha, que es análoga a las variables dependientes en el análisis estadístico tradicional, de alguna combinación con otros aspectos de los datos (variables predictoras, similares a variables independientes en el análisis estadístico tradicional).

En la Minería de Datos Educativos (MDE) o Educational Data Mining (EDM), los registradores y los regresores son los prototipos más comunes de modelos de predicción, y cada uno tiene varios subtipos.

En la minería de datos y la inteligencia artificial, los clasificadores y regresores tienen una amplia historia, que es explotado por la investigación EDM. El espacio de estimación en relación al conocimiento es de individualizada importancia dentro de EDM, y el desarrollo de esta área surge

mayoritariamente de las tradiciones de Modelado de Usuario, Inteligencia Artificial en Educación y Psicometría / Medición Educativa.

Los métodos de predicción se usan en muchas aplicaciones, habitualmente se utilizan para predecir cuál es el valor en contenidos en los que no se puede obtener directamente la información para esa construcción. Es particularmente útil para conducir en tiempo real la información, por ejemplo, para predecir el conocimiento que tiene un estudiante.

Se tiene conocimiento de varios estudios que aplican métodos de predicción entre los cuales podemos mencionar: los realizados por (Jones et al., 2001), donde se compara la exactitud relativa de los métodos de predicción del riesgo de enfermedad cardiovascular basados en ecuaciones derivadas del estudio cardíaco de Framingham en Birmingham. Así mismo, en el ámbito educativo se puede realizar el análisis de aprendizaje para la predicción del logro de los objetivos educativos, enfocados en la predicción del desempeño de los estudiantes. Predecir si el alumno alcanzará los resultados del sujeto en base a los resultados anteriores, permite a los profesores adaptar el diseño de aprendizaje del sujeto al proceso de enseñanza-aprendizaje (Fernández, Mucientes, B, & Lama, 2014).

En consecuencia, a lo investigado el autor afirma que uno de los métodos del AEI que se aproxima al LA, son los métodos de predicción, los cuales permiten al docente presagiar el comportamiento del estudiante, con el fin de retroalimentar y corregir el comportamiento del estudiante, además el docente, puede promover un aprendizaje personalizado y adaptativo, dependiendo de los resultados obtenidos a través de este método.

Los métodos de predicción se clasifican de acuerdo a (Fernández et al., 2014) en:

3.1.6.1.1.1. Clasificación

En los clasificadores de datos, la variable predicha puede ser binaria o categórica. También es conocido que algunos métodos populares de clasificación en los ámbitos educativos incluyen:

Árboles de decisión (buscar en LA)

(Nithyasri, Nandhini, & Chandra, 2011) define árbol de decisión como un popular clasificador de aprendizaje supervisado que no requiere ningún conocimiento o configuración de parámetros. Dado un dato de entrenamiento, se puede inducir un árbol de decisión. Desde un árbol de decisiones se puede crear fácilmente reglas sobre los datos. Al utilizar el árbol de decisiones, se puede predecir fácilmente la clasificación de los registros no vistos.

Árbol de decisión es una estructura de árbol jerárquico que se utiliza para clasificar las clases basadas en una serie de preguntas o reglas sobre los atributos de la clase. Los atributos de las clases pueden ser cualquier tipo de variables de valores binarios, nominales y cuantitativos, mientras que las

clases deben ser de tipo cualitativo categórico o binario u ordinal. Dado un dato de atributos junto con sus clases, un árbol de decisión produce una secuencia de reglas o series de preguntas que pueden ser usadas para reconocer la clase.

Los árboles de decisión son métodos de clasificación controlados que presentan una alta legibilidad puesto que los resultados del entrenamiento son un conjunto de sentencias del tipo "if-else" muy fáciles de interpretar. La idea primordial de este tipo de aprendizaje automatizado es interpretar el conjunto de entrenamiento como un conjunto de reglas que se deben aprender. (Pérez, 2014)

En definitiva, el actor concluye que una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y sigue el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

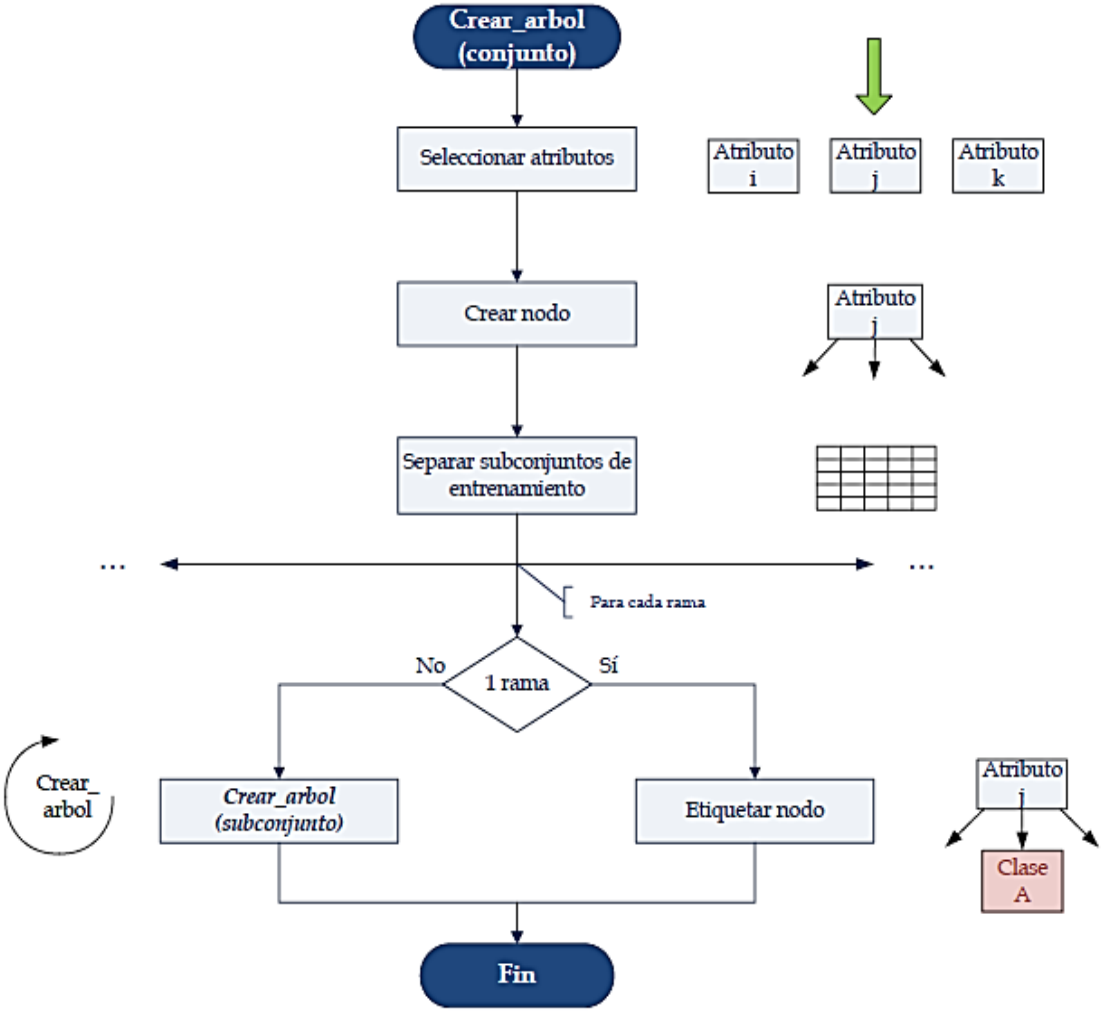
Entre las técnicas que frecuentemente se utilizan para la generación de árboles de decisión se tiene: algoritmos de inducción de árboles y poda de árboles de decisión, las mismas que fueron desarrollados por JR Quinlan.

a. Algoritmos de inducción de árboles.

Estos algoritmos se basan en los criterios de partición. La estrategia más utilizada por la mayoría de estos algoritmos es TDIDT (*Top-Down Induction of Decision Trees*), que no es más que establecer el árbol desde la raíz a las hojas.

Para realizar el árbol se parte del conjunto de entrenamiento completo del cual se selecciona el atributo que mejor divida a los datos. Mencionado atributo genera dos subconjuntos de entrenamiento sobre los que se aplica nuevamente el algoritmo. El proceso se repite hasta conseguir que cada subconjunto de entrenamiento esté compuesto por instancias de la misma clase. En la Figura 4 se observa el flujograma del algoritmo.

Figura 4. Flujograma de inducción de árboles de decisión.



Fuente: (Pérez, 2014)

b. Poda de árboles de decisión.

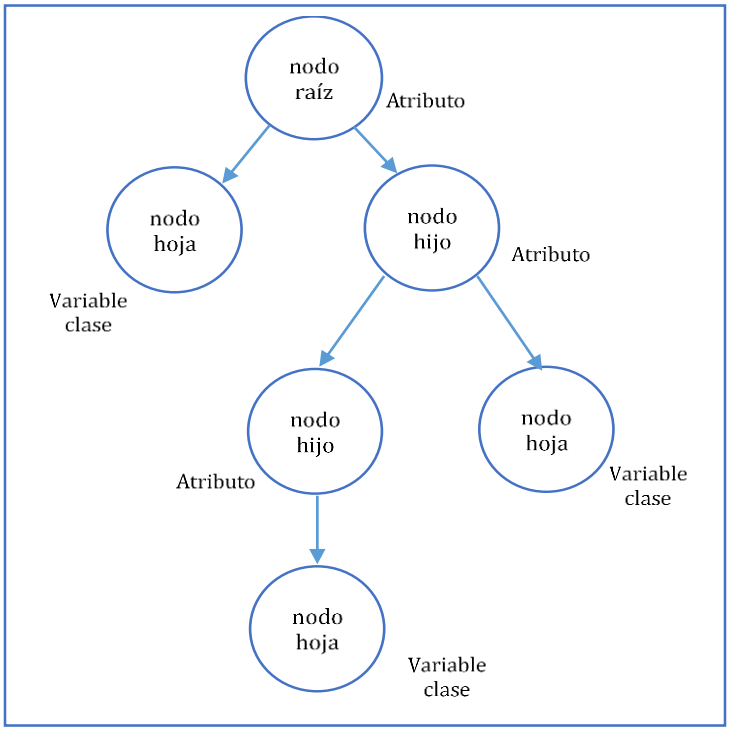
Un problema de consideración que presentan los árboles de decisión es el sobreajuste a los datos de entrenamiento, es decir, el modelo que genera presenta un comportamiento demasiado riguroso frente a los datos nuevos a clasificar. Este inconveniente se presenta en la clasificación de grandes cantidades de datos. Para realizar la clasificación del árbol, se extrae una parte de los datos y con ellos se genera un árbol de decisión. El sobreajuste provoca que se genere un árbol en el que el conjunto de información se clasifica correctamente, pero el resto de los datos no, puesto que se genera un árbol demasiado específico. Para solucionar este problema se debe eliminar las partes específicas del árbol y convertirlo es un modelo más generalista. El proceso mencionado se lo denomina poda y de acuerdo a la función y al momento en el que se aplica se tienen dos tipos de poda.

- ✓ **Pre-poda:** Se lo conoce al proceso que se realiza paralelamente a la generación del árbol. La ventaja primordial de este método, es la reducción del tiempo de cómputo del árbol, pero, se pierde la posibilidad de comparar los resultados de la poda con el árbol sin podar. El algoritmo CHAID (Kass, 1980) utiliza este tipo de poda.
- ✓ **Post-poda:** El corte se lo realiza una vez que se ha terminado de construir el árbol. A diferencia del proceso anterior, el árbol se construye en su totalidad por lo que el tiempo de cálculo es elevado, pero se dispone de la información completa del árbol sin podar. CART y C4.5 utilizan post-poda.

(Barrientos-Martínez et al., 2009), indica un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Es muy similar a los sistemas de predicción que se basan en reglas, que sirven para representar y categorizar una serie de condiciones que acontecen de forma sucesiva en la solución de un problema. Es el modelo de clasificación más utilizado y popular. El conocimiento obtenido durante el proceso de aprendizaje inductivo se puede representar mediante un árbol. Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo 2. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver (Ver Figura 5).

A partir de la descripción narrativa de un problema se construye el modelo, puesto que provee una visión gráfica de la toma de decisión, que detalla las variables que son evaluadas, las acciones que deben ser tomadas y el orden en el que la toma de decisión será efectuada. Cada vez que se ejecuta este tipo de modelo, sólo un camino será seguido, el cual depende del valor actual de la variable evaluada. Los valores que merecen tomar las variables para este tipo de modelos pueden ser discretos o continuos.

Figura 5. Árbol de decisión.



Fuente: (Barrientos-Martínez et al., 2009)

La utilización de árboles de decisión se lo ejemplifica en la investigación de (Martínez et al., 2009) , los mismos que son un modelo de clasificación utilizado en la inteligencia artificial, cuyo principal objetivo es el aporte visual a la toma de decisiones. Para probar el rendimiento en el proceso de clasificación de los árboles de decisión, se puede utilizar dos bases de datos que contienen información médica de pacientes reales. Esta información corresponde a la sintomatología que un médico especialista considera para el diagnóstico de cáncer de seno. Una de las bases de datos contiene 692 casos recopilados de las observaciones de un solo médico y la otra, contiene 322 casos recopilados de la observación de 19 especialistas. En general, se busca determinar la pertinencia de los árboles de decisión, es decir, si pueden ser una herramienta de apoyo para el diagnóstico médico. En base a los planteamientos se concluye que, a partir de un conjunto de datos aportados por un especialista en una disciplina, es posible tener en los árboles de decisión una herramienta de apoyo y ayuda confiable para el diagnóstico médico, al destacar que lo más importante es contar con un conjunto de datos consistente y confiable, ya que este tipo de herramientas están sometidas al conocimiento del experto que aportará la información. Por lo tanto, es necesario efectuar pruebas en otras especialidades médicas para encontrar el conjunto óptimo para la cimentación de este tipo de herramientas.

De acuerdo a lo investigado, los árboles de decisión son mapas de posibles resultados, de una serie de decisiones relacionadas, que permite al docente o la una institución educativa que lo aplique,

comparar posibles acciones entre sí, básicamente toma decisiones para la resolución de un problema, al buscar la mejor alternativa. Sin bien, puede ayudar en la toma de decisiones en el ámbito educativo, no es una herramienta útil cuando se tiene varias opciones, por lo que es una técnica poco utilizada en el ámbito educativo, por no manejar gran cantidad de datos, de acuerdo a la investigación realizada. Está técnica lleva tiempo, por trabaja con observaciones y construcciones lógicas, si se tiene en cuenta está investigación busca la técnica más óptima en tiempo y uso de memoria, por lo que no es útil, al llevarle tiempo al docente en la obtención de resultados.

Bosques al azar o aleatorios

Los estudios realizados por (Gislason, Benediktsson, & Sveinsson, 2006) se considera los bosques al azar para la clasificación de teledetección multisource y datos geográficos. En los últimos años se han propuesto varios métodos de clasificación de conjuntos. El clasificador Random Forest utiliza bagging, o agregación de bootstrap, para formar un conjunto de clasificación y regresión de árboles (classification and regression tree CART) como clasificadores. Además, busca sólo un subconjunto aleatorio de las variables para una división en cada nodo CART, con el fin de minimizar la correlación entre los clasificadores en el conjunto. Este método no es sensible al ruido ni al sobre entrenamiento, ya que el remuestreo no se basa en la ponderación. También, es computacionalmente mucho más ligero que los métodos basados en impulsar y algo más ligero que el empaquetamiento simple. En el trabajo se explora el uso del clasificador de bosque al azar para la clasificación de la cubierta terrestre. Se compara la precisión del clasificador de bosque aleatorio o azar con otros métodos de conjunto mejor conocidos sobre la teledetección multisources y los datos geográficos.

De acuerdo a (Vandamme, Meskens, & Superby, 2007) ejemplifican la utilización de las reglas de decisión, en el caso del fracaso académico entre los estudiantes universitarios de primer año ha alimentado durante mucho tiempo un gran número de debates. Es así que varios psicólogos educativos y estadísticos han tratado de entender para luego explicarlo e incluso preverlo. Esta investigación tiene como objetivo clasificar, al inicio del año académico a los estudiantes en tres grupos: los estudiantes de "bajo riesgo", que tienen una alta probabilidad de éxito; Los estudiantes de "riesgo medio", que pueden triunfar gracias a las medidas adoptadas por la universidad; y los estudiantes de "alto riesgo", que tienen una alta probabilidad de fracasar (o abandonar). Esta investigación describe la metodología y aporta las variables más significativas que se relacionan con el éxito académico entre todas las preguntas planteadas a 533 estudiantes universitarios de primer año durante el mes de noviembre del año académico 2003/04. Finalmente, se presentan los resultados del análisis como: discriminantes,

redes neuronales, bosques aleatorios y árboles de decisión orientados a predecir el éxito académico de los estudiantes.

Sobre las bases de las ideas expuesta, el autor noto que los tiempos de ejecución de bosques al azar son bastantes rápidos, capaces de tratar datos desbalanceados y desaparecidos, pero con una gran debilidad cuando se utiliza para regresión, no puede predecir más allá de la escala de datos de entrenamiento. Una gran ventaja que presenta está técnica es que maneja gran cantidad de datos, lo que es importante al momento de majear datos referentes al campo educativo.

Reglas de decisión

(Apté & Weiss, 1997) , describe el uso de las reglas de inducción en aplicaciones de minería de datos. De los métodos de clasificación y regresión que se han desarrollado en los campos del reconocimiento de patrones, estadísticas y aprendizaje de máquinas, que son de particular interés para la minería de datos, y utilizan representaciones simbólicas e interpretables. Las soluciones simbólicas pueden proporcionar un alto grado de comprensión de los límites de decisión que existen en los datos y de la lógica subyacente a ellos. Este aspecto hace que estas técnicas de minería predictiva sean particularmente atractivas en aplicaciones de minería de datos comerciales e industriales. Se presenta una sinopsis de algunas de las principales metodologías avanzadas de minería de árboles y reglas, así como algunos avances recientes.

En la investigación de (Galbraith, Stephenson, & Buckman, 1993) donde se realizó el estudio para confirmar las conclusiones de que las mujeres invocan una regla de decisión que es significativamente diferente de la de sus homólogos masculinos al hacer juicios de valor ético. Además, el estudio examina si la misma regla de decisión es utilizada por hombres y mujeres para todo tipo de situaciones éticas. Los resultados indican que los varones y las mujeres usan diferentes reglas de decisión al hacer evaluaciones éticas, aunque existen circunstancias en las que no hay diferencias significativas en las reglas de decisión que utilizan. Los resultados no sugieren que regla de decisión particular es utilizada por la mayoría de los hombres o mujeres en diferentes tipos de juicios éticos. Conociéndose que hay variedad en las reglas de decisión utilizadas por las mujeres que por los hombres.

De lo investigado, se puede notar, que se refiere exclusivamente a los procesos de decisión utilizados para evaluar la aceptabilidad o inaceptabilidad de ciertas acciones, además, la técnica de reglas de decisión es atractiva en aplicaciones de minería de datos comerciales e industriales, debe ser por ese motivo, al buscar investigaciones en el campo educativo, no se encontró, por ser más aplicado a otros campos.

Regresión escalonada

La regresión escalonada es una herramienta automatizada que se utiliza en las etapas exploratorias de la construcción de un modelo para identificar un subconjunto de predictores útil. El proceso añade de manera metódica la variable más significativa o elimina la menos significativa en cada paso que se ejecuta. (International Sales and Support, 2017)

Así, una compañía de asesoría en comercialización de viviendas recoge datos sobre la venta de viviendas del año anterior con la finalidad de pronosticar los futuros precios de venta. Al contar con más de 100 variables predictoras, encontrar modelos más significativos podría ser una tarea que requiere mucho tiempo. La función Regresión escalonada de Minitab genera automáticamente los modelos más significativos junto con R2, R2 ajustado, R2 pronosticado, S y Cp de Mallows, permite obtener un acertado punto de partida. Los procedimientos generales para la regresión escalonada son:

- La regresión escalonada estándar agrega y elimina predictores, según el requerimiento, en cada etapa o proceso. Minitab se detiene cuando todas las variables que no están en el modelo tienen valores p mayores que el valor específico alfa ingresado y cuando todas las variables del modelo tienen valores p menores o iguales al valor alfa especificado a eliminar.
- La selección hacia adelante inicia sin predictores en el modelo y Minitab agrega la variable más significativa para cada etapa. Minitab se detiene cuando todos los valores de las variables no están en el modelo y p es mayor que el valor alfa a entrar especificado.
- La selección hacia atrás inicia con todos los predictores en el modelo y Minitab elimina la variable menos significativa para cada etapa. Minitab se detiene cuando las variables del modelo tienen valores p menores o iguales al valor alfa a eliminar que se ha especificado.

De acuerdo a la investigación de (Henderson & Denison, 1989) se conoce que la regresión paso a paso se construye con un modelo de regresión a partir de un conjunto de variables predictoras postulantes, al ingresar y eliminar predictores - en forma escalonada - hasta que no haya justificación para ingresar o quitar más. El objetivo es terminar con un modelo de regresión razonable y útil. Hay una manera tangible de terminar con un modelo seguro que no se especifica – en caso de que el predictor candidato no incluya todas las variables que efectivamente predicen la respuesta. Esto forma una regla fundamental del procedimiento de regresión escalonada - la lista de variables predictoras candidatas debe incluir todas las variables que realmente predicen la respuesta; de lo contrario, probablemente no puede terminar en un modelo de regresión que está subespecificado, al poder ser tanto timador.

Al realizar la investigación, se pudo ver que la regresión escalonada, es una técnica que utiliza las etapas exploratorias de la construcción de modelos, elimina y agrega términos al modelo. Está técnica

al trabajar con estadística, es compleja, en especial para el docente que no cuente con estos conocimientos, lo cual hace difícil aplicarlo en el campo educativo, además se observó que no existen estudios o aplicaciones relacionadas al campo educativo. Esto se puede deber a que existe problemas de predicción cuando existen variables correlacionadas y al ajustar varios modelos, podrán los resultados no ser los óptimos al estudio realizado.

Regresión logística

La regresión logística facilita un método para modelar una variable de respuesta binaria, que toma valores de 1 y 0. Por ejemplo, se puede investigar cómo la muerte (1) o la supervivencia (0) de los pacientes puede ser predicha por el nivel de uno o más factores de marcadores metabólicos. Para ilustrar, considere una muestra de 2000 pacientes cuyos niveles de un marcador metabólico se han medido. Los datos son agrupados en categorías según el nivel de marcador metabólico, y se da la proporción de muertes en cada categoría. Las proporciones de muertes son estimaciones de las probabilidades de muerte en cada categoría. Propone que la probabilidad de muerte aumenta con el nivel del marcador metabólico, sin embargo, los resultados muestran una relación no lineal y la probabilidad de muerte cambia muy poco en los extremos alto o bajo del nivel del marcador. Este patrón es típico porque las proporciones no pueden estar fuera del rango de 0 a 1. La relación se puede describir como seguir una curva en forma de "S". (Bewick, Cheek, & Ball, 2005).

Los autores (Ohlmacher & Davis, 2003) efectuaron estudios que utilizan regresión logística, al considerar los deslizamientos en el terreno montañoso a lo largo de los ríos Kansas y Missouri. En el noreste de Kansas causaron daños a la propiedad durante la última década por millones de dólares. Para efectuar este problema, se ha utilizado el método estadístico denominado regresión logística múltiple para crear un mapa de amenaza de desprendimiento de tierra para Atchison, Kansas y áreas circundantes. Los datos incluyeron geología digitalizada, pendientes y deslizamientos de tierra, manipulados que utilizan ArcView GIS. La regresión logística relaciona las variables predictoras con la ocurrencia o no ocurrencia de deslizamientos de tierra dentro de los espacios geográficos mencionados y utiliza la relación para producir un mapa que muestra la probabilidad de deslizamientos futuros, dadas las pendientes locales y las unidades geológicas. Los resultados indicaron que la pendiente es la variable más importante para estimar el riesgo de deslizamiento en el área de estudio. Las unidades geológicas que consistían principalmente de lutitas, siltstone y arenisca eran las más susceptibles a los deslizamientos de tierra. Se consideró el tipo de suelo y las elevaciones que presentan, al descartar el análisis final, porque las variables no aumentaron significativamente el poder predictivo de la regresión logística. Los tipos de suelo estaban altamente aglutinados con las

unidades geológicas, y no existían relaciones significativas entre los deslizamientos de tierra y el aspecto de la pendiente.

En EDM, los clasificadores son típicamente aprobados mediante la validación cruzada, manteniéndose el conjunto de datos repetidamente y sistemáticamente, ser útil para probar la bondad del modelo.

La validación cruzada debe realizarse en múltiples niveles, en base al tipo de generalización deseada. Por ejemplo, es estándar el EDM para que los investigadores validen el nivel del estudiante y confirmar que el modelo trabaje para los nuevos docentes, para que los investigadores también validen en términos de poblaciones o de contenido de aprendizaje. Hay que considerar que la regresión por etapas y la regresión logística, a pesar de sus nombres, son clasificadoras en lugar de regresores. Algunas métricas comunes usadas para clasificadores incluyen A '/ AUC, kappa, precisión, y recuerdo. La exactitud, es a menudo popular en otros campos, pero no es sensible a las tasas de base y sólo debe utilizarse si forman parte de las tasas de base.

El autor de esta investigación, concluye que la regresión logística, es utilizada para predecir el resultado de una variable, en función de las variables independiente. Está técnica es útil cuando se necesita predecir la ausencia o presencia de una característica o resultado.

3.1.6.1.1.2. Regresión

La variable pronosticada es una variable continua en la regresión. El proceso más popular dentro de EDM es la regresión lineal, con árboles de regresión también muy populares.

Se puede identificar que un modelo producido por medio de este método es matemáticamente el mismo que la regresión lineal utilizado en las pruebas de significancia estadística, pero el método para seleccionar y validar el modelo en el uso de EDM de regresión lineal es totalmente diferente que en las pruebas de significación estadística. Son poco comunes en EDM los regresores como las redes neuronales y las máquinas de vectores de apoyo, que son prominentes en otros dominios de minería de datos. Se cree que esto se debe a que los altos grados de ruido y los múltiples factores explicativos en los dominios educativos a menudo conducen a que los algoritmos más conservadores sean más exitosos. Los regresores pueden ser validados al utilizar las mismas técnicas generales que en los clasificadores, a menudo se utiliza las métricas de correlación lineal o de error medio cuadrático - root mean squared error (RMSE).

En su investigación (Theobald & Freeman, 2014) ejemplifican el uso de la regresión lineal, indican que aunque los investigadores de la ciencia, la tecnología, la ingeniería y la educación de matemáticas utilizan actualmente varios métodos para analizar las ganancias de aprendizaje de los datos previos y

posteriores, los enfoques más comúnmente utilizados tienen deficiencias significativas. El principal de ellos es la incapacidad de distinguir si las diferencias en las ganancias de aprendizaje se deben al efecto de una intervención instructiva o las diferencias en las características del estudiante cuando los estudiantes no pueden ser asignados a grupos de control y tratamiento al azar. Al usar los resultados de pre y post-test de un curso introductorio de biología, ilustramos cómo los métodos actualmente en uso amplio pueden conducir a conclusiones erróneas y cómo la regresión lineal múltiple ofrece un marco efectivo para distinguir el impacto de una intervención instructiva del impacto de las características del estudiante En los resultados de la prueba. En general, recomendamos que los investigadores usen siempre modelos de regresión a nivel de estudiantes que controlan las posibles diferencias en la capacidad y preparación de los estudiantes para estimar el efecto de cualquier intervención instruccional no aleatoria sobre el desempeño estudiantil.

De estas evidencias, el autor concluye, que la técnica de regresión es poco común, al momento de utilizar en minería de datos educativos, sin embargo, son aplicable, al utiliza un modelo matemático para relacionar las variables dependientes entre variables dependientes. Es la técnica más utilizada para predecir varios rangos de fenómenos, como económicos hasta aspectos de comportamiento humano. Si bien es aplicable en la parte educativa, no es menos cierto por estos antecedentes no será la mejor técnica, por necesitar por trabajar estrictamente con relación entre variables, pero no se descarta si se logra identificar como una técnica similar entre el AEI y el LA.

3.1.6.1.1.3. Estimación del conocimiento latente

La estimación de conocimiento latente es importante en EDM por ser un caso especial e importante en la clasificación. De acuerdo a (Baker & Inventado, 2014), se indica que, en la estimación de conocimiento latente, el conocimiento de habilidades de un estudiante y conceptos específicos se evalúa de acuerdo a patrones de corrección en esas habilidades (y ocasionalmente otra información). El término "latente" hace referencia a la idea de que el conocimiento no es directamente insumable, debiéndose deducir del desempeño del estudiante. Inferir el conocimiento del estudiante es útil para varios objetivos, puede ser un aporte significativo a otros análisis, puede ser útil para decidir cuándo avanzar en el currículo o intervenir de otras maneras, es decir, es una información muy útil para los docentes.

Los modelos que se utilizan para valorar el conocimiento latente en el aprendizaje en línea generalmente difieren de los modelos psicométricos aplicados en las pruebas de papel o en las pruebas adaptativas de computadoras, como el conocimiento latente en el aprendizaje en línea es en sí dinámico. Los modelos utilizados para la valoración del conocimiento latente en EDM provienen de

dos fuentes: nuevas tomas de los enfoques psicométricos clásicos, e investigación sobre el modelaje del usuario / inteligencia artificial en la literatura educativa. Existe una variedad de algoritmos para la estimación de conocimiento latente. El algoritmo clásico es Bayes Nets para estructuras de conocimiento complejas, o Bayesian Knowledge Tracing para casos en los que cada problema o paso del problema está asociado principalmente con una sola habilidad determinada del momento en el que se encuentra. Últimamente, se sugiere un enfoque basado en la regresión logística, la evaluación de los factores de desempeño, puede ser eficaz para los casos en que múltiples habilidades son relevantes para un problema o problema de paso al mismo tiempo. El trabajo de Pardos y colegas ha determinado pruebas de que la combinación de múltiples enfoques a través de la selección de conjuntos puede ser más eficaz para grandes conjuntos de datos que los modelos únicos existentes.

El autor concluye, de acuerdo a investigación que la estimación del conocimiento latente, es una herramienta utilizado exclusivamente en la minería de datos educativos, donde permite al docente inferir en el conocimiento del estudiante, puede ser un aporte significativo de análisis, dentro y fuera del ámbito educativo. Está técnica, dentro de la educación puede ser útil para decidir cuándo avanzar en el currículo o intervenir de otras maneras, es decir, está técnica proporciona una información muy útil para los docentes

3.1.6.1.2. Minería de relaciones (Relationship Mining)

Descubrir relaciones entre variables en un conjunto de datos con un gran número de variables, es lo que deducen fundamentalmente en la minería de relaciones (Baker & Inventado, 2014). Esto quiere decir, que las variables están más asociadas con una sola variable de interés particular, o se puede intentar descubrir qué las relaciones entre dos variables pueden ser más fuertes.

En general, hay cuatro tipos de relación en minería frecuentemente utilizadas en EDM: minería de reglas de asociación, minería secuencial de patrones, minería de correlación y minería de datos causales. La extracción de reglas de asociación proviene del campo de la minería de datos, al ser el análisis de "canasta de mercado" utilizado en la minería de datos empresariales. La extracción secuencial de patrones también proviene de la minería de datos, con algunas variantes emergentes de la comunidad bioinformática. La minería de correlación ha sido una práctica en las estadísticas desde hace algún tiempo (y los métodos de análisis post hoc se produjeron en parte para hacer este tipo de método más válido). La minería de datos causales también proviene de la intersección de estadísticas y minería de datos

(Kaiser & Bodendorf, 2009) en su estudio analizan las minerías de relación en los foros online. Internet contiene un gran número de foros en línea donde los consumidores intercambian opiniones

de productos. Es importante que las empresas conozcan cómo los consumidores juzgan sus productos y cómo estas opiniones se difunden por interacciones a través de foros en línea. Con este conocimiento es posible reconocer riesgos y oportunidades. Sin embargo, la investigación de opinión clásica es muy lenta y sólo es posible hasta cierto punto.

Considerando la investigación, el autor indica que relación en minería, tiene una gran aceptación y sus técnicas (minería de reglas de asociación, minería secuencial de patrones, minería de correlación y minería de datos causales), son frecuentemente utilizadas en EDM, es así que se ha encontrado un sin número de investigaciones y aplicaciones de estas técnicas, a continuación, se detalla cada una de estas técnicas:

Minería de reglas de asociación

En la minería de reglas de asociación, el objetivo es encontrar reglas if-then de la forma que, si se encuentra algún conjunto de valores de variable, otra variable tendrá generalmente un valor específico. Por ejemplo, se puede encontrar una regla del formulario: SI el estudiante está frustrado OR tiene una meta más fuerte de aprender que de rendimiento LUEGO el estudiante frecuentemente pide ayuda.

Las reglas descubiertas por la minería de reglas de asociación revelan co-ocurrencias comunes en datos que hubieran sido difíciles de descubrir manualmente. La minería de reglas de asociación se ha utilizado para una variedad de aplicaciones en EDM. Por ejemplo, (Ben-Naim, Bain, & Marcus, 2009) encontraron reglas de asociación dentro de los datos de los estudiantes de una clase de ingeniería, al representar patrones de desempeño exitoso de los estudiantes, y (Merceron & Yacef, 2005) estudiaron qué errores de los estudiantes tienden a ir juntos. Además, muestran el uso de algoritmos de minería de datos puede ayudar a descubrir conocimientos pedagógicamente relevante contenido en bases de datos que se obtienen de sistemas educativos en la *Web*. Estos hallazgos pueden usarse para ayudar a los maestros a manejar su clase, entender el aprendizaje de sus estudiantes, reflexionar sobre su enseñanza y apoyar la reflexión del alumno proporciona retroalimentación proactiva a los estudiantes.

Dicho de otra manera y luego de revisar varios artículos sobre minería de reglas de asociación, como autor de esta investigación concluyo, que está técnica es una de las más utilizadas dentro de la minería de datos educativos, esto se podrá corroborar al realizar el análisis para la obtención de las técnicas similares entre el AEI y el LA. Por otra parte, está técnica es importante para los docentes, al descubrir hechos que ocurren dentro de un conjunto de datos, en este caso dentro del ámbito educativo, lo cual permite obtener los resultados deseados y de esta forma tomar las decisiones necesarias, para poder mejorar o reestructurar la actividad en estudio.

Minería de patrones secuenciales

El objetivo en la extracción secuencial de patrones, es encontrar asociaciones temporales entre eventos. Se presentan dos paradigmas que hallan patrones secuenciales: la extracción clásica de patrones secuenciales (Srikant & Agrawal, 1996), que es un caso especial de minería de reglas de asociación y motivo por lo cual (Lin, Keogh, Lonardi, & Chiu, 2003), analizan las explosiones paralelas de interés en la transmisión de datos y la minería de datos de series de tiempo han tenido sorprendentemente poca intersección. Esto es a pesar del hecho de que los datos de series de tiempo son típicamente datos de flujo continuo. La principal razón de esta aparente paradoja es el hecho de que la gran mayoría del trabajo sobre datos de transmisión explícitamente asume que los datos son discretos, mientras que la gran mayoría de los datos de series de tiempo es real valorada. Muchos investigadores han considerado transformar series de tiempo real valoradas en representaciones simbólicas, las mismas que permitirían a los investigadores aprovechar la riqueza de estructuras de datos y algoritmos de las comunidades de procesamiento de texto y bioinformática, además de permitir que los antiguos problemas de "solo lote" sean abordados por la comunidad de streaming. Mientras que muchas representaciones simbólicas de series de tiempo se han introducido durante las décadas pasadas, todos sufren de tres defectos fatales. En primer lugar, la dimensionalidad de la representación simbólica es la misma que la de los datos originales, y prácticamente, todos los algoritmos de minería de datos varían poco con la dimensionalidad. En segundo lugar, aunque las medidas de distancia pueden definirse en los enfoques simbólicos, estas medidas de distancia tienen poca correlación con las medidas de distancia definidas en las series temporales originales. Finalmente, la mayoría de estos enfoques simbólicos requieren que uno tenga acceso a todos los datos, antes de crear la representación simbólica. Esta última característica explícitamente frustra los esfuerzos por utilizar las representaciones con algoritmos de transmisión. En este trabajo se introduce una nueva representación simbólica de series de tiempo. La representación es única, ya que permite la reducción de la dimensionalidad / numeración, y también permite definir las medidas de distancia sobre el enfoque simbólico que las medidas de distancia correspondientes a los límites inferiores se definen en la serie original. La última característica es particularmente excitante porque permite ejecutar ciertos algoritmos de minería de datos en la representación simbólica manipulada eficientemente, produce resultados idénticos a los algoritmos que operan sobre los datos originales. Por último, la representación permite que los datos de valor real se conviertan en una forma de streaming, con sólo un tiempo infinitesimal y espacio superior.

En base a lo investigado, se concluye que la minería de patrones secuenciales es el proceso donde se obtiene relaciones entre ocurrencias secuenciales, para encontrar un orden específico en el que ocurren los eventos. Además, se puede afirmar que está técnica es eficaz, eficiente y escalable(Guevara Fuente de la Vega & Beltran Castañon, 2014), la misma puede ser utilizada dentro del campo educativo, por manejar gran cantidad de datos, esto permite obtener los mejores resultados. Una gran ventaja de ser escalable, es al manejar grupos de datos muy grandes, está técnica tiene la propiedad de aumentar la capacidad de trabajo. Es importante indicar que en la parte experimental se desarrolla un análisis de su uso y su aproximación entre el AEI y el LA, de lo cual depende para su uso dentro del diseño cuasi-experimental.

Minería de Correlación

En la minería de correlación, el objetivo es encontrar correlaciones lineales positivas o negativas entre variables. Este objetivo no es nuevo; es un objetivo bien conocido dentro de las estadísticas, donde ha surgido una literatura sobre cómo utilizar el análisis post hoc y / o la dimensionalidad de técnicas de reducción para evitar encontrar relaciones falsas. La minería de correlación se ha utilizado para estudiar la relación entre las actitudes de los estudiantes y los comportamientos de búsqueda de ayuda y estudiar la relación entre el diseño de sistemas de tutoría inteligente y si los estudiantes juegan el sistema (Arroyo & Woolf, 2005)

A continuación, el autor de esta investigación, indica que la técnica de correlación se puede entender como el grado de relación que existe entre dos variables, está técnica puede ser utilizada para estudiar la relación de actitudes y comportamiento de los estudiantes, con el fin de modificar el diseño del plan de tutoría que el docente lleva a cabo durante un período escolar. Siempre será importante ver alternativas de solución a los problemas en el campo educativo y está técnica puede ayudar al obtener resultados que ayuden en la formación académica del estudiante, posteriormente se analiza si es similar con las técnicas AEI.

Minería de Datos Causales

En la minería de datos causales, el objetivo es encontrar si un evento (o constructo observado), fue la causa de otro evento (o constructo observado) causal. (Spirtes, Glymour, & Scheines, 2000).

La minería de datos se distingue de la predicción en sus intentos de encontrar no sólo predictores sino relaciones causales reales, a través de los patrones de covarianza entre esas variables y otras variables en el conjunto de datos. La minería de datos causales en paquetes como TETRAD (Scheines, Spirtes, Glymour, Meek, & Richardson, 1998), indican que el proyecto TETRAD y el trabajo conexo en ciencias de la computación y estadísticas apuntan a aplicar esas normas al problema del uso de datos

y conocimientos básicos para hacer inferencias sobre la especificación de un modelo. Lo cual se ha utilizado en EDM para predecir qué factores llevarán a un estudiante a hacer mal en una clase (Fancsali, 2012) , presenta un método para buscar simultáneamente variables de nivel de estudiante construidas a partir de datos de registro de Tutor Cognitivo y modelos causales gráficos. Busca explicaciones causales de la conducta en Tutores Cognitivos, incluye un "juego del sistema" y comportamiento fuera de la tarea, selecciona las variables por su contribución a la estructura causal y al aprendizaje de la fuerza.

De aquí que el autor concluye, que el objetivo de la minería de datos causales es encontrar asociaciones temporales entre eventos. Un ejemplo práctico es al saber cuál es el interés por el aprendizaje de un estudiante, al aplicar la técnica está busca los eventos causales para que se dé el interés del estudiante. Está técnica es aplicada dentro del campo educativo, pero se debe tener en cuenta que trabaja con grafos, para lo cual se necesita saber interpretarlos, esto cuando se maneja gran cantidad de información.

3.1.6.1.3. Descubrimiento de estructuras

Los autores (Baker & Inventado, 2014), analizan a los algoritmos de descubrimiento de estructuras e intentan encontrar la estructura en los datos, es decir, tener una idea a priori de lo que se debe encontrar. De esta manera, este tipo de minería de datos contrasta fuertemente con los modelos de predicción, donde deben ser aplicados a un subconjunto de los datos, el cual se debe desarrollar antes de que ocurra un suceso. Los algoritmos de descubrimiento de estructuras comunes en los datos educativos incluyen algoritmos de agrupación, análisis de factores y descubrimiento de estructuras de dominio. El agrupamiento y el análisis de factores se han utilizado desde los primeros días del campo de la estadística, y fueron refinados y explorados más a fondo por las comunidades de minería de datos y aprendizaje de máquinas. El descubrimiento de la estructura del dominio surgió del campo de la psicometría / medición educativa.

A continuación, se detalla cada uno de las técnicas del descubrimiento de estructuras:

3.1.6.1.3.1. Clustering

En la agrupación (clustering), el objetivo es encontrar puntos de datos que se agrupan naturalmente, divide el conjunto de datos completo en un conjunto de grupos (Kaufman & Rousseeuw, 2009). El agrupamiento es particularmente útil en los casos en que las categorías más comunes dentro del conjunto de datos no se conocen de antemano. Si un conjunto de clústeres es óptimo, cada punto de datos de un clúster será en general más similar a los otros puntos de datos que los de otros clústeres. Los clusters se pueden crear en varios tamaños de poca diferencia. Por ejemplo, las escuelas podrían

agruparse (para investigar las similitudes y diferencias entre las escuelas), los estudiantes podrían agruparse (para investigar similitudes y diferencias entre los estudiantes), o las acciones de los estudiantes podrían agruparse (para investigar patrones de comportamiento) (Amershi & Conati, 2009) , además, presenta un marco de modelado de usuarios basado en datos que utiliza la clasificación no supervisada y supervisada para construir modelos de estudiantes para entornos exploratorios de aprendizaje. Se aplica el marco para construir modelos de estudiantes para dos entornos de aprendizaje diferentes y usar dos fuentes de datos diferentes (interfaz de sesión y datos de rastreo ocular). A pesar de las limitaciones debido al tamaño del conjunto de datos, se ofrece evidencia inicial de que el marco puede identificar automáticamente comportamientos significativos de interacción del estudiante y puede ser usado para construir modelos de usuario para la clasificación en línea de nuevos comportamientos estudiantiles en línea. También se muestra la transferibilidad del marco entre aplicaciones y tipos de datos. Por otra parte, los algoritmos de agrupación se dividen típicamente en dos categorías: enfoques jerárquicos como el agrupamiento aglomerado jerárquico - hierarchical agglomerative clustering (HAC) y enfoques no jerárquicos como k-Means, modelado de mezcla gaussiana (a veces denominado en EM-based clustering) y agrupación espectral. La diferencia clave es que los enfoques jerárquicos suponen que los clústeres se agrupan, mientras que los enfoques no jerárquicos asumen que los clústeres están separados entre sí.

Sobre las bases de las ideas expuestas, el autor indica que clustering es conocido como agrupamiento en el idioma español, es una técnica de minería de datos, la cual es aplicada en varios campos de exploración de datos, entre ellos se encuentra el campo educativo. Se basa en similitudes y diferencias entre variables o atributos. Está técnica es muy utilizada, por lo cual más adelante se realiza un analiza todos los algoritmos que aplica está técnica, como son k-Means, Cobweb entre otros.

3.1.6.1.3.2. Factor Análisis

Para el análisis de factores, el objetivo es encontrar variables que evidentemente se agrupan, divide el conjunto de variables (en oposición a los puntos de datos) en un conjunto de factores latentes (no directamente observables) (Kline, 2014). El análisis factorial se utiliza frecuentemente en la psicometría para validar o determinar escalas. En EDM, el análisis de factores se utiliza para la reducción de la dimensionalidad (por ejemplo, la reducción del número de variables), incluye en el preprocesamiento para reducir el potencial de superposición y para determinar meta-características.

Un ejemplo de su uso en EDM es el trabajo para determinar qué características de los sistemas de tutoría inteligente se agrupan… (véase Baker et al., 2009).

El análisis de factores incluye algoritmos tales como análisis de componentes principales y análisis de componentes principales de la familia exponencial.

De las evidencias anteriores, el autor indica que la técnica de factor de análisis parte de un conjunto de variables, las cuales presentan interrelaciones o características comunes las cuales son transcendentales, las mismas que no son observables de forma directa. Si bien son aplicadas en EDM, son usadas más en marketing, gestión de productos, investigaciones operativas, entre otras.

3.1.6.1.3.3. Descubrimiento de la Estructura del Dominio

Encontrar qué elementos se asignan a las habilidades específicas entre los estudiantes es el objetivo principal del descubrimiento de la estructura del dominio. El enfoque Q-Matrix para hacerlo es bien conocido en la psicometría (Tatsuoka, 1995). Se ha aplicado un trabajo considerable a este problema en EDM, para datos de prueba (T. Barnes, 2005) , es uno de los pocos investigadores talentosos que ha creado una herramientas para la instrucción asistida por ordenador y los sistemas de tutoría inteligente, la cual es una herramienta de alta calidad, eficaz, escalable, pero individualizadas para el aprendizaje a bajo costo. Muchas herramientas de aprendizaje crean modelos complejos de comportamiento estudiantil que requieren un tiempo extenso por parte de expertos en la materia, así como investigadores de ciencias cognitivas, para crear una ayuda eficaz.

Se puede utilizar una serie de algoritmos para el descubrimiento de la estructura del dominio, a partir de algoritmos puramente automatizados (Tiffany Barnes, Bitzer, & Vouk, 2005) , los autores dan a conocer que el método q-matrix, es útil para la extracción de datos y el descubrimiento del conocimiento, se compara con el análisis de factores y análisis de conglomerados en el análisis de catorce conjuntos de datos experimentales. Este método crea un modelo basado en matriz que extrae las relaciones latentes entre las variables binarias observadas. Los resultados muestran que el método q-matrix ofrece varias ventajas sobre el análisis factorial y el análisis de conglomerados para el descubrimiento del conocimiento. El método q-matrix puede realizar clustering totalmente sin supervisión, donde el número de clusters no se conoce de antemano. También produce mejores tasas de error que el análisis de factores, y es comparable en error al análisis de conglomerados. El método q-matrix también permite la interpretación automática de los conjuntos de datos. Estos resultados sugieren que el método q-matrix puede ser una herramienta importante en el descubrimiento automatizado de conocimiento.

Como complemento, el autor indica que el descubrimiento de la estructura del dominio, se utiliza en la psicometría, al realizar una medición de las funciones mentales de la persona, por lo que dentro

del análisis de datos educativo puede ayudar en los problemas de bajo rendimiento presente en la actualidad.

3.1.6.1.4. Descubrimiento con modelos

En el descubrimiento con modelos, es un modelo de un fenómeno se desarrolla a través de la predicción, agrupación o, en algunos casos, ingeniería del conocimiento (dentro de la ingeniería del conocimiento, el modelo desarrolla razonamiento humano en lugar de métodos automatizados).

Este modelo se utiliza como un componente en un segundo análisis o modelo, por ejemplo, en la predicción o la minería de la relación. El descubrimiento con modelos no es común en la minería de datos en general, pero se ve en alguna forma en muchos dominios de ciencias computacionales.

En el caso de EDM, es común usarlo cuando las predicciones de un modelo inicial (que representan variables predichas en el modelo original) se convierten en variables predictoras en un nuevo modelo de predicción. Por ejemplo, los modelos de predicción del aprendizaje vigoroso de los estudiantes generalmente han dependido de modelos de comportamientos meta-cognitivos estudiantiles (Baker et al., 2011), que a su vez dependieron de las evaluaciones del conocimiento latente de los estudiantes. Estas evaluaciones del conocimiento del estudiante a su vez han dependido de modelos de estructura de dominio.

A menudo, el descubrimiento con modelos aprovecha la generalización de un modelo de predicción a través de contextos. Por ejemplo, (Baker & Gowda, 2010), estudiaron cómo los comportamientos de los estudiantes asociados con la separación difieren entre los diferentes escenarios escolares. En este sentido, se investiga la variación en la frecuencia del comportamiento fuera de las tareas, el juego del sistema y el descuido en una escuela urbana, una escuela rural y una escuela suburbana en los Estados Unidos de América. Este análisis se lleva a cabo mediante la aplicación de detectores automatizados de estos comportamientos a los datos de los estudiantes que utilizan el mismo método cognitivo.

Las evidencias anteriores, permite al autor indicar que el descubrimiento de modelos, en el caso de EDM, es común usarlo cuando las predicciones de un modelo inicial, que representan variables predichas en el modelo original, el cual se utiliza como un componente clave en un nuevo análisis de EDM.

3.1.6.2. Según Papamitsiou y Economides

Además de los métodos sugeridos por Baker e inventado, éstos se complementan con los propuestos por Papamitsiou y Economides.

El objetivo de Papamitsiou y Economides (Z. Papamitsiou & Economides, 2014), es dar a conocer y comprender los conocimientos actuales sobre Learning Analytics (LA) y Educational Data Mining

(EDM) y su impacto en el aprendizaje adaptativo. Constituye una visión general de la evidencia empírica que respalda los objetivos clave de la posible adopción de LA / EDM en la planificación estratégica genérica de la educación. Los autores analizaron las preguntas de investigación, metodología y resultados de estos trabajos publicados y consecuentemente los clasificaron. Se utilizaron métodos no estadísticos para evaluar e interpretar los hallazgos de los estudios escogidos. Los resultados han destacado cuatro direcciones principales distintas de la investigación empírica LA / EDM, estas son: Análisis de redes sociales, minería de texto, visualización y estadísticas.

(Z. K. Papamitsiou, Terzis, & Economides, 2014), indican que predecir el desempeño del alumno es una tarea difícil y complicada para las instituciones, los instructores y los estudiantes. Las predicciones precisas del desempeño podrían conducir a mejores resultados de aprendizaje y un mayor logro de metas. A continuación, se exploran las capacidades predictivas del tiempo dedicado al estudiante a contestar (in-) correctamente cada pregunta de un cuestionario de evaluación de opción múltiple, junto con la puntuación final del mismo, en el contexto de la prueba computarizada. También se explora la correlación entre el factor tiempo-gastado (como se define aquí) y la meta-expectativa. Se presenta un estudio de caso e investigar el valor de la utilización de este parámetro como un factor de análisis de aprendizaje para mejorar la predicción de rendimiento durante las pruebas basadas en computadora. Los resultados iniciales son alentadores e indican que la dimensión temporal de la analítica del aprendizaje debe ser explorada más a fondo.

A continuación, se detalla cada uno de los métodos no estadísticos para evaluar e interpretar los hallazgos de los estudios escogidos, estos son: análisis de redes sociales, minería de textos, visualización y estadística

3.1.6.2.1. Análisis de Redes Sociales

El uso de análisis de contenido generados por computadora surgió para explorar lo que la gente discute (Fourniern, Kop, & Sitlia, 2011). Además, las entrevistas con un énfasis en la recuperación de eventos críticos se centran en las experiencias de los participantes para averiguar "por qué se habla como lo hacen" y mediante el análisis de redes sociales averiguar la dinámica de la red para ver "quién se comunica con quién". Esto parece una opción viable de los métodos de investigación.

El Análisis de redes sociales sería una forma de aprendizaje analítico, y un método cuantitativo, que podría aclarar cuáles son los nodos centrales de la red, en otras palabras, las personas en la red desempeñan funciones vitales de la conexión con la otra no conectada. También podría proporcionar información sobre la importancia de los "conectores" a otras redes, lo que sería importante para averiguar quiénes son los innovadores de la red, es decir, los que vincularían los flujos de información

vital. Argumenta que el uso de métodos cualitativos adicionales y que la etnografía virtual sería el método más apropiado de investigación cualitativa en redes de aprendizaje. El investigador se interesa en los procesos que tienen lugar, las perspectivas y entendimientos de las personas en el escenario, los detalles, el contexto, la emoción y las redes de relaciones sociales que se unen entre sí, destaca que en un entorno tecnológicamente rico, como Internet, la tecnología en sí y los artefactos que produce deben ser tomados en consideración en la etnografía "en línea", ya que éstos forman parte de la investigación y pueden influir en las interacciones humanas investigadas.

A medida que se generan grandes cantidades de datos en el aprendizaje en red en un entorno abierto, las herramientas computacionales de análisis e interpretación tendrán que desempeñar un papel en la investigación. Algunos abogan por un enfoque de método mixto en la investigación educativa como "las teorías" que se tienen, y la formación que se ha recibido, afectan de manera crítica a los datos que se recopilan y las lentes que se elige al examinar tales datos. Sostienen que el uso de más de un método en la investigación aumentará su potencia según (Boyd & Crawford, 2012), ellos son científicos sociales que investigan "Big Data" destacan algunas otras preocupaciones metodológicas, especialmente al analizar Big Data recogido en redes en línea: 1) los datos más grandes no siempre son mejores que obtenidos en otras investigaciones como la fiabilidad que dependerá en gran medida de las estrategias de muestreo utilizadas; 2) es necesario tomar precauciones porque no todos los datos se crean por igual; 3) lo que la gente hace es de importancia limitada a menos que se pregunte a la gente por qué hicieron lo que hicieron; 4) argumenta que los investigadores cualitativos no son los únicos que interpretan los datos, que también los investigadores cuantitativos lo hacen; al disipar el mito de que "son investigadores cualitativos quienes están en el negocio de interpretar historias e investigadores cuantitativos que están en el negocio de producir hechos". La interpretación como parte del análisis es la más difícil de cualquier análisis de datos, grande o pequeño. Boyd (Boyd & Crawford, 2012) quisiera ver a expertos de la computadora que trabajen junto con los científicos sociales para evitar falacias en interpretaciones.

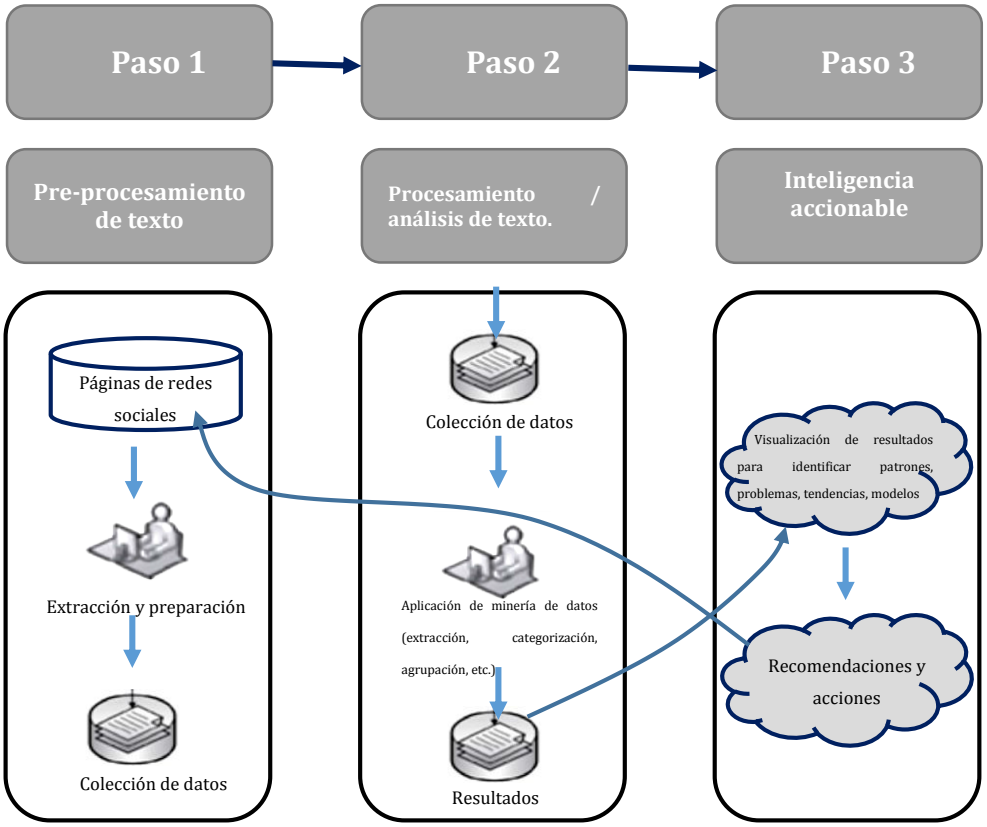
Los temas tratados, permite al autor consolidar está información, al indicar que el análisis de redes sociales es una técnica, que busca predecir el comportamiento de una red social, además de estudiar las relaciones de varias personas entre sí, mediante el protocolo de comunicaciones de la capa de enlace ARP, para lo cual existe variar aplicaciones Sprout Social, Coogle Analytics, entre otros. Esto es fundamental dentro del EDM, debido a que el docente puede monitorear las actividades realizadas por sus estudiantes, al trabajar en una plataforma como la de Moodle.

3.1.6.2.2. Minería de texto

Antes de examinar minería de texto, se debe tener en cuenta que surge de las redes sociales y el uso de plataformas educativas, las cuales se maneja publicaciones, mensajes de texto, entre otras. Si bien no utilizan la totalidad de docentes, estas herramientas (Edmodo, miaulas, Facebook, entre otras) son utilizadas, las mismas que son un apoyo en la parte académica, de igual forma al momento de evaluar al estudiante. Por todas estas razones es importante analizar la información que transita en la red, con el fin de proponer mejoras en su uso. Por lo tanto, una de las técnicas que pueden ayudar al momento de realizar un análisis de datos masivos en la red, es la minería de datos. Para una mejor comprensión se detalla a continuación algunas investigaciones.

Las redes sociales han sido adoptadas por muchas empresas. Cada vez más empresas utilizan las herramientas de medios sociales como Facebook y Twitter para ofrecer diversos servicios e interactuar con los clientes. Como resultado, una gran cantidad de contenido generado por el usuario está disponible gratuitamente en los sitios de redes sociales. Para aumentar la ventaja competitiva y evaluar efectivamente el entorno competitivo de las empresas, éstas necesitan monitorear y analizar no sólo el contenido generado por el cliente en sus propios sitios de redes sociales, sino también la información textual en los sitios de medios sociales de sus competidores. En un esfuerzo por ayudar a las empresas a comprender cómo realizar un análisis social de los medios de comunicación y transformar los datos de las redes sociales en conocimiento para los tomadores de decisiones y los vendedores electrónicos, este (He, Zha, & Li, 2013) describe un estudio de caso en profundidad que aplica la minería de texto para analizar contenido de texto no estructurado en Facebook y sitios de Twitter de las tres mayores cadenas de pizza: Pizza Hut, Domino's Pizza y Papa John's Pizza. Los resultados revelan el valor del análisis competitivo de las redes sociales y el poder de la minería de textos como una técnica eficaz para extraer valor comercial de la gran cantidad de datos de medios sociales disponibles. También se ofrecen recomendaciones para ayudar a las empresas a desarrollar su estrategia de análisis competitivo de medios sociales. (He et al., 2013) propone la siguiente metodología que se puede representar en la siguiente gráfica:

Figura 6. Proceso de minería de texto para contenido de medios sociales.



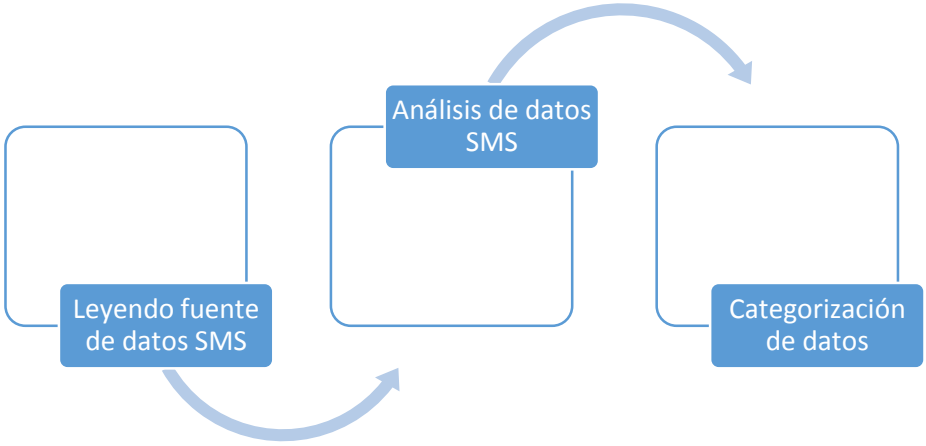
Fuente: (He et al., 2013)

La figura 6 numera los principales pasos para el proceso de minería de texto utilizado. Al seguir tres pasos (pre-procesamiento, aplicación de minería de texto y evaluación de los resultados de la minería y reconocimiento de información accionable), se puede identificar nuevos conocimientos, incluidos patrones, temas y datos de redes sociales recopilados. A menudo, la aplicación de la minería de texto a conjuntos de datos requiere una evaluación continua y un refinamiento para lograr los mejores resultados.

Por otra parte, (Leong, Lee, & Mak, 2012) explora la posible aplicación de la minería sentimental para analizar textos de servicio de mensajes cortos (SMS) en la evaluación de la enseñanza. La preparación de datos implica la lectura, análisis y categorización de los textos SMS. Se desarrollaron tres modelos: el modelo base, el modelo ''corregido'' que ajusta los errores de ortografía y el modelo ''sentimiento'' que se extiende el modelo ''corregido'' mediante la realización de la minería sentimiento. Un criterio de "interés" selecciona el modelo de "sentimiento" desde el cual se disciernen

los sentimientos de los estudiantes hacia la conferencia. También se identifican dos tipos de textos SMS incompletos y se determinan las implicaciones de su eliminación para el análisis.

Figura 7. Minería de SMS: preparación de datos.



Fuente: (Leong et al., 2012)

El proceso de preparación de datos de textos SMS se compone de 3 fases, como se muestra en la Figura 7. son:

1. **Leer el cuerpo del SMS o la colección de textos SMS.**
2. **Análisis de los textos SMS.**

   Part of speech (POS) tagging: Cada término se etiqueta con una etiqueta POS. En este análisis, el marcado POS es un paso intermedio que se lleva a cabo para identificar conceptos.

   Derivación: La inflexión se refiere al cambio de forma en las palabras para marcar el género, el número o el tiempo. Por ejemplo, la palabra raíz prevenir tiene las siguientes inflexiones: impide, evita y previene. Derivación (Stemming) es el proceso de asignar todas las variantes a su palabra raíz.

   Excluir lista: En este análisis, los pronombres, las partículas y las preposiciones (excepto de) no se extraen durante el análisis, ya que estos POS suelen ser redundantes en la minería de texto. Para los conceptos adicionales que son redundantes, se rellenan en la lista de exclusión.

   Tipo/Extracción de entidad: Para cada concepto extraído, también se identifica un tipo o entidad. Un tipo se define como un agrupamiento semántico de conceptos. Los tipos incluyen conceptos de nivel superior, palabras y calificadores positivos y negativos, calificadores contextuales, nombres, lugares, organizaciones, entre otros. Si el concepto no pertenece a ninguno de los tipos definidos, entonces se escribe como Desconocido. Los tipos ayudan a agrupar conceptos en análisis de enlaces de texto para ofrecer visualización de relaciones interesantes entre conceptos.

3. **Categorización del texto:** Con base en los conceptos definidos para cada categoría, cada texto SMS puede pertenecer a ninguna categoría, a una categoría o varias categorías.

En el proceso de preparación de datos, se observa que algunos de estos textos SMS son incompletos. En el conjunto de datos se pueden distinguir dos tipos de textos SMS incompletos. La incompletitud de tipo 1 implica un mensaje incompleto debido a las limitaciones en el número máximo de caracteres que se pueden almacenar para cada texto SMS en el sistema de retroalimentación en línea. Por otro lado, la incompletitud tipo 2 implica que el encuestado envíe una sola letra alfabética en lugar de un mensaje completo. Esto puede o no ser una calificación otorgada.

Sobre las bases de las ideas expuestas minería de textos, el autor indica, si bien es una técnica eficaz para extraer valor comercial de la gran cantidad de datos de medios sociales disponibles, también puede ser aplicada en la EDM, al permitir categorizar los textos SMS, si bien no es eficaz porque solo lee mensajes cortos (SMS) y su uso es restringido a la evaluación de la enseñanza. Si bien en la parte de resultados de esta investigación se realiza un análisis técnico se ve que es muy poco aplicable en el campo educativo, lo cual afirma al obtener resultados que se aprecia con este estudio.

3.1.6.2.3. Visualización

(Clow & Makriyannis, 2011) proponen iSpot analysed (iSpot analizado): aprendizaje participativo y de reputación. Presentan un análisis de la actividad en iSpot, el cual es un sitio *web* que apoya el aprendizaje participativo sobre la vida silvestre a través de redes sociales. Un sistema de reputación sofisticado y novedoso proporciona retroalimentación sobre la experiencia científica de los usuarios, permite a los usuarios rastrear su propio aprendizaje y el de otros, en un contexto de aprendizaje informal. Encontramos una desigual distribución de la actividad, característica de las redes sociales, y evidencia del funcionamiento del sistema de reputación para ampliar la contribución de los expertos acreditados. Sostenemos que existe un potencial considerable para aplicar dicho sistema de reputación en otros contextos de aprendizaje participativo.

Aumentar la motivación de los estudiantes y ayudarlos a reflexionar sobre sus procesos de aprendizaje es un factor importante para el aprendizaje de la investigación analítica. (Santos, Govaerts, Verbert, & Duval, 2012) presenta una investigación sobre el desarrollo de un tablero que permite auto-reflexión en las actividades y la comparación con los compañeros. Describimos los resultados de la evaluación de cuatro iteraciones de una metodología de investigación basada en el diseño que evalúan la usabilidad, el uso y la utilidad de diferentes visualizaciones. Se describen las lecciones aprendidas de las diferentes evaluaciones realizadas durante cada iteración.

Además, estas evaluaciones ilustran que el tablero es una herramienta útil para los estudiantes. Sin embargo, se necesitan más investigaciones para evaluar el impacto en el proceso de aprendizaje.

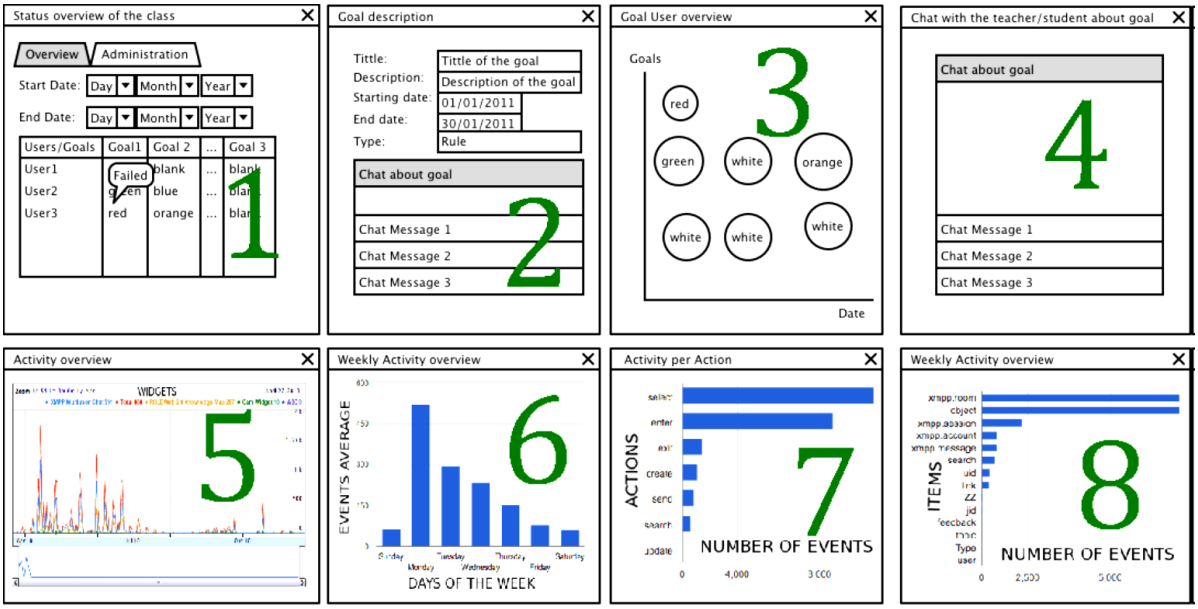
La metodología de investigación basada en el diseño se ha aplicado para llevar a cabo esta investigación. Esta metodología se basa en el prototipado rápido para evaluar las ideas en frecuentes ciclos de iteración corta. El enfoque permite recopilar tanto cualitativos y cuantitativos de datos de evaluación durante todo el proceso de diseño de *software*.

En las dos primeras iteraciones, desarrolladas un prototipo en papel y un prototipo digital. La evaluación de las iteraciones recogidas datos cualitativos de entrevistas y observaciones de los usuarios de 15-30 minutos utilizan el pensamiento en voz alta protocolo.

Seis maestros y asistentes de enseñanza participaron en la primera iteración y 5 en la segunda iteración. Las evaluaciones con estos participantes son útiles para recopilar los requisitos e identificar posibles problemas de usabilidad con las técnicas de interacción.

La tercera y cuarta iteración se llevan a cabo mediante una metodología mixta de evaluación de la investigación con cuestionarios y preguntas abiertas. En estas iteraciones, se realiza las evaluaciones con 36 y 10 estudiantes, respectivamente. Estos cuestionarios se centraron en aspectos concretos de la aplicación y permitieron el análisis estadístico de los datos de evaluación.

Figura 8. Prototipo de papel - Paper prototype



Fuente: (Santos et al., 2012)

De las evidencias anteriores, el autor concluye que la visualización de datos permite conocer ideas e informaciones complejas de forma clara, precisa y eficiente. Esta técnica es aplicable al momento de motivar a los estudiantes y ayudarlos a reflexionar sobre sus procesos de aprendizaje, además de

permitir comparar la información con otros estudiantes. En la actualidad varias herramientas están a disposición, como *visually, thinglink, carto builder,* entre otras.

3.1.6.2.4. Estadísticas

(Aaten, Van den Heuvel-Panhuizen, & Elia, 2011), fue quien investigó las capacidades imaginarias de toma de perspectiva "imaginary perspective taking" (IPT) de los niños de kindergarten, examinan su capacidad para imaginar si un objeto es visible desde otro punto de vista (IPT Competencia 1) y cómo un objeto se ve desde otro punto de vista. El número de participantes fue de 308 niños de entre 4 y 5 años de edad en los Países Bajos. Se desarrolló y administró a los niños una prueba de papel y lápiz de varias tareas pictóricas de toma de perspectiva. Los resultados muestran que la competencia 2 del IPT es más difícil que la competencia 1 del IPT, y que ambas competencias se desarrollan durante los años del kindergarten. Además, para ambas competencias IPT 1 y 2 se encontró una relación positiva con el rendimiento de las matemáticas, mientras que no se encontró diferencia de género en ninguna de las competencias del IPT.

Las competencias de los niños de kindergarten determinan la visibilidad y la apariencia de los objetos como se ve desde otro punto de vista. En particular, se examina cómo se relacionan estas habilidades, cómo se desarrollan y cómo se relacionan con el rendimiento de género y con las matemáticas.

En cuanto a la capacidad de los niños de kindergarten para determinar si un objeto es visible desde un punto de vista diferente (IPT Competencia 1), se obtuvo una tasa de éxito global de 73% mientras que la tasa de éxito en los elementos que requieren la capacidad de determinar la apariencia de un objeto de un diferente punto de vista (Competencia 2 del IPT) fue significativamente menor, es decir, el 34%. Este resultado indica que los ítems de competencia 2 de IPT son más exigentes para los niños que los ítems de competencia 1 de IPT (Hipótesis 1), lo cual está alineado con estudios previos.

El Análisis Estadístico Implicativo demostró que el éxito en varios de los ítems de competencia 2 del IPT implicaba éxito tanto en otros ítems de competencia 2 de IPT como en ítems de competencia 1 de IPT. La otra dirección, una relación implicativa de un elemento competencia 1 del IPT y un elemento competencia 2 del IPT, no se encontró. Esto sugiere que el desarrollo de la competencia 1 del IPT precede al desarrollo de la competencia 2 del IPT. Esta conclusión está de acuerdo a que los niños se desempeñaron mejor en los ítems de Competencia 1 del IPT que en los ítems competencia 2 del IPT.

En educación los niños de la edad del jardín durante su infancia es un período muy importante en muchos campos. Esto resultó ser también el caso de las competencias del IPT. En su estudio se encontró que ambas competencias IPT aumentan significativamente de K1 a K2.

El nivel de rendimiento de las matemáticas está significativamente relacionado con sus competencias de IPT. Esto está acorde con los resultados anteriores sobre la relación entre la capacidad espacial y el rendimiento de las matemáticas. Sin embargo, en el estudio la relación entre el rendimiento de las matemáticas y el IPT fue más fuerte para la Competencia 1 que para la competencia 2. La investigación futura podría explorar las causas de esta diferencia y si cambia con el tiempo.

Por otro lado (Elia, Özel, Gagatsis, Panaoura, & Özel, 2016), investiga las concepciones de valor absoluto (VA) de los estudiantes, su desempeño en diversos ítems de AV, sus errores en estos ítems y las relaciones entre las concepciones de los estudiantes y su desempeño y errores. El Espacio de Trabajo Matemático - Mathematical Working Space (MWS) se utiliza como un marco para estudiar el trabajo matemático de los estudiantes sobre VA y los obstáculos que dificultan su trabajo en Turquía y Chipre. Se llevó a cabo un estudio comparativo entre los dos países, mediante el cual se obtuvo una comprensión más profunda del MWS personal de los estudiantes sobre VA. En particular, se realizó una encuesta en Turquía, tras una encuesta similar en Chipre, en la que se evaluó el rendimiento de los estudiantes de secundaria mediante una prueba. Los hallazgos mostraron una discrepancia en la concepción de VA más prevalente en cada país, indican las diferencias en la referencia y el MWS adecuado entre los dos países. Para Turquía, la concepción de VA como distancia de 0, que fue la definición más ampliamente utilizada, dio un soporte positivo a la solución de los ítems que implican el razonamiento discursivo. Este no fue el caso de Chipre, en el que la concepción más frecuente de VA fue «número sin signo». Un análisis de los errores de los estudiantes turcos reveló una distinción entre los errores en la génesis discursiva de los estudiantes y la génesis semiótica, que eran una consecuencia de obstáculos didácticos o epistemológicos que intervinieron en el MWS personal de los estudiantes. Luego (Zilková, Guncaga, & Kopácová, 2015) analiza la educación en Eslovaquia. El currículo de matemáticas para la educación primaria en Eslovaquia se define por el Programa Nacional de Educación, también conocido como "NEP", (Instituto Nacional de Educación, 2009). Se llama Matemáticas y Trabajo con Información. Proporciona objetivos para la educación primaria y estándares matemáticos mínimos que los estudiantes de matemáticas deben adquirir. La geometría representa una parte muy pequeña del currículo en Eslovaquia. Estudiantes eslovacos en los estudios 2007 y 2011, obtuvieron una puntuación más baja en las preguntas de las pruebas de geometrías que sus homólogos internacionales en la UE y la OCDE. Esto puede ser causado por el reciente Programa

de Educación Nacional para la educación primaria que ha reducido significativamente el currículo geométrico. Así, (Scholtzová, 2014) analiza la educación geométrica en la enseñanza primaria en Eslovaquia y menciona que los estudiantes llegan a la siguiente etapa de educación con un rango limitado de conocimientos geométricos.

La reducción de la geometría en el plan de estudios (así como el número de horas) en Eslovaquia ha cambiado la actitud de los maestros hacia la enseñanza de la geometría en la educación primaria. Por lo general, los profesores dedican muy poco tiempo a la geometría en la educación primaria y, por lo tanto, pueden hacerlo muy ampliamente. Se supone que esta situación mejorará significativamente si y sólo si habrá un cambio de actitud sobre la geometría en los profesores y futuros profesores.) También, (Scholtzová, 2014) anota pocas cosas sobre matemáticas (geometría). "La educación de los futuros docentes de las escuelas primarias debe incorporar el máximo de factores determinantes de la realidad legislativa y educativa". Por lo tanto, se considera la formación superior de maestros geométricos para la educación primaria como uno de los puntos de partida para abordar esta situación.

Hoy en día, la preparación matemática de los profesores de educación primaria en Eslovaquia se realiza en seis universidades. Cada universidad tiene su propio enfoque hacia la educación matemática. Algunas facultades prefieren enseñar enfoques didácticos en la educación matemática, mientras que otras enfatizan los componentes matemáticos profesionales. En general, se ha observado una disminución del nivel de conocimientos geométricos en los profesores de educación primaria.

Con respecto a la estadística de datos masivos, el autor acota que los datos son objetos y no solo mediciones, que requieren modelos complejos para su interpretación, más aún cuando se trabaja con grandes cantidades de datos, el cual se ha convertido en un fenómeno, porque cada día crece más la cantidad de datos a nivel mundial, por lo que importante buscar la técnica más óptima para la obtención de resultados y reducir los tiempos de respuesta. Este análisis de los métodos y sus diferentes técnicas del AEI que se aproximan al LA, es importante al momento de implementar el diseño cuasi-experimental, para obtener la técnica óptima en tiempo y uso de memoria.

3.2. Estado del Arte

Para fundamentar la investigación se ha determinado la necesidad de establecer enfoques presentados por varios autores sobre el Learnign Analytics y el Análisis Estadístico Implicativo, estudios que se detallan a continuación:

Las investigaciones de (Becker, 2013), indican que la migración de las aulas tradicionales a los entornos de aprendizaje en línea está en pleno efecto. En este proceso de estos cambios, es necesario considerar un nuevo enfoque para el aprendizaje de la analítica. El LA se refiere al proceso de

recolección y estudio de datos que se usan para tomar decisiones de instrucción que apoyen el éxito del estudiante. En LA, los "datos de uso" pueden referirse a una amplia gama de información producida por la población observada. Las herramientas y la tecnología necesarias para estudiar los análisis de aprendizaje empiezan a simplificarse, gracias al LA, al permitir interpretar grandes cantidades de información producida dentro del ámbito educativo.

Por medio del uso del método más apropiado se puede reducir tiempos en la obtención de resultados en el análisis de datos (evaluaciones, rendimiento, uso de tecnología, entre otros) dentro del campo educativo, para lo cual la presente investigación permite obtener la técnica similar más óptima entre AEI y LA, para alcanzar tiempos mínimos en el análisis de grandes cantidades de datos educativos.

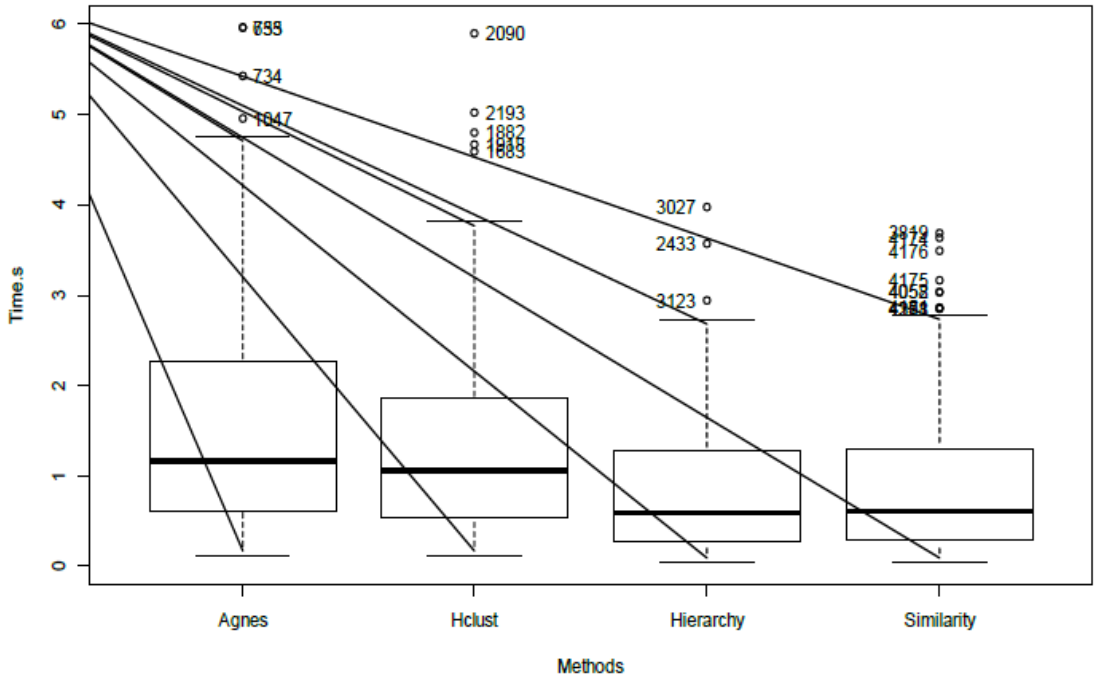
Al revisar conceptos de (Boyd & Crawford, 2012); (Berland, Baker, & Blikstein, 2014) se considera que el construccionismo puede ser un poderoso marco para enseñar contenidos complejos a los novatos. En el centro del construccionismo se considera que permitir a los estudiantes construir artefactos creativos que requieran, ser el contenido complejo para funcionar. Esos estudiantes tendrán la oportunidad de aprender este contenido de manera contextualizada y personalmente significativa. Se investigó la relevancia de un conjunto de enfoques ampliamente denominados "minería de datos educativos" o "análisis de aprendizaje" (EDM) para ayudar a proporcionar una base para la investigación cuantitativa por muchos investigadores en el paradigma del aprendizaje construccionista que no abandone la riqueza vista como esencial.

Un apoyo significativo y potencial en la investigación de la EDM ser útil tanto a los investigadores que trabajan activamente en la tradición construccionista, así como a las comunidades más amplias. Tales colaboraciones tienen el potencial de mejorar la capacidad de los investigadores construccionistas al aportar ricas inferencias sobre el aprendizaje y los estudiantes

(Chatti, Dyckhoff, Schroeder, & Thüs, 2012) manifiesta que existe un creciente interés en aprender analítica en Tecnología de Aprendizaje Mejorado (TEL). En general, la analítica de aprendizaje se ocupa del desarrollo de métodos que aprovechan los conjuntos de datos educativos para apoyar el proceso de aprendizaje. Learning analytics (LA) es un campo multidisciplinario que involucra el aprendizaje automático, la inteligencia artificial, la recuperación de información, estadísticas y visualización. LA es también un campo en el que convergen varias áreas relacionadas de investigación en TEL. Estos incluyen análisis académicos, investigación de acción, minería de datos educativos, sistemas de recomendación y aprendizaje adaptativo personalizado.

Es importante el análisis realizado por los autores (R. A. Pazmiño-Maji, F. J. García-Peñalvo, & M. Conde-González, 2017), quienes indican que Learning Analytics ha sido y es una tecnología emergente en educación; la cantidad de investigación sobre el análisis de aprendizaje aumenta cada año. La integración de nuevas herramientas de código abierto, métodos de análisis y otras opciones de cálculo son importantes. Aspectos a tomar en cuenta para nuestra investigación es la comparación de árboles jerárquicos en el Análisis Estadístico Implicativo (AEI) y algunos clústeres jerárquicos en Learning Analytics. Además de usar un diseño cuasi-experimental con datos binarios aleatorios. Una comparación es sobre el tiempo que lleva evaluar la función para ejecutar los cuatro algoritmos de cluster: árbol de cohesión (AEI), árbol de similitud (AEI), agnes (cluster R ackage) y hclust (R base function). Este análisis, al mismo tiempo proporciona un clúster jerárquico alternativo utilizado en el Análisis Estadístico Estadístico que es posible usar en Learning Analytics (LA). Además, proporciona un R-program comparativo utilizado e identifica investigaciones futuras sobre el rendimiento del *software*. Toda esta información: la comparación de árboles jerárquicos en el Análisis Estadístico Implicativo (AEI) con clústeres jerárquicos en Learning Analytics, el diseño cuasi-experimental aplicado y el programa R de comparativo, ayuda en el desarrollo de nuestra investigación. Los resultados de esta investigación se pueden notar en la siguiente figura:

Figura 9. Diferencia significativa entre el tiempo.



Fuente: (R. A. Pazmiño-Maji et al., 2017)

Al existir una diferencia significativa entre el tiempo, pero el árbol de cohesión (callHierarchyTree: Hierarchy) y el árbol de similitud (callSimilarityTree: Similarity) de los algoritmos parecen similares y más pequeños en el tiempo que el hclust y el agnes. La diferencia entre los tiempos para evaluar las funciones para ejecutar el árbol de cohesión, el árbol de similitud, los algoritmos de agnes y hclust es altamente significativa (valor de p <2.2e-16) y se necesitan de 2 a 2 comparaciones en el futuro (usando los R-packages dunn.test, conver.test, PCMCR, u otros), al no ser el tiempo el mismo en al menos uno de los algoritmos (Agnes, Hclust, Hierarchy o Similarity). Al necesitar una comparación de dos a dos, es importante nuestro planeamiento, y así dar continuidad a esta investigación, al realizar un análisis de tiempos y uso de memoria entre las técnicas clustering del LA (dendro_diana, dendro_variable, hclust_vector) y AEI (callHierarchyTree: Hierarchy) y el árbol de similitud (callSimilarityTree: Similarity). Al utilizar otros algoritmos se comprobará si en verdad los algoritmos Hierarchy y Similarity son los algoritmos más óptimos, en el manejo de datos masivos, todo enfocado a los datos masivos trabajados dentro de la educación.

Por otra parte, en otro artículo (Rubén A Pazmiño-Maji et al., 2016) indican que LA es la medición, recopilación, análisis e informes de datos sobre estudiantes y sus contextos, a los efectos de comprender y optimizar el aprendizaje y los entornos en que ocurre. Al estudia la aproximación de la teoría del Análisis Estadístico Implicativo (AEI) a Learning Analytics (LA). Con el fin, de crear un marco de aproximación basado en la definición, las etapas y los métodos utilizados en LA, además de proporcionar temas referentes a la aproximación del AEI a LA, también proporciona el enfoque de porcentajes por categoría e identifica un número de investigaciones futuras, y al ser nuestra investigación enfocada a las técnicas similares entre el AEI y LA, muestra un marco de definiciones que ayudará a la investigación a desarrollar.

Ya desde el punto de vista educativo se puede identificar varios desafíos y oportunidades de investigación en el área de LA en relación a la educación y su aporte al análisis de datos, algunos de ellos se detallan a continuación:

En la investigación de (Drachsler, Dietze, Herder, d'Aquin, & Taibi, 2014) indica que el LAK Data Challenge 2014 continúa los esfuerzos al estimular la investigación en los campos en evolución Learning Analytics (LA) y Educational Data Mining (EDM). Basado en una serie de actividades del proyecto LinkedUp, tener como objetivo generar nuevos conocimientos y análisis sobre las disciplinas de LA y EDM.

Actualmente existen varios grupos de investigación sobre el LA, lo que permite mejorar el análisis de datos educativos, con el objetivo de obtener resultados reales y a corto plazo.

(Ferguson, 2016), menciona que el análisis de aprendizaje es un área significativa del aprendizaje que ha surgido durante la última década. Este análisis de campo, comienza con un examen de los aspectos tecnológicos, educativos y políticos, factores que han impulsado el desarrollo de la analítica en entornos educativos. También indica el surgimiento de la analítica del aprendizaje, incluye sus orígenes en el siglo XX, el desarrollo de la analítica basada en datos, el surtimiento de perspectivas de aprendizaje enfocadas y la influencia de las preocupaciones económicas nacionales.

Las relaciones entre el análisis de aprendizaje y el análisis académico, permiten la obtención de resultados reales por medio del uso de medios tecnológicos.

Un análisis integral lo realiza en su libro (Mayor, 2015), donde indica que la cantidad de datos en el mundo aumenta exponencialmente a medida que pasa el tiempo. Se estima que la cantidad total de datos producidos en 2020 será de 20 zettabytes, es decir, 20 mil millones de terabytes. Las organizaciones gastan mucho esfuerzo y dinero en recolectar y almacenar datos, y aun así, la mayoría no se analiza en absoluto o no se analiza correctamente. Una razón para analizar datos es predecir el futuro, es decir, producir conocimiento accionable.

A partir del análisis de estas definiciones, se ve el propósito principal del autor, el cual es mostrarle cómo obtener resultados con algoritmos razonablemente simples.

El análisis hecho por (Orús, Peydró, & Gregori, 2013) menciona que el análisis estadístico implicativo es uno de los métodos de análisis de datos, concebido en el campo de la Didáctica de las Matemáticas, que pretende desvelar relaciones de causalidad entre las variables estudiadas. Por ello se considera interesante que los docentes conozcan esta herramienta y la puedan utilizar en su práctica educativa, ya que permite confirmar o refutar algunas creencias del profesor, así como descubrir otro tipo de relaciones que hasta el momento no le fueran evidentes.

En Análisis Estadístico Implicativo es una la herramienta estadística como la existencia y las posibilidades del fondo documental de cara a la formación de los estudiantes.

Los estudios realizados por (Siemens, 2013), indica que el Learning Analytics (LA) ha llamado la atención de académicos, investigadores y administradores. Este interés está motivado por la necesidad de comprender mejor la enseñanza, el aprendizaje, el "contenido inteligente", la personalización y la adaptación. Aún en las primeras etapas de la investigación y la implementación, varias organizaciones (Society for Learning Analytics Research y la International Educational Data Mining Society) se han formado para fomentar una comunidad de investigación en torno al papel de la analítica de datos en la educación.

Se han aportado tecnologías y metodologías para el desarrollo de análisis, modelos analíticos, la importancia de aumentar las capacidades analíticas en las organizaciones y los modelos para implementar análisis en entornos educativos.

(Zamora & Díaz, 2008), en su trabajo de investigación revela las posibles relaciones de similitud, implicación y cohesión entre el rendimiento académico de estudiantes provenientes de preuniversitarios que ingresan a las carreras de Matemática y Ciencia de la Computación y el rendimiento que muestran en las asignaturas de corte matemático y de Programación que reciben en el primer año de las mencionadas carreras, al dar continuidad a la investigación comenzada por Zamora y Díaz, en el 2008. El rendimiento fue analizado a través del índice de ingreso a la Educación Superior, la nota del curso introductorio universitario y las notas obtenidas en el primer año de la carrera en asignaturas de corte matemático y en Programación (para computadoras). Además (Zamora-Matamoros, Díaz-Silvera, & Portuondo-Mallet, 2015) ofrece algunos conceptos fundamentales del análisis estadístico implicativo para el caso de variables modales y se propone un índice para establecer la similaridad entre dos variables modales, así como expresiones para el cálculo de la tipicalidad y contribución de los individuos a las clases que se forman en la clasificación. Con el objetivo de ilustrar la técnica presentada, se aplica a dos juegos de datos, uno binario, el cual permite mostrar numéricamente la coincidencia de las fórmulas presentadas con las existentes para el caso de variables binarias, y otro modal con más de dos modalidades.

La exploración de información sobre Learnign Analytics y Análisis Estadístico Implicativo permiten adoptar el mejor método, para que el docente pueda obtener resultados al optimizar tiempo y proceso.

Capítulo 4

Metodología

4.1. Paradigmas de investigación

Para el establecimiento de un análisis formal, es necesario el escogimiento de una metodología clara y definida. Por lo tanto, el paradigma de investigación es cuantitativo porque el tipo de diseño utilizado es cuasi-experimental:

- ✓ El tiempo de estudio es longitudinal.
- ✓ El colectivo de estudio lo conforman 383 bases de datos binarias, que es la muestra a utilizar en el cuasi-experimento, la cual es un muestreo aleatorio simple con parámetro de interés la media.
- ✓ La población lo conforman la información de nombre, filas, columnas, total de datos, tiempo y memoria.
- ✓ La amplitud de estudio es un muestreo.

Además de realizar los siguientes procesos para la obtención de resultados:

- ✓ Medición objetiva.
- ✓ Generación de datos.
- ✓ Análisis estadístico.
- ✓ Manejo de *software* estadístico.
- ✓ Comprobación de tiempo de procesamiento.
- ✓ Uso de espacio de memoria.

En base al diseño cuasi-experimental con el uso de *software* estadístico (R, RStudio, Rchic), a través del manejo de datos generados aleatoriamente y con la ayuda de una técnica de exploración de datos (algoritmo cluster), se podrá comparar el AEI y el LA en relación al uso de las técnicas de exploración de datos educativos.

4.2. Método(s) aplicado(s)

Al momento de identificar las técnicas similares entre el Análisis Estadístico Implicativo y el Learning Analytic. se aplica un método el cual permite cumplir de forma secuencial y específica,

aspectos que permita un análisis eficaz de la información, la cual se pretende analizar. El método manejado se conoce como el método de estudio de similitud entre modelos y estándares (MSSS), fue propuesto por unos estudiantes de la Universidad Politécnica de Madrid (Calvo-Manzano et al., 2008). Este método ha sido adaptado para identificar las técnicas similares que se estudia en este proyecto de investigación.

Por otra parte y de acuerdo a (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2003) y (Rusu, 2011), el método explicativo supera la simple descripción de concepciones, fenómenos o de relación entre conceptos; es decir, son encaminados a responder fenómenos físicos, sociales a causa de eventos. Este método se centra en exponer por que sucede y en qué condiciones se muestra el fenómeno, además de observar las relaciones que se pueden dar entre variables.

Dentro de la investigación el método explicativo busca el porqué de los hechos mediante el establecimiento de la relación causa – efecto entre el entre los algoritmos similares a LA y AEI y las funciones espacio y tiempo a través del diseño cuasi-experimental de la ingeniería de *software*.

Dentro de método experimental se puede mencionar el:

Método general: Diseño cuasi-experimental.

Con base a los diseños cuasi-experimentales de investigación sobre la enseñanza realizado por (Campbell & Stanley, 1966), se realiza el diseño cuasi-experimental de la ingeniería de *software*, el cual consta de los siguientes pasos:

1. Generación de la base de datos informática (variables v1, v2 ,v3; las misma que almacenan números binarios)
2. Determinación de variables dependientes, factores, variables intervinientes,
3. Definición del diseño cuasi-experimental 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.

Método específico: Experimentación en la ingeniería de *software*, con base a los diseños cuasi-experimentales sobre la enseñanza realizado por Donald Campbell y Julian Stanley.

4.3. Materiales y herramientas

Los materiales utilizados para la comprobación de estudio comparativo del Análisis Estadístico Implicativo y el Learning Analytics en relación al uso de las técnicas de exploración de datos educativos, se divide de acuerdo a la tabla 2.

Tabla 2. Materiales informáticos (hardware – *software*)

Requisitos	Computadora 1	Computadora 2	Computadora 3
Procesador	Core I7	Core I7	Core I7
Velocidad	2,2 Ghz	2,2 Ghz	2,2 Ghz
Memoria RAM	8Gb	8Gb	8Gb
Sistema Operativo	MAC OS 10 64 bits	Windows 8 64 bits	Linux – Ubuntu 16.04 64 bits
Aplicaciones	✓ Software R (versión 3.4.1) ✓ RStudio (Versión 1.0.143) ✓ RCHIC (Versión 0.24)	✓ Software R (versión 3.4.1) ✓ RStudio (Versión 1.0.143) ✓ RCHIC (Versión 0.24)	✓ Software R (versión 3.4.1) ✓ RStudio (Versión 1.0.143) ✓ RCHIC (Versión 0.24)

Fuente: Elaboración propia

4.3.1. Software R

R es un entorno de *software* libre y lenguaje de programación con un enfoque al análisis estadístico. R es una implementación de *software* libre del lenguaje S pero con soporte de alcance estático. Se trata de uno de los lenguajes más utilizados en la investigación por la comunidad estadística, ser además muy popular en el campo de la minería de datos. R es parte del sistema GNU y se distribuye bajo la licencia GNU GPL.

Compila y ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS (ver apéndice A.1.).

4.3.2. RStudio:

RStudio es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. Está disponible para Windows, Mac y Linux (ver apéndice A.2.).

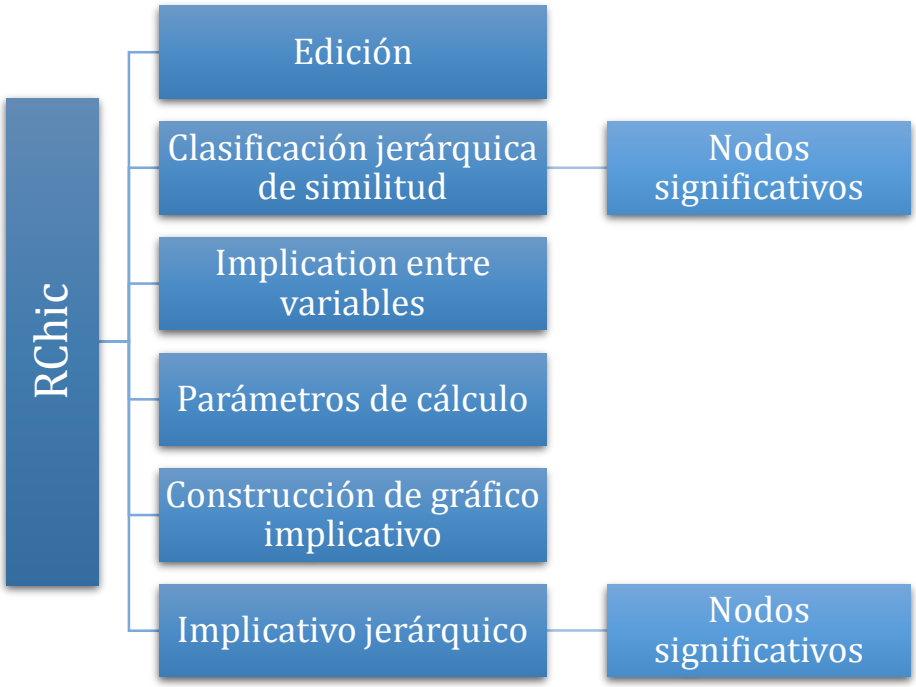
Dentro de este entorno integrado para R, se puede instalar los paquetes disponibles por R, uno de los cuales es importante en el desarrollo del diseño cuasi experimental es Rchic. Para instalar los paquetes es necesario seguir los pasos que se detalla en el apéndice A.3.

4.3.3. RChic

Programa que permite realizar análisis estadísticos, similitudes y cohesiones entre variables, trabaja bajo el entorno de RStudio (ver apéndice A.3.).

CHIC permite el uso de la mayoría de los métodos definidos en el marco de la UPS (Participación de Estadísticas Implicativo). Su objetivo es buscar las implicaciones más relevantes entre las variables en un conjunto de datos. Para ello, se propone organizar las implicaciones como una jerarquía Cohesiva (orientado) o gráfico implicativo. Además, proporciona una jerarquía de similitudes (no dirigidos) en base a las similitudes de las variables. En este artículo se describe la historia, las características y el uso de CHIC. (Couturier & Almouloud, 2009)

Figura 10. Diagrama de flujo del *software* RChic



Fuente: (Couturier & Almouloud, 2009)

Rchic trabaja con Jerarquía de las similitudes y Cohesiva jerárquica, calcula el conjunto de todas las reglas basadas en parámetros seleccionados por el usuario, es posible construir una jerarquía de estas normas. Esta jerarquía puede parecerse método de clasificación orientado o no depende del tipo de cálculo seleccionado "similitud o la participación." Sin embargo, maneras de construir cada una de estas jerarquías tienen ciertas similitudes. En la siguiente regla se llama una clase, se agrega dos variables en su forma más simple. En cada nivel de clasificación, Rchic elige la clase que tiene la más

alta cohesión (semejanza o participación). Luego, en cada paso, Rchic calcula un conjunto de nuevas clases de las clases de la jerarquía.

Para crear una nueva clase, una clase existente se agregan, ya sea con una variable que no ha sido agregada por el momento o con otra clase en la jerarquía.

Sin embargo, cada par de variables se agregan dos clases debe tener una intensidad válida. Por ejemplo, la clase de entrenamiento ((a, b), c) requiere que las clases (a, c) y (b, c) tener una buena cohesión (con la implicación) o son similares (con análisis similitudes). La clase ((a, b), c) representa la regla (a) b)) c con el análisis implicativo y representa el hecho de que un implica B y la clase ((a, b), c) tiene una buena cohesión y la clase (a, b) es similar a c con el análisis de similitudes.

4.3.4. Otras herramientas

R cuenta con varios paquetes, los cuales deben ser instalados posterior a la instalación de R y Rstudio, a continuación, se detalla los paquetes a ser utilizados para el desarrollo de del diseño cuasi-experimental.

4.3.4.1. Microbenchmark

Los autores (Mersmann, Beleites, Hurling, & Friedman, 2014) indican que el microbenchmark sirve como una sustitución más precisa de la expresión system.time (replicate (1000, expr)). Se intenta medir con precisión sólo el tiempo que se tarda en evaluar expr. Para lograr esto, las funciones de temporización precisas de sub-milisegundo (supuestamente nanosegundos) que se utilizan en la mayoría de los sistemas operativos modernos. Además, todas las evaluaciones de las expresiones se realizan en código C para minimizar cualquier sobrecarga. Esta función sólo está diseñada para micro-benchmarking pequeños fragmentos de código fuente y para comparar sus características de rendimiento relativo. Finalmente nos proporciona infraestructura para medir y comparar con precisión el tiempo de ejecución de las expresiones R.

Uso

Microbenchmark (..., list = NULL, times = 100L, unidad, check = NULL,

Control = list ())

Argumentos

...	Expresiones de benchmark..
Lista	Lista de expresión no evaluada de benchmark..
Veces	Número de veces para evaluar la expresión.
Unidad	Unidad predeterminada utilizada en summary y print.

Check Función para comprobar si las expresiones son iguales. Por defecto, NULL omite comprobar.

Control Lista de argumentos de control.

4.3.4.2. Ggplot2

Los desarrolladores (Wickham & Chang, 2016) del paquete ggplot2, definen como un paquete que permite crear visualizaciones de datos elegantes al usar gramática de gráficos. Usted proporciona los datos, le dice a 'ggplot2' cómo mapear las variables a la estética, qué primitivas gráficas usar y se ocupa de los detalles.

Una de las funciones a utilizar es el autoplot el cual usa el paquete ggplot2 para dibujar un gráfico particular para un objeto de una clase particular en un solo comando.

Uso

autoplot (objeto, ...)

Argumentos

Object: un objeto, cuya clase determinará el comportamiento de autoplot

4.3.4.3. Cluster

Los autores (Maechler et al., 2017) desarrollaron este paquete con el fin de aplicar el método de análisis de conglomerados.

Las funciones a utilizar del paquete cluster, dentro del experimento cuasi-experimental está:

- ✓ **Diana - Divisive Analysis Clustering**

Está función calcula una agrupación jerárquica divisiva del conjunto de datos que devuelve un objeto de la clase diana. Probablemente sea único en el cálculo de una jerarquía divisiva, mientras que la mayoría de los otros *software* para la agrupación jerárquica son aglomerativos.

El diana-algoritmo construye una jerarquía de agrupamientos, al comenzar con un grupo grande que contiene todas las n observaciones. Los grupos se dividen hasta que cada grupo contiene solo una observación. En cada etapa, se selecciona el clúster con el diámetro más grande. (El diámetro de un grupo es la mayor diferencia entre dos de sus observaciones). Para dividir el clúster seleccionado, el algoritmo primero busca su observación más dispar (es decir, que tiene la mayor diferencia de promedio con las otras observaciones del clúster seleccionado).

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)

Código implementado:

dendro_diana<-function(x)

```
{
 df<-read.csv(x, sep = ";")
 dend <- diana(df, metric = "euclidean", stand = TRUE)
 plot(dend)
 return(dend)
}
```

Técnicas clustering

Las técnicas cluster permiten extraer caracterizaciones, o permitirán predecir características, o deducir relaciones útiles, a lo que se denomina como agrupación (clustering). Algunos de los algoritmos más comunes son: Cobweb, EM, y Kmeans. Al ser este último el más utilizado (Malbernat, Clemens, Varela, & Urrizaga, 2015).

Tabla 3. Caracterización de las técnicas clustering

Técnicas clustering	**Situación didáctica**	**Justificación**	**Herramientas informáticas**	**Artículo**
K-means (método basado en particiones)	Factibilidad de incorporar actividades virtuales según las competencias docentes aplican técnicas de Data Mining, en relación con la Preparación y la Actitud para la modalidad virtual, los docentes pueden clasificarse como Innovadores, Indiferentes y Refractarios,	Permite correlaciones significativas, patrones y tendencias que se obtienen al examinar grandes volúmenes de datos almacenados en repositorios, empleado tanto tecnologías de reconocimiento de patrones como técnicas estadísticas y matemáticas.	Weka (Waikato Environment for Knowledge Analysis), PSPP (Perfect Statistics Professionally Presented) y RapidMiner versión Starter.	Aplicación de Técnicas de Data Mining en Gestión de Docentes de Educación Superior. (Malbernat et al., 2015)

	Uso de los medios sociales para explorar el impacto académico.	Permite agrupar y visualizar los patrones de impacto representados por una subselección de indicadores, lo que permite derivar reglas simples para predecir con exactitud la membresía de clones k-means.	R Project for Statistical Computing	Altmetrics en la naturaleza: El uso de las redes sociales para explorar el impacto académico. (Priem, Piwowar, & Hemminger, 2012)
Algoritmo EM (Expectation Maximization).	Agrupar a los estudiantes, en primer lugar, a partir de los datos de una síntesis de uso de Moodle y/o las calificaciones finales de los alumnos en un curso.	Se utilizó este algoritmo por ser un algoritmo de clustering bien conocido y además, no requiere que el usuario especifique el número de grupos.	DM WEKA (Witten y Frank, 2005) Esperanza-Maximización (EM)	Applying data mining to discover common learning routes in Moodle (Bogarín Vega et al., 2015)
CobWeb	Clasificación automática de textos consiste en colocar un documento dentro de un grupo de clases previamente definidas.	Enfoque de Aprendizaje Computacional que utilizan características léxicas	No da a conocer.	Método Semisupervisado para la Clasificación Automática de Textos de Opinión. (Arredondo, 2009)

Fuente: Elaboración propia

Los creadores del paquete cluster (Maechler et al., 2017) para el análisis de Cluster, parten de la extendió el original de Peter Rousseeuw, Anja Struyf y Mia Hubert, basado en Kaufman y Rousseeuw (1990) `Finding Groups in Data '.

R tiene una sorprendente variedad de funciones para el análisis de conglomerados. En esta sección, describiré uno de los muchos enfoques: partición. Si bien no hay mejores soluciones para el problema de determinar el número de racimos que se extraen, a continuación, se presentan el enfoque de particionamiento.

- ✓ **Particionamiento**

K-means clustering es el método de partición más popular. Requiere que el analista especifique el número de clústeres a extraer. Una gráfica de la suma de cuadrados de los grupos dentro del número de racimos extraídos puede ayudar a determinar el número apropiado de conglomerados. El analista busca una curva en la gráfica similar a una prueba de scree en el análisis factorial. (Everitt & Hothorn).

4.3.4.4. Fastcluster

El desarrollador de fastcluster (Müllner, 2017), implementa un paquete dos en uno que proporciona interfaces tanto para R como para Python. Implementa rutinas jerárquicas y aglomeración de rutinas de agrupamiento. Parte de la funcionalidad está diseñada como un reemplazo directo para 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 ofrecen los paquetes existentes.

La función a implementar en nuestro diseño cuasi-experimental de este paquete es:

- ✓ **hclust.vector - Agrupación jerárquica y aglomerativa rápida de datos vectoriales.**

La función hclust.vector proporciona agrupamiento cuando la entrada es datos vectoriales. Utiliza algoritmos de ahorro de memoria que permiten el procesamiento de conjuntos de datos más grandes que hclust. Los métodos "ward", "centroid" y "median" requieren metric = "euclidean" y agrupan el conjunto de datos con respecto a las distancias euclidianas. Para el agrupamiento de ligamiento "único", se puede elegir cualquier medida de desemejanza. Actualmente, las mismas métricas se implementan como proporciona la función dist.

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

Código implementado:

```
hclust_vector<-function(x)
```

```
{
 df<-read.csv(x, sep = ";")
 #hc <- hclust.vector(df, "cen")
 hc<-hclust.vector(df, method="single", members=NULL, metric='euclidean', p=NULL)
 plot(hc)
 return(hc)
}
```

4.3.4.5. CluMix

Los autores del paquete CluMix (Hummel, Edelmann, & Kopp-Schneider, 2017), proporcionan utilidades para agrupar sujetos y variables de tipos de datos mixtos. Las similitudes entre las variables se pueden evaluar de dos maneras:

1. Mediante la combinación de medidas de asociación apropiadas para diferentes pares de tipos de datos.
2. Basado en correlación de distancia. Alternativamente, las variables también pueden agruparse por el enfoque 'ClustOfVar'.

La característica principal del paquete es la generación de un mapa de calor de datos mixtos. Para visualizar similitudes entre sujetos o variables, se puede dibujar un mapa de calor de la matriz de distancia correspondiente. Las asociaciones entre variables pueden explorarse mediante un "diagrama de confusión", que permite la detección visual de posibles factores de confusión, colineales o sustitutos para algunas variables de interés primario. Las matrices de distancia y los dendrogramas para sujetos y variables se pueden derivar y usar para otras visualizaciones y aplicaciones.

La función a utilizar de este paquete se detalla a continuación:

- ✓ **Dendro.variables - Variables dendrogram**

Está función permite la agrupación de variables i) basada en la similitud al usar medidas de asociación, ii) basada en la similitud al usar la correlación de distancia, o iii) mediante el enfoque ClustOfVar, que utiliza el análisis de componentes principales para datos mixtos.

Uso: dendro.variables(data, method = c("associationMeasures", "distcor", "ClustOfVar"), linkage="ward.D2", associationFun = association, check.psd = TRUE)

Código implementado:

```
dendro_variables<-function(x)
      {
       df<-read.csv(x, sep = ";")
```

```
  dend1 <- dendro.variables(df, method="distcor")
  plot(dend1)#opcional
  return(dend1)
 }
```

4.3.4.6. Arules

El paquete desarrollado por (Hahsler et al., 2018) proporciona la infraestructura para representar, manipular y analizar datos y patrones de transacciones (conjuntos de elementos frecuentes y reglas de asociación). También proporciona implementaciones en C de los algoritmos de minería de asociación Apriori y Eclat.

Las funciones a utilizar son las siguientes:

- ✓ **Apriori**

Es un conjunto de elementos frecuentes, reglas de asociación o hiper bordes de asociación que utilizan el algoritmo Apriori.

El algoritmo Apriori emplea la búsqueda a nivel de conjuntos de elementos frecuentes. Incluye la implementación de un árbol de prefijos y clasificación de elementos).

Método:

- Deje k = 1
- Genera conjuntos de elementos frecuentes de longitud 1
- Repite hasta que no se identifiquen nuevos conjuntos de elementos frecuentes
  - Genera conjuntos de elementos de longitud de longitud (k + 1) a partir de la longitud k conjuntos de elementos frecuentes
  - Poda conjuntos de elementos candidatos que contienen subconjuntos de longitud k que son infrecuentes
  - Cuenta el apoyo de cada candidato al escanear la DB
  - Elimina candidatos que son poco frecuentes, al dejar solo aquellos que son frecuentes

La implementación de apriori, es a traves de uso de Hash Tree.

Uso

apriori (data, parameter = NULL, appearance = NULL, control = NULL)

Código implementado:

```
 met_apriori<-function(x){
  d<-read.csv(x, sep = ";")
  df<-as.matrix(d)
```

```
ma<-apriori(df, parameter = list(supp = 0.5, maxlen = 3, conf = 0.6, target = "rules"))
plot(ma)
return(ma)
}
```

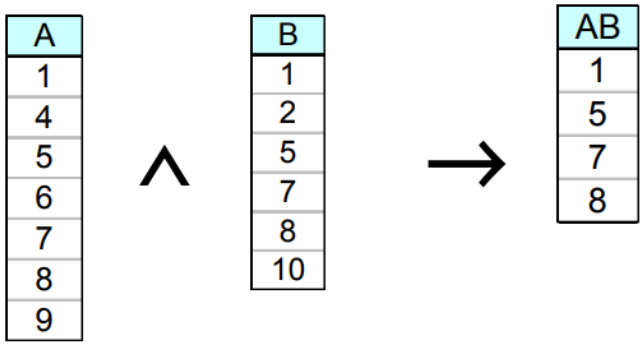
- ✓ **Eclat**

De acuerdo (Hurairah, 2014), el ste algoritmo usa operaciones simples de intersección para el agrupamiento de clase de equivalencia junto con el recorrido de retícula ascendente.

Generación

- ✓ Determinar el soporte de cualquier k-conjunto de elementos mediante la intersección de listas de dos de sus subconjuntos (k-1).

Figura 11. Generación - algoritmo eclat



Fuente: (Hurairah, 2014)

- ✓ 3 enfoques transversales: de arriba hacia abajo, de abajo hacia arriba e híbridos
- ✓ Ventaja: recuento de soporte muy rápido
- ✓ Desventaja: las listas intermedias pueden convertirse también grande para la memoria

La generación frecuente de conjuntos de elementos es computacionalmente costosa.

Algoritmo ECLAT

- ✓ Agrupación de clase de equivalencia y de abajo hacia arriba Lattice Traversal- ECLAT
- ✓ Método para la generación frecuente de conjuntos de elementos.
- ✓ Busca de manera DFS.
- ✓ Representar los datos en formato vertical.

Uso:

eclat(data, parameter = NULL, control = NULL)

Código implementado:

```
met_eclat<-function(x){
  d<-read.csv(x, sep = ";")
  df<-as.matrix(d)
  rules_ec <- eclat(df,parameter = list(supp = 0.5, maxlen = 3))
  rules_ec
  plot(rules_ec)
  return(rules_ec)
 }
```

- ✓ **Weclat**

Esta implementación utiliza combinaciones optimizadas de tidlist y ponderaciones de transacción para implementar la extracción de reglas de asociación ponderada (WARM).

Uso:

```
weclat(data, parameter = NULL, control = NULL)
```

Código implementado:

```
met_weclat<-function(x){
d<-read.csv(x, sep = ";")
df<-as.matrix(d)
rules_weclat <- weclat(df, parameter = list(support = 0.5, maxlen = 3), control = list(verbose = TRUE))
plot(rules_weclat)
return(rules_weclat)
}
```

4.3.4.7. Factoextra

Los creadores (Kassambara & Mundt, 2016) del paquete factoextra, definen como un paquete R que facilita la extracción y visualización de la salida de análisis exploratorios de datos multivariados, al incluir:

- ✓ Análisis de Componentes Principales (PCA), que se utiliza para resumir la información contenida en un continuo (es decir, cuantitativo) datos multivariantes mediante la reducción de la dimensionalidad de los datos sin perder la información importante.

- ✓ Correspondence Analysis (CA), que es una extensión del análisis de componentes principales adecuado para analizar una gran tabla de contingencia formada por dos variables cualitativas (o datos categóricos).
- ✓ Multiple Correspondence Analysis (MCA), que es una adaptación de CA a una tabla de datos que contiene más de dos variables categóricas.
- ✓ Análisis de Factores Múltiples (MFA) dedicado a conjuntos de datos donde las variables están organizadas en grupos (variables cualitativas y / o cuantitativas).
- ✓ Análisis Jerárquico de Factores Múltiples (HMFA): Una extensión del AMF en una situación donde los datos están organizados en una estructura jerárquica.

Contiene también funciones para simplificar algunos pasos de análisis de agrupación y proporciona una visualización de datos elegante basada en 'ggplot2'

4.3.4.8. ClValid

Los autores(Brock, Pihur, Datta, & Datta, 2011)indican que el paquete R clValid contiene funciones para validar los resultados de un análisis de agrupación. Existen tres tipos principales de medidas de validación de clúster disponibles, \ internal ", \ stability" y \ biological ". el usuario puede elegir entre nueve algoritmos de agrupación en los paquetes R existentes, incluye mapas jerárquicos, K-means, self-organizing maps (SOM), y clustering basado en modelos. Además, proporcionan una función para realizar el algoritmo de árbol auto-organizable (SOTA) método de agrupación.

Medidas de Validación

El paquete clValid ofrece tres tipos de validación de clúster, \ internal ", \ stability" y \ biological". Las medidas de validación interna toman sólo el conjunto de datos y la partición de clústeres como entrada y usan información intrínseca en los datos para evaluar la calidad del agrupamiento.

Las medidas de estabilidad son una versión especial de las medidas internas. Ellos evalúan la consistencia de un resultado de agrupación comparándolo con los grupos obtenidos después de cada columna se elimina, uno a la vez. La validación biológica evalúa la capacidad de un algoritmo de agrupamiento para producir clusters biológicamente significativos. Se tiene medidas para investigar tanto la homogeneidad biológica como la estabilidad de los resultados del agrupamiento.

a. Medidas internas

Para la validación interna, se selecciona medidas que reflejan la compacidad, la conexión y la separación de las particiones del clúster. La conectividad se relaciona con la medida en que las observaciones se colocan en el mismo grupo que sus vecinos más cercanos en el espacio de datos, y aquí se mide por la conectividad (Handl et al., 2005). La compacidad evalúa la homogeneidad del clúster, por lo general se observa la varianza intragrupo, mientras que la separación cuantifica el grado

de separación entre los grupos (usualmente mide la distancia entre los centroides del grupo). Puesto que la compacidad y la separación demuestran tendencias opuestas (la compacidad aumenta con el número de racimos, pero la separación disminuye), los métodos populares combinan las dos medidas en una sola puntuación. El Dunn Index (Dunn, 1974) y el Silhouette Width (Rousseeuw, 1987) son ambos ejemplos de combinaciones no lineales de compacidad y separación, y con la conectividad comprenden las tres medidas internas disponibles en clValid. Los detalles de cada medida se dan a continuación, y para una buena visión general de las medidas internas en general ver Handl et al. (2005).

b. Medidas de estabilidad

Las medidas de estabilidad comparan los resultados de la agrupación basada en los datos completos a la agrupación basada en la eliminación de cada columna, una a la vez. Estas medidas funcionan especialmente bien si los datos están altamente correlacionados, lo que suele ocurrir en los datos genómicos de alto rendimiento. Las medidas incluidas son la proporción media de no superposición (APN), la distancia media (AD), la distancia media entre las medias (ADM) y la medida del mérito (FDA) (Datta y Datta, 2003; Yeung et al. , 2001). En todos los casos el promedio se toma sobre todas las columnas eliminadas, y todas las medidas deben ser minimizadas.

c. Biológico

La validación biológica evalúa la capacidad de un algoritmo de agrupamiento para producir clusters biológicamente significativos. Una aplicación típica de la validación biológica es en los datos de microarrays, donde las observaciones corresponden a genes (donde \ genes "podrían ser marcos de lectura abiertos (ORFs), etiquetas de secuencia expresa (ESTs), análisis en serie de etiquetas de expresión génica (SAGE) Existen dos medidas disponibles, el índice de homogeneidad biológica (BHI) y el índice de estabilidad biológica (BSI), ambos presentados originalmente en Datta y Datta (2006).

4.4. Población y muestra

Población. - El tamaño de la población es 100000 datos binarios (0 y 1) generados aleatoriamente por el *software* RStudio, formados por una combinación de observaciones y variables de tipo binario (rendimiento, evaluaciones, uso de tecnología).

Muestra: Por el gran tamaño de la población, se escogió trabajar con una muestra significativa.

Tamaño de la muestra y fórmula:

Se consideró la siguiente fórmula para el cálculo del tamaño de la muestra:

Fórmula 1. Cálculo de la muestra.

(1)

$$n = \frac{S^2}{\frac{E^2}{Z^2_{\frac{\acute{\alpha}}{2}}} + \frac{S^2}{N}}$$

Para la aplicación de la fórmula se utilizó los siguientes parámetros: Desviación Estándar = 1; α = 5%; Z = 1,96; E= 10%; N = 100000 y se generó un tamaño de la muestra de 382,675 que aproximadamente es 383.

Tipo de muestreo: Muestreo aleatorio simple con parámetro de interés la media.

Capítulo 5

Resultados

5.1. Identificación de técnicas similares entre el Análisis Estadístico Implicativo y el Learning Analytics.

Como se detalló en el capítulo 4, el método a utilizar para verificar las técnicas similares es el método conocido como método de estudio de similitud entre modelos y estándares (MSSS)(Calvo-Manzano et al., 2008).

El análisis que realiza (Gasca Hurtado, 2010), indica que el método MSSS precisa pasos mínimos, a través de fases. Estas fases son:

1. Seleccionar posibles estándares y modelos a analizar.
2. Seleccionar o definir el modelo de referencia.
3. Seleccionar el o los procesos que se van a analizar.
4. Establecer el nivel de detalle.
5. Crear una plantilla de correspondencia.
6. Identificar las similitudes entre los modelos.
7. Presentar resultados obtenidos.

Al ser estas fases generales, han sido adaptadas para el estudio e investigación de este proyecto, a continuación, se muestra la metodología determinada para identificar las técnicas similares entre el Análisis Estadístico Implicativo y el Learning Analytics.

5.1.1. Metodología generada a partir de MSSS

En esta sección se muestra la implementación del método MSSS para identificar las técnicas similares entre el AEI y LA.

Como se ha visto el método MSSS está integrado por siete pasos generales, sin embargo, para llevar a cabo el estudio de las técnicas del AEI y LA, fue necesario adaptar dichos pasos de tal forma que se pudiera aplicar a dicho ámbito.

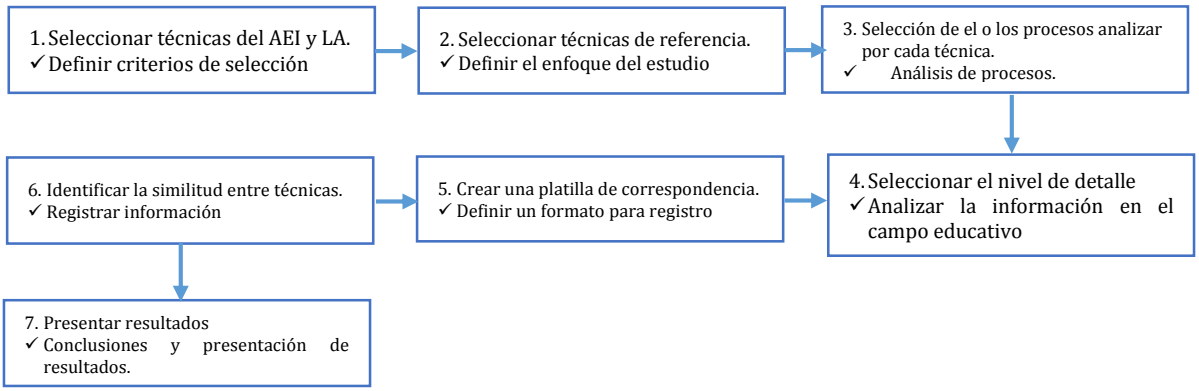
Tabla 4. Adaptación del MSSS como método para identificar las técnicas similares entre AEI y LA

Fases	Método de estudio de similitud entre modelos y estándares	Método de estudio para identificar técnicas similares entre AEI y LA
1	Seleccionar posibles estándares y modelos a analizar.	Seleccionar técnicas del AEI y LA.
2	Seleccionar o definir el modelo de referencia.	Seleccionar o definir las técnicas de referencia.
3	Seleccionar el o los procesos que se van a analizar.	Seleccionar el o los procesos que se van a analizar.
4	Establecer el nivel de detalle.	Establecer el nivel de detalle.
5	Crear una plantilla de correspondencia.	Crear una plantilla de correspondencia.
6	Identificar las similitudes entre los modelos.	Identificar la/las similitudes entre las técnicas.
7	Presentar resultados obtenidos.	Presentar resultados obtenidos.

Fuente:(Calvo-Manzano et al., 2008)

La adaptación del método MSSS para la identificación de técnicas similares de AEI y LA está compuesto por siete fases, los cuales se muestra la figura 12, para una mejor interpretación.

Figura 12. Pasos para identificar las técnicas similares de AEI y LA

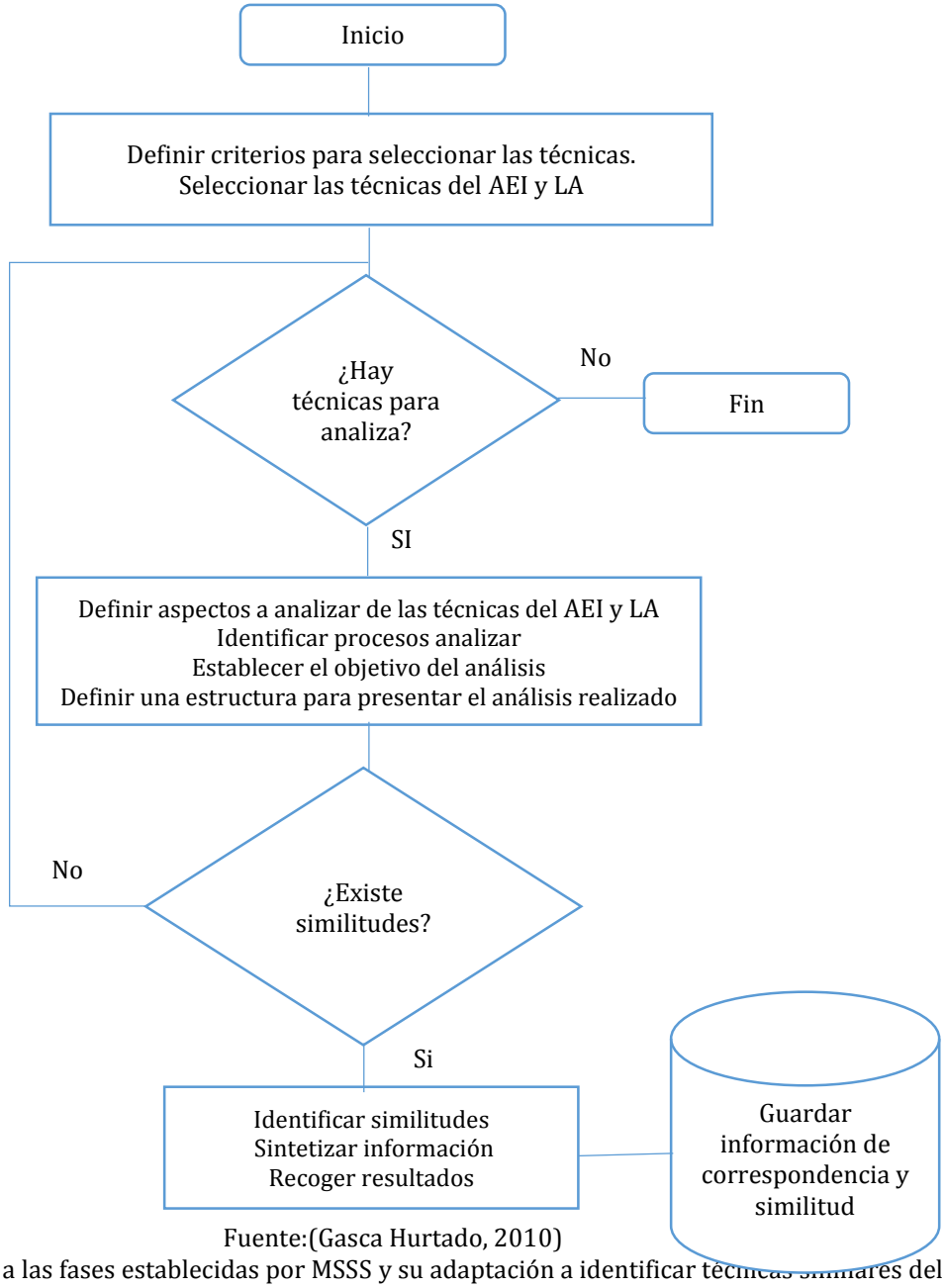


Fuente: Elaboración propia

5.1.2. Adaptación de MSSS para identificar las técnicas similares entre AEI y LA

En la figura 13 se muestra el diagrama de flujo que se establece para llevar a cabo el procedimiento de selección e identificación de las técnicas similares de AEI y LA. el mismo que se obtiene a partir de los pasos generales de MSSS(Calvo-Manzano et al., 2008).

Figura 13.MSSS para Identificar las técnicas similares entre el AEI y LA.



Fuente:(Gasca Hurtado, 2010)

De acuerdo a las fases establecidas por MSSS y su adaptación a identificar técnicas similares del AEI y LA, se detalla el proceso a realizar en cada fase para su posterior ejecución y obtención de resultados.

5.1.2.1. Seleccionar técnicas del AEI y LA

Para seleccionar las técnicas tanto del AEI y LA, se analizará algunos criterios para la adopción o selección de las técnicas a analizar en este proyecto son:

a. Todas aquellas técnicas usadas tanto en AEI y LA.

b. Aquellas técnicas que sean más utilizadas dentro del AIE y LA.
c. Aquellas técnicas que tengan su información disponible.

5.1.2.2. Seleccionar o definir las técnicas de referencia del AEI y LA

La utilizar los criterios antes mencionados, y consider el análisis realizado por (Rubén A. Pazmiño-Maji et al., 2016), donde los autores realizan una aproximación del AEI a los métodos LA, quienes enfatizan en criterios como:

- ✓ Aproximación de AEI a la definición de LA y fuente de datos
- ✓ Aproximación AEI a las etapas LA
- ✓ Aproximación AEI a los métodos LA: Clasificación Baker e Inventado
- ✓ AEI Aproximación a los métodos LA: Clasificación de Papamitsiou y Economides

5.1.2.3. Selección de el o los procesos analizar de cada una de las técnicas.

De las técnicas seleccionadas se analizará aspectos como:

- ✓ Alcance o enfoque de la aplicación de la técnica.
- ✓ Filosofía o principios básicos y parámetros de la técnica.
- ✓ Estructura de los pasos o fases que componen la técnica.

Estos aspectos permitirán comparar o identificar las similitudes entre las técnicas del AEI y LA

5.1.2.4. Nivel de detalle del uso de las técnicas seleccionas en el campo educativo.

Los procesos que se van analizar están determinados por el ámbito de las técnicas del AEI y LA. Se analizarán los procesos y elementos de las técnicas, con el fin de determinar aspectos del AEI y LA que cada uno de ellos contempla, dentro del campo educativo.

5.1.2.5. Establecer el objetivo del análisis y establecer una plantilla de correspondencia.

Con este análisis se pretende determinar las características de las técnicas aplicadas al AEI y LA. Por lo tanto, se comparan las técnicas, y se presentan en tablas donde se comparan las características más relevantes de cada uno de ellos.

Esta comparación permitirá determinar cuáles de las técnicas se considerarán relevantes en el ámbito de estudio que se trata.

5.1.2.6. Identificar las similitudes entre las técnicas y definir la estructura para presentar el análisis realizado.

La estructura por medio de la cual se presentará el análisis e identificación de técnicas similares del AEI y LA, está determinada por los aspectos a analizar que se definieron antes. La descripción de las similitudes se define por medio del análisis de los pasos anterior, resumidos en una tabla que permita identificar de mejor forma las similitudes presentadas de las técnicas del AEI y LA.

5.1.2.7. Conclusiones y presentación de resultados finales.

Estas actividades se presentan al finalizar la descripción de cada una de las técnicas. La forma de presentación de los resultados es por medio de tablas: una tabla donde se comparan las técnicas y otra tabla donde se comparan las características más representativas de cada uno.

Con esta metodología definida se realizó el estudio que permitió determinar los resultados que se muestran a continuación.

5.1.3. Resultados sobre técnicas similares entre el AEI y LA

A continuación, se presentan los resultados de la similitud de las técnicas AEI y LA. Estos resultados, tal como se ha mencionado, se presentan en tablas. Se presentan conforme a la estructura de estudio determinada por ámbitos.

5.1.3.1. Fase 1: Selección de técnicas del AEI y LA

En base al artículo de (Rubén A. Pazmiño-Maji et al., 2016) (ver tabla 5), quienes realizan una revisión de los artículos (paper reference) relacionados a las técnicas del AEI aproximadas al LA y a los estudios realizados en el capítulo 3 (marco teórico), se estudia cada una de las técnicas, las cuales se listan en la tabla 6, las cuales se consideró desde el campo educativo y su influencia en la toma de decisiones, al igual se menciona las técnicas utilizadas por el AEI (ver tabla 7).

Tabla 5. Aproximación AEI a los métodos LA

SUBCATEGORY OF LA METHODS	PAPER REFERENCE (See References)	FREQ (%)
BAKER AND INVENTADO CLASSIFICATION		
Classification		0
Regression		0
Latent Knowledge Estimation		0
Association Rule Mining	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [56]	23 95.8%
Sequential Pattern Mining		0
Correlation Mining		0
Causal Data Mining	[54]	1 4.2%
Clustering	[2], [9], [14], [17], [57], [28], [32], [40], [54]	9 37.5%
Factor Analysis		0
Domain Structure Discovery		0
Discovery with models		0
PAPAMITSIOU AND ECONOMIDES CLASSIFICATION		
Social Network Analysis		0
Text Mining		0
Visualization		0
Statistics	[1], [17], [57], [31], [36]	5 20.8%

Fuente: (Rubén A. Pazmiño-Maji et al., 2016)
La tabla muestra una revisión de artículos científicos, de las técnicas del AEI aproximadas al LA.

Tabla 6. Técnicas del AEI aproximadas al LA

Según Baker e Inventado			
Técnicas	**Tipos**	**Referencia**	**Tópico de estudio de la referencia**
Clasificación	✓ Árboles de decisión	(Nithyasri et al., 2011)	Estructura de árbol jerárquico como clasificador de aprendizaje.
	✓ Bosques al azar o aleatorios.	(Vandamme et al., 2007)	Analiza el fracaso académico, al presentar los resultados en bosques aleatorios.
	✓ Regresión escalonada.	(International Sales and Support, 2017)	Regresión escalonada como herramienta automatizada de predicción.
	✓ Regresión logística	(Bewick et al., 2005)	Analiza la muerte o supervivencia de un grupo de pacientes.
Regresión	-	-	

Estimación del conocimiento latente	-	-	
Minería de relaciones (Relationship mining)	✓ Minería de reglas de asociación.	(Ben-Naim et al., 2009)	Encontrar reglas de asociación dentro de los datos de los estudiantes de una clase.
	✓ Minería de patrones secuenciales.	(Lin et al., 2003) (Srikant & Agrawal, 1996)	Paradigmas de extracción clásica de patrones secuenciales.
	✓ Minería de correlación.	(Arroyo & Woolf, 2005)	Relación entre el diseño de sistemas de tutoría inteligentes.
	✓ Minería de datos causales.	(Fancsali, 2012)	Factores que llevan a un estudiante hacer mal las cosas.
Descubrimiento de estructuras	✓ Clustering	(Amershi & Conati, 2009)	Ejemplos de agrupaciones en el ámbito educativo.
	✓ Factor análisis	-	
	✓ Descubrimiento de la estructura de domino	-	
	✓ Descubrimiento de modelos	-	
Según Papamitsiou y Economides			
Técnicas	**Tipos**	**Referencia**	**Tópico de estudio de la referencia**
Análisis de redes sociales	✓	-	
Minería de texto	✓	(He et al., 2013)	Analiza contenido de texto no estructurado en Facebook y twitter.
Visualización	✓	(Santos et al., 2012)	Desarrollo de un tablero que permite auto-reflexión, mediante la visualización.
Estadística	✓	(Elia et al., 2016)	Relaciones entre las concepciones de los estudiantes y su desempeño.

Fuente:(Rubén A. Pazmiño-Maji et al., 2016)

Tabla 7. Técnicas del Análisis Estadístico Implicativo

Técnicas	Tipos	Referencia	Tópico de estudio de la referencia
Según Regis Gras: (Gras et al., 2008)	Cohesión e implicación	(Raphaël Couturier & Pazmiño, 2016)	El AEI para poder analizar el comportamiento general de la población (ESPOCH).

Según Israel Lerman. (Lerman, 1981)	Similaridad	(Zamora & Díaz, 2008)	Investiga las posibles relaciones de similaridad, implicación y cohesión, entre el rendimiento académico de estudiantes provenientes de preuniversitarios que ingresan a las carreras de Matemática y Ciencia de la Computación y el rendimiento que muestran en las asignaturas de corte matemático y de Programación que reciben en el primer año de las mencionadas carreras.

Fuente: Elaboración propia

5.1.3.2. Fase 2: Selección de técnicas referenciales

Con lo antes mencionado (ver tabla 5) se puede dar cuenta que la regla de asociación de minería, clustering y minería de datos causales son las técnicas más utilizadas LA, y al hacer referencia o al tener un alcance con AEI, se puede encontrar las características relevantes de cada una de ellas.

Tabla 8. Características de las técnicas seleccionadas.

LA / ASI	Técnicas	Características	Referencia
Learning Analytics	Association Rule Mining (Minería de reglas de asociación)	✓ Asociación de reglas ✓ Algoritmos: Apriori, ECLAT y FP-crecimiento ✓ Medidas de Interés ✓ Aplicaciones ✓ Regla de asociación Minería con R ✓ Eliminación de la redundancia ✓ Reglas de Interpretación ✓ Visualización de reglas de asociación ✓ Lecturas adicionales y recursos en línea	(Zhao & Bhowmick, 2015)
	Clustering	✓ Algoritmos: o Agrupamiento por particiones (k-Means, PAM/CLARA/CLARANS, BFR) o Métodos basados en densidad (DBSCAN, Optics, DenClue)	(Berzal)

		○ Clustering jerárquico (Diana/Agnes, BIRCH, CURE, Chameleon, ROCK) ✓ Se obtiene como resultado final un conjunto de agrupamientos. ✓ Minimizar distancia intra-cluster (cohesión) ✓ Maximizar distancia inter-cluster (separación) ✓ Ordenar los datos en la matriz de similitud con respecto a los clusters ✓ Escoger una muestra y aplicar un método jerárquico.	
	Causal Data Mining (Minería de Datos Causales)	✓ Relaciones predictivas en los datos. ✓ Estudia el comportamiento ✓ Se agrupan de acuerdo a su comportamiento. ✓ Encontrar relaciones causales en los datos. ✓ Estudia el comportamiento causal.	(Li et al., 2013)
Análisis estadístico implicativo	Según Regis Gras: (Gras et al., 2008) Cohesión e implicación	✓ Descubrir y estructurar en forma de reglas. ✓ Uso de variables (binario, modal, numérico, intervalo, difuso, vectorial) ✓ Estructuras: grafo implicativo, jerarquía orientada ✓ La visualización de los resultados, así como su interpretación, se facilita con el *software* C.H.I.C. (Clasificación Jerárquica Implicativa y Cohesiva).	(Radès, 2015)
	Según Israel Lerman. (Lerman, 1981) Similaridad	✓ Formalización de una cuasi-regla de similaridad. ✓ Niveles de confianza ✓ Ley de Poisson de parámetro $\frac{n_a n_{\bar{b}}}{n}$. ✓ Crea un subconjunto aleatorio de transacciones binarias ✓ Método de extracción	(Gras & Kuntz, 2009)

Fuente: Elaboración propia

La importancia de las técnicas en el manejo de información, permite ejemplifica alguno de los ejemplos prácticos de estas técnicas.

✓ Minería de reglas de asociación

Como se menciona en el capítulo 3, está técnica tiene como objeto encontrar reglas y asociarlas para que revelen relaciones comunes de datos, que sean difíciles de descubrir manualmente, con esto detalles se puede citar algunos ejemplos que aplican esta técnica:

Tabla 9. Ejemplo - Minería de reglas de asociación.

Técnica	Situación didáctica	Objetivo	Referencias
Minería de reglas de asociación	Investigar las creencias, actitudes, comprensión e intenciones de los estudiantes de Electricidad del Instituto Tecnológico de Kozani, Grecia, para estimar los factores que determinan el grado de comprensión del tema de las construcciones electrónicas.	Medir la actitud de los estudiantes de las Facultades de Tecnología de los Institutos Tecnológicos hacia las construcciones electrónicas	(S. D. Anastasiadou, Anastasiadis, Vandikas, & Angeletos, 2011)
	Desarrollar un análisis basado en la teoría de la AEI que permita a un investigador en ciencias sociales suprimir pseudo-implicaciones sin interés a priori en el análisis.	Revelan vínculos entre las diferentes capacidades matemáticas que los estudiantes pueden tener. Si un estudiante puede resolver un problema de tipo A, ¿significa esto que será capaz de resolver un problema?	(Delacroix & Boubekki, 2014)

Fuente: (S. D. Anastasiadou et al., 2011) (Delacroix & Boubekki, 2014)

✓ Clustering

El propósito de esta técnica es encontrar datos, los cuales se puedan agrupar, divide un conjunto de datos en un conjunto de grupos, para esto esta técnica utiliza algoritmos de agrupación (k-means, EM - Expectation Maximization), los que permiten visualizar patrones de impacto.

Tabla 10. Ejemplo - Clustering.

Técnica	Situación didáctica	Objetivo	Referencias
Clustering	Modelo teórico para el cambio sistémico en lo que respecta al aprendizaje y la enseñanza de las matemáticas	Establecer cambios en la comunidad a través de relaciones, prácticas y procedimientos duraderos a largo plazo	(Kortenkamp & Ladel, 2014)
	Comprender mejor el desempeño de los niños de jardín en la toma de perspectiva imaginaria al examinar si pueden imaginar lo que es visible desde un punto de vista particular	Establecer el desempeño de los niños de jardín.	(Van den Heuvel-Panhuizen, Elia, & Robitzsch, 2015)

Fuente: (Kortenkamp & Ladel, 2014) (Van den Heuvel-Panhuizen et al., 2015)

✓ Minería de datos causales

Esta técnica se va basa en encontrar un evento a través de la observación y ver si fue causa de otro evento, por ahí el nombre causal.

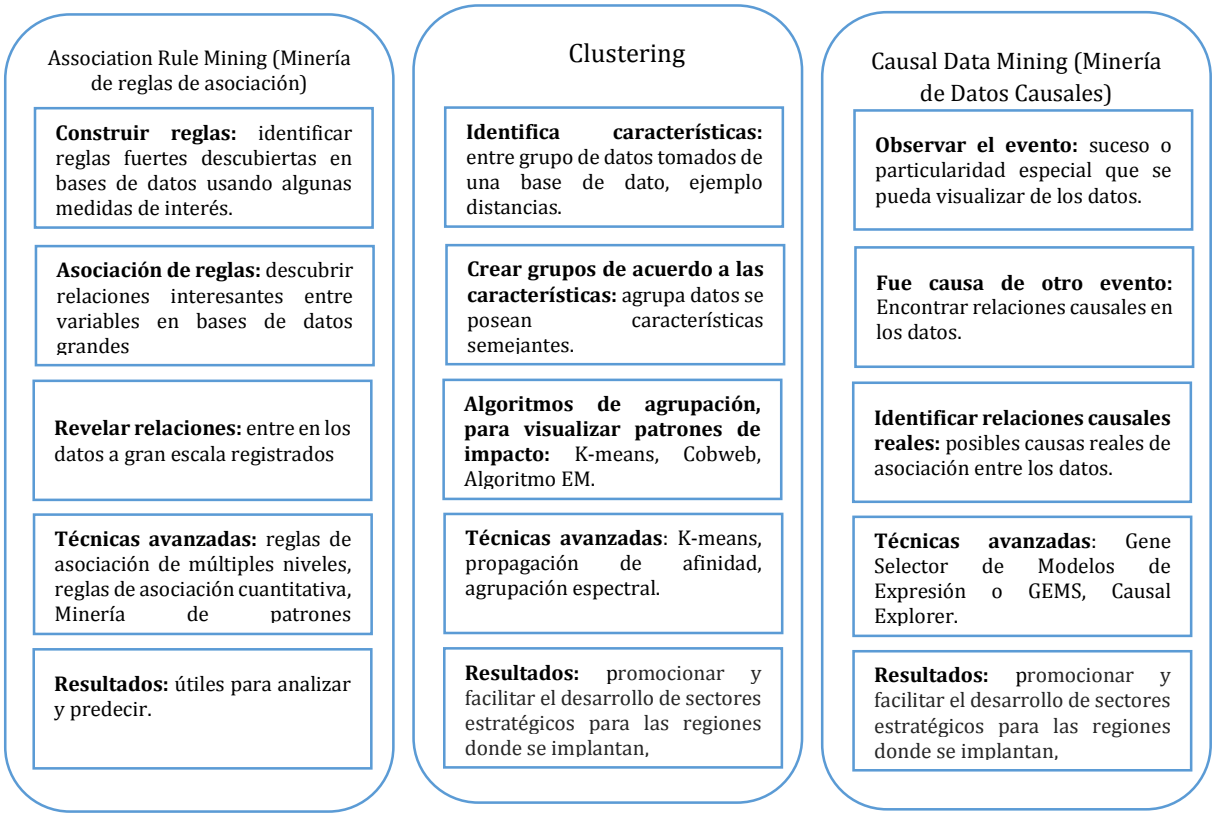
Tabla 11. Ejemplo – minería de datos causales.

Técnica	Situación didáctica	Objetivo	Referencias
Minería de datos causales	Comprender mejor el desempeño de los niños de jardín en la toma de perspectiva imaginaria, al examinar si pueden imaginar lo que es visible desde un punto de vista particular	Establecer el desempeño de los niños de jardín.	(Van den Heuvel-Panhuizen et al., 2015)

Fuente: (Van den Heuvel-Panhuizen et al., 2015)

Luego de analizar las técnicas más relevantes, se puede resumir algunas de sus características en la figura.

Figura 14. Técnicas del LA seleccionadas para análisis.



Fuente: Elaboración propia

Por otra parte, al tener solo dos técnicas del AEI, se trabajará con ellas, estas son: cohesión e implicación según Regis Gras (Gras et al., 2008) y similaridad según Israel Lerman (Lerman, 1981)

5.1.3.3. Fase 3: Selección de los procesos analizar

En esta sección se ve los procesos y manejo de información con un enfoque técnico, de acuerdo a las características analizadas en la fase 2 (ver tabla 8) al considerar los principios básicos, parámetros, con la finalidad de comparar los procesos de minería de reglas de asociación, clustering y minería de datos causales del LA y cohesión e implicación y similaridad del AEI. La tabla 12 permite comparar los procesos realizados o utilizados por las técnicas en análisis.

Tabla 12. Comparativa de procesos de las técnicas

Técnicas del AEI y LA	**Learning Analytics**			**Análisis Estadístico Implicativo**	
Procesos	**Association Rule Mining.** (Minería de reglas de asociación)	**Clustering**	**Causal Data Mining** (Minería de Datos Causales)	**Cohesión e implicación** (Gras et al., 2008)	**Similaridad** (Lerman, 1981)
Análisis de datos	SI	Si	Si	SI	Si
	Construye reglas	Identifica características	Observa eventos	Forma reglas	Similaridad
Proceso de agrupación de datos	Si	Si	Si	Si	Si
	Asocia reglas semejantes	Crea grupos de acuerdo a sus características similares	Relaciona si un evento fue causado por otro.	Asocia reglas	Asocia variables similares
Uso de la estadística	Si	Si	Si	Si	Si
	Cálculos muestra	Cálculos muestra	Cálculos muestra	Subconjunto aleatorio de transacciones binarias	Ley de Poisson
Forma de proporcionar resultados	Si	Si	Si	Si	Si
	Revela relaciones	Revela patrones de impacto	Revela relaciones causales reales	Grafo implicativo, jerarquía orientada	Jerarquización de similaridad
Manejo de grandes volúmenes de datos	Si	Si	Si	Si	Si
	Reportes – resúmenes -BD	Reportes – resúmenes -BD	Reportes – resúmenes -BD	Reportes – resúmenes -BD	Reportes – resúmenes -BD
Manejo de muestreo para grandes cantidades de información	Si	Si	Si	Si	Si
	Muestreo aleatorio de acuerdo a la cantidad de información	Muestreo aleatorio de acuerdo a la cantidad de información	Muestreo aleatorio de acuerdo a la cantidad de información	Muestreo aleatorio de acuerdo a la cantidad de información	Muestreo aleatorio de acuerdo a la cantidad de información

Fuente: Elaboración propia

5.1.3.4. Fase 4: Nivel de detalle del uso de las técnicas seleccionas en el campo educativo.

A continuación, se detalla el enfoque, objetivo y relación de las técnicas seleccionadas, dentro del contexto educativo, con el propósito de fortalecer lo investigado y analizado en el capítulo 3 (marco

teórico) con referencia a las técnicas seleccionadas. Si nos preguntan ¿Porqué un enfoque educativo?, es debido a que se necesita identificar las técnicas similares del AEI y LA, a través de la exploración de datos educativos. Al tener esto como punto de partida la tabla 13 muestra un análisis de la investigación realizada, dentro de contexto educativo con el uso de las técnicas regla de asociación de minería, clustering y minería de datos causales del LA y cohesión e implicación y similaridad del AEI.

Tabla 13. Detalle de las técnicas del AEI y LA con enfoque educativo.

Técnicas	Nombre	Enfoque	Objetivo	Relación con la adquisición y carencias
Learning Analytics	Association Rule Mining. (Regla de Asociación de Minería)	Educativo	Realizar un proceso de obtención de reglas y asociarlas con el propósito de obtener información que permita la toma de decisiones en el ámbito educativo.	Se presenta como un método general para el proceso. Carece de una secuencia de procesos. Es un método que indica los aspectos que se deben tener en cuenta para el proceso dentro del LA sin el diseño de un proceso específicas.
	Clustering	Educativo	Ofrecer cluster o grupo de datos asociados por características comunes, que ayuden en tomar decisiones frente a un problema educativo.	Es una secuencia de pasos generales, basado en agrupación con nivel de complejidad. Es una propuesta importante para él LA. Carece de un enfoque especial para la interpretación de resultados, inserta un nivel de complejidad alto, que dificulta la implementación en pequeños entornos.
	Causal Data Mining (Minería de Datos Causales)	Educativo	Establecer causas de un evento y encontrar posibles eventos causales de otros eventos, con la finalidad de obtener resultados de las causas que originan un problema.	Es una ilustración por medio del cual se centra en obtener los eventos causales, he aquí la importancia dentro del LA. Carece de un enfoque general, por lo que dificultaría su adaptación a otros ámbitos, en especial en el área donde es difícil obtener causales.

Análisis estadístico implicativo	Cohesión e implicación (Gras et al., 2008)	Educativo	Establece cualidades entre variables, con el propósito de obtener una clasificación jerárquica implicativa y cohesiva a través de reglas.	Presenta un método general en base a las variables ejecutar un proceso, mediante el manejo de reglas que permitan determinar aspectos de cohesión e implicación dentro del AEI, además con una herramienta (CHIC) muy eficaz a la hora de ejecutar este método.
	Similaridad (Lerman, 1981)	Educativo	Formula cuasi-regla de similaridad, en base a los niveles de confianza, con el objetivo de obtener similaridad entre las variables, esto en base a la ley de Poisson.	La ilustración se da por medio de la formulación de cuasi-reglas de similaridad lo cual se centra en encontrar niveles de confianza, con la finalidad de encontrar similitudes entre variables, he aquí la importancia dentro del AEI. Posee un enfoque general, por lo que es de gran ayuda dentro de cualquier área.

Fuente: Elaboración propia

5.1.3.5. Fase 5: Crear una plantilla de correspondencia

En base a las características técnicas de funcionamiento se realiza una tabla de correspondencia (tabla 14), la cual permite visualizar la correspondencia de cada característica con regla de asociación, clustering y minería de datos causales del LA y cohesión e implicación y similaridad del AEI. Lo que ayuda en la identificación de características similares trabajadas por las tres técnicas.

Tabla 14. Plantilla de correspondencia

Técnicas AEI y LA	OBJETIVO ESPECIFICO, CORTO Y COMPARABLE	PRERREQUISITOS					INPUTS										PROCESS					OUTPUTS							Inconvenientes
		1	2	3	4	5	Números Binarios	Números Enteros	Números Reales	Números Imaginarios	Vectores	Caracteres	Texto	Imagen	Video	Otro	Software1	Opciones Adicional1	Software2	Opciones Adicional2	Software3	Opciones Adicional3	Números	Vectores	Matrices	Relaciones	Tablas	Gráficos	Otros	
Association Rule Mining. (Regla de Asociación de Minería)	Obtención de reglas, con el objeto de asociar la información	Transacción	Elementos	Cobertura (Número de instancias predichas)			1	1	1			1					A priori algorithm	Road map					1				1	1		
Clustering	Utiliza distancias agrupa sujetos homogéneos entre sí en clusters Heterogéneos	Datos Numéricos					1	1	1								SPSS	Tabla Anova	R	Aproximación número	Weka	Gráfico Puntos	1				1		Gráfico de clusters	Necesita especificar al inicio de número de clusters
Clustering jerárquico	Muestra agrupaciones superpuestas mediante dendogramas	Datos Numéricos					1	1	1								SPSS	Matriz de Proximidades	R	Aproximación número clusters	Weka	Gráficos comparativos	1				1	1		No permite modificar fácilmente las variables para visualizar en el gráfico
Causal Data Mining (Minería de Datos Causales)	Establecer causas de un evento y posibles eventos causales de otros eventos	Predictive										1	1				Tetrad IV	Statistical data					1				1	1		
Según Regis Gras: (Gras et al., 2008)	Cohesión e implicación	Datos numéricos	Binario, modal,				1	1	1	1	1						CHIC	Grafo implicativo,					1				1	1		

Según Israel Lerman. (Lerman, 1981)	Similaridad	Datos numéricos	Binario, modal,				1	1	1	1	1						CHIC	Ley de Posson					1				1			

Fuente: Elaboración propia

Una vez realizado el análisis de las características técnicas de funcionamiento, se llega a la comparación práctica y de enfoque de las técnicas, las que se ve reflejadas partes del análisis de correspondencia.

Tabla 15. Correspondencia práctica y de enfoque de las técnicas analizadas.

Técnicas del AEI y LA	**Learning Analytics**			**Análisis Estadístico Implicativo**	
Características	**Association Rule Mining. (Regla de Asociación de Minería)**	**Clustering**	**Causal Data Mining (Minería de Datos Causales)**	**Cohesión e implicación (Gras et al., 2008)**	**Similaridad Según Israel Lerman. (Lerman, 1981)**
Definición de prácticas para AEI y LA	✓	✓	✓	✓	✓
AEI y LA con un enfoque general	✓	✓	✓	✓	✓
AEI y LA con un enfoque específico a otras áreas		✓			
Prácticas con AEI y LA enfocadas a la educación	✓	✓	✓	✓	✓

Fuente: Elaboración propia

5.1.3.6. Fase 6: Identificar la/las similitudes entre las técnicas

Luego de un profundo análisis de las técnicas seleccionadas, se sigue las fases de MSSS (Calvo-Manzano et al., 2008) y las técnicas del AEI aproximadas al LA (Rubén A. Pazmiño-Maji et al., 2016), nos ha permitido identificar las similitudes entre las técnicas: regla de asociación de minería, clustering y minería de datos causales, las que reflejan (ver tabla16) su similitud entre las técnicas del AEI y las técnicas del LA, en función del contexto educativo.

Tabla 16. Comparativa de características entre las técnicas LA y AEI

Técnicas	**Learning Analytics**	**Análisis Estadístico Implicativo**
Características	Medición de datos	Estructuración de datos
	Recopilación de datos	Interrelacionan sujetos y variables

	Análisis de datos	Extracción de reglas inductivas entre las variables
	Presentación de datos	Presentación de resultados

Fuente: Elaboración propia

Tabla 17. Similitud de las técnicas del AEI y LA

Técnicas del AEI y LA	**Learning Analytics**			**Análisis Estadístico Implicativo**	
Características	Association Rule Mining. (Regla de Asociación de Minería)	**Clustering**	Causal Data Mining (Minería de Datos Causales)	**Cohesión e implicación** (Gras et al., 2008)	**Similaridad** Según Israel Lerman. (Lerman, 1981)
AEI y LA con un enfoque general	✓	✓	✓	✓	✓
AEI y LA con un enfoque específico	✓	✓		✓	✓
Contiene la identificación de procesos/eventos/características como fase inicial	✓	✓	✓	✓	✓
Técnicas recomendadas para el AEI y LA	✓	✓		✓	✓

Fuente: Elaboración propia

5.1.3.7. Fase 7: Conclusiones y resultados del estudio

Las técnicas del LA y AEI, son modelos reconocidos en el ámbito estadístico, son el punto de partida para generar iniciativas que permiten verificar su uso y utilidad dentro del ámbito educativo.

LA es una técnica que tiene como finalidad mejorar los procesos de enseñanza aprendizaje y dar respuesta a las necesidades de los estudiantes, técnicas que se aplican en diferentes ámbitos educativos. Por otra parte, el objetivo del AEI contempla la estructuración de datos, interrelacionan sujetos y variables, extracción de reglas inductivas entre las variables y, a partir de la contingencia de estas normas, la explicación y en consecuencia una determinada previsión en distintos ámbitos: psicología, sociología, biología, etc. De la misma forma, al tratamiento de variables binarias (por ejemplo, descriptores), se añaden progresivamente variables modales, frecuenciales y, recientemente, de variables-intervalo y variables difusas.

Las aproximaciones del AEI a los métodos LA realizadas por (Rubén A. Pazmiño-Maji et al., 2016), son indispensables en la investigación, mismas dan a conocer las técnicas más relevantes del LA, como

son Association Rule Minimg, Causal Data Mining y Clustering, de acuerdo a la clasificación de las técnicas realizadas por Baker e Inventado, esto conjuntamente con las técnicas del AEI que son cohesión e implicación y similaridad, realizadas por Regis Gras y Israel Lerman, respectivamente.

De las técnicas del LA, se seleccionó Association Rule Minimg, Causal Data Mining y Clustering, por ser más utilizadas y encontrar información disponible, como se puede notar en la tabla 5, y las técnicas de cohesión e implicación y similaridad del AEI, por ser únicas.

Una de las características de mayor importancia en el desarrollo del experimento son los algoritmos de reglas de asociación (apriori, ECLAT, WECLAT) las cuales son utilizadas por la técnica de minería de reglas de asociación, ya que permiten verificar asociaciones entre variables que trabajan de diferente forma; conjuntamente con las técnicas clustering (hclust.vector, dendro.variables, diana), estas dos técnicas del LA, se suma la del AEI cohesión e implicación (callHierarchyTree, callSimilarityTree, implicativeGraph). Todas las características mencionadas permiten a través del uso del *software* R verificar la velocidad de procesamiento y uso de memoria, al utilizar los comandos necesarios para realizar su aplicación.

Luego de analizar cada una de las tablas se puede visualizar que existen similitudes entre las técnicas utilizadas en el LA y el AEI, como:

- Definición de prácticas a través del uso de comando con el uso de *software* R, dentro de LA: reglas de asociación (apriorio, eclat, weclat), clustering ((hclust.vector, dendro.variables, diana) y por AEI: (callHierarchyTree, callSimilarityTree, implicativeGraph).
- Tienen un enfoque general, ya que puede ser aplicada en diferentes ámbitos científicos, sociales, educativos, etc.
- En la práctica, tanto las técnicas LA y AEI pueden ser aplicadas dentro del ámbito educativo, los cuales permiten ver las asociaciones e implicaciones entre variables (sexo, edad, rendimiento, aprovechamiento).

Se puede comparar entre medición de datos LA y estructura de datos AEI, ya que el LA trabaja con distancias entre las variables y el AEI utiliza una estructura a través de similitudes entre variables. LA recopila datos y los agrupa según distancias entre variables, el AEI interrelaciona las variables, de esto se puede notar que son formas diferentes, pero a la vez similares porque permiten relacionar las variables. En cuanto a la forma de analizar los datos el LA interpreta las distancias y el AEI formula reglas inductivas entre variables, al proporcionar los dos métodos resultados que pueden ser visualizados a través de dendrogramas, gráficos relacionales, agrupaciones que puedan ser interpretados.

5.2. Elaborar el diseño cuasi-experimental a utilizar mediante la ingeniería de *software*.

En base a los diseños cuasi-experimentales de investigación sobre la enseñanza realizado por Donald Campbell y Julian Stanley, se realiza el diseño cuasi-experimental de la ingeniería de *software*, el cual consta de los siguientes pasos:

5.2.1. Generación de la base de datos informática

A continuación, se presenta el código fuente detalladamente el cual genera los datos aleatoriamente,

Código generador de datos aleatorios.

```
nVar<- round(runif(1),2)*100
nFilas<- round(runif(1),3)*1000
DataBase<-replicate(nVar, round(runif(nFilas),0))
rownames(DataBase)<-paste('S',1:nFilas,sep='')
colnames(DataBase)<-c(';V1',paste('V',2:nVar,sep=''))
f<-paste('_D',toString(i),'.csv',sep="_")
T<-paste('_T',toString(i),'.csv',sep="_")
```

Round: redondea los valores en su primer argumento al número de decimales especificado (valor predeterminado 0).

Uso: round(x, digits = 0)

Argumentos:

X: vector numérico.

Digits: número entero que indica la cantidad de lugares decimales (redondos) o dígitos significativos (signif) que se utilizarán.

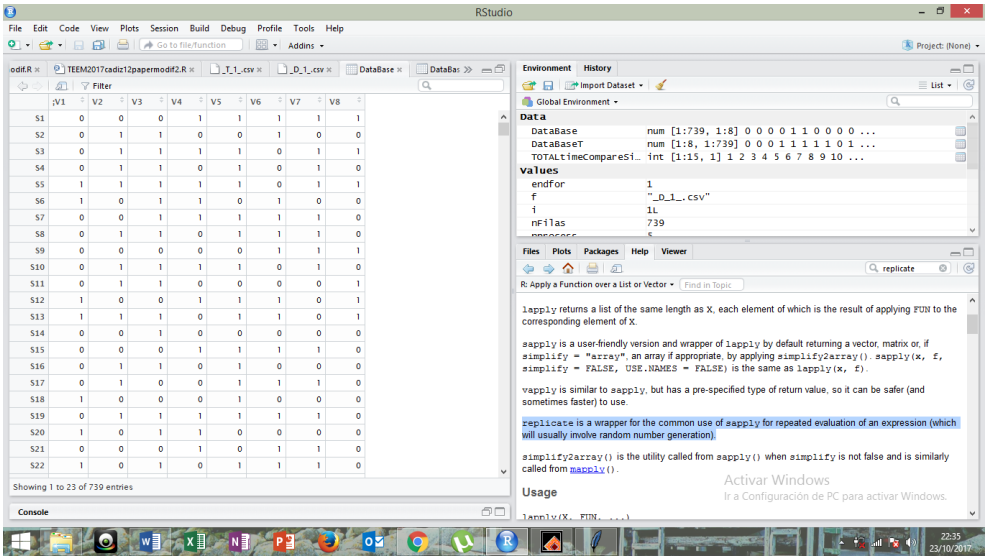
nVar: Genera número de variables a utilizar en la ejecución del programa.

nFilas: Genera número de datos binarios a utilizar.

Replicate: es un contenedor para el uso común de sapply (versión amigable para el usuario y contenedor por defecto que devuelve un vector o matriz) para la evaluación repetida de una expresión (que normalmente implicará la generación de números aleatorios).

DataBase: Base de datos generada aleatoriamente, la misma que conta de variables y datos binarios.

Figura 15. Ejemplo de datos aleatorios generados.



Fuente: Elaboración propia

Paste: Este comando concatena vectores de cadenas u objetos similares a vectores que contienen cadenas.

Ejemplo: rownames(DataBase)<-paste('S',1:nFilas,sep='')

colnames(DataBase)<-c(';V1',paste('V',2:nVar,sep=''))

f: Genera un archive .csv el cual contiene la base de datos, la cual almacena las variables y los datos binarios generados aleatoriamente.

Ejemplo: f<-paste('_D',toString(i),'.csv',sep="_")

Figura 16. Almacenamiento datos generados.

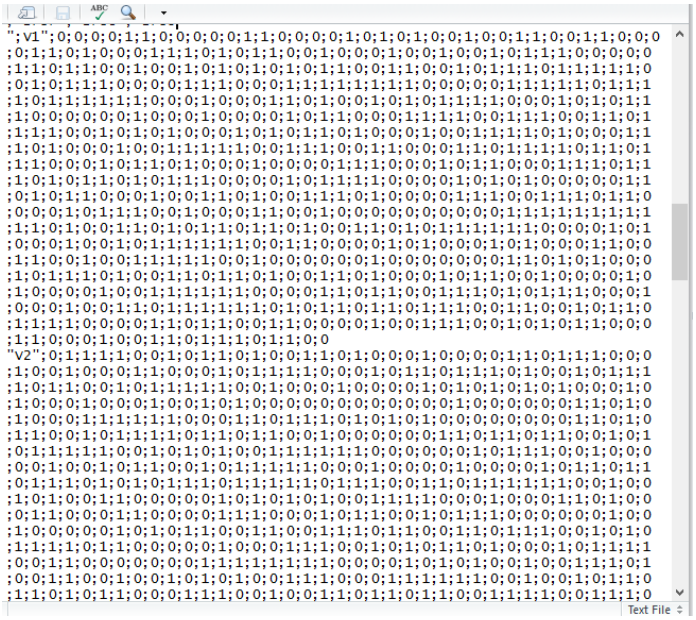
```
;V1;V2;V3;V4;V5;V6;V7;V8
S1;0;0;0;1;1;1;1;1
S2;0;1;1;0;0;1;0;0
S3;0;1;1;1;1;0;1;1
S4;0;1;1;0;1;0;1;0
S5;1;1;1;1;1;0;1;1
S6;1;0;1;1;0;1;0;0
S7;0;0;1;1;1;1;1;0
S8;0;1;1;0;1;1;1;0
S9;0;0;0;0;0;1;1;1
S10;0;1;1;1;1;0;1;0
S11;0;1;1;0;0;0;0;1
S12;1;0;0;1;1;1;0;1
S13;1;1;1;0;1;1;0;1
S14;0;0;1;0;0;0;0;0
S15;0;0;0;1;1;1;1;0
S16;0;1;1;0;1;0;0;0
S17;0;1;0;0;1;1;1;0
S18;1;0;0;0;1;0;0;0
S19;0;1;1;1;1;1;1;0
S20;1;0;1;1;0;0;0;0
S21;0;0;0;1;0;1;1;0
S22;1;0;1;0;1;1;1;0
S23;0;1;1;1;1;0;1;0
S24;0;0;0;1;0;1;0;1
```

Fuente: Elaboración propia

t: Genera un archive .csv el cual contiene la base de datos, agrupados cada variable con sus datos binarios.

Ejemplo: T<-paste('_T',toString(i),'.csv',sep="_")

Figura 17. Base de datos agrupados por variables.



Fuente: Elaboración propia

Estos archivos generados, posteriormente serán útiles al momento de ejecutar las funciones del AEI (hclust.vector, dendro.variables y diana), para comprobar cual se ejecuta más rápido y utiliza más espacio de memoria.

5.2.2. Determinación de variables dependientes, factores, variables intervinientes

Variable independiente: Número y tipo de variables, número de casos, sistema operativo y las técnicas similares de exploración de datos utilizadas tanto en el Análisis Estadístico Implicativo (AEI) como en el Learning Analytics (LA).

Variable dependiente: Espacio de memoria ocupado (en megabytes) y el tiempo de ejecución (en minutos).

5.2.3. Definición del diseño cuasi-experimental a utilizar

En este trabajo se realizará un análisis comparativo de los tiempos de procesamiento y espacio de memoria entre las técnicas de agrupación y las técnicas de búsqueda de reglas de asociación similares al Análisis Estadístico Implicativo y Learnign Analytics. Se consideran como factores el número y tipo de variables, número de casos, algoritmos y sistema operativo (variables independientes) y como variables dependientes las funciones espacio y tiempo. La hipótesis por demostrar es que existe diferencia significativa (función espacio y/o función tiempo) entre los algoritmos similares a LA y AEI.

Para el estudio se utilizan dos computadores con el mismo microprocesador y con los sistemas operativos Windows 10 y Ubuntu 16.04 y MacOS Sierra Versión 10.12.5.

Todos los computadores y sistemas operativos trabajan con el *software* estadístico libre R y el entorno de desarrollo integrado libre (IDE) RStudio. Para demostrar las hipótesis se plantea un experimento de tipo RGXO, (notación de Campbell y Stanley) donde RG representa el grupo experimental conformado por datos generados aleatoriamente, X son los tratamientos y O es la observación luego del tratamiento, no se tiene grupo de control ni tampoco una pre-prueba. Luego de la comprobación de los supuestos de normalidad, homocedasticidad e independencia se empleará un test paramétrico ANOVA (probar hipótesis a través de analizar la variabilidad) con cinco factores, se trabajará con un nivel de significancia del 95%.

5.2.4. Análisis del tipo de datos

La demostración de la hipótesis sobre la diferencia entre las funciones espacio y tiempo de los algoritmos de agrupación y de búsqueda de reglas de asociación utilizados en AEI y LA, permitirá que el docente pueda resolver más rápidamente los problemas educativos que utilizan datos masivos y de diferente tipo (binarios, modales, numéricos), al seleccionar los algoritmos más óptimos. Esto implica que los docentes obtendrán los resultados con mayor rapidez (por utilizar la mejor técnica en el análisis de datos masivo que se obtuvo del cuasi-experimento), al ser de gran beneficio para poder actuar y tomar las medidas necesarias dentro del ámbito educativo. Los datos para los cuales se realizarán las pruebas serán aquellos que son utilizados en la mayoría de casos en el ámbito educativo, es decir de tipo binario (si/no, verdadero/falso, admite/no admite, con propiedad/sin propiedad, aprueba/desaprueba, admitido/no admitido, promovido/no promovido, motivado/desmotivado), modal (bajo/medio/alto, malo/regular/bueno/muy bueno/sobresaliente, nada de acuerdo/poco de acuerdo / ni en acuerdo ni en desacuerdo / muy de acuerdo / completamente de acuerdo) y numérico (notas, rendimiento, asistencia, número de faltas), para nuestro diseño cuasi-experimental serán generados aleatoriamente, mediante líneas de código en *software* R.

5.2.5. Selección de la prueba estadística a utilizar

ANOVA no paramétrico, se utiliza en la comprobación de una hipótesis para comparar las medianas de dos o más poblaciones cuyas distribuciones tienen la misma forma e igual varianza. Otro factor que nos indica que se puede aplicar está prueba es que se utiliza cuando cada muestra es mayor a 20.

Al desear comparar la velocidad de procesamiento y uso de memoria, entre los métodos de jerarquización (cluster) y reglas de asociación, además de tener una muestra mayor a 20 (muestra 383), ANOVA no paramétrica es la más adecuada para comprobar nuestra hipótesis.

5.2.6. Comprobación de supuestos

Antes de iniciar la elaboración del diseño cuasi experimental, es muy importante tener en cuenta aspectos como:

5.2.6.1. Selección de equipos informáticos

Para la obtención de equipos similares para la aplicación del diseño cuasi-experimental se siguió el siguiente procedimiento:

5.2.6.1.1. Análisis de componentes influyentes en la velocidad de procesamiento de un computador.

De acuerdo al estudio realizado sobre los componentes que brindan la velocidad de procesamiento a un computador, según (Patterson & Hennessy, 2004), estos son: el procesador, memoria RAM, tarjeta gráfica y en menor proporción el disco duro. El procesador a utilizar es (Intel, 2017), el mismo que es el cerebro del computador, realiza toda las tareas de procesamiento de información, además de contener a la memoria cache, la cual ayuda a mejorar la velocidad de procesamiento. La memoria RAM es la memoria principal la cual almacena temporalmente datos y programas que se están utilizado en un instante de tiempo. La tarjeta gráfica por su parte es la que procesa los datos provenientes del procesador y los convierte en información que sea visible a través del monitor. Si bien no es necesario almacenar una gran cantidad de información, será importante la velocidad de lectura y escritura de información lo cual es medido por revoluciones por minuto (rpm), que viene dado por los fabricantes.

5.2.6.1.2. Búsqueda de equipos similares

Los altos costos en equipos de gran rapidez de procesamiento, se puso de manifiesto, es así que se optó por escoger una gama media alta, los cuales son computadores con procesadores Core I5 (HP y Apple)

5.2.6.1.3. Análisis de parámetros de los equipos (ver tabla 18, 19 y 20)

Tabla 18. Cuadro comparativo nivel hardware

CUADRO COMPARATIVO NIVEL HARDWARE					
Características de comparación		**MacBook Pro (13Inch, Late 2011)**	**Notebook HP - 14-am012la**	**Similaridad**	**Justificación**
r o	**Microprocesador**	Intel Core i5	Intel® Core™ i5-6200U	100%	Iguales modelo de microprocesadores

	Generación	Segunda (Sandy Bridge - 2011)	Sexta (Skylake - 2015)	0%	Fecha de lanzamiento
	Número de núcleos	Doble núcleo	2 núcleos	100%	Poseen doble núcleo
	Velocidad de reloj	2.4 GHz	2.4GHZ	100%	Los dos procesadores utilizan la misma velocidad de procesamiento.
	Cache	3 MB de caché	3 MB de caché	100%	La memoria auxiliar de alta velocidad destinada a realizar copias de archivos para acceder más rápidamente, son iguales.
Memoria	**Capacidad**	4 GB dos módulos SO-DIMM de 2 GB	4 GB (1 x 4 GB)	100%	Igual capacidad
	Familia de memoria RAM	SDRAM (Synchronous Dynamic Random-Access Memory)	SDRAM (Synchronous Dynamic Random-Access Memory)	100%	Igual familia de RAM
	Tecnologías de memoria de acceso aleatorio	DDR3	DDR4	90%	Las diferencias es que puedes encontrar memorias DDR4 a mayores velocidades, pero en tema de rendimiento, por ejemplo, DDR3 2133 mhz vs DDR4 2133 mhz, existe una mínima diferencia de rendimiento
	Velocidad de memoria del reloj	1.333 MHz.	2133 MHz	65%	Mayor velocidad de transferencia de

					información la de 2133MHz.
Compatibilidad con gráficos y video	**Gráficos**	HD Graphics 3000 de Intel con 384 MB de SDRAM DDR3 compartida con la memoria principal	Intel HD Graphics 515 (24 EUs) 2104MB, compartida 128MB	80%	Intel HD 3000 (Desktop V1 1.1 GHz) - 392nd / 544 Intel HD 515 (Mobile Skylake) - 278th / 544
	Video	Cámara FaceTime HD, puerto Thunderbolt con compatibilidad para vídeo DVI, VGA, DVI de doble canal y HDMI (requiere adaptadores que se venden por separado)	Integrada compatibilidad VGA, HDMI	100%	Maneja VGA y HDMI, las más importantes
Almacenamiento	**Capacidad**	500GB	1TB	50%	HP posee el doble de capacidad de almacenamiento, pero no afecta la velocidad de procesamiento.
	Tipo de conexión	Serial ATA	Serial ATA	100%	Similar tipo de conexión del HD
	Velocidad de reloj	5.400 rpm	5.400 rpm	100%	Igual velocidad de lectura y escritura de datos
Unidad Óptica		SuperDrive a 8x (DVD±R de doble capa, DVD±RW y CD-RW)	Unidad óptica no incluida	0%	La notebook HP no viene incluido
Pantalla		Panorámica brillante de 13,3 pulgadas (en	WLED HD SVA BrightView de	92%	Al ser la HP de 14' va a tener mayor resolución

	diagonal) retroiluminada por LED con 1.280 por 800 píxeles de resolución	14" en diagonal (1366 x 768)		
Redes	Ethernet 10/100/1000BASE-T (Gigabit)	LAN Ethernet 10/100 BASE-T integrada	100%	Similar conexión, velocidad de trasmisión y acceso a la red
Redes inalámbricas	Conexión inalámbrica Wi-Fi (basada en la norma 802.11n del IEEE);2 Bluetooth 2.1 + EDR (Enhanced Data Rate)	Combinación 802.11b/g/n (1x1) y Bluetooth® 4.0	100%	Conexión inalámbrica de acuerdo a la norma 802.11b
Audio	Altavoces estéreo con refuerzo de graves, micrófono omnidireccional y entrada combinada de audio/auriculares (salida digital compatible)	DTS Studio Sound™ con 2 altavoces	90%	Cumplen la misma función pero con mayor salida de audio, no influye en la velocidad de procesamiento
Puertos USB	Dos puertos USB 2.0 (hasta 480 Mb/s)	1 USB 3.0; 2 USB 2.0	80%	La Notebook HP posee un puerto adicional USB 3.o, pero esto no influye en el procesamietno
Ranuras para tarjetas	SDXC (Extreme Capacity)	Lector de tarjetas SD multiformato	80%	Mac posee lector de alta capacidad, pero cumplen la misma función de uso.

Fuente: Elaboración propia

Tabla 19. Cuadro comparativo nivel *software*

CUADRO COMPARATIVO NIVEL SOFTWARE					
Características de comparación		**MacBook Pro (13Inch, Late 2011)**	**Notebook HP - 14-am012la**	**Similaridad**	**Justificación**
Sistema operativo	**Edición**	MacOS Sierra Versión 10.12.5	Windows 10 / Ubuntu 16.04	0%	Sistemas operativos diferentes
	Fabricante	Apple	Microsoft / Canonical Ltd. / Fundación Ubuntu	0%	Diferentes fabricantes
	Tipo	64 bits	64 bits	100%	Igual manera administra la información.
Software a utilizar	**R**	Versión 3.4.1	Versión 3.4.1	100%	Igual versión
	RStudio	Versión 1.0.143	Versión 1.0.143	100%	Igual versión
	Rchic	Versión 0.25	Versión 0.25	100%	Igual versión

Fuente: Elaboración propia

Luego de realizar un análisis de las características mostradas (ver tabla 18, 19 y 20), se toma en cuenta las que influyen al momento de realizar el experimento, es decir las que proporcionan la velocidad y procesamiento de la información, las mismas que deben ser similares. Por lo antes mencionado se toma en cuenta los componentes hardware, los mismo que son los que influyen en la velocidad de procesamiento, estos aspectos son:

Tabla 20. Cálculo de similaridad de equipos computacionales a utilizar

Características de comparación		MacBook Pro (13Inch, Late 2011)	Notebook HP - 14-am012la	Similaridad
Procesador	**Microprocesador**	Intel Core i5	Intel® Core™ i5-6200U	100%
	Generación	Segunda (Sandy Bridge - 2011)	Sexta (Skylake - 2015)	No influye
	Número de núcleos	Doble núcleo	2 núcleos	100%
	Velocidad de reloj	2.4 GHz	2.4GHZ	100%
	Cache	3 MB de caché	3 MB de caché	100%
Media de similaridad entre procesadores				100%
Memoria	**Capacidad**	4 GB dos módulos SO-DIMM de 2 GB	4 GB (1 x 4 GB)	100%
	Familia de memoria RAM	SDRAM (Synchronous Dynamic Random-Access Memory)	SDRAM (Synchronous Dynamic Random-Access Memory)	100%
	Tecnologías de memoria de acceso aleatorio	DDR3	DDR4	90%
	Velocidad de memoria del reloj	1.333 MHz.	2133 MHz	65%
Media de similaridad entre memorias				**88.75%**
	Gráficos	HD Graphics 3000 de Intel	Intel HD Graphics 515	80%

Compatibilidad con gráficos y vídeo		con 384 MB de SDRAM DDR3 compartida con la memoria principal	(24 EUs) 2104MB, compartida 128MB	
	Video	Cámara FaceTime HD, puerto Thunderbolt con compatibilidad para vídeo DVI, VGA, DVI de doble canal y HDMI (requiere adaptadores que se venden por separado)	Integrada compatibilidad VGA, HDMI	No influye
Media de similaridad entre tarjetas gráficas				**80%**
Almacenamiento	**Capacidad**	500GB	1TB	No influye
	Tipo de conexión	Serial ATA	Serial ATA	No influye
	Velocidad de reloj	5.400 rpm	5.400 rpm	100%
Media de similaridad entre HD				**100%**
Suma Total				**368.75**
Media de similaridad componentes influyentes en velocidad y procesamiento de un computador:				92.19%

Fuente: Elaboración propia

Luego del análisis de los componentes de cada uno de los equipos computacionales a utilizar **MacBook Pro (13Inch, Late 2011) y Notebook HP - 14-am012la,** se puede dar cuenta que existe un 92,19% de similaridad de sus componentes, todo esto del análisis de tecnología y fabricante utilizada en los componentes que brindan la velocidad de procesamiento a un computador, los cuales son procesador, tarjeta gráfica y memoria entre los más importantes respectivamente.

5.2.7. Ejecución del experimento

Una vez verificado la similaridad de la velocidad de procesamiento entre los equipos a utilizar MacBook Pro (13Inch, Late 2011) y Notebook HP - 14-am012la, se procede a su desarrollo, que se detalla a continuación:

5.2.7.1. Instalación sistemas operativos *software*

En equipo informático Notebook HP -14-am012la – Core I5, se realizó una instalación dual, al contar este equipo con un sistema operativo Windows 10 de 64 bits, se instaló el sistema operativo Ubuntu 16.04 de 64 bits, y tener los dos sistemas operativos en perfecto funcionamiento para ser ejecutados sobre los mismos componentes hardware. Por otra parte, solo se verificó el estado de la MacBook Pro, la cual posee un sistema operativo original MacOS Sierra Versión 10.12.5

5.2.7.2. Instalación y configuración herramientas *software* estadísticas

Se procede a la instalación de las herramientas *software* R y RStudio (ver apéndice A), en los tres sistemas operativos, Windows, Ubuntu y Mac; a considerar el proceso respectivo en cada uno.

5.2.7.3. Manejo y uso de funciones R

Se procede a realizar un análisis de las diferentes funciones que nos permitan, comprobar velocidad de procesamiento y uso de memoria, mediante la herramienta *software* R. En la sección materiales y herramientas se puede ver el funcionamiento de cada una de las funciones más importantes utilizadas para la implementación de los algoritmos que utilizan métodos cluster, para el cálculo de la velocidad como para el uso de la memoria. Estas funciones son:

Learning Analytics

- ✓ Cluster (diana)
- ✓ Fastcluster (hclust_vector)
- ✓ CluMix (dendro_variables)

Análisis estadístico implicativo

- ✓ Rchic
  - o callHierarchyTree
  - o callSimilarityTree

Velocidad de procesamiento

- ✓ Microbenchmark: Mide la velocidad de procesamiento de un grupo de funciones.
- ✓ ClValid : Validación estadística y biológica de los resultados de la agrupación
- ✓ Factoextra: Ofrece algunas funciones fáciles de usar para extraer y visualizar la salida de análisis de datos multivariantes

Uso memoria

- ✓ Gc: Imprime estadísticas de uso de memoria

De igual forma las funciones necesarias para la implementación del algoritmo para reglas de asociación, para el cálculo de la velocidad y uso de memoria, las cuales son:

Learning Analytics

- ✓ Aureles
  - o Apriori
  - o Eclat
  - o Weclat
- ✓ ggplot2
  - o autoplot
  - o boxplot
  - o plot

Análisis estadístico implicativo

- ✓ Rchic
  - o implicativeGraph

Velocidad de procesamiento

- ✓ Microbenchmark: Mide la velocidad de procesamiento de un grupo de funciones.
- ✓ ClValid : Validación estadística y biológica de los resultados de la agrupación
- ✓ Factoextra: Ofrece algunas funciones fáciles de usar para extraer y visualizar la salida de análisis de datos multivariantes

Uso memoria

- ✓ Gc: Imprime estadísticas de uso de memoria

5.2.7.4. Diseño e implementación del algoritmo en *software* R

Una vez realizada la investigación de las diferentes funciones, se procede al diseño de los algoritmos (ver apéndice C1, C2, C3):

a. Algoritmo calculo velocidad de procesamientos método cluster
b. Algoritmo calculo uso memoria método cluster
c. Algoritmo calculo velocidad de procesamientos y uso de memoria reglas de asociación

5.2.7.5. Ejecución del algoritmo y recolección de información

Una vez terminada la aplicación se procede a la ejecución de cada uno de los algoritmos, los mismo que arrojan los resultados en archivos .csv (ver apéndice C.4). Cada ejecución genera un archivo, el

cual contiene 3 repeticiones por cada función. De acuerdo a la muestra cada algoritmo se ejecutó 383 veces, con un total de 1532 archivos generados como resultados, distribuidos de la siguiente forma:

Tabla 21. Recolección de información

Métodos	**Archivos generados**		
	Velocidad de procesamiento	**Uso memoria**	**Total**
Jerarquización (cluster)	383	383	**766**
Reglas de asociación	383	383	**766**
Total	**766**	**766**	**1532**

Fuente: Elaboración propia

5.2.7.6. Agrupación de datos por sistemas operativos, método y velocidad - uso de memoria

A continuación de contar con los 1532 archivos .csv, se procede a clasificar la información en un solo archivo Excel de la siguiente forma:

Tabla 22. Agrupación de datos

Métodos	**Archivos generados**		
	Velocidad de procesamiento	**Uso memoria**	**Total**
Windows	1	1	**2**
Ubuntu	1	1	**2**
Mac	1	1	2
Total	**3**	**3**	**6**

Fuente: Elaboración propia

5.2.7.7. Análisis de datos

Las hipótesis estadísticas que se demostrará bajo normalidad según en test de Anderson-Darling, test de hipótesis de Kruskal- Wallis y su respectivo post test.

Para demostrar las hipótesis se planteó un cuasi-experimento en la ingeniería de *software* de tipo RGXO1. Donde RG representa el grupo aleatorio del grupo experimental (tanto inter como intra grupos), X representa el tratamiento que en este caso son los 3 métodos de cluster jerárquicos Learning Analytics hclust.vector, dendro.variables y diana y las 2 técnicas AEI implementadas mediante las funciones callHierarchyTree y callSimilarityTree. Y por otra parte 3 métodos para reglas de asociación LA: apriori, eclat y weclat y una 1 técnica AEI implementada mediante la función implicativeGraph. Se trabajó un nivel de significancia del 95%. La variable dependiente fue el espacio de memoria ocupado (en kilobytes) que es de tipo numérico.

La hipótesis estadística por demostrar para cluster jerárquicos se muestra a continuación:

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4 = \tilde{\mu}_5$$
$$H_1: \exists i, j \in \{1,2,3,4,5\} / \tilde{\mu}_i \neq \tilde{\mu}_j$$

y para reglas de asociación es:

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4$$
$$H_1: \exists i, j \in \{1,2,3,4\} / \tilde{\mu}_i \neq \tilde{\mu}_j$$

En base a la parte teórica, se procede a recolectar los datos, mediante el diseño cuasi-experimental y comprobación de la hipótesis, mediante los siguientes procesos:

5.2.7.7.1. Pruebas de normalidad

Para aplicar las pruebas de normalidad, es necesario realizar un diseño descriptivo de los datos, que parten de la información obtenida de la agrupación de datos por sistema operativo, método y velocidad – uso de memoria; los cuales se muestran a continuación:

Clustering – Uso de memoria

```
OPERATING.SYSTEM   FILE.NAME         NDATA            ROWS           COLUMNS         REPETITIONS
 MAC    :5745     out34  :   51   Min.   :   55   Min.   :  8.00   Min.   :   5.0   Min.   :1
 Ubuntu :5745     out1   :   45   1st Qu.: 9884   1st Qu.: 31.00   1st Qu.: 273.0   1st Qu.:1
 Windows:5745     out10  :   45   Median :22440   Median : 53.00   Median : 498.0   Median :2
                  out100 :   45   Mean   :27796   Mean   : 54.09   Mean   : 506.7   Mean   :2
                  out101 :   45   3rd Qu.:40700   3rd Qu.: 78.00   3rd Qu.: 750.0   3rd Qu.:3
                  out102 :   45   Max.   :96723   Max.   :100.00   Max.   :1000.0   Max.   :3
                  (Other):16959
            CLUSTER.METHODS    NDATA.1        MEMORY.MB
 dendro_diana(T)    :3447   Min.   :   55   Min.   :103.0
 dendro_variables(T):3447   1st Qu.: 9884   1st Qu.:145.0
 hclust_vector(T)   :3447   Median :22440   Median :249.0
 hrarchy(f)         :3447   Mean   :27796   Mean   :225.2
 simlrty(f)         :3447   3rd Qu.:40700   3rd Qu.:280.0
                            Max.   :96723   Max.   :384.0
```

Clustering – Velocidad

```
ï..OPERATING.SYSTEM   FILE.NAME         NDATA            ROWS           COLUMNS         REPETITIONS
 MAC    :5745        out34  :   51   Min.   :  136   Min.   :  8.0   Min.   :  4.0   Min.   :1
 Ubuntu :5745        out1   :   45   1st Qu.: 7750   1st Qu.: 33.0   1st Qu.:200.0   1st Qu.:1
 Windows:5745        out10  :   45   Median :20468   Median : 54.0   Median :437.0   Median :2
                     out100 :   45   Mean   :26057   Mean   : 55.7   Mean   :461.9   Mean   :2
                     out101 :   45   3rd Qu.:39711   3rd Qu.: 79.0   3rd Qu.:719.0   3rd Qu.:3
                     out102 :   45   Max.   :89965   Max.   :100.0   Max.   :999.0   Max.   :3
                     (Other):16959
           CLUSTER.METHODS    NDATA.1            TIME              TIME2
 dendro_diana(T)    :3447   Min.   :  136   1113548125:    2   0,069639651:    2
 dendro_variables(T):3447   1st Qu.: 7750   120845603 :    2   0,090296282:    2
 hclust_vector(T)   :3447   Median :20468   124768374 :    2   0,120845603:    2
 hrarchy(f)         :3447   Mean   :26057   126953247 :    2   0,124768374:    2
 simlrty(f)         :3447   3rd Qu.:39711   127674826 :    2   0,126953247:    2
                            Max.   :89965   129031625 :    2   0,127674826:    2
                                            (Other)   :17223   (Other)    :17223
```

Reglas de asociación – Uso de memoria

```
ï..OPERATING.SYSTEM   FILE.NAME         NDATA            ROWS          COLUMNS     REPETITIONS    CLUST
ER.METHODS
 MAC    :4596        out34  :   42   Min.   :  420   Min.   : 4.0   Min.   : 23   Min.   :1    met_apr
iori:3447
 Ubuntu :4596        out1   :   36   1st Qu.: 9568   1st Qu.:37.0   1st Qu.:253   1st Qu.:1    met_ASI
    :3447
 Windows:4596        out10  :   36   Median :25308   Median :56.0   Median :493   Median :2    met_ecl
at  :3447
                     out100 :   36   Mean   :27935   Mean   :55.5   Mean   :488   Mean   :2    met_wec
lat :3447
                     out101 :   36   3rd Qu.:39930   3rd Qu.:75.0   3rd Qu.:703   3rd Qu.:3
                     out102 :   36   Max.   :93248   Max.   :99.0   Max.   :996   Max.   :3
                     (Other):13566

    NDATA.1        MEMORY.MB        X             X.1
 Min.   :  420   166,3  :  220   Mode:logical   Mode:logical
 1st Qu.: 9568   162,4  :  206   NA's:13788     NA's:13788
 Median :25308   161,6  :  199
 Mean   :27935   156,6  :  182
```

```
 3rd Qu.:39930   158,8  :  170
 Max.   :93248   156,2  :  167
                 (Other):12644
```

Reglas de asociación – Velocidad

```
OPERATING.SYSTEM   FILE.NAME         NDATA            ROWS          COLUMNS       REPETITIONS    CLUST
ER.METHODS
 MAC    :4596     out34  :   42   Min.   :  420   Min.   : 4.00   Min.   : 23.0   Min.   :1    met_apr
iori:3447
 Ubuntu :4596     out1   :   36   1st Qu.: 9568   1st Qu.:37.00   1st Qu.:253.0   1st Qu.:1    met_ASI
   :3447
 Windows:4596     out10  :   36   Median :25308   Median :56.00   Median :493.0   Median :2    met_ecl
at  :3447
                  out100 :   36   Mean   :28001   Mean   :55.55   Mean   :488.5   Mean   :2    met_wec
lat :3447
                  out101 :   36   3rd Qu.:39930   3rd Qu.:75.00   3rd Qu.:703.0   3rd Qu.:3
                  out102 :   36   Max.   :93248   Max.   :99.00   Max.   :996.0   Max.   :3
                  (Other):13566

    NDATA.1             TIME.1             TIME2
 Min.   :  420   101095877  :    3   0,071094183:    3
 1st Qu.: 9568   102155402  :    3   0,071677013:    3
 Median :25308   109699048  :    3   0,072651681:    3
 Mean   :28001   110393493  :    3   0,073238557:    3
 3rd Qu.:39930   110786774  :    3   0,075933341:    3
 Max.   :93248   12262194239:    3   0,076484072:    3
                 (Other)    :13770   (Other)    :13770
```

5.2.7.7.1.1. Factores de prueba de la hipótesis - Sistema Operativo óptimo

Una vez recolectado la información sobre el uso de memoria y tiempos de velocidad de procesamiento en los sistemas operativos (Windows, Ubuntu y Mac) de acuerdo al agrupamiento de datos (ver tablas 21 y 22), se procede al análisis de sistema operativo más óptimo para la ejecución de estos procesos, para lo cual se realiza un análisis de varianza (ANOVA) no paramétrico, con el uso del *software* R y las funciones t2way y bwtrim. La primera función (t2way) calcula un ANOVA de dos vías para medios recortados con efectos de interacción y la segunda función (bwtrim) calcula un ANOVA de sujetos bidireccionales dentro de los medios recortados. Permite una variable entre sujetos y una dentro de la variable de sujetos.

Ho: No se observa diferencia entre los datos de ocupación de memoria y velocidad de procesamiento en los SO (Windows, Ubuntu, Mac).

H1: Se observa diferencia entre los datos de ocupación de memoria y velocidad de procesamiento en los SO (Windows, Ubuntu, Mac).

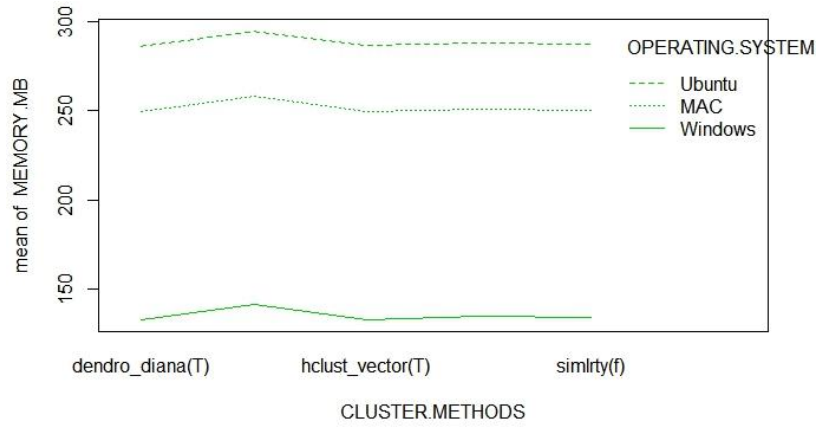
Clustering - memoria

```
t2way(MEMORY.MB ~ OPERATING.SYSTEM*CLUSTER.METHODS, data = df)
Call:
t2way(formula = MEMORY.MB ~ OPERATING.SYSTEM * CLUSTER.METHODS,
    data = df)

                                      value p.value
OPERATING.SYSTEM                339882.0970   0.001
CLUSTER.METHODS                    743.1692   0.001
OPERATING.SYSTEM:CLUSTER.METHODS     2.5033   0.962
```

```
> bwtrim(MEMORY.MB ~ OPERATING.SYSTEM*CLUSTER.METHODS, OPERATING.SYSTEM
 , data = df)
```

```
Call:
bwtrim(formula = MEMORY.MB ~ OPERATING.SYSTEM * CLUSTER.METHODS,
    id = OPERATING.SYSTEM, data = df)

                                        value p.value
CLUSTER.METHODS                       108.3924  0.0000
OPERATING.SYSTEM                    312375.1788  0.0000
CLUSTER.METHODS:OPERATING.SYSTEM         0.5636  0.8084
```

Figura 18. Uso de memoria de los SO, al aplicar técnicas clustering



Fuente: Elaboración propia

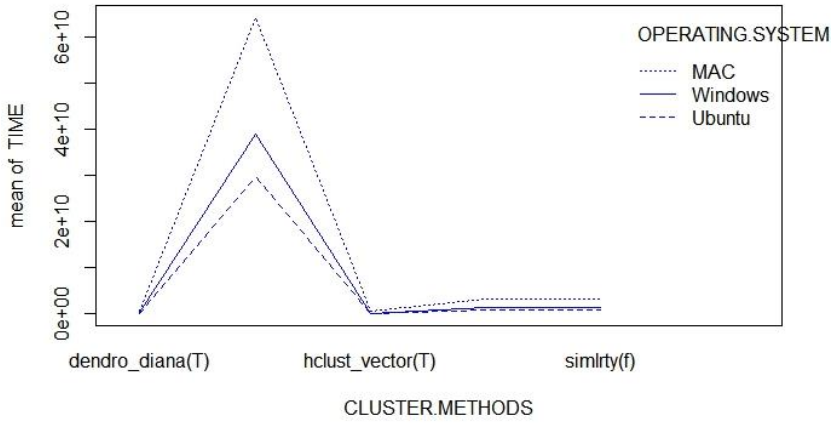
Al ejecutar el algoritmo 383 en cada sistema operativo, para obtener el uso de memoria, de acuerdo a la muestra, se puede visualizar (ver figura 18), los resultados obtenidos al aplicar ANOVA no paramétrica, donde se puede notar que el SO Ubuntu es el que más memoria utiliza para los procesos y el que menos utiliza es el SO Windows, donde se debe tener en cuenta que a mayor uso de memoria menor tiempos de respuesta.

Clustering - velocidad

```
> t2way(TIME ~ OPERATING.SYSTEM*CLUSTER.METHODS, data = df)
Call:
t2way(formula = TIME ~ OPERATING.SYSTEM * CLUSTER.METHODS, data = df)

                                      value p.value
OPERATING.SYSTEM                    140.8029   0.001
CLUSTER.METHODS                   11346.4923   0.001
OPERATING.SYSTEM:CLUSTER.METHODS   2949.4151   0.001
> bwtrim(TIME ~ OPERATING.SYSTEM*CLUSTER.METHODS, OPERATING.SYSTEM , da
ta = df)
Call:
bwtrim(formula = TIME ~ OPERATING.SYSTEM * CLUSTER.METHODS, id = OPERAT
ING.SYSTEM,
    data = df)
                                        value p.value
```

```
CLUSTER.METHODS                    2747.7487       0
OPERATING.SYSTEM                     65.0336       0
CLUSTER.METHODS:OPERATING.SYSTEM    344.0993       0
```

Figura 19. Velocidad de procesamiento de los SO, al aplicar técnicas clustering



Fuente: Elaboración propia

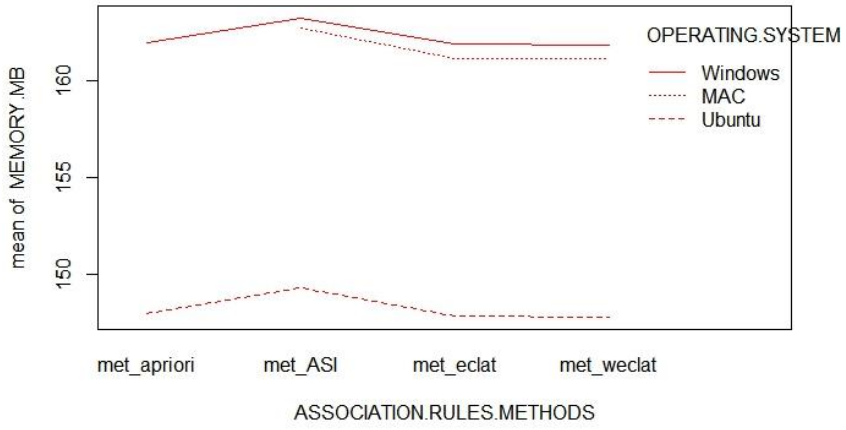
Al ejecutar el algoritmo 383 en cada sistema operativo, para obtener los tiempos de respuesta (velocidad de procesamiento), de acuerdo a la muestra, se puede visualizar (ver figura 19), los resultados obtenidos al aplicar ANOVA no paramétrica, se puede notar que el SO Ubuntu es el que procesa más rápido, al presentar en menor tiempo los resultados, todo esto gracias a mayor uso de memoria, la cual es asignada por el sistema operativo.

Reglas de asociación - memoria

```
> t2way(MEMORY.MB ~ OPERATING.SYSTEM*ASSOCIATION.RULES.METHODS, data =
df)
Call:
t2way(formula = MEMORY.MB ~ OPERATING.SYSTEM * ASSOCIATION.RULES.METHOD
S,
    data = df)
                                              value p.value
OPERATING.SYSTEM                          11282.8290   0.001
ASSOCIATION.RULES.METHODS                    46.5219   0.001
OPERATING.SYSTEM:ASSOCIATION.RULES.METHODS     0.3869   0.999
> bwtrim(MEMORY.MB ~ OPERATING.SYSTEM*ASSOCIATION.RULES.METHODS, OPERAT
ING.SYSTEM , data = df)
Call:
bwtrim(formula = MEMORY.MB ~ OPERATING.SYSTEM * ASSOCIATION.RULES.METHO
DS,
    id = OPERATING.SYSTEM, data = df)
                                            value p.value
ASSOCIATION.RULES.METHODS                 14.9668  0.0000
OPERATING.SYSTEM                        4934.7371  0.0000
```

```
ASSOCIATION.RULES.METHODS:OPERATING.SYSTEM     0.0598   0.9992
```

Figura 20. Uso de memoria de los SO, al aplicar técnicas de reglas de asociación



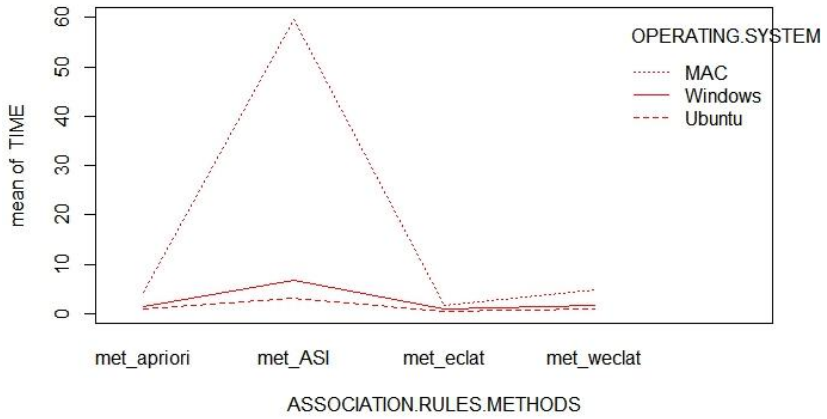
Fuente: Elaboración propia

Al ejecutar el algoritmo 383 en cada sistema operativo, para obtener el uso de memoria, de acuerdo a la muestra, se puede visualizar (ver figura 20), los resultados obtenidos al aplicar ANOVA no paramétrica, donde se puede notar que el SO Ubuntu es el que menos memoria utiliza para los procesos, se concluye que el SO Ubuntu es administra de mejor manera los procesos para su ejecución, como se verá en la figura 19, es el que tarda menos en entregar los resultados.

Reglas de asociación - velocidad

```
> t2way(TIME ~ OPERATING.SYSTEM*ASSOCIATION.RULES.METHODS, data = df)
Call:
t2way(formula = TIME ~ OPERATING.SYSTEM * ASSOCIATION.RULES.METHODS,
    data = df)
                                           value p.value
OPERATING.SYSTEM                         15817.33   0.001
ASSOCIATION.RULES.METHODS                20208.68   0.001
OPERATING.SYSTEM:ASSOCIATION.RULES.METHODS 15660.17   0.001
> bwtrim(TIME ~ OPERATING.SYSTEM*ASSOCIATION.RULES.METHODS, OPERATING.S
YSTEM , data = df)
Call:
bwtrim(formula = TIME ~ OPERATING.SYSTEM * ASSOCIATION.RULES.METHODS,
    id = OPERATING.SYSTEM, data = df)
                                           value p.value
ASSOCIATION.RULES.METHODS                6716.114       0
OPERATING.SYSTEM                         7872.953       0
ASSOCIATION.RULES.METHODS:OPERATING.SYSTEM 2594.067       0
```

Figura 21. Velocidad de procesamiento de los SO, al aplicar técnicas de reglas de asociación



Fuente: Elaboración propia

Al ejecutar el algoritmo 383 en cada sistema operativo, para obtener los tiempos de respuesta (velocidad de procesamiento), se puede visualizar (ver figura 21), los resultados obtenidos al aplicar ANOVA no paramétrica, donde se puede notar que el SO Ubuntu es el que presenta los resultados en menor tiempo.

Ho: No se observa diferencia entre los datos de ocupación de memoria y velocidad de procesamiento en los SO (Windows, Ubuntu, Mac).

H1: Se observa diferencia entre los datos de ocupación de memoria y velocidad de procesamiento en los SO (Windows, Ubuntu, Mac).

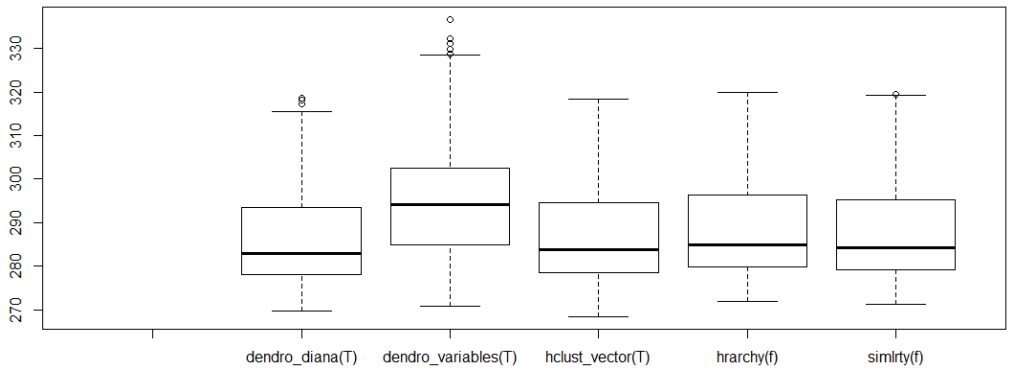
De los resultados obtenidos se rechaza categóricamente la hipótesis nula (Ho), y se acepta la hipótesis alternativa (H1), al observa diferencia entre los datos de ocupación de memoria y velocidad de procesamiento en los SO (Windows, Ubuntu, Mac). De esta forma se concluye que el sistema operativo que mejor aprovecha los recursos del computador y que entrega más rápido, en menor tiempo los resultados es el SO Ubuntu.

De esta forma se procede a realizar la hipótesis de normalidad, con los datos arrojados por el SO Ubuntu, para verificar la técnica más optima entre las técnicas del análisis estadístico implicativo (AEI) y learning analytics (LA).

5.2.7.7.1.2. Hipótesis de Normalidad Técnica clustering – uso de memoria

Se procedió a realizar un gráfico de cajas y alambres comparativo para cada uno de los 5 métodos analizados, éste se muestra a continuación en la Figura 22.

Figura 22. Gráfico comparativo de cajas y alambres



Fuente: Elaboración propia

La tabla 23, muestra un cuadro comparativo entre las medidas de centralización, dispersión y el tamaño de muestra utilizado para cada uno de los métodos analizados.

Tabla 23. Cuadro comparativo entre métodos cluster su media, desviación estándar y el tamaño de muestra

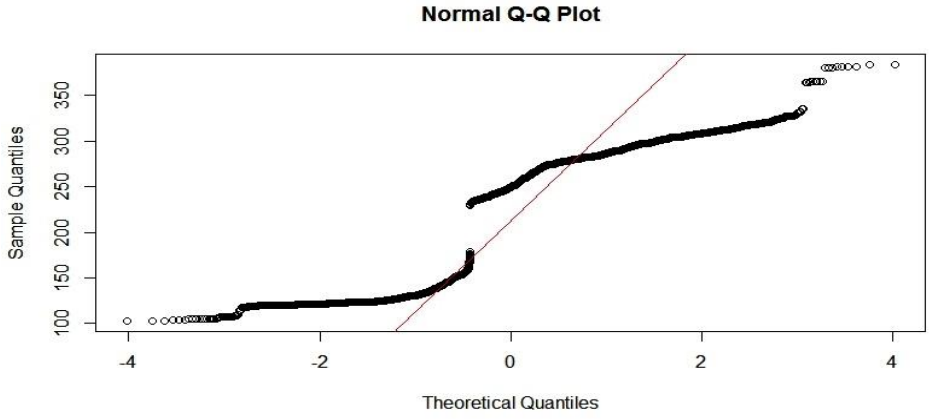
	dendro_diana (T)	**dendro_variables(T)**	**hclust_vector (T)**	**Hrarchy (f)**	**Simlrty (f)**
Centralización	10,9833599	12,7246268	10,96933278	11,1253011	11,056894
Dispersión					
Tamaño	3447	3447	3447	3447	3447

Fuente: Elaboración propia

Comprobación de supuestos

Para determinar el test apropiado a utilizar se procedió a la comprobación de los supuestos. A continuación, se muestra la gráfica de cuartiles que nos dio una idea gráfica sobre la normalidad de los datos sobre la memoria ocupada por los diferentes métodos.

Figura 23. Gráfico de cuartiles



Fuente: Elaboración propia

Las hipótesis estadísticas sobre normalidad se muestran a continuación:

Ho: No se observa diferencia entre los datos de ocupación de memoria y la distribución normal

H1: Se observa diferencia entre los datos de ocupación de memoria y la distribución normal

Para su demostración se utilizaron los test de Anderson-Darling y Crames-von Mises, los resultados se muestran a continuación:

Tabla 24. Resultados de los test de Normalidad

TEST DE NORMALIDAD	ESTADÍSTICO	P-VALUE
Anderson-Darling	A = 1224	< 2.2e-16
Cramer-von Mises	W = 212.03	=7.37e-10

Fuente: Elaboración propia

En todos los test se puede observar que se rechaza la hipótesis nula.

Prueba de hipótesis

La hipótesis estadística por demostrar se muestra a continuación:

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4 = \tilde{\mu}_5$$
$$H_1: \exists i, j \in \{1,2,3,4,5\} / \tilde{\mu}_i \neq \tilde{\mu}_j$$

Se utilizó el test de hipótesis no paramétrico suma de rangos para muestras independientes de Kruskal-Wallis, los resultados entregados por la función kruskal.test(x, y) del paquete estándar del *software* R fueron los siguientes :

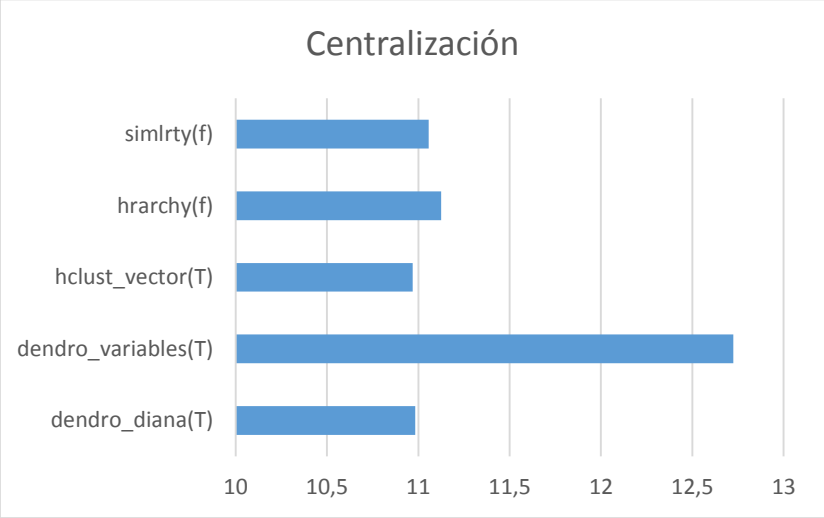
```
        Kruskal-Wallis rank sum test
data:  x and Y
Kruskal-Wallis chi-squared = 387.59, df = 4, p-value < 2.2e-16
```

Para determinar los grupos de homogeneidad se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes. Se obtuvieron las siguientes salidas:

```
                   dendro_diana(T) dendro_variables(T) hclust_vector(T)
dendro_variables(T) < 2e-16        -                   -
hclust_vector(T)    0.12737        < 2e-16             -
hrarchy(f)          1.2e-07        < 2e-16             0.00021
simlrty(f)          0.00609        < 2e-16             0.12737
           hrarchy(f)
dendro_variables(T) -
hclust_vector(T)    -
hrarchy(f)          -
simlrty(f)          0.04282
```

con las salidas respectivas se elaboró la gráfica de la Figura 22.

Figura 24. Grupos de homogeneidad – clustering memoria



Fuente: Elaboración propia

Discusión

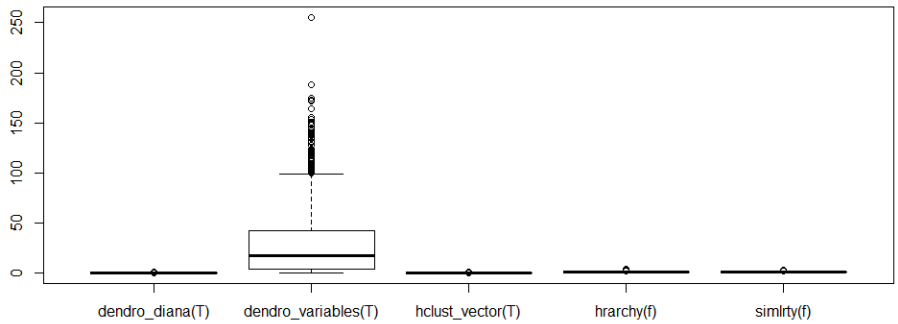
El gráfico de cajas y alambres nos muestra una homogeneidad en la dispersión de los 3 métodos de Learning Analytics y también en los de las técnicas del Análisis estadístico Implicativo, pero curiosamente las técnicas AEI son más homogéneas. En cuanto a las medidas de centralización se puede observar que aparentemente la ocupación de memoria es similar entre los 5 métodos, con una aparente mayor ocupación del método dendro.variables. Antes de realizar la prueba de hipótesis, se procedió a comprobar sus supuestos. Se realizó la prueba de normalidad de Anderson-Darling que nos dio un p-valor de 2.2e-16 indicándonos que se debe rechazar la hipótesis nula y que por tanto los datos no han sido extraídos de una población normal, este resultado se corrobora con el gráfico de cuartiles que muestra un gran alejamiento de la distribución de datos a los cuartiles teóricos. El no cumplimiento de este supuesto es suficiente para optar por los test no paramétricos. El test no paramétrico seleccionado es el test de hipótesis no paramétrico de suma de rangos para muestras independientes de Kruskal-Wallis que nos entrega un valor de chi cuadrado y un p-valor de 2.2e-16, que nos indica con un alto valor de significancia que se rechaza la hipótesis nula y por lo tanto al menos un par de los 5 métodos clusters son diferentes. Con el objetivo de determinar la relación entre los pares se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes con el cual se ha construido el gráfico de homogeneidad de medidas de centralización que nos indica que hay cuatro grupos bien definidos, de los cuales nos llama la atención el método dendro.variables que es el que más ocupación de memoria tiene y además nos indica que los métodos callSimilarityTree y callHierarchyTree (simlrty, hrarchy) del análisis estadístico Implicativo y los métodos, dendro.diana y hclust_vector de Learning Analytics son los que

menos ocupación de memoria tienen, definiéndose como el más óptimo hclust_vector con 10,96933278

5.2.7.7.1.3. Clustering – velocidad

Se procedió a realizar un gráfico de cajas y alambres comparativo para cada uno de los 5 métodos analizados, éste se muestra a continuación en la Figura 25.

Figura 25. Gráfico comparativo de cajas y alambres



Fuente: Elaboración propia

La tabla 25, muestra un cuadro comparativo entre las medidas de centralización, dispersión y el tamaño de muestra utilizado para cada uno de los métodos analizados.

Tabla 25. Cuadro comparativo entre métodos cluster su media, desviación estándar y el tamaño de muestra

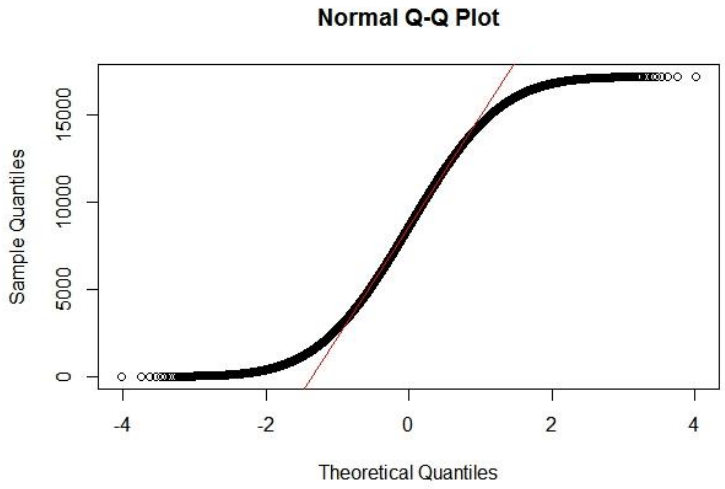
	dendro_diana (T)	**dendro_variables (T)**	**hclust_vector (T)**	**Hrarchy (f)**	**Simlrty (f)**
Centralización	0,08301572	34,9213604	0,059625645	0,44441203	0,47690769
Dispersión					
Tamaño	3447	3447	3447	3447	3447

Fuente: Elaboración propia

Comprobación de supuestos

Para determinar el test apropiado a utilizar se procedió a la comprobación de los supuestos. A continuación, se muestra la gráfica de cuartiles que nos dio una idea gráfica sobre la normalidad de los datos sobre la memoria ocupada por los diferentes métodos.

Figura 26. Gráfico de cuartiles



Fuente: Elaboración propia

Las hipótesis estadísticas sobre normalidad se muestran a continuación:

Ho: No se observa diferencia entre los tiempos de respuesta (velocidad de procesamiento) y la distribución normal

H1: Se observa diferencia entre los tiempos de respuesta (velocidad de procesamiento) y la distribución normal

Para su demostración se utilizaron los test de Anderson-Darling y Crames-von Mises, los resultados se muestran a continuación:

Tabla 26. Resultados de los test de Normalidad

TEST DE NORMALIDAD	ESTADÍSTICO	P-VALUE
Anderson-Darling	A = 191.3	< 2.2e-16
Cramer-von Mises	W = 26.226	=7.37e-10

Fuente: Elaboración propia

En todos los test se puede observar que se rechaza la hipótesis nula.

Prueba de hipótesis

La hipótesis estadística por demostrar se muestra a continuación:

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4 = \tilde{\mu}_5$$
$$H_1: \exists i, j \in \{1,2,3,4,5\} / \tilde{\mu}_i \neq \tilde{\mu}_j$$

Se utilizó el test de hipótesis no paramétrico suma de rangos para muestras independientes de Kruskal-Wallis, los resultados entregados por la función kruskal.test(x, y) del paquete estándar del *software* R fueron los siguientes :

```
        Kruskal-Wallis rank sum test
data:  x and y
```

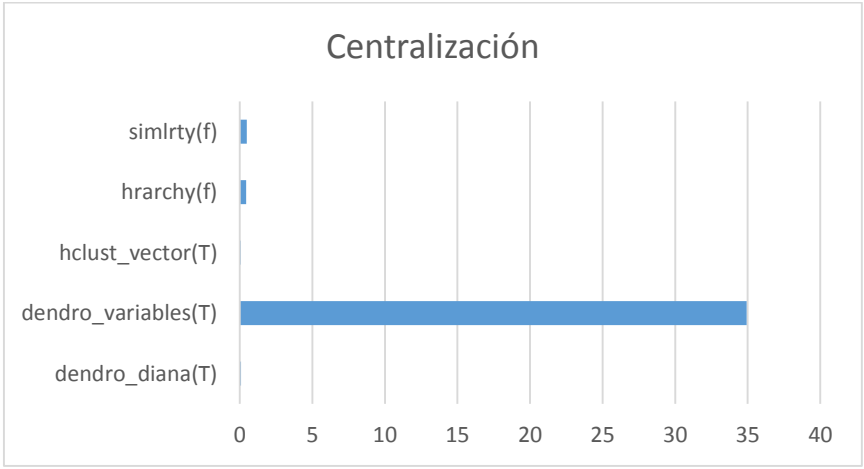
```
Kruskal-Wallis chi-squared = 4613.3, df = 4, p-value < 2.2e-16
```

Para determinar los grupos de homogeneidad se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes. Se obtuvieron las siguientes salidas:

```
                     dendro_diana(T) dendro_variables(T) hclust_vector(T)
dendro_variables(T)       110.71236                  NA               NA
hclust_vector(T)           21.82916           132.54152               NA
hrarchy(f)                 65.37835            45.33401         87.20751
simlrty(f)                 67.07620            43.63616         88.90536
                     hrarchy(f)
dendro_variables(T)          NA
hclust_vector(T)             NA
hrarchy(f)                   NA
simlrty(f)             1.697845
```

con las salidas respectivas se elaboró la gráfica de la Figura 25.

Figura 27. Grupos de homogeneidad – clustering velocidad



Fuente: Elaboración propia

Discusión

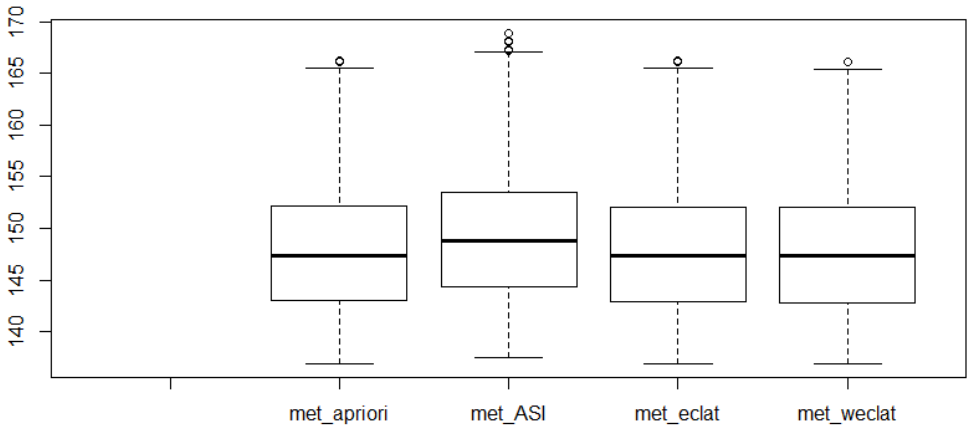
El gráfico de cajas y alambres muestra homogeneidad en la dispersión de los métodos de Learning Analytics (dendro_diana, dendro_variable, hclust_vector) y también en los de las técnicas del análisis estadístico Implicativo callHierarchyTree y callSimilarityTree (hrarchy, simlrty), además se puede ver que las técnicas AEI son más homogéneas. Por otra parte, al analizar las medidas de centralización se puede observar que los tiempos de respuesta (velocidad de procesamiento), las técnicas de AEI son similares entre ellas y de igual forma las técnicas de LA son similares, pero se ve que la técnica clustering dendro_variable tiene un aparente mayor tiempo de respuesta, lo cual concuerda con el análisis realizado en la técnica clustering de acuerdo al uso de memoria. Se procedió a comprobar supuestos, antes de realizar la prueba de hipótesis. La prueba de normalidad de Anderson-Darling, se hizo, y se tiene como p-valor 2.2e-16 indicándonos que se debe rechazar la hipótesis nula y que por tanto los datos no han sido extraídos de una población normal, lo cual es corroborado por el resultado

del gráfico de cuartiles, el cual muestra un gran alejamiento de la distribución de datos a los cuartiles teóricos. El no cumplimiento de este supuesto es suficiente para optar por los test no paramétricos. El test no paramétrico seleccionado es el test de hipótesis no paramétrico de suma de rangos para muestras independientes de Kruskal-Wallis que nos entrega un valor de chi cuadrado y un p-valor de 2.2e-16, que nos indica con un alto valor de significancia que se rechaza la hipótesis nula y por lo tanto al menos un par de los 5 métodos clusters son diferentes. Con el objetivo de determinar la relación entre los pares se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes con el cual se ha construido el gráfico de homogeneidad de medidas de centralización que nos indica que hay cuatro grupos bien definidos, de los cuales nos llama la atención el método dendro.variables que es el que tarde más en entregar los resultados, es decir tiene mayor tiempo de respuesta y además nos indica que los métodos callHierarchyTree y callSimilarityTree (hrarchy, simlrty) del análisis estadístico Implicativo los métodos, dendro_diana y hcluster_vector de Learning Analytics son los que menos ocupación de memoria tienen, al ser la más optima hcluster_vector la cual entrega resultados en el menor tiempo con 0,059625645, y menos óptima dendro_variable.

5.2.7.7.1.4. Reglas de asociación – uso de memoria

Se procedió a realizar un gráfico de cajas y alambres comparativo para cada uno de los 4 métodos analizados, éste se muestra a continuación en la Figura 28.

Figura 28. Gráfico comparativo de cajas y alambres



Fuente: Elaboración propia

La tabla 27, muestra un cuadro comparativo entre las medidas de centralización, dispersión y el tamaño de muestra utilizado para cada uno de los métodos analizados.

Tabla 27. Cuadro comparativo entre métodos reglas de asociación su media, desviación estándar y el tamaño de muestra

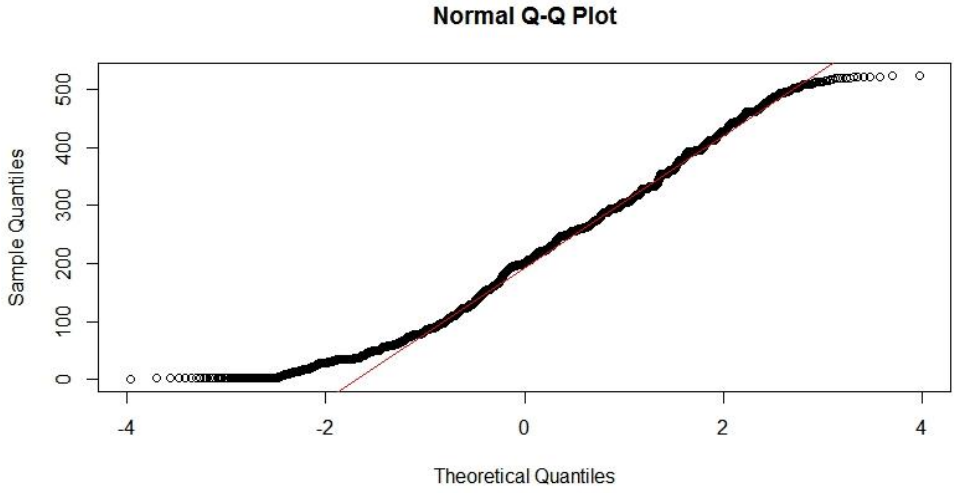
	met_apriori(T)	met_eclat(T)	met_weclat(T)	met_ASI(f)
Centralización	6,30144075	6,28895568	6,27645716	6,33584713
Dispersión				
Tamaño	3447	3447	3447	3447

Fuente: Elaboración propia

Comprobación de supuestos

Para determinar el test apropiado a utilizar se procedió a la comprobación de los supuestos. A continuación, se muestra la gráfica de cuartiles la cual da una idea gráfica sobre la normalidad de los datos sobre la memoria ocupada por los diferentes métodos.

Figura 29. Gráfico de cuartiles



Fuente: Elaboración propia

Las hipótesis estadísticas sobre normalidad se muestran a continuación:

Ho: No se observa diferencia entre los datos de ocupación de memoria y la distribución normal

H1: Se observa diferencia entre los datos de ocupación de memoria y la distribución normal

Para su demostración se utilizaron los test de Anderson-Darling y Crames-von Mises, los resultados se muestran a continuación:

Tabla 28. Resultados de los test de Normalidad

TEST DE NORMALIDAD	ESTADÍSTICO	P-VALUE
Anderson-Darling	A = 51.244	< 2.2e-16
Cramer-von Mises	W = 6.6416	=7.37e-10

Fuente: Elaboración propia

En todos los test se puede observar que se rechaza la hipótesis nula.

Prueba de hipótesis

La hipótesis estadística por demostrar se muestra a continuación:

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4$$
$$H_1: \exists i, j \in \{1,2,3,4\} / \tilde{\mu}_i \neq \tilde{\mu}_j$$

Se utilizó el test de hipótesis no paramétrico suma de rangos para muestras independientes de Kruskal-Wallis, los resultados entregados por la función kruskal.test(x, y) del paquete estándar del *software* R fueron los siguientes :

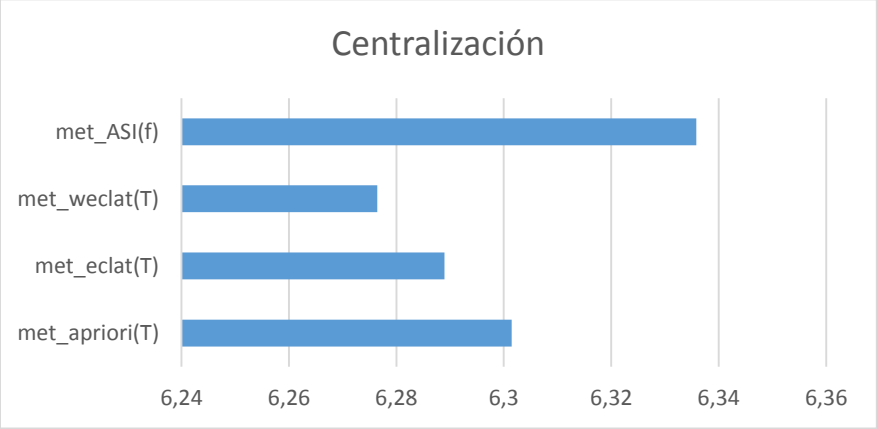
```
        Kruskal-Wallis rank sum test
data:  x and g
Kruskal-Wallis chi-squared = 45.14, df = 3, p-value =
8.641e-10
```

Para determinar los grupos de homogeneidad se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes. Se obtuvieron las siguientes salidas:

```
           met_apriori met_ASI met_eclat
met_ASI    2.7e-06     -       -
met_eclat  0.78        1.4e-07 -
met_weclat 0.78        3.2e-08 0.78
```

con las salidas respectivas se elaboró la gráfica de la Figura 28.

Figura 30. Grupos de homogeneidad – reglas de asociación memoria



Fuente: Elaboración propia

Discusión

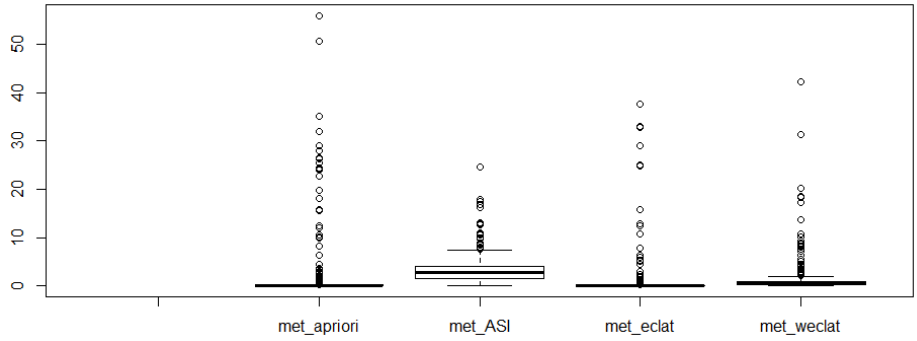
La homogeneidad en la dispersión, de acuerdo a los resultados del gráfico de cajas y alambres nos muestra que los 3 métodos de Learning Analytics (met_apriori, met_eclat y met_weclat) y también en el de las técnicas del Análisis estadístico Implicativo (met_ASI), son homogéneas curiosamente. Además, en lo referente a las medidas de centralización se observar que la ocupación de memoria es similar entre los 4 métodos, con una aparente mayor ocupación de memoria del método met_ASI. Al

comprobar los supuestos, antes de ejecutar la prueba de hipótesis. Se realizó la prueba de normalidad de Anderson-Darling que nos dio un p-valor de 2.2e-16 indicándonos que se debe rechazar la hipótesis nula y que por tanto los datos no han sido extraídos de una población normal, este resultado se corrobora con el gráfico de cuartiles que muestra un gran alejamiento de la distribución de datos a los cuartiles teóricos. El no cumplimiento de este supuesto es suficiente para optar por los test no paramétricos. El test no paramétrico seleccionado es el test de hipótesis no paramétrico de suma de rangos para muestras independientes de Kruskal-Wallis que nos entrega un valor de chi cuadrado y un p-valor de 8.641e-16, que nos indica con un alto valor de significancia que se rechaza la hipótesis nula y por lo tanto al menos un par de los 4 métodos de reglas de asociación son diferentes. Con el objetivo de determinar la relación entre los pares se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes con el cual se ha construido el gráfico de homogeneidad de medidas de centralización que nos indica que hay cuatro grupos bien definidos, de los cuales nos llama la atención el método met_ASI que es el que más ocupación de memoria tiene y además nos indica que las técnicas met_apriori, met_eclat y met_weclat de Learning Analytics son los que menos ocupación de memoria tienen, por lo cual se concluye que el más óptimo es el método weclat de LA con 6,27645716.

5.2.7.7.1.5. Reglas de asociación – velocidad

Se procedió a realizar un gráfico de cajas y alambres comparativo para cada uno de los 4 métodos analizados, éste se muestra a continuación en la Figura 31.

Figura 31. Gráfico comparativo de cajas y alambres



Fuente: Elaboración propia

La tabla 29, muestra un cuadro comparativo entre las medidas de centralización, dispersión y el tamaño de muestra utilizado para cada uno de los métodos analizados.

Tabla 29. Cuadro comparativo entre métodos reglas de asociación su media, desviación estándar y el tamaño de muestra

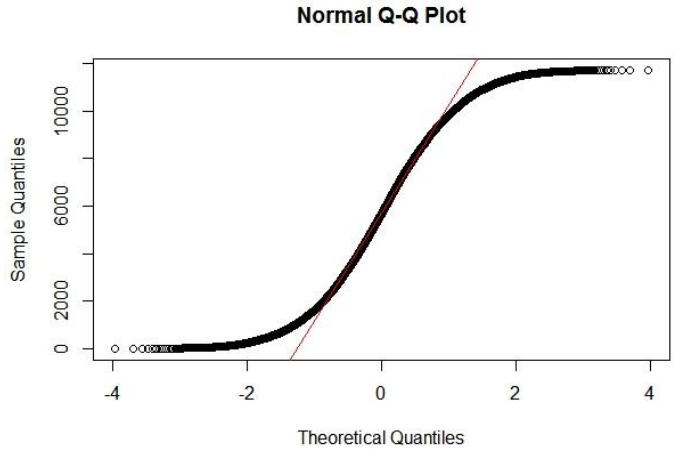
	met_apriori(T)	**met_eclat(T)**	**met_weclat(T)**	**met_ASI(f)**
Centralización	4,37071349	2,96676779	2,50341311	2,24905934
Dispersión				
Tamaño	3447	3447	3447	3447

Fuente: Elaboración propia

Comprobación de supuestos

Para determinar el test apropiado a utilizar se procedió a la comprobación de los supuestos. A continuación, se muestra la gráfica de cuartiles que nos dio una idea gráfica sobre la normalidad de los datos sobre la memoria ocupada por los diferentes métodos.

Figura 32. Gráfico de cuartiles



Fuente: Elaboración propia

Las hipótesis estadísticas sobre normalidad se muestran a continuación:

Ho: No se observa diferencia entre los tiempos de respuesta (velocidad de procesamiento) y la distribución normal

H1: Se observa diferencia entre los tiempos de respuesta (velocidad de procesamiento) y la distribución normal

Para su demostración se utilizaron los test de Anderson-Darling y Crames-von Mises, los resultados se muestran a continuación:

Tabla 30. Resultados de los test de Normalidad

TEST DE NORMALIDAD	**ESTADÍSTICO**	**P-VALUE**
Anderson-Darling	A = 189.1	< 2.2e-16
Cramer-von Mises	W = 26.797	=7.37e-10

Fuente: Elaboración propia

En todos los test se puede observar que se rechaza la hipótesis nula.

Prueba de hipótesis

La hipótesis estadística por demostrar se muestra a continuación:

$$H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3 = \tilde{\mu}_4$$
$$H_1: \exists i, j \in \{1,2,3,4\} / \tilde{\mu}_i \neq \tilde{\mu}_j$$

Se utilizó el test de hipótesis no paramétrico suma de rangos para muestras independientes de Kruskal-Wallis, los resultados entregados por la función kruskal.test(x, y) del paquete estándar del *software* R fueron los siguientes :

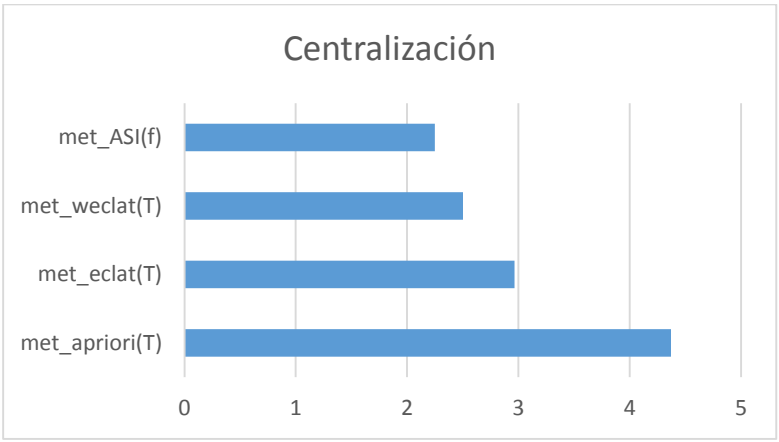
```
        Kruskal-Wallis rank sum test
data:  x and y
Kruskal-Wallis chi-squared = 2732.4, df = 3, p-value < 2.2e-16
```

Para determinar los grupos de homogeneidad se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes. Se obtuvieron las siguientes salidas:

```
          met_apriori  met_ASI met_eclat
met_ASI       64.23740       NA        NA
met_eclat     10.85084 75.08824        NA
met_weclat    27.73366 36.50374   38.5845
```

con las salidas respectivas se elaboró la gráfica de la Figura 31.

Figura 33. Grupos de homogeneidad – reglas de asociación velocidad



Fuente: Elaboración propia

Discusión

El gráfico de cajas y alambres muestra homogeneidad en la dispersión de los métodos de Learning Analytics (met_apriori, met_eclat y weclat) y la técnica del análisis estadístico Implicativo (met_ASI "implicativeGraph"). Por otra parte, al analizar las medidas de centralización se puede observar que los tiempos de respuesta (velocidad de procesamiento), de las técnicas de LA son similares entre ellas, y estas a su vez con la técnica de AEI, pero se ve que la técnica de reglas de asociación con su método met_apriori tiene un aparente mayor tiempo de respuesta. Se procedió a comprobar supuestos, antes

de realizar la prueba de hipótesis. Se realizó la prueba de normalidad de Anderson-Darling que nos dio un p-valor de 2.2e-16 indicándonos que se debe rechazar la hipótesis nula y que por tanto los datos no han sido extraídos de una población normal, este resultado se corrobora con el gráfico de cuartiles que muestra un gran alejamiento de la distribución de datos a los cuartiles teóricos. El no cumplimiento de este supuesto es suficiente para optar por los test no paramétricos. El test no paramétrico seleccionado es el test de hipótesis no paramétrico de suma de rangos para muestras independientes de Kruskal-Wallis que nos entrega un valor de chi cuadrado y un p-valor de 2.2e-16, que indica con un alto valor de significancia que se rechaza la hipótesis nula y por lo tanto al menos un par de los 4 métodos de reglas de asociación son diferentes. Con el objetivo de determinar la relación entre los pares se utilizó la posprueba de Kruskal-Conover con el método de ajuste del p-valor de Holm para comparación de muestras independientes con el cual se ha construido el gráfico de homogeneidad de medidas de centralización que nos indica que hay cuatro grupos bien definidos, de los cuales nos llama la atención el método met_apriori que es el que tarda más en entregar los resultados, es decir tiene mayor tiempo de respuesta y además nos indica que la técnica callSimilarityTree (método implicativeGraph) del análisis estadístico Implicativo y los métodos met_eclat y weclat de Learning Analytics son los que tiene menores tiempos de respuesta, por lo cual se deduce que la más optima met_ASI (AEI) con 2,24905934, el cual entrega resultados en el menor tiempo, y menos óptima met_apriori (LA) con 4,37071349.

5.2.8. Conclusiones sobre las hipótesis

De acuerdo al planteamiento de las hipótesis, éstas permiten realizar comparaciones de supuestos a ser comprobados, en la una se niega y la otra afirma haber una clara diferencia observable entre los datos y la distribución normal, con una aceptabilidad y confiabilidad óptima, que mediante la ejecución permite emitir un juicio veraz.

Se observa todas las pruebas de normalidad y el p-valor se puede concluir que se rechaza la hipótesis nula y, por tanto, se acepta la hipótesis alterna que dice: que existe una clara diferencia observable entre los datos y la distribución normal, es decir, hay diferencia entre los datos del tiempo de ejecución y la distribución normal correspondiente.

Se observa los factores de pruebas de uso de memoria y tiempos de respuesta al ejecutar los algoritmos, de acuerdo a la muestra, se concluye que el sistema operativo Ubuntu es el que administra de mejor manera los recursos informativos como memoria y procesamiento, al obtener menores tiempos de respuesta en la ejecución de procesos, tanto al ejecutar los algoritmos de uso de memoria y velocidad, para las técnicas clustering con los métodos de LA (dendro_diana, dendro_variable-

hclust_vector) y AEI callHierarchyTree y callSimilarityTree (harrchy, simlrty); y las técnicas de reglas de asociación con los métodos de LA(met_apriori, met_eclat, met_weclat) y AEI (implicativeGraph).

Existe homogeneidad en el uso de memoria entre las técnicas clustering similares de LA (dendro_diana, dendro_variable, hclust_vector) y AEI callHierarchyTree y callSimilarityTree (hrarchy, simlrty), curiosamente tanto las técnicas de AEI y las técnicas LA son similares entre ellas, definiéndose como el método más óptimo hclust_vector de LA y la menos óptima dendro_variable.

Existe homogeneidad en el tiempo de ejecución entre las técnicas clustering similares de LA (dendro_diana, dendro_variable, hclust_vector) y AEI callHierarchyTree y callSimilarityTree (hrarchy, simlrty), curiosamente tanto las técnicas de AEI y las técnicas LA son similares entre ellas, definiéndose como la más óptima por tener menor tiempo de respuesta hclust_vector de LA y la menos óptima pero no menos importante dendro_variable.

Existe homogeneidad en el uso de memoria entre las técnicas reglas de asociación similares de LA (met_aprori, met_eclat, met_weclat) y EAI (met_ASI "implicativeGraph"), curiosamente tanto las técnicas de LA y AEI son similares entre ellas, definiéndose como la más óptima weclat de LA por ocupar menos memoria, pero no implica esto que sea el que tenga menores tiempo de respuesta y el que ocupa mayor memoria es met_ASI (implicativeGraph).

Existe homogeneidad en el tiempo de ejecución entre las técnicas reglas de asociación similares de LA (met_aprori, met_eclat, met_weclat) y EAI (met_ASI "implicativeGraph"), al ser similares entre ellas, definiéndose como la más óptima por tener menor tiempo de respuesta met_ASI (implicativeGraph) de AEI y la menos óptima pero no menos importante met_apriori. Además, vale indicar a mayor uso de memoria, menor tiempo de respuesta, el met_ASI utiliza mayor memoria, pero tiene menor tiempo de respuesta de los procesos.

5.3. Identificar las técnicas más óptimas.

Los autores (Moreno García, Quintales, García Peñalvo, & Polo Martín, 2004) indican que los principales inconvenientes de los algoritmos de reglas de asociación son la obtención de reglas no interesantes, un gran número de reglas descubiertas y un bajo rendimiento del algoritmo. Como

respuesta a estas observaciones, se han realizado algunos esfuerzos para mejorar el rendimiento de los algoritmos ARM mediante el uso de ontologías.

Estudios realizados por los autores (Moreno García, Segrera, & Batista, 2005), indican que la mejora de los algoritmos de las reglas de asociación es el tema de muchos trabajos en la literatura, se ha hecho poca investigación sobre su aspecto de clasificación y no se toma en cuenta los principales inconvenientes de los algoritmos de reglas de asociación que son:

- ✓ Obtención de reglas no interesantes
- ✓ Gran cantidad de reglas descubiertas
- ✓ Bajo rendimiento del algoritmo

Los autores (Yang, Liu, & Fu, 2010), indican que al ser las reglas de asociación ampliamente utilizadas, es necesario estudiar muchos problemas, uno de los cuales son los conjuntos de datos generalmente más grandes y multidimensionales, y el rápido crecimiento del conjunto de datos. La memoria del procesador único y los recursos de la CPU son muy limitados, lo que hace que el rendimiento del algoritmo sea ineficiente.

He aquí la importancia de nuestra investigación al momento de escoger la técnica más óptima para el manejo de datos masivos. Las reglas de asociación tienen problemas como: reglas irrelevantes, bajo rendimiento en sus algoritmos, memoria del procesador y recursos del CPU limitados.

Todo esto se genera por el manejo de gran cantidad de información lo que hace que los algoritmos utilicen gran cantidad de recursos del computador, y al no ser suficientes los resultados a obtener son más lentos, lo que hace que este método en la actualidad sea uno de los menos óptimos.

Está investigación permite capturar datos de la velocidad y uso de memoria de los diferentes métodos, de lo cual se obtuvo los resultados esperados y que concuerdan con otros investigadores, los algoritmos de reglas de asociación (a priori, eclat y weclat) son las que utilizan mayores recursos del CPU, lo que hace que los lentos los procesos sean más lentos.

El análisis realizado por los autores (Moreira Huilcapi & Lojan Cueva, 2016), indican que Linux (Ubuntu) es muy superior frente a Windows y Mac OS, donde se destaca el gran rendimiento de memoria, y capacidad para realizar trabajos de tipos enteros.

Tabla 31. Comparación entre Windows, Linux y Mac Os

	Windows 7		**Ubuntu 15.0**		**Mac Os Yosemite 10.0**	
	Single-Core Score	**Multi-Core Score**	**Single-Core Score**	**Multi-Core Score**	**Single-Core Score**	**Multi-Core Score**
Integer Score	1512	3747	1633	4059	1422	3582
Floating Point Score	1861	4511	1865	4530	1742	4316
Memory Score	2114	2310	2740	4054	1759	3722

Total Score	1772	3765	1947	4054	1759	3722

Fuente: (Moreira Huilcapi & Lojan Cueva, 2016)

En base a esta información, se analizan los resultados obtenidos (ver tabla 32), al ejecutar nuestro algoritmo.

Tabla 32. Tiempo de ejecución del algoritmo

Técnicas	Windows	Ubuntu	Mac OS
Clustering	8.430220	6.282168	14.353253
Reglas de asociación	2.641932	1.328199	17.591907

Fuente: Elaboración propia

Al conocer que el sistema operativo más rápido es Ubuntu concuerdan con los resultados obtenidos en la fase experimental de nuestra investigación, donde se puede ver que efectivamente nuestro algoritmo se ejecuta con mayor velocidad en este sistema operativo. Ubuntu permite una mayor distribución de los procesos durante la ejecución del algoritmo.

En la tabla 31, el cual es un análisis previo (Moreira Huilcapi & Lojan Cueva, 2016), realizado en una tesis se enfoca en el rendimiento de la memoria a mayor trabajo de la memoria mayor rapidez, a menor trabajo menor rapidez, en la tabla 32 muestra la velocidad promedio de ejecución de los algoritmos de velocidad en los tres sistemas operativos.

Luego de un preámbulo de análisis, se puede resumir en la siguiente tabla comparativa de resultados, la cual nos permitirá identificar las técnica más óptimas, al momento de trabajar con grandes cantidades de datos, dentro del ámbito educativo.

Tabla 33. Comparativo de resultados de técnicas AEI y LA

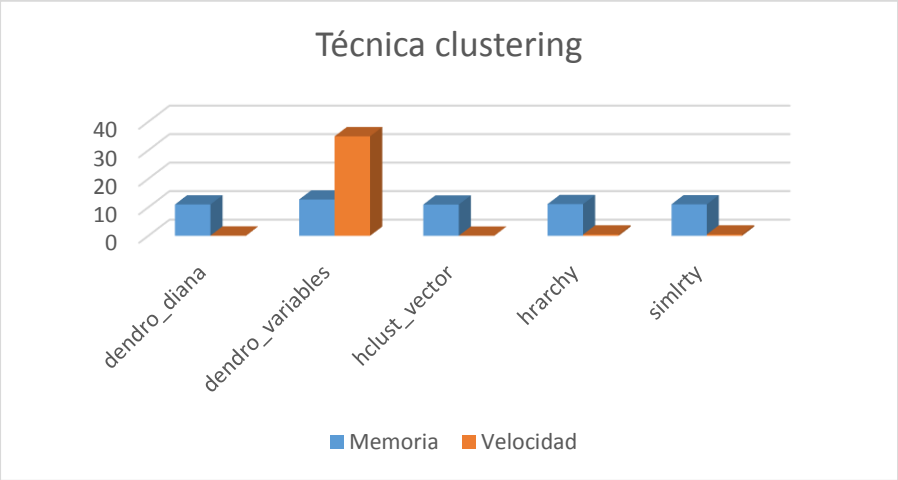
Técnicas de análisis de datos	CLUSTERING			REGLAS DE ASOCIACIÓN		
	Métodos	Memoria	Velocidad	Métodos	Memoria	Velocidad
Learning Analytics	**dendro_diana(T)**	10,9833599	0,08301572	**met_apriori(T)**	6,30144075	4,37071349
	dendro_variables	12,7246268	34,9213604	**met_eclat(T)**	6,28895568	2,96676779
	hclust_vector(T)	10,96933278	0,059625645	**met_weclat(T)**	6,27645716	2,50341311
Análsis estadístico implicativo	**hrarchy(f)**	11,1253011	0,44441203	**met_ASI(f)**	6,33584713	2,24905934
	simlrty(f)	11,056894	0,47690769			

Fuente: Elaboración propia

La tabla 33, muestra los resultados obtenidos de la ejecución del cuasi-experimento, lo cual concuerda con el análisis de (Moreira Huilcapi & Lojan Cueva, 2016), que indica que a mayor

rendimiento de memoria menor tiempo de respuesta de los procesos. Está información es utilizada para diagramar los datos para una mejor interpretación de los mismos.

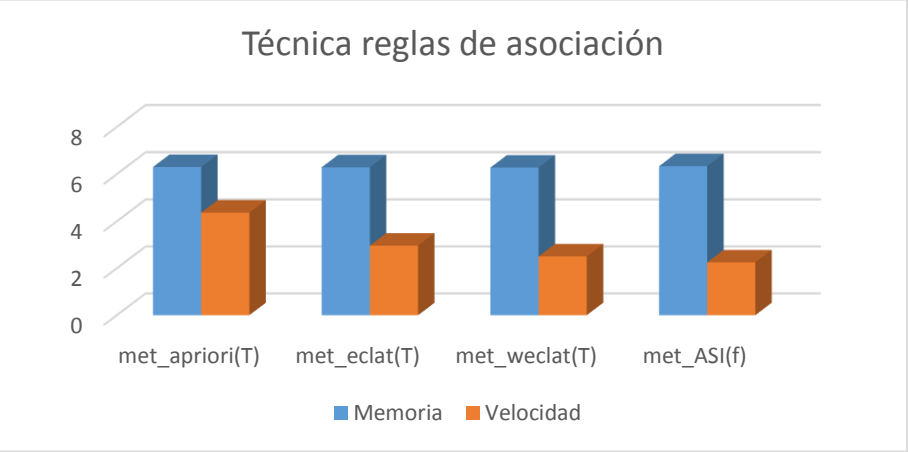
Figura 34. Técnicas óptimas similares AEI y LA – método clustering



Fuente: Elaboración propia

Como se puede observar el método que entrega los resultados en menor tiempo o tiene menor tiempo de respuesta es el método hclust_vector (0,059625645), de learning analytics, con un uso de memoria (10,96933278) razonable, esto en cuanto a LA, con clustering (jerarquización y no jerarquización).

Figura 35. Técnicas óptimas similares de AEI y LA – métodos de reglas de asociación



Fuente: Elaboración propia

Sobre reglas de asociación, se puede observar que método met_ASI (implicativeGraph) de AEI, es el óptimo, al tener un tiempo de respuesta de 2,24905934 y un uso de memoria de 6,33584713.

Luego de este análisis se puede concluir que se concuerda con estudios con autores como (Yang et al., 2010), indican que al ser las reglas de asociación ampliamente utilizadas, es necesario estudiar muchos problemas, como la memoria del procesador único y los recursos de la CPU son muy limitados, lo que hace que el rendimiento del algoritmo sea ineficiente. Esto es verificado y ratificado por nuestro cuasi-experimento, en donde se puede notar que clustering es superior a reglas de asociación. Al tomar las técnicas más óptima tanto de clustering (hclust_vector) como de reglas de asociación (met_ASI – impliactiveGraph), el método hclust_vector tiene menor tiempo de respuesta al momento de trabajar con datos masivos, ya que tiene un tiempo de respuesta de 0,059625645. Se puede también indicar que, al momento de utilizar reglas de asociación, el mejor método aplicar es el met_ASI (implicativeGraph) de AEI, todo esto según los resultados estadísticos del diseño cuasi-experimental y la comprobación de nuestra hipótesis, dicho método tiene un tiempo estimado de respuesta de 2,24905934.

Capítulo 6

Conclusiones y Recomendaciones

6.1. Conclusiones

Existen técnicas similares de agrupación (clustering) entre LA y AEI, estas son dendro_variable, dendro_diana y hclust_vector de Learning Analytics y callHierarchyTree y callSimilarityTree de AEI; y las técnicas similares de reglas de asociación entre LA y AEI, son: apriori, eclat, weclat de LA e implicativeGraph de AEI.

El diseño cuasi- experimental, dio como resultado, que el sistema operativo Ubuntu presenta mejor administración de los recursos, como la asignación de procesos a memoria, de donde se concluye que a mayor uso de memoria menor tiempos de respuesta, el autor de esta investigación concuerda con los autores (Moreira Huilcapi & Lojan Cueva, 2016).

Existe homogeneidad en el uso de memoria entre las técnicas clustering similares de LA (dendro_diana, dendro_variable, hclust_vector) y AEI callHierarchyTree y callSimilarityTree (hrarchy, simlrty), curiosamente tanto las técnicas de AEI y las técnicas LA son similares entre ellas, definiéndose como el método más óptimo hclust_vector de LA y la menos óptima dendro_variable.

Existe homogeneidad en el tiempo de ejecución entre las técnicas clustering similares de LA (dendro_diana, dendro_variable, hclust_vector) y AEI callHierarchyTree y callSimilarityTree (hrarchy, simlrty), curiosamente tanto las técnicas de AEI y las técnicas LA son similares entre ellas, definiéndose como la más óptima por tener menor tiempo de respuesta hclust_vector de LA y la menos óptima pero no menos importante dendro_variable.

Existe homogeneidad en el uso de memoria entre las técnicas reglas de asociación similares de LA (met_aprori, met_eclat, met_weclat) y AEI (met_ASI “implicativeGraph”), curiosamente tanto las técnicas de LA y AEI son similares entre ellas, definiéndose como la más óptima weclat de LA por

ocupar menos memoria, pero no implica esto que sea el que tenga menores tiempo de respuesta y el que ocupa mayor memoria es met_ASI (implicativeGraph).

Existe homogeneidad en el tiempo de ejecución entre las técnicas reglas de asociación similares de LA (met_aprori, met_eclat, met_weclat) y AEI (met_ASI "implicativeGraph"), al ser similares entre ellas, definiéndose como la más óptima por tener menor tiempo de respuesta met_ASI (implicativeGraph) de AEI y la menos óptima pero no menos importante met_apriori. Además, vale indicar a mayor uso de memoria, menor tiempo de respuesta, el met_ASI utiliza mayor memoria, pero tiene menor tiempo de respuesta de los procesos.

6.2. Recomendaciones

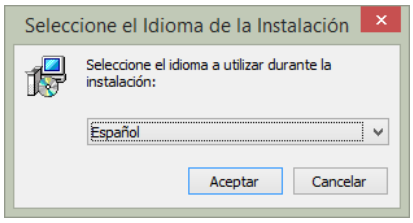
Se recomienda difundir este material, con el propósito que los docentes apliquen las técnicas AEI y LA, con la finalidad de fortalecer la educación ecuatoriana, mediante el análisis de datos masivos.

Que el departamento de investigación de la Pontificia Universidad Católica Sede Ambato, fortalezcan esta investigación, aplicar una de las técnicas óptimas que se hace alusión en este documento.

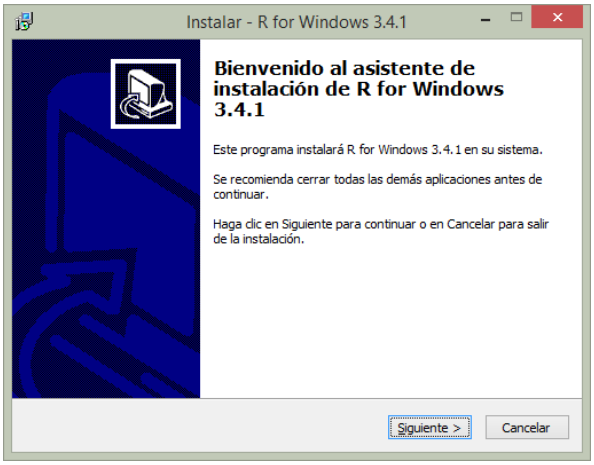
Apéndice A

A.1. Instalación *software* R

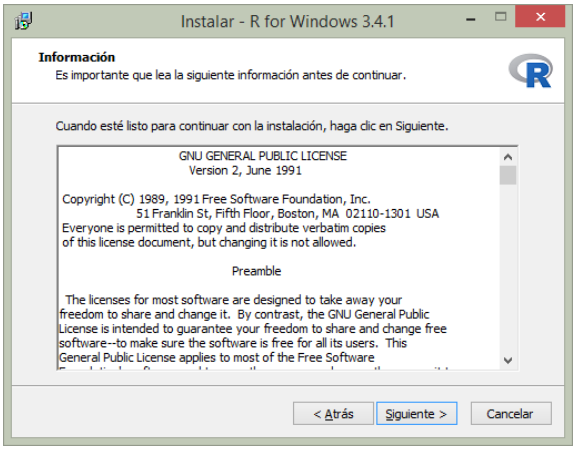
1. Descargar *software* R de la dirección:
   https://cran.r-project.org/mirrors.html
2. Seleccionar el idioma.



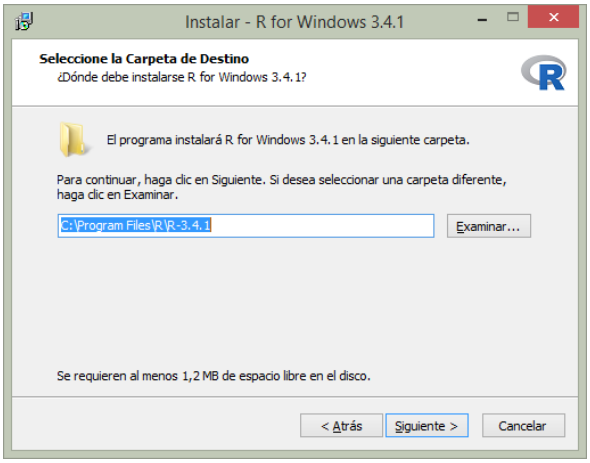
3. Pulsar siguiente para continuar con la instalación.



4. Leer la licencia y continuar con la instalación.



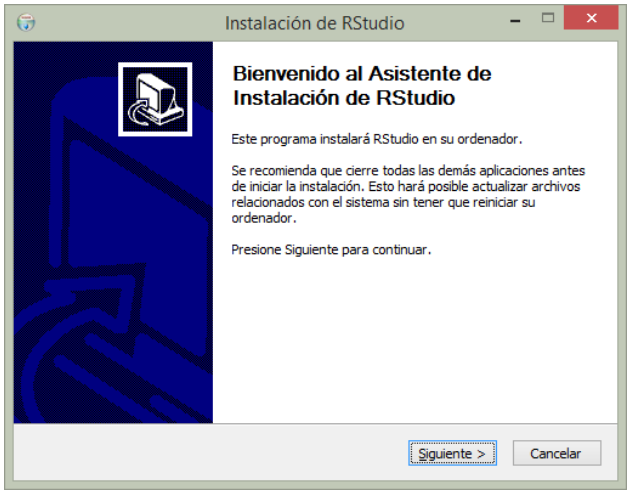
5. Seleccionar el directorio de instalación y continuación dar clic en siguiente.



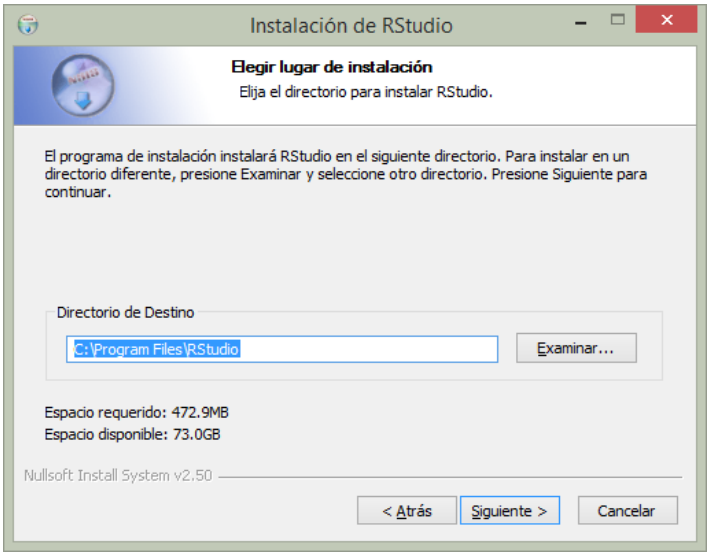
6. Finalmente dar clic en el botón finalizar.

A.2. Instalación RStudio

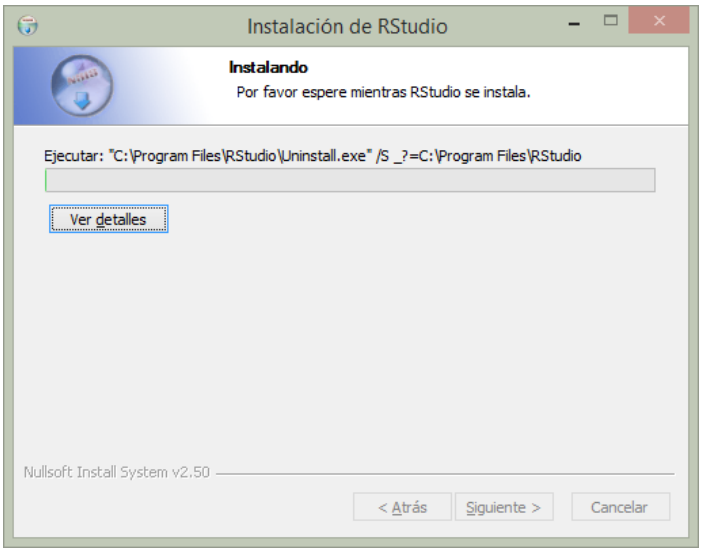
1. Descargar RStudio desde el siguiente enlace: https://www.rstudio.com/products/rstudio/download/
2. Leer las recomendaciones y posteriormente dar clic en el botón siguiente para continuar.



3. Seleccionar el directorio de instalación.



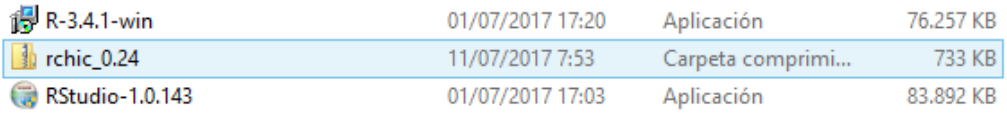
4. Proceso de instalación.



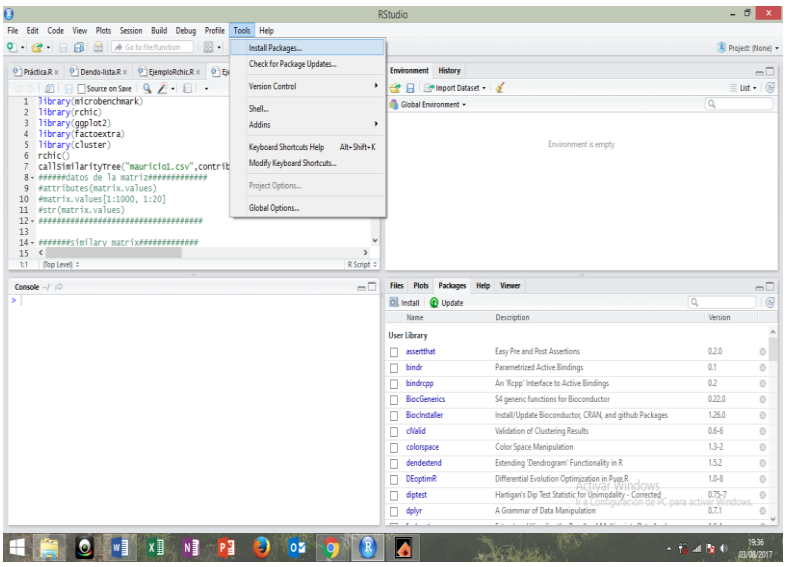
5. Finalmente dar clic en finalizar.

A.3. Instalación paquetes R

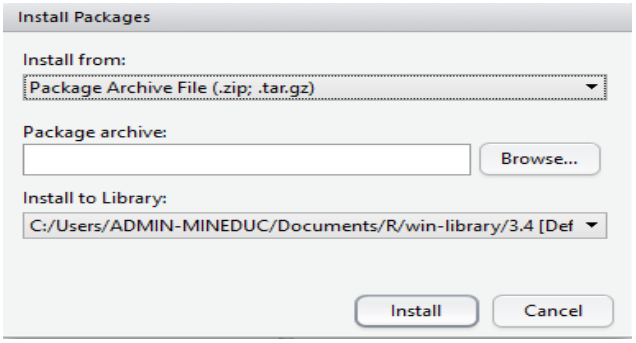
1. Descarga el paquete Rchic de la dirección electrónica:
   https://www.r-project.org/



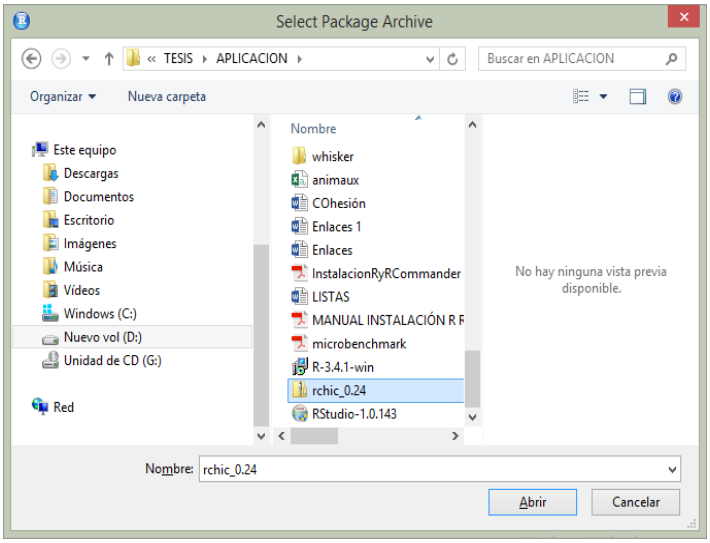
2. Una vez descargado, desde RStudio se escoge de la barra de herramientas la opción Tools, donde se despliega un menú, a continuación, se selecciona Installar Packages.



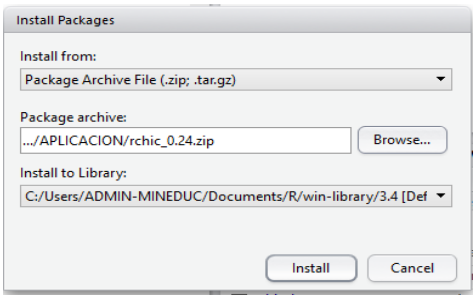
3. A continuación, nos aparecerá una venta denominada Install Packages, donde se da clic en el botón Browse



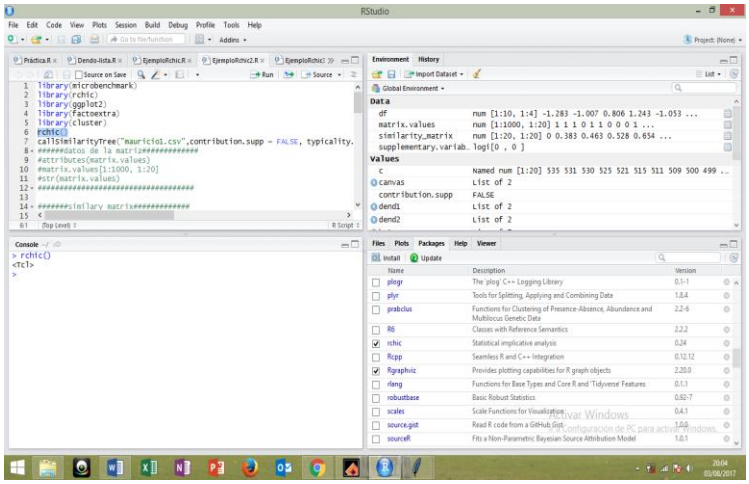
4. Luego se debe buscar el archivo RChic, donde se ubicó, selecciona el archivo y da clic en el botón Abrir.



5. Posterior a ellos, se verá el archivo Rchic listo para ser instalado, una vez visualizado la siguiente venta se procede a dar clic en el botón Install.



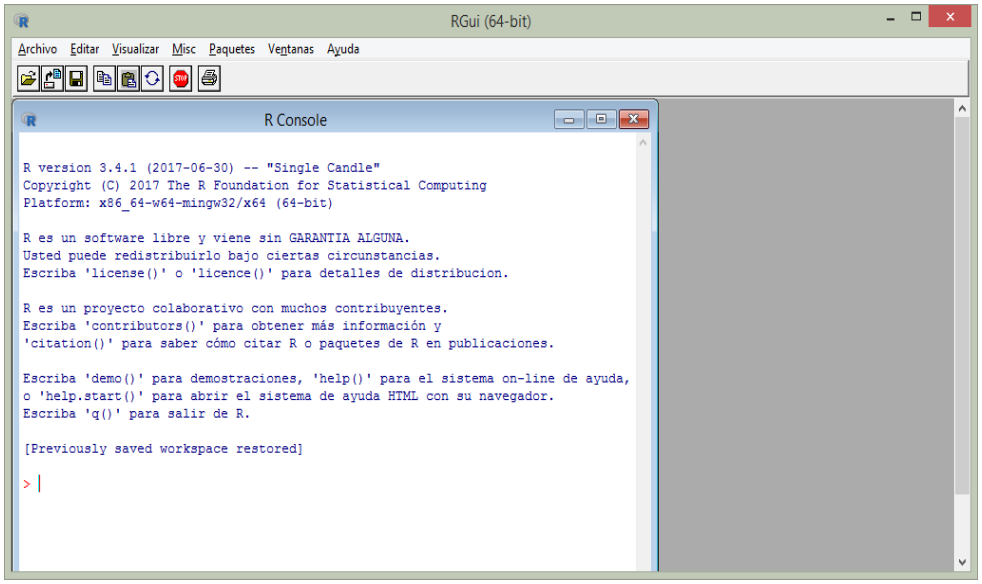
6. Una vez finalizado la instalación se puede visualizar el paquete RChic, junto a los demás paquetes, además para poder comprobar su correcto funcionamiento se lo puede hacer a través del comando rchic().



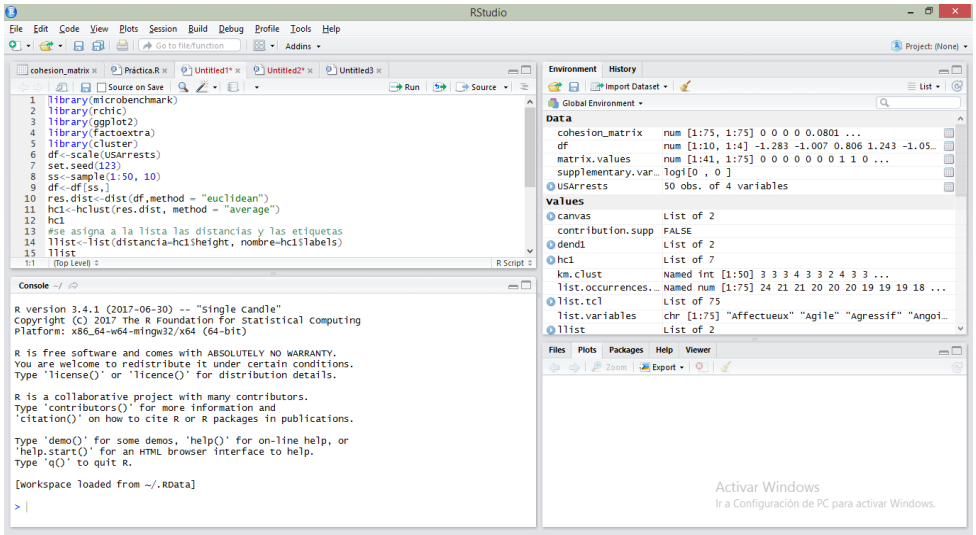
Apéndice B

B.1.Interfaces *software*

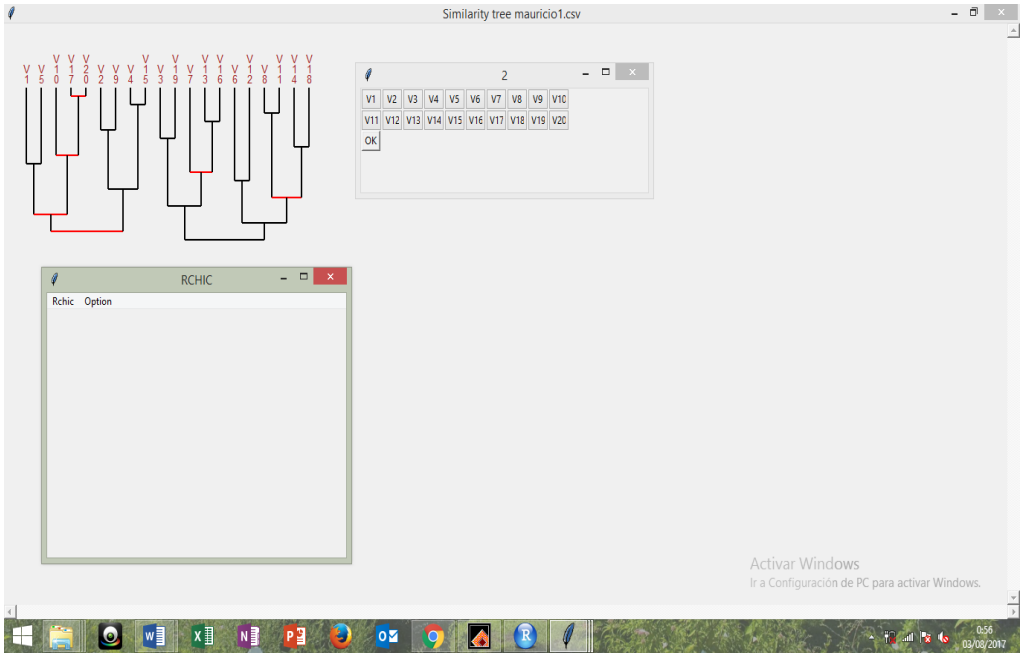
B.1.1.Interfaz R



B.1.2. Interfaz RStudio



B.1.3.Ejecución ejemplo Rchic



Apéndice C

C.1. Algoritmo calculo velocidad de procesamientos - método cluster

```
library("microbenchmark");
library("rchic");
library("ggplot2");
library("factoextra");
library("cluster")
library("fastcluster")
library("CluMix")
library("dplyr")
library("replyr")
rchic()
hclust_vector<-function(x){
 df<-read.csv(x, sep = ";")
 #hc <- hclust.vector(df, "cen")
 hc<-hclust.vector(df, method="single", members=NULL, metric='euclidean', p=NULL)
```

```
 plot(hc)#opcional
 return(hc)
}
dendro_variables<-function(x){
 df<-read.csv(x, sep = ";")
 dend1 <- dendro.variables(df, method="distcor")
 plot(dend1)#opcional
 return(dend1)
}
dendro_diana<-function(x){
 df<-read.csv(x, sep = ";")
 dend <- diana(df, metric = "euclidean", stand = TRUE)
 plot(dend)#opcional
 return(dend)
}
hrarchy<-function(x){
 hr<-callHierarchyTree (x,contribution.supp=FALSE,
                typicality.supp=FALSE,
                computing.mode=3,verbose=FALSE)
 return (hr)
}
simlrty<-function(x){
 st<-callSimilarityTree(x,contribution.supp=FALSE,
                typicality.supp=FALSE,verbose=FALSE)
 return (st)
}
nprocess<-5;rep<-3;endfor<-1
TOTALtimeCompareSimilarityMatrix<-matrix(1:(rep*nprocess),nrow=(rep*nprocess),ncol=1)
for (i in 1:endfor)
{
 nVar<- round(runif(1),2)*100
 nFilas<- round(runif(1),3)*1000
```

```
DataBase<-replicate(nVar, round(runif(nFilas),0))
rownames(DataBase)<-paste('S',1:nFilas,sep='')
colnames(DataBase)<-c(';V1',paste('V',2:nVar,sep=''))
f<-paste('_D',toString(i),'.csv',sep="_")
T<-paste('_T',toString(i),'.csv',sep="_")
write.table(DataBase, file=f, sep=";" ,quote = FALSE)
DataBaseT<-t(DataBase)
write.table(DataBaseT, file = T ,sep=";")
timeCompareSimilarity<-microbenchmark(hclust_vector(T), dendro_variables(T),
dendro_diana(T), simlrty(f), hrarchy(f),times=rep,unit="ms",control=list("inorder"))
timeCompareSimilarityMatrix<-as.matrix(timeCompareSimilarity)
TOTALtimeCompareSimilarityMatrix<-
cbind(TOTALtimeCompareSimilarityMatrix,timeCompareSimilarityMatrix)
colnames(TOTALtimeCompareSimilarityMatrix)[2*i+1]<-
paste(toString(nVar),'x',toString(nFilas),paste('_D',toString(i),sep="_"))
write.table(TOTALtimeCompareSimilarityMatrix,file = "out.csv",sep=";",row.names = FALSE)
}
```

C.2. Algoritmo calculo uso memoria - método cluster

```
library("microbenchmark");
library("rchic");
library("ggplot2");
library("factoextra");
library("cluster")
library("fastcluster")
library("CluMix")
library("dplyr")
library("replyr")
rchic()
hclust_vector<-function(x){
df<-read.csv(x, sep = ";")
#hc <- hclust.vector(df, "cen")
hc<-hclust.vector(df, method="single", members=NULL, metric='euclidean', p=NULL)
```

```
  plot(hc)#opcional
  return(hc)
 }
 dendro_variables<-function(x){
  df<-read.csv(x, sep = ";")
  dend1 <- dendro.variables(df, method="distcor")
  plot(dend1)#opcional
  return(dend1)
 }
 dendro_diana<-function(x){
  df<-read.csv(x, sep = ";")
  dend <- diana(df, metric = "euclidean", stand = TRUE)
  plot(dend)#opcional
  return(dend)
 }
 hrarchy<-function(x){
  hr<-callHierarchyTree (x,contribution.supp=FALSE,
                typicality.supp=FALSE,
                computing.mode=3,verbose=FALSE)
  return (hr)
 }
 simlrty<-function(x){
  st<-callSimilarityTree(x,contribution.supp=FALSE,
                typicality.supp=FALSE,verbose=FALSE)
  return (st)
 }
 nprocess<-5;rep<-3;endfor<-3
 TOTALtimeCompareSimilarityMatrixM<-
matrix(1:(rep*nprocess),nrow=(rep*nprocess),ncol=1)
 for (i in 1:endfor)
 {
  nVar<- round(runif(1),2)*100
```

```
nFilas<- round(runif(1),3)*1000
DataBase<-replicate(nVar, round(runif(nFilas),0))
rownames(DataBase)<-paste('S',1:nFilas,sep='')
colnames(DataBase)<-c(';V1',paste('V',2:nVar,sep=''))
f<-paste('_D',toString(i),'.csv',sep="_")
T<-paste('_T',toString(i),'.csv',sep="_")
write.table(DataBase, file=f, sep=";",quote = FALSE)
DataBaseT<-t(DataBase)
write.table(DataBaseT, file = T ,sep=";")
fc<-paste(toString(nVar),'x',toString(nFilas))
hc<-gc(verbose = hclust_vector(T), reset = TRUE)
hc1<-gc(verbose = hclust_vector(T), reset = TRUE)
hc2<-gc(verbose = hclust_vector(T), reset = TRUE)
dendv<-gc(verbose = dendro_variables(T), reset = TRUE)
dendv1<-gc(verbose = dendro_variables(T), reset = TRUE)
dendv2<-gc(verbose = dendro_variables(T), reset = TRUE)
dendd<-gc(verbose = dendro_diana(T), reset = TRUE)
dendd1<-gc(verbose = dendro_diana(T), reset = TRUE)
dendd2<-gc(verbose = dendro_diana(T), reset = TRUE)
s<-gc(verbose = simlrty(f), reset = TRUE)
s1<-gc(verbose = simlrty(f), reset = TRUE)
s2<-gc(verbose = simlrty(f), reset = TRUE)
hr<-gc(verbose = hrarchy(f), reset = TRUE)
hr1<-gc(verbose = hrarchy(f), reset = TRUE)
hr2<-gc(verbose = hrarchy(f), reset = TRUE)
a1<-hc[1,2]
b1<-hc1[1,2]
c1<-hc2[1,2]
a2<-dendv[1,2]
b2<-dendv1[1,2]
c2<-dendv2[1,2]
a3<-dendd[1,2]
```

```
b3<-dendd1[1,2]
c3<-dendd2[1,2]
a4<-s[1,2]
b4<-s1[1,2]
c4<-s2[1,2]
a5<-hr[1,2]
b5<-hr1[1,2]
c5<-hr2[1,2]
UsoMemoriaMatriz<-
matrix(c("hcluster_vector","hcluster_vector","hcluster_vector","dendro_variable","dendro_variable
","dendro_variable","dendro_diana","dendro_diana","dendro_diana","simlrty","simlrty","simlrty","h
rarchy","hrarchy","hrarchy", a1,b1,c1, a2,b2,c2, a3,b3,c3, a4,b4,c4,a5,b5,c5), nrow = 15, ncol = 2)
colnames(UsoMemoriaMatriz)[2]<-paste(toString(nVar),'x',toString(nFilas))
colnames(UsoMemoriaMatriz)[1]<-"Memoria"
timeCompareSimilarityMatrixM<-as.matrix(UsoMemoriaMatriz)
TOTALtimeCompareSimilarityMatrixM<-
cbind(TOTALtimeCompareSimilarityMatrixM,timeCompareSimilarityMatrixM)
write.table(TOTALtimeCompareSimilarityMatrixM,file = "out.csv",sep=";",row.names = FALSE)
}
```

C.3. Algoritmo calculo velocidad de procesamientos y uso de memoria - reglas de asociación

```
library("microbenchmark");
library("rchic");
library("ggplot2");
library("factoextra");
library("cluster")
library("fastcluster")
library("CluMix")
library("dplyr")
library("replyr")
library("arules")
library("arulesViz")
```

```
library("vcd")
library("Rgraphviz")
library("readr")
rchic()

#Método ASI
met_ASI<-function(x){
  callSimilarityTree(x,contribution.supp=FALSE, typicality.supp=FALSE,verbose=FALSE)
 sm<-similarity_matrix
 implicativeGraph(sm, list.variables = list.variables, computing.mode = 1,complete.graph = 0)
 }
#Método Apriori
met_apriori<-function(x){
 d<-read.csv(x, sep = ";")
 df<-as.matrix(d)
 #hc <- hclust.vector(df, "cen")
 ma<-apriori(df, parameter = list(supp = 0.5, maxlen = 3, conf = 0.6, target = "rules"))
 plot(ma)#opcional
 #plot(ma, method="graph", control=list(type="items"))
 #plot(ma, method="paracoord", control=list(reorder=TRUE))
 return(ma)
}
#Método eclat
met_eclat<-function(x){
 d<-read.csv(x, sep = ";")
 df<-as.matrix(d)
 rules_ec <- eclat(df,parameter = list(supp = 0.5, maxlen = 3))
 rules_ec
 #t_ec<-inspect(rules_ec)
 plot(rules_ec)
 #plot(rules_ec, method="graph", control=list(type="items"))
 #plot(rules_ec, method="paracoord", control=list(reorder=TRUE))
```

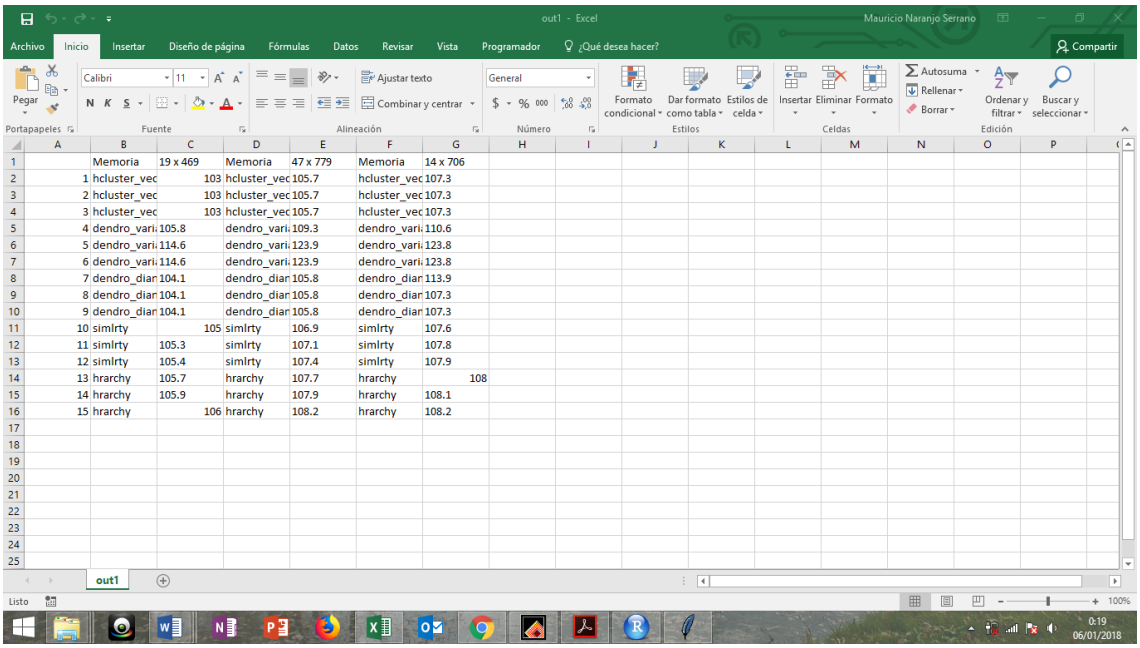
```
 return(rules_ec)
}
#Método weclat
met_weclat<-function(x){
 d<-read.csv(x, sep = ";")
 df<-as.matrix(d)
 rules_weclat <- weclat(df, parameter = list(support = 0.5, maxlen = 3), control = list(verbose =
TRUE))
 #t_ins<-inspect(rules_weclat)
 #rules_weclat
 plot(rules_weclat)
 #plot(rules_weclat, method="graph", control=list(type="items"))
 #plot(rules_weclat, method="paracoord", control=list(reorder=TRUE))
 return(rules_weclat)
}
hclust_vector<-function(x){
 df<-read.csv(x, sep = ";")
 #hc <- hclust.vector(df, "cen")
 hc<-hclust.vector(df, method="single", members=NULL, metric='euclidean', p=NULL)
 plot(hc)#opcional
 return(hc)
}
dendro_variables<-function(x){
 df<-read.csv(x, sep = ";")
 dend1 <- dendro.variables(df, method="distcor")
 plot(dend1)#opcional
 return(dend1)
}

dendro_diana<-function(x){
 df<-read.csv(x, sep = ";")
 dend <- diana(df, metric = "euclidean", stand = TRUE)
```

```
  plot(dend)#opcional
  return(dend)
 }
 hrarchy<-function(x){
  hr<-callHierarchyTree (x,contribution.supp=FALSE,
                typicality.supp=FALSE,
                computing.mode=3,verbose=FALSE)
  return (hr)
 }
 simlrty<-function(x){
  st<-callSimilarityTree(x,contribution.supp=FALSE,
                typicality.supp=FALSE,verbose=FALSE)
  return (st)
 }
 nprocess<-4;rep<-3;endfor<-3
 TOTALtimeCompareSimilarityMatrix<-matrix(1:(rep*nprocess),nrow=(rep*nprocess),ncol=1)
 TOTALtimeCompareSimilarityMatrixM<-
matrix(1:(rep*nprocess),nrow=(rep*nprocess),ncol=1)
 for (i in 1:endfor)
 {
  nVar<- round(runif(1),2)*100
  nFilas<- round(runif(1),3)*1000
  DataBase<-replicate(nVar, round(runif(nFilas),0))
  rownames(DataBase)<-paste('S',1:nFilas,sep='')
  colnames(DataBase)<-c(';V1',paste('V',2:nVar,sep=''))
  f<-paste('_D',toString(i),'.csv',sep="_")
  T<-paste('_T',toString(i),'.csv',sep="_")
  write.table(DataBase, file=f, sep=";" ,quote = FALSE)
  DataBaseT<-t(DataBase)
  write.table(DataBaseT, file = T ,sep=";")
  nc<-ncol(DataBaseT)
  nf<-nrow(DataBaseT)
```

```
timeComparesReglasAso<-microbenchmark(met_apriori(T),met_eclat(T), met_weclat(T), met_ASI(f), times=rep,unit="ms",control=list("inorder"))
autoplot(timeComparesReglasAso)
boxplot(timeComparesReglasAso)
timeCompareSimilarityMatrix<-as.matrix(timeComparesReglasAso)
TOTALtimeCompareSimilarityMatrix<-cbind(TOTALtimeCompareSimilarityMatrix,timeCompareSimilarityMatrix)
colnames(TOTALtimeCompareSimilarityMatrix)[2*i+1]<-paste(toString(nVar),'x',toString(nFilas),paste('_D',toString(i),sep="_"))
write.table(TOTALtimeCompareSimilarityMatrix,file = "outv.csv",sep=";",row.names = FALSE)
#USO MEMORIA
ma<-gc(verbose = met_apriori(T), reset = TRUE)
ma1<-gc(verbose = met_apriori(T), reset = TRUE)
ma2<-gc(verbose = met_apriori(T), reset = TRUE)
me<-gc(verbose = met_eclat(T), reset = TRUE)
me1<-gc(verbose = met_eclat(T), reset = TRUE)
me2<-gc(verbose = met_eclat(T), reset = TRUE)
mw<-gc(verbose = met_weclat(T), reset = TRUE)
mw1<-gc(verbose = met_weclat(T), reset = TRUE)
mw2<-gc(verbose = met_weclat(T), reset = TRUE)
mASI<-gc(verbose = met_ASI(f), reset = TRUE)
mASI1<-gc(verbose = met_ASI(f), reset = TRUE)
mASI2<-gc(verbose = met_ASI(f), reset = TRUE)
a1<-ma[1,2]
b1<-ma1[1,2]
c1<-ma2[1,2]
a2<-me[1,2]
b2<-me1[1,2]
c2<-me2[1,2]
a3<-mw[1,2]
b3<-mw1[1,2]
c3<-mw2[1,2]
```

```
a4<-mASI[1,2]
b4<-mASI1[1,2]
c4<-mASI2[1,2]
UsoMemoriaMatriz<-
matrix(c("met_apriori","met_apriori","met_apriori","met_eclat","met_eclat","met_eclat","met_weclat","met_weclat","met_weclat","met_ASI","met_ASI","met_ASI", a1,b1,c1, a2,b2,c2, a3,b3,c3, a4,b4,c4), nrow = 12, ncol = 2)
colnames(UsoMemoriaMatriz)[2]<-paste(toString(nVar),'x',toString(nFilas))
colnames(UsoMemoriaMatriz)[1]<-"Memoria"
timeCompareSimilarityMatrixM<-as.matrix(UsoMemoriaMatriz)
TOTALtimeCompareSimilarityMatrixM<-
cbind(TOTALtimeCompareSimilarityMatrixM,timeCompareSimilarityMatrixM)
write.table(TOTALtimeCompareSimilarityMatrixM,file = "outm.csv",sep=";",row.names = FALSE)
}
```

C.4. Resultados en archivo .csv



	A	B	C	D	E	F	G
1		Memoria	19 x 469	Memoria	47 x 779	Memoria	14 x 706
2	1	hcluster_vec	103	hcluster_vec	105.7	hcluster_vec	107.3
3	2	hcluster_vec	103	hcluster_vec	105.7	hcluster_vec	107.3
4	3	hcluster_vec	103	hcluster_vec	105.7	hcluster_vec	107.3
5	4	dendro_vari	105.8	dendro_vari	109.3	dendro_vari	110.6
6	5	dendro_vari	114.6	dendro_vari	123.9	dendro_vari	123.8
7	6	dendro_vari	114.6	dendro_vari	123.9	dendro_vari	123.8
8	7	dendro_diar	104.1	dendro_diar	105.8	dendro_diar	113.9
9	8	dendro_diar	104.1	dendro_diar	105.8	dendro_diar	107.3
10	9	dendro_diar	104.1	dendro_diar	105.8	dendro_diar	107.3
11	10	simlrty	105	simlrty	106.9	simlrty	107.6
12	11	simlrty	105.3	simlrty	107.1	simlrty	107.8
13	12	simlrty	105.4	simlrty	107.4	simlrty	107.9
14	13	hrarchy	105.7	hrarchy	107.7	hrarchy	108
15	14	hrarchy	105.9	hrarchy	107.9	hrarchy	108.1
16	15	hrarchy	106	hrarchy	108.2	hrarchy	108.2

Referencias

Aaten, A. B., Van den Heuvel-Panhuizen, M., & Elia, I. (2011). *Kindergartners' perspective taking abilities.* Paper presented at the Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education.

Alcalde, I. (2015). Learning Analytics: el big data de la educación, from https://www.ignasialcalde.es/learning-analytics-el-big-data-de-la-educacion/

Amershi, S., & Conati, C. (2009). Combining unsupervised and supervised classification to build user models for exploratory. . *Journal of Educational Data Mining, 1*, 18-71.

Anastasiadou, S., & Gagatsis, A. (2005). *Attitudes des étudiants, futurs enseignants de l'école primaire grecque à l'égard de la statistique* Paper presented at the Troisièmes Rencontres Internationales – Terzo Convegno Internazionale Palermo-Italy.

Anastasiadou, S. D., Anastasiadis, L., Vandikas, I., & Angeletos, T. (2011). Implicative Statistical Analysis and Principal Components Analysis in Recording Students' Attitudes to Electronics and Electrical Construction Subjects. *International Journal of Technology, Knowledge & Society, 7*(1).

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems, 13*(2), 197-210. doi: http://dx.doi.org/10.1016/S0167-739X(97)00021-6

Arredondo, N. P. A. (2009). Método Semisupervisado para la Clasificación Automática de Textos de Opinión.

Arroyo, I., & Woolf, B. (2005). *Inferring learning and attitudes from a Bayesian Network of log file data.* . Paper presented at the Proceedings of the 12th International Conference on Artifi cial Intelligence in Education.

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. *Eric*, 17.

Baker, R., Gowda, S., & Corbett, A. (2011). *Towards predicting future transfer of learning.* Paper presented at the Artificial intelligence in education, Heidelberg, Germany: Springer.

Baker, R. S., & Gowda, S. M. (2010). *An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools.* Paper presented at the Educational Data Mining 2010.

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). New York, NY: Springer New York.

Barnes, T. (2005). *The Q-matrix method: Mining student response data for knowledge*. Paper presented at the Proceedings of the American Association for Artificial Intelligence 2005 Educational Data Mining Workshop.

Barnes, T., Bitzer, D., & Vouk, M. (2005). Experimental Analysis of the Q-Matrix Method in Knowledge Discovery. In M.-S. Hacid, N. V. Murray, Z. W. Raś & S. Tsumoto (Eds.), *Foundations of Intelligent Systems: 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, May 25-28, 2005. Proceedings* (pp. 603-611). Berlin, Heidelberg: Springer Berlin Heidelberg.

Barrientos-Martínez, R. E., Cruz-Ramírez, N., Acosta-Mesa, H. G., Rabatte-Suárez, I., Gogeascoechea-Trejo, M. d. C., Pavón-León, P., & Blázquez-Morales, S. L. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la Universidad Veracruzana, 9*(2), 19-24.

Becker, B. (2013). Learning analytics: Insights into the natural learning behavior of our students. *Behavioral & Social Sciences Librarian*, 63-67.

Ben-Naim, D., Bain, M., & Marcus, N. (2009). *A User-Driven and Data-Driven Approach for Supporting Teachers in Reflection and Adaptation of Adaptive Tutorials*. Paper presented at the Proceedings of the 2nd International Conference on Educational Data Mining, Sydney, Australia.

Berland, M., Baker, R., & Blikstein, P. (2014). Educational Data Mining and Learning Analytics: Applications to Constructionist Research. *Technology, Knowledge and Learning*, 205-220.

Berzal, F. Clustering basado por particiones. *línea]. Available: http://elvex. ugr. es/idbis/dm/slides/41% 20Clustering*.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. [journal article]. *Critical Care, 9*(1), 112. doi: 10.1186/cc3045

Bogarín Vega, A., Romero Morales, C., & Cerezo Menéndez, R. (2015). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. [Base de datos; aprendizaje; estudiante; red de información]. *2015, 5*(1), 20. doi: 10.21071/edmetic.v5i1.4017

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662-679.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*.

Calvo-Manzano, J., Cuevas, G., Muñoz, M., & San Feliu, T. (2008). Process similarity study: Case study on project planning practices based on CMMI-DEV v1. 2. *EuroSPI 2008 Industrial Proceedings*.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*.

Campbell John P., & G., O. D. (2007). Academic Analytics *Educause Quarterly*, 1-20.

Clow, D., & Makriyannis, E. (2011). *iSpot analysed: participatory learning and reputation*. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.

Couturier, & Almouloud. (2009). *Historique et fonctionnalités de CHIC*.

Couturier, R. (2015). Validation of RCHIC and some facts about RCHIC.

Couturier, R. (2018). Rchic, from http://members.femto-st.fr/raphael-couturier/en/rchic

Couturier, R., & Pazmiño, R. (2016). Use of Statistical Implicative Analysis in Complement of Item Analysis. *International Journal of Information and Education Technology, 6*(1), 39-43. doi: 10.7763/ijiet.2016.v6.655

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 318-331.

Chatti, M. A., Lukarov, V., Thüs, H., Muslim, A., Yousef, A. M. F., Wahid, U., . . . Schroeder, U. (2014). Learning Analytics: Challenges and Future Research Directions. *eleed, 10*(1).

Delacroix, T., & Boubekki, A. (2014). An application of multiple behavior SIA for analyzing data from student examsApplications multiples de l´ ASI pour l´ analyse des données des examens d´ étudiants. *Educação Matemática Pesquisa, 16*(3).

Drachsler, H., Dietze, S., Herder, E., d'Aquin, M., & Taibi, D. (2014). The learning analytics & knowledge (LAK) data challenge 2014. 289-290. doi: 10.1145/2567574.2567630

Elia, I., Özel, S., Gagatsis, A., Panaoura, A., & Özel, Z. E. Y. (2016). Students' mathematical work on absolute value: focusing on conceptions, errors and obstacles. *ZDM, 48*(6), 895-907.

Elias, T. (2011). Learning analytics. *Learning*.

Everitt, B. S., & Hothorn, T. *A Handbook of Statistical Analyses Using R* (Second Edition ed.).

Fancsali, S. (2012). *Variable construction and causal discovery for cognitive tutor log data: Initial results.* . Paper presented at the Proceedings of the 5th Conference on Educational Data Mining

Ferguson, R. (2014). Learning Analytics: drivers, developments and challenges. [Educational Technology; Academic Analytics; Action Analytics; Educational Data Mining; Learning

Analytics; Social Learning Analytics; Technology Enhanced Learning (TEL)]. *2014, 22*(3), 10. doi: 10.17471/2499-4324/183

Ferguson, R. (2016). *Learning analytics: drivers, developments and challenges*.

Fernández, M., Mucientes, M., B, V., & Lama, M. (2014, 22-25 Oct. 2014). *Learning analytics for the prediction of the educational objectives achievement.* Paper presented at the 2014 IEEE Frontiers in Education Conference (FIE) Proceedings.

Fidalgo, Á. (2012). Learning Analytics (Analíticas de Aprendizaje). Qué, cómo y para qué., from https://innovacioneducativa.wordpress.com/2012/11/10/learning-analytics-analiticas-de-aprendizaje-que-como-y-para-que/

Fisher, D., & Langley, P. (1985). Approaches to Conceptual Clustering. *Defense Technical Information Center*.

Fourniern, H., Kop, R., & Sitlia, H. (2011). *The value of learning analytics to networked learning on a personal learning environment*. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.

Galbraith, S., Stephenson, & Buckman, H. (1993). Decision rules used by male and female business students in making ethical value judgments: Another look. [journal article]. *Journal of Business Ethics, 12*(3), 227-233. doi: 10.1007/bf01686450

Gasca Hurtado, G. P. (2010). Estudio de similitud del proceso de gestión de riesgos en proyectos de outsourcing de software: utilización de un método. *Revista Ingenierías Universidad de Medellín, 9*(17), 119-130.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300. doi: http://dx.doi.org/10.1016/j.patrec.2005.08.011

Gras, R., & Kuntz, P. (2009). El Analisis Estadistico Implicativo (ASI) en respuesta a problemas que le dieron origen *Teoria y Aplicaciones del Analisis Estadistico Implicativo, Eds : P.Orus, L.Zemora, P.Gregori* (pp. 3-51): Universitat Jaume-1, Castellon.

Gras, R., Suzuki, E., Guillet, F., & Spagnolo, F. (2008). *Statistical implicative analysis: Theory and applications*: Springer.

Guevara Fuente de la Vega, K., & Beltran Castañon, C. (2014). *Descubrimiento de Patrones Secuenciales con Factores de Cantidad*. Paper presented at the VI Congreso Internacional de Computación y Telecomunicaciones.

http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/680/COMTEL-2014-177-184.pdf?sequence=1&isAllowed=y

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., Johnson, I., & Borgelt, C. (2018). Mining Association Rules and Frequent Itemsets.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33*(3), 464-472. doi: 10.1016/j.ijinfomgt.2013.01.001

Henderson, D. A., & Denison, D. R. (1989). Stepwise Regression in Social and Psychological Research. *Psychological Reports, 64*(1), 251-257. doi: doi:10.2466/pr0.1989.64.1.251

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2003). *Metodología de la investigación* (Vol. 707): México: McGraw-Hill.

Hummel, M., Edelmann, D., & Kopp-Schneider, A. (2017). Clustering and Visualization of Mixed-Type Data.

Hurairah, D. A. (2014). CIS527: Data Warehousing, Filtering, and Mining, from https://es.slideshare.net/wanaezwani/apriori-and-eclat-algorithm-in-association-rule-mining

Intel, C. (2017). Procesadores Intel from https://www.intel.la/content/www/xl/es/homepage.html

International Sales and Support, M. I. (2017). Información básica de regresión escalonada, from http://support.minitab.com/es-mx/minitab/17/topic-library/modeling-statistics/regression-and-correlation/basics/basics-of-stepwise-regression/

Iurato, G. (2012). *The implicative statistical analysis: an interdisciplinary paradigm*. Retrieved from https://hal.archives-ouvertes.fr/hal-00750049

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). NMC Horizon Report: Edición sobre Educación Superior 2016: The New Media Consortium.

Jones, A. F., Walker, J., Jewkes, C., Game, F. L., Bartlett, W. A., Marshall, T., & Bayly, G. R. (2001). Comparative accuracy of cardiovascular risk prediction methods in primary care patients. *Heart, 85*, 37-43.

Kaiser, C., & Bodendorf, F. (2009). *Opinion and Relationship Mining in Online Forums*. Paper presented at the Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01.

Kass, G. V. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Applied Statistics.

Kassambara, A., & Mundt, F. (2016). Factoextra: extract and visualize the results of multivariate data analyses.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344).

Kline, P. (2014). *An Easy Guide to Factor Analysis*: Taylor & Francis.

Kons, D. (2016). Learning Analytics en la educación. De primaria a bachillerato, from http://comunidad.konseye.es/

Kortenkamp, U., & Ladel, S. (2014). Flexible use and understanding of place value via traditional and digital tools. *RESEARCH REPORTS KNO–PI, 33*.

Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications, 39*(3), 2584-2589. doi: 10.1016/j.eswa.2011.08.113

Lerman, I. C. (1981). Classification et Analyse Ordinale des Données. *Dunond*.

Li, J., Le, T. D., Liu, L., Liu, J., Jin, Z., & Sun, B. (2013). *Mining causal association rules.* Paper presented at the Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on.

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). *A symbolic representation of time series, with implications for streaming algorithms*. Paper presented at the Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, San Diego, California.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., . . . Gonzalez, J. (2017). "Finding Groups in Data'': Cluster Analysis Extended Rousseeuw et al.

Malbernat, L. R., Clemens, M. P., Varela, A. E., & Urrizaga, M. (2015). *Aplicación de técnicas de data mining en gestión de docentes de educación superior.* Paper presented at the XVII Workshop de Investigadores en Ciencias de la Computación (Salta, 2015).

Martínez, R., Cruz, N., Acosta, H., Rabatte, I., Gogeascoechea, M., Pavón, P., & Blázquez, S. (2009). Decision trees as a tool in the medical diagnosis. *Revista médica de la Universidad Veracruzana*.

Mayor, E. (2015). *Learning Predictive Analytics with R*. Birmingham: Published by Packt Publishing Ltd.

Merceron, A., & Yacef, K. (2005). *Educational Data Mining: a Case Study*. Paper presented at the Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology.

Mersmann, O., Beleites, C., Hurling, R., & Friedman, A. (2014). Microbenchmark: Accurate Timing Functions. *URL http://CRAN. R-project. org/package= microbenchmark. R package version*, 1.4-2.

Montes, M. D., & Ursini, S. (2014). CHIC en el análisis de las actitudes hacia las matemáticas en estudiantes de secundariaCHIC: analyzing middle school student's attitudes towards mathematics. *Educação Matemática Pesquisa, 16*(3).

Moreira Huilcapi, R., Gabriel, & Lojan Cueva, E. L. (2016). Análisis comparativo del rendimiento del procesador en equipos computacionales con sistemas operativos windows, linux y mac os en empresas informáticas. .

Moreno García, M., Segrera, S., & Batista, V. (2005). *Association Rules: Problems, solutions and new applications Abstract*.

Moreno García, M. a. N., Quintales, L. A. M., García Peñalvo, F. J., & Polo Martín, M. J. (2004). Building knowledge discovery-driven models for decision support in project management. *Decision Support Systems, 38*(2), 305-317. doi: https://doi.org/10.1016/S0167-9236(03)00100-3

Moscoso-Zea, O., & Luján-Mora, S. (2016, 15-18 June 2016). *Educational data mining: An holistic view.* Paper presented at the 2016 11th Iberian Conference on Information Systems and Technologies (CISTI).

Müllner, D. (2017). Fast Hierarchical Clustering Routines for R and Python.

Nithyasri, B., Nandhini, K., & Chandra, E. (2011). Classification Techniques in Education Domain. *International Journal of Computer, Mathematical Sciences and Applications 5*, 15-23.

Ohlmacher, G. C., & Davis, J. C. (2003). Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering Geology, 69*(3–4), 331-343. doi: https://doi.org/10.1016/S0013-7952(03)00069-3

Orús, M. d. P., Peydró, L., & Gregori, P. (2013). El centro de recursos CRDM-Guy Brousseau y el análisis estadístico implicativo como herramienta en la formación de profesores. *Revista de didáctica de la Estadística*, 221-228.

Osuna Alarcón, M. R., & De La Cruz Gómez, E. (2010). Los sistemas de gestión de contenidos en Información y Documentación *Revista General de Información y Documentación, 20 (2010) 67-100*.

Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical EvidenceLiterature. *Educational Technology & Society*.

Papamitsiou, Z. K., Terzis, V., & Economides, A. A. (2014). Temporal learning analytics for computer based testing. 31-35. doi: 10.1145/2567574.2567609

Patterson, D. A., & Hennessy, J. L. (2004). *Estructura y diseño de computadores: interficie circuitería-programación* (Vol. 1): Reverté.

Pazmiño-Maji, R., García-Peñalvo, F., & Conde-González, M. (2017). *Statistical Implicative Analysis aproximation to KDD and Data Mining: A systematic and mapping review in Knowledge Discovery Database framework*.

Pazmiño-Maji, R. A., García-Peñalvo, F. J., & Conde-González, M. (2016). *Approximation of statistical implicative analysis to learning analytics: a systematic review*. Paper presented at the Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain.

Pazmiño-Maji, R. A., García-Peñalvo, F. J., & Conde-González, M. (2017). *Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics*. Paper presented at the Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, Cádiz, Spain.

Pazmiño-Maji, R. A., García-Peñalvo, F. J., & Conde-González, M. A. (2016). Approximation of statistical implicative analysis to learning analytics. 355-376. doi: 10.1145/3012430.3012540

Pazmiño, R. (2014). Aproximación al Análisis Estadístico Implicativo desde sus Aplicaciones Educativas.

Pazmiño, R., García-Peñalvo, F. J., Coutrier, R., & Conde-González, M. (2015). Statistical implicative analysis for educational data sets: 2 analysis with RCHIC.

Pérez, D. (2014). *Sistema de inteligencia embebida con autoaprendizaje basado en una arquitectura de árbol de decisión dinámico y adaptativo.* Universidad Politécnica de Madrid.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint arXiv:1203.4745*.

Radès, I. s. d. é. t. d. (2015). *A.S.I. 8*. Paper presented at the 8th International Meeting Statistical Implicative Analysis, Tunise.

Rusu, C. (2011). Metodología de la Investigación. *Obtenido de http://zeus. inf. ucv. cl/~ rsoto/cursos/DII711/Cap1_DII711. pdf*.

Santos, J. L., Govaerts, S., Verbert, K., & Duval, E. (2012). *Goal-oriented visualizations of activity tracking: a case study with engineering students*. Paper presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia, Canada.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD Project: Constraint Based Aids to Causal Model Specification. *Multivariate Behavioral Research, 33*(1), 65-117. doi: 10.1207/s15327906mbr3301_3

Scholtzová, I. (2014). Determinants of primary mathematics education–a national and international context. *Acta Mathematica 17*, 15.

Siemens, G. (2012). *Learning analytics: envisioning a research discipline and a domain of practice*. Paper presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia, Canada.

Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist*.

Siemens, G., & Baker, R. S. J. d. (2012). *Learning analytics and educational data mining: towards communication and collaboration*. Paper presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia, Canada.

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE review*, 46(45), 30.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*.

Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In P. Apers, M. Bouzeghoub & G. Gardarin (Eds.), *Advances in Database Technology — EDBT '96: 5th International Conference on Extending Database Technology Avignon, France, March 25–29, 1996 Proceedings* (pp. 1-17). Berlin, Heidelberg: Springer Berlin Heidelberg.

Steiner, C., UoB, D. M., UoB, M. J., Türker, A., Drnek, M., & Kickmeier-Rust, M. (2014). LA and EDM Approaches.

Tatsuoka, K. K. (1995). *Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. Cognitively diagnostic assessment*.

Theobald, R., & Freeman, S. (2014). Is It the Intervention or the Students? Using Linear Regression to Control for Student Characteristics in Undergraduate STEM Education Research. *CBE-Life Sciences Education, 13*(1), 41-48. doi: 10.1187/cbe-13-07-0136

Urdaneta, U. R. (sn). Statistix for windows

Van den Heuvel-Panhuizen, M., Elia, I., & Robitzsch, A. (2015). Kindergartners' performance in two types of imaginary perspective-taking. *ZDM, 47*(3), 345-362.

Van Harmelen, M., & Workman, D. (2012). Analytics for learning and teaching. *CETIS Analytics Series, 1*(3), 1-40.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics, 15*(4), 405-419. doi: 10.1080/09645290701409939

Wickham, H., & Chang, W. (2016). Create Elegant Data Visualisations Using the Grammar of Graphics.

Wolfgang, G., & Hendrik, D. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Journal of Educational Technology & Society, 15*(3), 42-57.

Yang, X. Y., Liu, Z., & Fu, Y. (2010, 23-25 June 2010). *MapReduce as a programming model for association rules algorithm on Hadoop.* Paper presented at the The 3rd International Conference on Information Sciences and Interaction Sciences.

Zamora-Matamoros, L., Díaz-Silvera, J. R., & Portuondo-Mallet, L. (2015). Fundamental Concepts on Classification and Statistical Implicative Analysis for Modal Variables. *Revista Colombiana de Estadística*, 335-351.

Zamora, L., & Díaz, J. (2008). Aplicación del análisis estadístico implicativo al estudio del rendimiento académico de estudiantes de primer año de las carreras de matemática y ciencia de la computación. *Série Estatística.-Universidade do Estado do Rio de Janeiro*, 1-17.

Zamora, L., Gregori, P., & Orús, P. (2009). *Conceptos Fundamentales del Análisis Estadístico Implicativo (ASI) y su Soporte Computacional CHIC.* Universitat Jaume I de Castellón, España.

Zhao, Y., & Bhowmick, S. S. (2015). Association Rule Mining with R. *A Survey Nanyang Technological University, Singapore*.

Zilková, K., Guncaga, J., & Kopácová, J. (2015). *(Mis)Conceptions about Geometric Shapes In Pre-Service Primary Teachers* (Vol. 8(1)).

Glosario

Learning Analytics (LA): es el uso de BigData para proporcionar inteligencia accionable para los estudiantes y docentes. (Ferguson, 2016)

Análisis Estadístico Implicativo (AEI): Un método de análisis de datos para la búsqueda de causalidades. (Gras et al., 2008)

ANOVA: es un modelo de análisis de varianza, permite comparar dos poblaciones de dos muestras. Utiliza supuestos de normalidad, homogeneidad de las varianzas, entre otros. (Urdaneta, sn)

CHIC: *software* para clasificación jerárquica implicativa y cohesiva. (Iurato, 2012)

Content Management Systems (CMS): son herramientas de gestión de la información, las cuales brindan soluciones globales a institución y organización. Estas herramientas han recibido la denominación de Sistemas de Gestión de Contenidos. (Osuna Alarcón & De La Cruz Gómez, 2010)

Educational Data Mining (EDM): se centra actualmente en descubrir patrones significativos en los datos educativos. (Amershi & Conati, 2009)

Dendrogramas: es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud.(International Sales and Support, 2017)

RCHIC: *software* AEI multiplataforma. (Pazmiño et al., 2015)

Data Mining (DM): técnicas y algoritmos que ayudan a la extracción de conocimientos desde la base de datos, para la toma de decisiones. (Moscoso-Zea & Luján-Mora, 2016)

Classification and regression tree (CART): conjunto de clasificación y regresión de árboles. (Gislason et al., 2006)

Root mean squared error (RMSE): métricas de correlación lineal o de error medio cuadrático.

Hierarchical agglomerative clustering (HAC): algoritmo de agrupamiento aglomerado jerárquico. (Amershi & Conati, 2009)

Imaginary perspective taking (IPT): capacidades imaginarias de toma de perspectiva de los niños de jardín.

Mathematical Working Space (MWS): marco para estudiar el trabajo matemático de los estudiantes. (Elia et al., 2016)

Technology Enhanced Learning (TEL): analítica en tecnología de aprendizaje mejorado. (Chatti et al., 2012)

LAK Data Challenge: conferencias que proporcionan acceso a metadatos estructurados de publicaciones de investigación en el campo del análisis de aprendizaje. (Drachsler et al., 2014)

Método de estudio de similitud entre modelos y estándares (MSSS): es un método de estudio de similitud desarrollado por la Universidad Politécnica de Madrid (Calvo-Manzano et al., 2008)