

PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”

DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.



CONFERENCIA INTERNACIONAL DE PROCESAMIENTO DE
LA INFORMACIÓN (CIPI 2019)

**Entorno para el minado de datos en ambientes educativos
utilizando la herramienta Spark**

*Environment for data mining in educational environments using the
Spark tool*

MSc. Neysa Baldoquin Alonso¹, Dr. C. Juan Pedro Febles Rodríguez², MSc. Ángel
Alberto Vázquez Sánchez³

1- Neysa Baldoquin Alonso. Universidad de las Ciencias Informáticas, Cuba.

neysa@uci.cu

2- Juan Pedro Febles Rodríguez. Universidad de las Ciencias Informáticas, Cuba.

febles@uci.cu

3- Ángel Alberto Vázquez Sánchez. Universidad de las Ciencias Informáticas, Cuba.

aavazquez@uci.cu

Resumen:

El desarrollo de las tecnologías de la información y las comunicaciones ha abierto una gama de posibilidades en el ámbito educativo. Uno de los mayores desafíos que afrontan las instituciones educativas es el crecimiento exponencial de los datos educativos y la utilización de los mismos para mejorar las decisiones administrativas. La aplicación de la minería de datos en la educación es concebida con el desarrollo de métodos para explorar los datos que provienen de entornos educativos. Para la exploración de dichos datos, existen técnicas novedosas que permiten realizar grandes procesamientos de cómputo utilizando hardware de pocas prestaciones y técnicas de minería de datos distribuida. Sobre esta base, existen modelos desarrollados para estimar y predecir cómo y cuánto aprenden los estudiantes, los cuales necesitan un entorno que brinde soporte a los procesos que se ejecutan dentro de estos modelos para garantizar el procesamiento de los datos de forma eficiente. El objetivo de la presente

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”

DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.



investigación es desarrollar un entorno distribuido para contribuir al uso de técnicas de minería de datos distribuida de forma eficiente en modelos de estimación del conocimiento latente aplicados a datos educacionales masivos. Como resultado de la investigación que se llevó a cabo, se obtuvo un entorno distribuido capaz de aglutinar bases de datos heterogéneas que son utilizadas para el procesamiento eficiente de datos a través de técnicas de minería de datos educacionales. A través de las diversas pruebas realizadas, se validó la eficiencia con que cuenta dicho entorno para su explotación.

***Abstract:** The development of information and communication technologies has opened a range of possibilities in the educational field. One of the biggest challenges facing educational institutions is the exponential growth of educational data and their use to improve administrative decisions. The application of data mining in education is conceived with the development of methods to explore the data that come from educational environments. For the exploration of such data, there are novel techniques that allow large computational processing using low-performance hardware and distributed data mining techniques. On this basis, there are models developed to estimate and predict how and how much students learn, which need an environment that provides support to the processes that run within these models to ensure the processing of data efficiently. The objective of this research is to develop a distributed environment to contribute to the use of Distributed Data Mining Techniques in an efficient way in Latent Knowledge Estimation Models applied to massive educational data. As a result of the research that was carried out, a distributed environment was obtained capable of agglutinating heterogeneous databases that are used for the efficient processing of data through educational data mining techniques. Through the various tests carried out, the efficiency of this environment for its exploitation was validated.*

Palabras Clave: Datos educativos, Minería de datos; Minería de datos educativos; Entorno distribuido; Eficiencia.

Keywords: Educational data, Data mining; Educational data mining; Distributed environment; Efficiency.

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

**PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”**

**DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.**



1. Introducción

Uno de los mayores desafíos que afrontan las instituciones educativas es el crecimiento exponencial de los datos educativos y la utilización de los mismos para mejorar las decisiones administrativas (Cristobal Romero, 2013). Constantemente usuarios tales como: instructores, estudiantes y administradores de cursos interactúan con sistemas educativos informatizados como son las aulas inteligentes, sistemas *e-learning*, LMS (*Learning Management System* o Sistema de Gestión de Aprendizaje) entre otros. Los mismos generan un gran cúmulo de información que es registrada en bases de datos educacionales. A partir de la generación de estas bases de datos son aplicadas técnicas de minería de datos, las cuales funcionan a través de algoritmos de inteligencia artificial, haciendo posible la identificación de patrones de comportamiento para la estimación del conocimiento latente (S. Aghabozorgi, 2014). La Minería de Datos (MD), se plantea como un área del conocimiento que tiene como objetivo la aplicación de técnicas para procesar y analizar grandes volúmenes de información con los que se puede descubrir patrones repetitivos, tendencias, o reglas que expliquen el comportamiento de los datos en un determinado contexto (S.L. González-Ruiz, 2015). La aplicación de minería de datos en la educación es un campo emergente, conocido como Minería de Datos Educativos (MDE), concebido con el desarrollo de métodos para explorar los tipos únicos de datos que provienen de entornos educativos. Se utiliza para entender mejor cómo los estudiantes identifican los entornos en los que aprenden, mejorar los resultados educativos y a comprender y explicar los fenómenos educativos (Cristobal Romero, 2013). Las técnicas de minería de datos utilizadas en el ámbito educativo, se han empleado con éxito para crear modelos predictivos en cuanto al rendimiento de los estudiantes (Carlos Márquez Vera, 2012). Participar e inferir en el conocimiento generado por los estudiantes a partir de técnicas computacionales puede ser de gran utilidad (Corbett, 2014). El área de la estimación de conocimiento latente (ECL) es de particular importancia dentro de la MDE, en la estimación del conocimiento latente, el dominio de un estudiante sobre determinadas habilidades y

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

**PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”**

**DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.**



conceptos es evaluado por sus patrones de corrección de esas habilidades. El término “latente” es referido a la idea de que el conocimiento no puede ser medido directamente, sino inferido a través del rendimiento del estudiante. Incrementar el conocimiento de los estudiantes es la principal meta de la educación, es por ello que, si el conocimiento puede ser inferido, se puede determinar de qué forma aumenta la capacidad de aprendizaje del estudiante, y por tanto ejecutar decisiones pedagógicas automáticas (Corbett, 2014). Para la estimación del conocimiento latente son aplicados diversos algoritmos de minería de datos, los mismos requieren numerosos cálculos complejos. La minería de datos distribuida (MDD) tiene como objetivo extraer información útil de los datos ubicados en sitios heterogéneos (Dillip Kumar Swain, 2017) (Chikhale, 2016) (Martínez, 2015), siguiendo una estrategia de minería descentralizada que difiere de la estrategia centralizada (S.Urmela, 2017). Los sistemas distribuidos están concebidos desde el punto de vista de las necesidades que las aplicaciones plantean, los cuales están basadas en realizar tareas de cómputo de alto rendimiento con hardware de bajas prestaciones (Lafuente, 2015). En este sentido, el Grupo de Investigación de Tecnologías Educativas (GITAE) de la Universidad de las Ciencias Informáticas (UCI) se ha propuesto determinar cómo estimar el conocimiento latente a partir de datos educacionales masivos, haciendo uso de técnicas de minería de datos distribuidas de forma eficiente. Es necesario tener en cuenta para arribar a la solución de dicha problemática el contexto en que se basa la educación cubana, teniendo en cuenta los recursos, tanto físicos como humanos, que se destinan con el fin de la informatización de nuestro país. Para dar cumplimiento a la problemática planteada se establece como objetivo de la presente investigación: Desarrollar un entorno distribuido para contribuir al uso de técnicas de minería de datos distribuida de forma eficiente en modelos de estimación del conocimiento latente aplicados a datos educacionales masivos.

2. Metodología

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”



DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.

- ✓ **Experimento:** Sirvió como base para validar el cumplimiento del objetivo planteado. Fue desarrollado en un ambiente informático real para poder hacer constar la eficiencia de la propuesta de solución.

3. Resultados y discusión

3.1 Modelo para la estimación del conocimiento latente en datos educacionales masivos

Teniendo en cuenta el objetivo planteado en la presente investigación para dar cumplimiento a la problemática, se diseñó un modelo para la estimación de conocimiento latente en datos educacionales masivos el cual se muestra en la Figura 1.

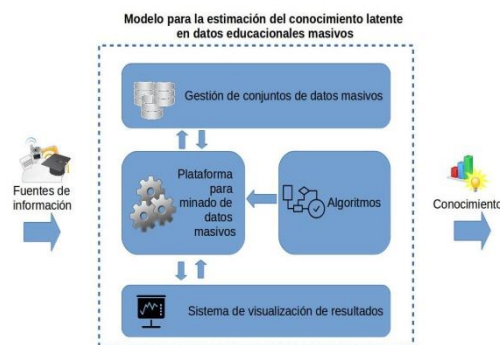


Figura 1: Modelo para la estimación del conocimiento latente en datos educacionales masivos.

El modelo está compuesto por cuatro fases, dando inicio en la entrada de diversas fuentes de información educacionales. Los datos educacionales almacenados son agrupados para formar un gran conjunto, el cual es tratado en el primer proceso que conforma el modelo de la Figura 1, que describe el proceso de gestión de conjunto de datos masivos. Posteriormente los datos obtenidos, se trasladan hacia el segundo proceso, denominado plataforma para el minado de datos masivos, la cual interactúa directamente con los algoritmos de minería de datos a utilizarse, en la que los resultados del procesamiento son trasladados al sistema de visualización. De esta manera a través de técnicas predictivas es obtenida la estimación del conocimiento. La plataforma para el minado de datos masivos, ocupa gran relevancia dentro del modelo, ya que es la encargada de recibir los datos del proceso anterior y realizar el agrupamiento de dichos

PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTIFICA INTERNACIONAL
“II CCI UCLV 2019”



DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.

datos. Posteriormente, se aplican dentro de la plataforma los algoritmos, asegurando a su vez que los mismos se ejecuten de forma eficiente teniendo en cuenta el hardware disponible. Finalmente, es obtenida la estimación del conocimiento, la cual es visualizada en el siguiente proceso del modelo descrito anteriormente.

3.2 Arquitectura distribuida para el entorno desarrollado

A partir de la utilización de la minería de datos distribuida en el desarrollo del entorno, se estableció una arquitectura (Figura 2).

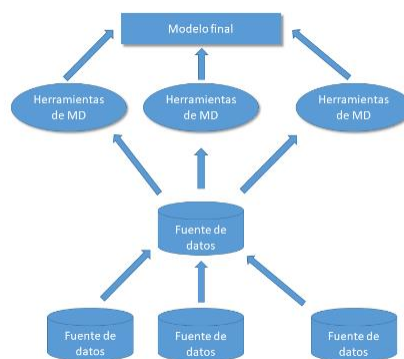


Figura 2: Arquitectura generada para el entorno distribuido para el minado de datos en ambientes educativos masivos.

Para el primer nivel se tuvo en cuenta la distribución de diferentes fuentes de datos educacionales, para en un segundo nivel realizar su unificación a través de una vista integrada de datos. A partir de la vista integrada de datos confeccionada se ejecuta el tercer nivel, en el cual se realizará la implementación de los algoritmos que funcionan dentro del entorno. Los algoritmos utilizados constituyen modificaciones para ajustarlos a un contexto de *Big Data* y lograr la eficiencia de los resultados que son obtenidos. Estos algoritmos son: Rastreo Bayesiano del Conocimiento (*Bayesian Knowledge Tracing*, BKT), Teoría de Respuesta al Ítem (*Item Response Theory*, IRT) y Algoritmo de Factor de Rendimiento (*Performance Factor Algorithm*, PFA), algoritmos propios para estimar conocimiento latente (Toledano, 2018). Los resultados obtenidos por son expuestos en un modelo global en el cuarto nivel, haciendo uso de las llamadas curvas de aprendizaje, las cuales infieren el conocimiento estimado que ostenta un estudiante

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.

sobre un tema en particular, es decir, el estudiante tendrá la posibilidad de visualizar por tres algoritmos diferentes el conocimiento estimado que posee (Perlich, 2011).

3.3 Flujo de procesos del entorno distribuido para el minado de datos educativos masivos

Para la confección del entorno se identificaron las fuentes de datos que se utilizan como entrada para el proceso de minado que se realiza dentro del entorno y los datos de salida que necesita generar el mismo. La Figura 3 representa el flujo de procesos del entorno distribuido.

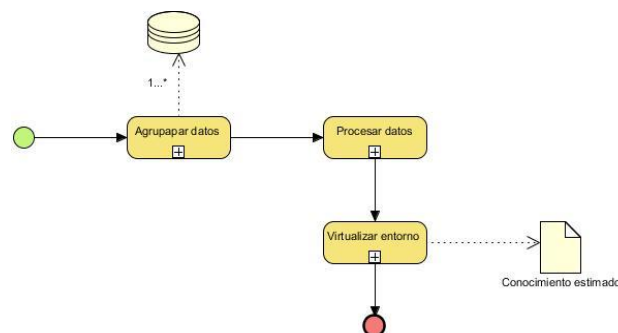


Figura 3: Diagrama de flujo: Entorno distribuido para el minado de datos en ambientes educativos masivos.

3.3.1 Agrupar datos

Para el proceso de agrupamiento de datos se hará uso de la herramienta Hadoop, basado en el sistema Hadoop Distributed File System o Hadoop HDFS, que contribuyen al proceso de agrupamiento de datos masivos. El inicio del proceso, se realiza a partir de la entrada de bases de datos distribuidas provenientes del primer proceso que describe el modelo para la estimación del conocimiento latente en datos educativos masivos (Figura 1), las cuales son accedidas a través de una dirección URL, realizándose el proceso de forma virtual para la conformación de los *datasets*. Para realizar el proceso de consulta desde los nodos trabajadores al directorio HDFS, los nodos acceden a la dirección URL proporcionada por el *master* que hace referencia al algoritmo que se desea utilizar en conjunto con el *dataset* seleccionado. Cada nodo trabajador puede acceder a la URL proporcionada por el *master*, para a partir de la misma realizar el

PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”

DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.



procesamiento distribuido de los datos. El resultado del procesamiento ejecutado por cada nodo trabajador, es almacenado en un fichero ubicado en el HDFS en el nodo *master*, el cual es accedido a través de la dirección URL.

3.3.2 Procesar datos

Teniendo en cuenta el fichero obtenido en el proceso anterior, se procede al procesamiento de dichos datos. Sobre la base de la aplicación del modelo propuesto en la Figura 1 para un contexto de *Big Data*, fue utilizada la herramienta Spark utilizando la librería MLlib, para la creación de los llamados RDD (Datasets Distribuidos Resistentes), apoyado en el método `textFile`. Dicho método crea un RDD a través del archivo de texto que es leído del directorio almacenado en el HDFS y realiza transformaciones a partir de las modificaciones realizadas a los algoritmos identificados BKT, IRT y PFA. Posteriormente de realizadas las transformaciones correspondientes, es obtenido un fichero que contiene el conocimiento estimado almacenado en el directorio HDFS. El fichero obtenido con el conocimiento estimado proveniente de los nodos trabajadores es almacenado en el nodo *master*.

3.3.3 Virtualizar entorno

La virtualización ofrece características importantes en la creación del entorno. La misma propicia un mejor aprovechamiento y ahorro del hardware disponible, lo que conlleva a una reducción de costes en cuanto a mantenimiento y administración. Contribuye además a aligerar los procesos de minado de datos que se realizan, contribuyendo con la eficiencia dentro del entorno. Docker fue la herramienta destinada para el proceso de virtualización, el mismo funciona a través de contenedores aislados. Para este proceso fue utilizada una imagen de la herramienta de software libre Spark en su versión 2.1.1, la cual se encargará del proceso de minado de datos. El nodo *master*, realiza el proceso de clusterización a través de tres nodos trabajadores, que a su vez ejecutan Docker Engine en un contenedor con la imagen de Spark. Fue puesto en práctica además Docker Swarm, que controla un grupo de hosts de Docker y lo expone como un solo host virtual, el mismo realiza la orquestación del *cluster*, posibilitando

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”

DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.



mediante su utilización un entorno controlado y administrado por sí mismo. Cada nodo trabajador será supervisado constantemente por el nodo *master*, de forma tal que se conozca el estado real de cada nodo trabajador.

4. Experimento para evaluar la eficiencia del entorno distribuido

Para la realización de este método experimental fue puesto en funcionamiento el entorno desarrollado para el minado de datos con datos experimentales sobre un total de 19 iteraciones. Sobre la base de los mismos datos, también fueron minados datos fuera del entorno, lo que permitió que a través de los resultados arrojados en ambos casos que se pudiera establecer un análisis y comparación de los mismos para evaluar la eficiencia en cada caso. Debido a que aún no se cuenta con un *dataset* con datos educacionales masivos de las diversas plataformas educativas y aulas inteligentes puestas en funcionamiento por la UCI, se tomó como base para el experimento un *dataset* con un tamaño de 2000x50, el cual es representado en un vector de respuesta dicotómico de X filas con Y columnas, lo que representa que X es el número de estudiantes, mientras que Y es la cantidad de ítems evaluativos de la muestra (Meyer, 2015). El hardware utilizado para el proceso de prueba estuvo dado por un total de tres nodos destinados a ser trabajadores y tres nodos ejecutados como *master*. Dentro de las características de los nodos trabajadores se enuncia que el procesador utilizado fue un Intel Celeron, con una velocidad de 2,8 GH, dos núcleos, 4GB de memoria RAM y como sistema operativo fue utilizado Ubuntu 16-0.4 LTS 64 bit. Para el caso de los nodos *masters* se utilizó en cada caso un Core i7 como tipo de procesador, con una velocidad de 2,8 GH, de 8 núcleos, 8 GB de memoria RAM e igual sistema operativo que el caso anterior.

4.1 Prueba de Ajuste de Datos No Censurados - FUERA DEL ENTORNO

Datos/Variable: FUERA ENTORNO;19 valores con rango desde 60226,0 a 62377,0. Distribuciones ajustadas fuera del entorno: media = 60926,3; desviación estándar = 593,663. Este análisis muestra los resultados de ajustar una distribución normal a los datos de la variable FUERA ENTORNO.

4.1.1 Pruebas de Normalidad para FUERA DEL ENTORNO

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

**PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”**



**DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.**

Tabla 1: Pruebas de normalidad realizadas Fuera del entorno. Fuente: Elaboración propia.

Prueba	Estadístico	Valor-P
Chi-Cuadrado	27,1053	0,00134377
Estadístico W de Shapiro-Wilk	0,779897	0,000344473

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que la variable FUERA ENTORNO proviene de una distribución normal con 95% de confianza.

4.2 Prueba de Ajuste de Datos No Censurados - DENTRO DEL ENTORNO

Datos/Variable: DENTRO ENTORNO; 19 valores con rango desde 41080,0 a 44238,0. Distribuciones ajustadas dentro del entorno: media = 41950,7; desviación estándar = 1029,77. El análisis muestra los resultados de ajustar una distribución normal a los datos de la variable DENTRO ENTORNO.

4.2.1 Pruebas de Normalidad para DENTRO DEL ENTORNO

Tabla 2: Pruebas de normalidad Fuera del entorno. Fuente: Elaboración propia.

Prueba	Estadístico	Valor-P
Chi-Cuadrado	41,0	0,00000500194
Estadístico W de Shapiro-Wilk	0,687168	0,0000122205

Debido a que el valor-P más pequeño de las pruebas realizadas es menor a 0,05, se puede rechazar la idea de que la variable DENTRO ENTORNO proviene de una distribución normal con 95% de confianza.

4.3 Resumen estadístico

Tabla 3: Resumen estadístico. Fuente: Elaboración propia.

	FUERA ENTORNO	DENTRO ENTORNO
Recuento	19	19
Promedio	60926,3	41950,7
Mediana	60769,0	41521,0

**PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”**



**DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.**

Los intervalos de confianza del 95,0% para la media de la variable FUERA ENTORNO: 60926,3 +/- 286,137 [60640,1; 61212,4], mientras que los intervalos de confianza del 95,0% para la media de la variable DENTRO ENTORNO: 41950,7 +/- 496,335 [41454,4; 42447,1]. Se ejecuta una prueba -t para comparar las medias de las dos muestras y se construyen los intervalos, o cotas de confianza para cada media y para la diferencia entre las mismas. De interés particular es el intervalo de confianza para la diferencia entre las medias, el cual se extiende desde 18417,6 hasta 19533,4. Puesto que el intervalo no contiene el valor 0, existe una diferencia estadísticamente significativa entre las medias de las dos muestras, con un nivel de confianza del 95,0%.

Prueba t para comparar medias

Hipótesis nula: $media1 = media2$ Hipótesis Alt.: $media1 \neq media2$

Sin suponer varianzas iguales: $t = 69,5857$; $valor-P = 0,0$. Se rechaza la hipótesis nula para $\alpha = 0,05$. La prueba que se presenta, se ha construido para determinar si la diferencia entre las dos medias es igual a 0,0 versus la hipótesis alterna de que la diferencia no es igual a 0,0. Puesto que el valor -P calculado es menor que 0,05, se puede rechazar la hipótesis nula en favor de la alterna.

4.4 Comparación de Medianas

Se ejecuta la prueba W de Mann-Whitney, prueba no paramétrica para comparar las medianas de dos muestras. Debido a que el valor-P es menor que 0,05, existe una diferencia estadísticamente significativa entre las medianas con un nivel de confianza del 95,0%. Mediana de muestra 1: 60769,0; Mediana de muestra 2: 41521,0.

Prueba W de Mann-Whitney (Wilcoxon) para comparar medianas:

Hipótesis Nula: $mediana1 = mediana2$; Hipótesis Alt.: $mediana1 \neq mediana2$

Rango Promedio de muestra 1: 29,0; Rango Promedio de muestra 2: 10,0

$W = 0,0$; $valor-P = 1,48277E-7$. Se rechaza la hipótesis nula para $\alpha = 0,05$, lo que significa que las medianas difieren significativamente.

Se demuestra que los intervalos de confianza del 95,0% para la media de la variable FUERA ENTORNO: 60926,3 +/- 286,137 [60640,1; 61212,4] y los intervalos de

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

**PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”**

**DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.**



confianza del 95,0% para la media de la variable DENTRO ENTORNO: 41950,7 +/- 496,335[41454,4; 42447,1], se puede concluir que dentro del entorno la eficiencia dada por la velocidad de procesamiento de los datos minados, es 1/3 más eficiente que los datos que son minados fuera del entorno.

4. Conclusiones

Finalizada la investigación se arriban a las siguientes conclusiones generales:

- ✓ Los procesos de gestión de datos, procesamiento de datos y virtualización del entorno, asociados a la arquitectura y modelo propuesto, permiten dar cumplimiento a las características de transparencia, escalabilidad y tolerancia a fallos, y garantizan la eficiencia dentro del entorno.
- ✓ Los métodos científicos utilizados para la validación demuestran la eficiencia del entorno y su factibilidad para ser utilizado en modelos de estimación del conocimiento latente aplicados a datos educacionales masivos.

5. Referencias bibliográficas

- Carlos Márquez Vera, C. R. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *Vol. 7*(Núm. 3).
- Corbett, R. S. (2014). Assessment of robust learning with educational data mining. *9*.
- Cristobal Romero, S. V. (2013). Data mining in education. *WIRES Data Mining and knowledge discovery, 3*, 12-27.
- Dillip Kumar Swain, S. M. (Mayo de 2017). Privacy Preservation in Distributed Data Mining Using Secured Multiparty Computation . *International Journal of Scientific Research Engineering & Technology (IJSRET), 6*(Issue 5), 494-499.
- Lafuente, A. (2015). Introducción a los sistemas distribuidos. En A. Lafuente. UPV/EHU .
- Martínez, J. J. (2015). Análisis de minería de datos distribuida con Weka Parallel en computadoras con múltiples procesadores físicos y lógicos. *Economía y Administración (E&A), 6*(2), 154-165.
- Meyer, J. P. (Abril de 2015). jMetrik: Open-Source Software for Psychometric Analysis. *AMERICAN PSYCHOLOGICAL ASSOCIATION, 2-10*.
- Perlich, C. (Enero de 2011). Learning Curves in Machine Learning. *IBM Research Report*.

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu

**PLANTILLA OFICIAL PARA LA PRESENTACIÓN DE TRABAJOS
II CONVENCION CIENTÍFICA INTERNACIONAL
“II CCI UCLV 2019”**

**DEL 23 AL 30 DE JUNIO DEL 2019.
CAYOS DE VILLA CLARA. CUBA.**



- S. Aghabozorgi, H. M. (2014). An Approachable Analytical Study on Big Educational Data Mining.
- S.L. González-Ruiz, I. G.-G.-B.-M. (2015). Algoritmos de clasificación y redes neuronales en la observación automatizada de registros. *Servicio de Publicaciones de la Universidad de Murcia*, 15(1).
- S.Urmela, M. (Marzo de 2017). Approaches and Techniques of Distributed Data Mining : A Comprehensive Study. *International Journal of Engineering and Technology (IJET)*, 9(1).
- Toledano, O. G. (2018). Adaptación del algoritmo Teoría de Respuesta al Ítem para la estimación del conocimiento latente sobre Datos Educativos. *III Conferencia Científica Internacional UCIENCIA 2018*. La Habana: Ediciones Futuro.

Información de contacto
convencionuclv@uclv.cu
www.uclv.edu.cu