Survey of academic data using data mining tools

Leticia Laura Ochoa, Mg[1], Karina Rosas Paredes, Mg[1], César Baluarte Araya, Dr[1] [1]Universidad Católica de Santa María, Perú, llaura@ucsm.edu.pe, kparedes@ucsm.edu.pe, cbaluart@ucsm.edu.pe

***Abstract*– To support the decision-making of the authorities and academic coordinators of educational entities as well as analyze and explore the academic performance of the students in order to reduce the level of dropout and re-enrollment of the same courses, in this work, an academic data survey is conducted through basic statistical tests. Furthermore, histograms that show the frequency of notes are created. In addition, charts of probability distribution are prepared to analyze the behavior of the notes according to the curve that is formed. Next, a matrix of correlations that allows to find the relationship between the courses is created. The main components examined to generate the main plane, where the students are shown with a circle of correlations of the courses. Additionally, the If-means algorithm of data mining is employed for grouping students according to their academic performance using data mining tools, e.g., Weka and R. This allows us to obtain timely knowledge for improving the teaching–learning process.**

***Keywords*– Academic data survey, academic performance, data mining, teaching–learning.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2017.1.1.116
ISBN: 978-0-9993443-0-9
ISSN: 2414-6390

Exploración de Datos Académicos Utilizando Herramientas de Minería de Datos

Leticia Laura Ochoa, Mg[1], Karina Rosas Paredes, Mg[1], César Baluarte Araya, Dr[1]
[1]Universidad Católica de Santa María, Perú, llaura@ucsm.edu.pe, kparedes@ucsm.edu.pe, cbaluart@ucsm.edu.pe

***Resumen*– Con el objetivo de dar soporte a la toma decisiones a las autoridades y coordinadores académicos de entidades educativas, analizar y explorar el rendimiento académico de los alumnos para disminuir el nivel de deserción y reincidencia de matrículas en los mismos cursos, en este trabajo se realiza la exploración de datos académicos mediante pruebas estadísticas básicas, así como la creación de histogramas que muestran la frecuencia de notas, gráficos de distribución de probabilidad para ver el comportamiento de las notas según la curva que se forma, matriz de correlaciones que permite encontrar la relación entre los cursos, análisis de componentes principales para generar el plano principal donde se muestran los alumnos con el círculo de correlaciones de los cursos y la aplicación del algoritmo K-means de minería de datos para la agrupación de alumnos según su rendimiento académico utilizando las herramientas de minería de datos Weka y R, que permiten descubrir conocimiento oportuno para mejorar el proceso de enseñanza-aprendizaje.**

***Palabras claves*– Exploración de datos académicos, rendimiento académico, minería de datos, enseñanza-aprendizaje.**

I. Introducción

Las entidades educativas para asegurar e incrementar la calidad en la educación y disminuir la deserción y reincidencia de matrículas, requieren el uso de herramientas de minería de datos que permitan explorar los datos académicos y descubrir conocimiento oportuno sobre el rendimiento de sus alumnos para tomar acciones de reforzamiento en los mismos.

La minería de datos forma parte integral del descubrimiento de conocimientos en bases de datos (KDD por sus siglas en inglés, Knowledge Discovery in Database). Tal como lo dice su nombre, el proceso de KDD se refiere al proceso global de conversión de los datos en bruto, dentro de las bases de datos, en información útil [1]. Esta tecnología emergente combina el análisis estadístico, el aprendizaje automático, y gestión de base de datos para extraer información de sistemas de bases de datos de gran tamaño. La minería de datos requiere la integración de varias tecnologías [2]. La minería de datos educacional (MDE) permite responder preguntas sobre qué sabe realmente un estudiante y cómo está aprendiendo. De esta manera, la MDE permite descubrir información útil que ayuda a los docentes y responsables de las instituciones educativas en determinar la manera más pertinente para guiar a sus estudiantes, maximizando su aprendizaje [3].

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2017.1.1.116
ISBN: 978-0-9993443-0-9
ISSN: 2414-6390

La MDE emerge como un paradigma orientado para la generalización de modelos, tareas, métodos y algoritmos para la exploración de datos que provienen de un contexto educativo; asimismo tiene como función encontrar, analizar patrones que caractericen los comportamientos en base a sus logros, evaluaciones y el dominio de contenido de conocimiento que tienen los alumnos en los diversos mecanismos de aprendizaje-enseñanza que hoy en día son otorgados en las diversas instituciones públicas y privadas con el objetivo de generar modelos educativos en los cuales puedan fomentar nuevas técnicas o herramientas que puedan analizar e incrementar el nivel participativo de los estudiantes sobre los sistemas de aprendizaje-enseñanza [4].

Weka (Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato) es una plataforma de software para aprendizaje automático y minería de datos escrito en Java. WEKA es un software libre distribuido bajo licencia GNU GPL. Contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades [5].

R es un lenguaje de programación y entorno de software de código abierto para computación y gráficos estadísticos. Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásicas, clasificación, agrupación en clústeres, etc. [6]. R Commander es una interfaz tipo ventana que cubre la mayor parte de los análisis estadísticos más habituales en unos menús desplegables a los que estamos bastante acostumbrados, ya que la mayoría de los programas que utilizamos en cualquier sistema operativo son de este tipo. Podemos decir que es una manera de manejar R sin necesidad de aprender su código o casi nada de él, lo cual lo hace bastante práctico cuando se está aprendiendo a usarlo [7]. Rattle es una interfaz gráfica de usuario para la minería de datos en R, el objetivo es proporcionar una interfaz sencilla e intuitiva que permita al usuario cargar rápidamente datos de un archivo CSV (o mediante ODBC), transformar y explorar los datos, construir y evaluar modelos, y exportar modelos como PMML (Predictive Modelling Markup Language) [8].

El uso de herramientas de minería de datos en las entidades educativas da una ventaja competitiva ya que tienen implementados algoritmos que integran diferentes técnicas de inteligencia artificial, cálculos matemáticos, estadísticos,

sistemas de bases de datos; que permiten encontrar conocimiento útil a los coordinadores para mejorar el proceso de enseñanza-aprendizaje.

II. Metodología

El desarrollo del presente trabajo consta de 6 pasos:

1) *Recolección de datos académicos:* Los datos académicos iniciales fueron proporcionados por la Escuela Profesional de Ingeniería de Sistemas [9] de la Universidad Católica de Santa María [10], los cuales fueron extraídos del Sistema Académico de Ingreso de Notas.
2) *Selección de registros y atributos:* Se seleccionaron los registros correspondientes a alumnos del primer año y atributos principales como código de curso, nombre de la asignatura, código del alumno, nombre del alumno, promedio.
3) *Transformación y limpieza de datos:* La herramienta empleada para limpiar los datos y realizar el proceso de Extracción, Transformación y Carga (ETL) fue Pentaho [11]. Se utilizó para la des-normalización de fila, tratamiento de valores nulos, conversión de tipos, omitir código y nombre de los alumnos.
4) *Exploración de datos académicos:* Incluye la realización de pruebas estadísticas básicas, como la creación de tablas de frecuencia, gráficos de distribución, matriz de correlaciones.
5) *Seleccionar y aplicar la técnica de minería de datos:* Se selecciona y aplica la técnica de minería de datos de clustering para realizar la agrupación de alumnos en dos clusters o grupos, el cual permite separar y explorar los datos de los alumnos con promedios finales desaprobados y con promedios finales aprobados.
6) *Interpretación de datos:* Mediante los resultados obtenidos por las herramientas de minería de datos utilizadas se procede a la interpretación de datos.

III. Resultados y Discusión

En este trabajo se realizó la exploración de datos académicos correspondientes a alumnos del primer año de la carrera profesional de Ingeniería de Sistemas utilizando las herramientas Weka y R.

Las variables seleccionadas para la exploración y análisis se muestran en la Tabla 1.

TABLA 1
Variables Usadas para la Exploración

VARIABLE	DESCRIPCIÓN
CODIGO	Código del Alumno
CALC_DIFE	Cálculo Diferencial
DESA_HUMA	Desarrollo Humano
COMU_ORAL_ESCR	Comunicación Oral y Escrita
FUND_PROG	Fundamentos de Programación
PROG_I	Programación I
ESTR_DISC_I	Estructuras Discretas I

El archivo de notas con el que se trabajo tiene la extensión .CSV, el cual es un documento en formato abierto sencillo para representar datos en forma de tabla, en la que las columnas se separan por comas y las filas por saltos de línea [12].

A. *Resumen de las Variables*

Se exploró los resúmenes de las variables de forma individual por cada curso utilizando la herramienta Weka. En la Fig. 1 se muestra el resumen del curso de Desarrollo Humano, donde se observa la nota mínima y máxima del curso así como el promedio y desviación estándar.

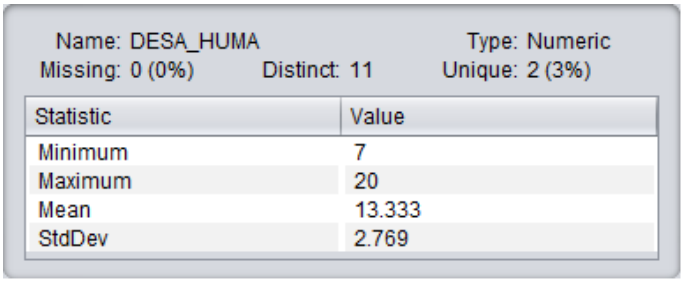

Fig. 1 Resumen del curso de desarrollo humano.

De igual manera se pudo realizar utilizando el paquete R Commander y el lenguaje de programación R como se puede observar en la Fig. 2, donde se muestra la nota mínima, primer cuartil, mediana, promedio, tercer cuartil y nota máxima del curso.

```
> summary(notas$DESA_HUMA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   11.00   13.00   13.33   15.00   20.00
```

Fig. 2 Resumen del curso de desarrollo humano.

B. *Histogramas*

Un histograma es una representación gráfica de una variable en forma de barras. Se utilizan para variables continuas o para variables discretas, con un gran número de datos, y que se han agrupado en clases [13].

Los histogramas se utilizaron para observar la frecuencia de notas de los diferentes cursos, se pudo generar con la herramienta Weka como se observa en la Fig. 3, a través del cual se puede interpretar que hay cuatro alumnos que obtuvieron notas bajas en el curso de Desarrollo Humano comprendidas entre 7 y 9, así como tres alumnos que obtuvieron notas altas comprendidas entre 18 y 20.

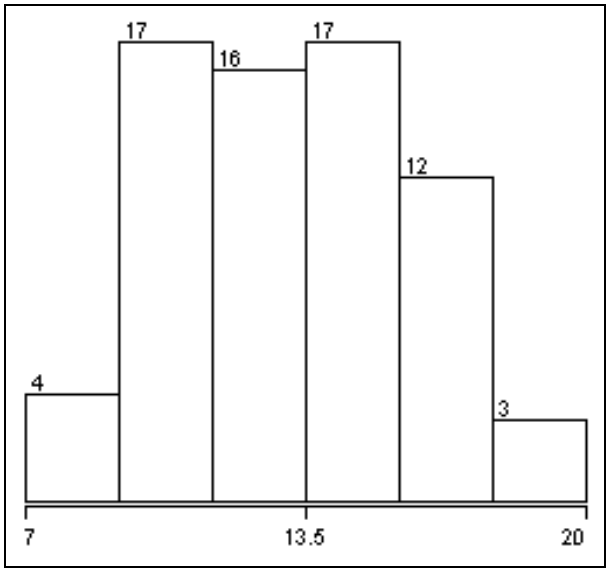

Fig. 3 Histograma del curso de desarrollo humano.

Mediante el paquete R Commander también se pudo generar histogramas de forma individual por cada curso. En la Fig. 4 se muestra la exploración del curso de Cálculo Diferencial, en donde se observa que las notas tienen mayor frecuencia entre 08 y 15.

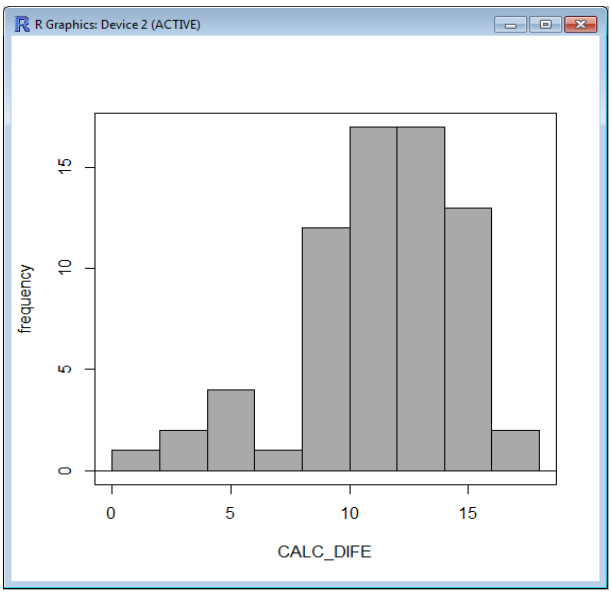

Fig. 4 Histograma del curso de cálculo diferencial.

C. Gráficos de Distribución

Para realizar los gráficos de distribución de probabilidad se utilizó el paquete Rattle y el lenguaje de programación R.

En la Fig. 5 se muestra la distribución de probabilidad del curso de Desarrollo Humano, donde podemos observar que se genera una distribución normal, lo que indica que las notas de los alumnos en ese curso se concentran con mayor frecuencia cerca del promedio y con menor frecuencia a medida que se acercan a los extremos que corresponden a las notas 0 (cero) y 20 (Veinte), por lo que se considera que el comportamiento de las notas es normal.

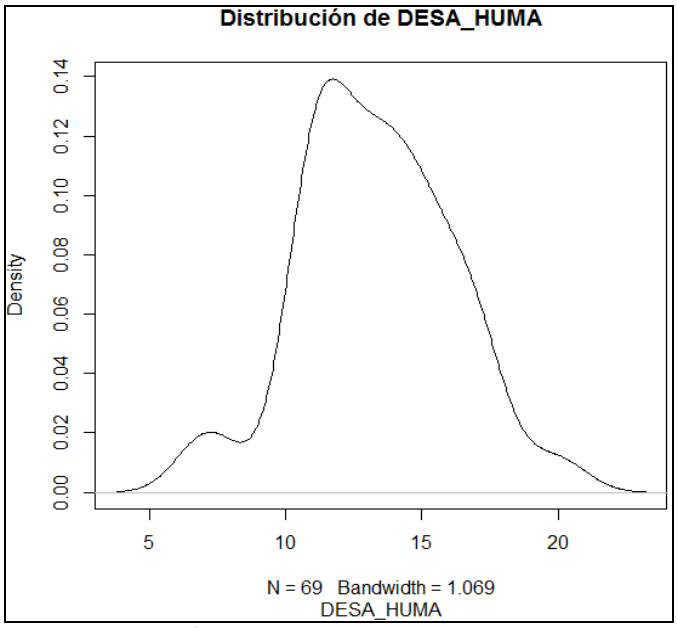

Fig. 5 Distribución de probabilidad del curso de desarrollo humano.

Cuando la distribución no es normal, es posible que sea necesario tomar medidas de acción porque se puede encontrar la campana de Gauss sesgada a la izquierda representando mayor frecuencia de notas desaprobatorias en ese curso.

D. Diagramas de Cajas

Se utilizó para identificar datos atípicos empleando el paquete R Commander.

En la Fig. 6 se puede observar que el alumno de código 59 presenta una nota atípica en el curso de Cálculo Diferencial por debajo de 04 (cuatro).

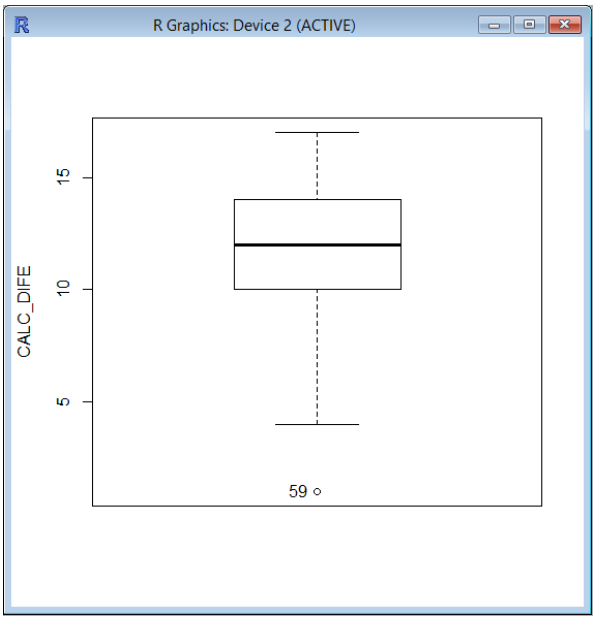

Fig. 6 Diagrama de caja para el curso de cálculo diferencial.

E. Gráficos de Tallos y Hojas

Se realizó el gráfico de tallos y hojas para mostrar la frecuencia de cada nota comprendida entre la nota mínima y máxima del curso, para ello se utilizó el paquete R Commander y el lenguaje de programación R.

En la Fig. 7 se muestra la cantidad de alumnos que obtuvieron determinadas notas. Se observa que hay dos alumnos que obtuvieron la nota máxima veinte (20), así como tres alumnos que obtuvieron la nota mínima siete (7).

```
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 69
   3      7 | 000
   4      8 | 0
          9 |
         10 |
  21     11 | 00000000000000000
  28     12 | 0000000
  (9)    13 | 000000000
  32     14 | 000000000
  23     15 | 00000000
  15     16 | 00000
  10     17 | 0000000
   3     18 | 0
         19 |
   2     20 | 00
```

Fig. 7 Gráfico de tallos y hojas del curso de desarrollo humano.

F. Matriz de Correlaciones

La matriz de correlaciones permite determinar la relación entre dos variables, es una matriz simétrica de n filas y n columnas, donde el elemento Aij de la matriz es igual al elemento Aji, en la cual i representa el número de fila y j el número de columna de la matriz, en la diagonal de la matriz la correlación siempre es 1. Si el elemento de una matriz tiene el valor de 0 quiere decir que no existe correlación entre esas dos variables y si el valor se aproxima a 1 indica que la correlación es alta y positiva, es decir, están fuertemente correlacionadas.

A través del paquete R Commander y el lenguaje de programación R, se generó la matriz de correlación.

En la Tabla 2 se puede observar que el curso de PROG_I (Programación I) está altamente correlacionado con los cursos de FUND_PROG (Fundamentos de Programación) con 0.81 y ESTR_DISC_I (Estructuras Discretas I) con 0.80. Esto quiere decir por ejemplo que los alumnos que tienen buen rendimiento en el curso de Programación I también lo tendrán en los cursos de Fundamentos de Programación y Estructuras Discretas I, ya que están altamente correlacionados.

TABLA 2
VALORES DE LA MATRIZ DE CORRELACIÓN

	CD	**DH**	**CO**	**FP**	**PR**	**ED**
CD	1.00	0.59	0.73	0.77	0.73	0.75
DH	0.59	1.00	0.67	0.58	0.58	0.65
CO	0.73	0.67	1.00	0.63	0.64	0.70
FP	0.77	0.58	0.63	1.00	0.81	0.78
PR	0.73	0.58	0.64	0.81	1.00	0.80
ED	0.75	0.65	0.70	0.78	0.80	1.00
CD: CALC_DIFE, DH: DESA_HUMA, CO: COMU_ORAL_ESCR, FP: FUND_PROG, PR: PROG_I, ED: ESTR_DISC_I						

La matriz de correlación también se puede observar gráficamente como se muestra en la Fig. 8, en este caso mientras los círculos sean más azules y grandes la correlación es alta y positiva, por lo que se puede decir que el curso de Programación I tiene alta correlación con los cursos de Fundamentos de Programación y Estructuras Discretas I.

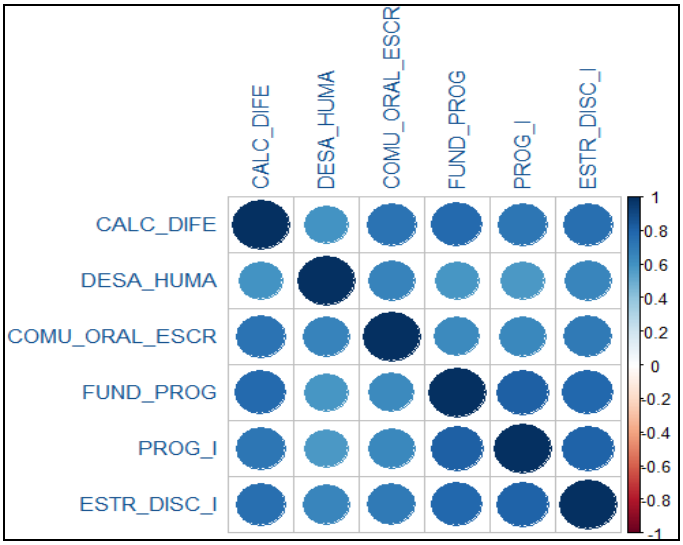

Fig. 8 Matriz de correlación.

G. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible [14].

Se utilizó el paquete Rattle y el lenguaje de programación R para generar el plano de componentes principales con el círculo de correlaciones.

En la Fig. 9 se muestra el análisis de componentes principales, donde se puede apreciar a partir de los ángulos que forman las variables, que los cursos CALC_DIFE (Cálculo Diferencial) y ESTR_DISC_I (Estructuras Discretas I) están altamente correlacionadas así como FUND_PROG (Fundamentos de Programación) y PROG_I (programación I). También se puede observar que los códigos que están más alejado de los cursos como son 62, 57, 54 entre otros, corresponden a los alumnos de bajo rendimiento, así como los códigos que están más cerca corresponden a los alumnos de alto rendimiento, verificando con el registro de notas se pudo comprobar que el código 55 corresponde al más alto promedio, seguido del código 19 y 4.

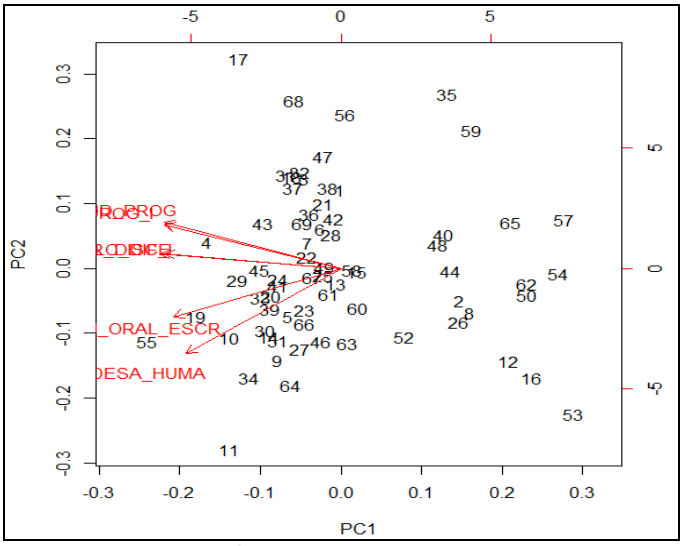

Fig. 9 Análisis de componentes principales.

H. Agrupación o Clustering

Permite obtener grupos o conjuntos en donde se incorpore elementos similares extraídos de las clases del dominio dado [15].

Se realizó la agrupación de alumnos en dos clusters o grupos para separar y explorar los datos de los alumnos con promedios finales desaprobados (Grupo 1) y con promedios finales aprobados (Grupo 2). Para ello se utilizó el algoritmo K-means en la herramienta Weka y el entorno de software y lenguaje de programación R, dando los mismos resultados.

En la Fig. 10 se muestra los resultados de la agrupación en Weka, en donde se observa dos grupos que están representados por cluster 0 y cluster 1. Se puede observar también que el cluster 0 está conformado por 17 instancias o registros y el cluster 1 por 52 instancias, que son la cantidad de alumnos desaprobados y aprobados respectivamente.

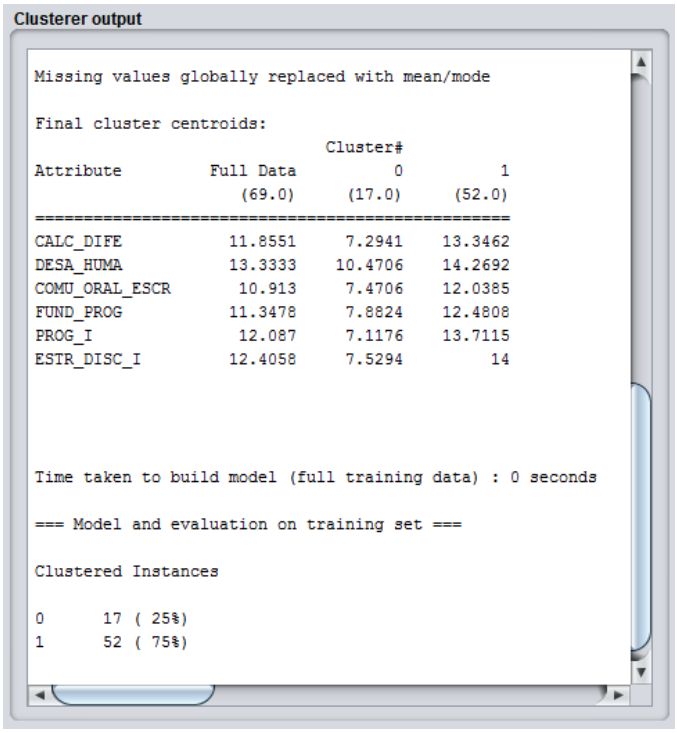

Fig. 10 Resultados de la agrupación en weka.

En la Fig, 11 se muestra los resultados de la agrupación en R, que muestra los promedios de cada grupo en los diferentes cursos y los 69 registros con su respectivo grupo al que pertenecen.

```
> grupos_Kmeans<-kmeans(notas, 2,iter.max=50, nstart=50)
> grupos_Kmeans
K-means clustering with 2 clusters of sizes 17, 52

Cluster means:
  CALC_DIFE DESA_HUMA COMU_ORAL_ESCR FUND_PROG    PROG_I ESTR_DISC_I
1  7.294118  10.47059       7.470588  7.882353  7.117647    7.529412
2 13.346154  14.26923      12.038462 12.480769 13.711538   14.000000

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 2  1  2  2  2  2  2  1  2  2  2  1  2  2  2  1  2  2  2  2  2  2  2
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
 2  2  1  2  2  2  2  2  2  2  2  1  2  2  2  2  1  2  2  2  1  2  2
47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
 2  1  2  1  2  1  1  1  2  2  1  2  1  2  2  1  2  2  1  2  2  2  2

Within cluster sum of squares by cluster:
[1] 571.7647 911.5769
 (between_SS / total_SS =  60.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Fig. 11 Resultados de la agrupación en R.

En la Fig. 12 se observa que el Cluster 0 está representado por los promedios desaprobados en los diferentes cursos y el Cluster 1, por los promedios aprobados.

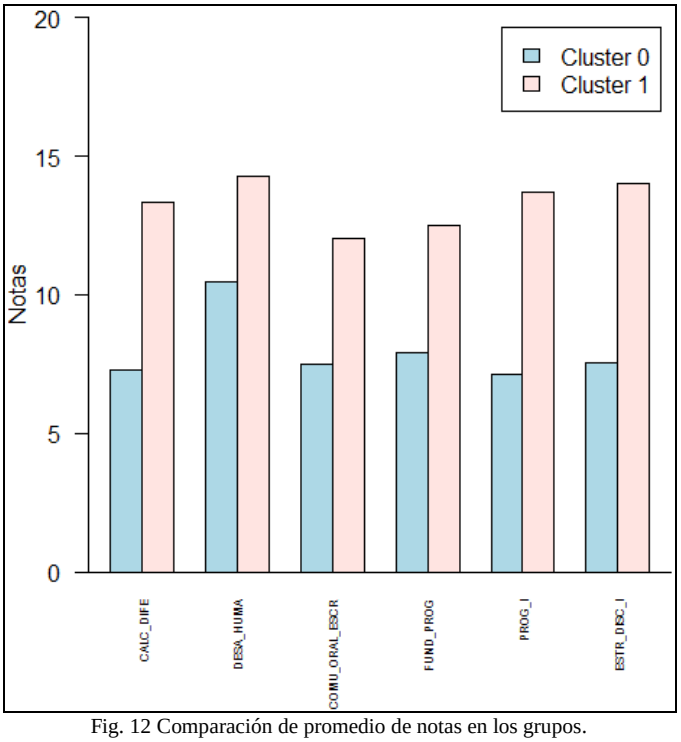

Fig. 12 Comparación de promedio de notas en los grupos.

En la Fig. 13 se observa gráficamente que el 25% de la cantidad de alumnos pertenecen al Cluster 0 que son los alumnos con promedios desaprobados y el 75% son los alumnos con promedios aprobados.

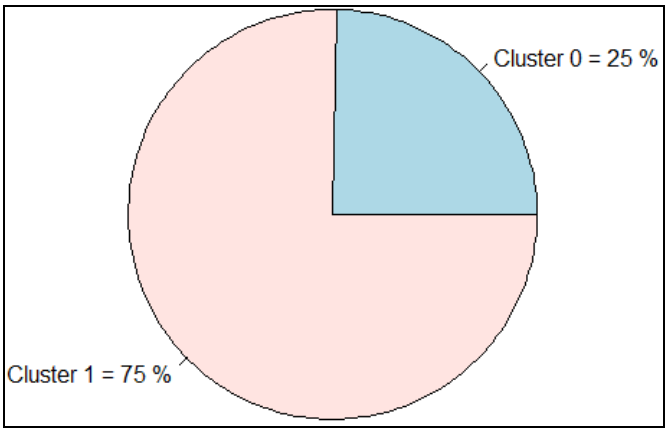

Fig. 13 Porcentajes de cantidad de alumnos en cada grupo.

En la Tabla 3 se observa que los alumnos que pertenecen al Cluster 0 tienen promedios bajos y en la mayoría de casos desaprobados, sobre los cuales se debe enfocar esfuerzos personalizados para mejorar su rendimiento y evitar su posible deserción.

TABLA 3
PRIMEROS REGISTROS DE ALUMNOS CON SUS RESPECTIVOS GRUPOS AL QUE PERTENECEN

COD.	CD	DH	CO	FP	PR	ED	GRUPO
1	13	11	11	12	12	13	Cluster1
2	9	11	9	10	5	8	Cluster0
3	15	12	11	13	13	14	Cluster1
4	16	17	12	15	17	17	Cluster1
5	14	14	14	12	12	14	Cluster1
6	12	13	11	13	12	14	Cluster1
7	14	13	12	13	12	13	Cluster1
8	9	11	9	8	7	7	Cluster0
…							

CD: CALC_DIFE, DH: DESA_HUMA, CO: COMU_ORAL_ESCR, FP: FUND_PROG, PR: PROG_I, ED: ESTR_DISC_I

IV. CONCLUSIONES

Se realizó la exploración de datos académicos que resulta de utilidad a las autoridades y coordinadores de las entidades educativas para examinar la frecuencia y comportamiento de las notas, encontrar relación entre los cursos, detectar datos atípicos y agrupar a los alumnos según su rendimiento. Dicho estudio resulta favorable ser aplicado después de las evaluaciones para detectar los cursos donde se deben tomar medidas correctivas, así como alumnos sobre los cuales se debe prestar atención con el objeto de mejorar su rendimiento, y de este modo reducir la deserción y reincidencia de matrícula. Se presentó resúmenes de los datos académicos, así como representaciones gráficas que hacen más fácil su comprensión, exploración y análisis a través de herramientas de minería de datos como Weka y R. Se sugiere, que con los resultados obtenidos, se refuerce el aprendizaje de los alumnos, por ejemplo mediante una enseñanza personalizada.

Como trabajo futuro se recomienda investigar la exploración de datos académicos y aplicación de técnicas de minería de datos educacionales sobre datos no estructurados como páginas de internet, para analizar el aprendizaje de los estudiantes en plataformas virtuales.

REFERENCIAS

[1] P. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, New York, EEUU: Pearson Education, 2006.

[2] B. Thuraisingham, "A Primer for Understanding and Applying Data Mining," *IT Professional*, vol. 2, no. 1, pp. 28-31, 2000.

[3] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135-146, 2007.

[4] A. Ballesteros, D. Sánchez and R. García, "Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo," *Latin-American Journal of Physics Education*, vol. 7, no. 4, pp. 662-668, 2013.
[5] D. Borao, "Incidencia del ruido en los datos de test sobre la precisión de modelos de clasificación y regresión," Tesis de Máster Universitario en Ingeniería del Software, Métodos Formales y Sistemas de Información, Universidad Politécnica de Valencia, España, 2013.
[6] F. García, "Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA)," Trabajo Fin de Máster Universitario en Estadística Aplicada, Universidad de Granada, España, 2013.
[7] A. Sáez, Métodos Estadísticos con R y R Commander (2010), Dpto de Estadística e Investigación Operativa, Universidad de Jaén, https://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf, Revisado en Octubre del 2016.
[8] CRAN, Package rattle, https://CRAN.R-project.org/package=rattle, Revisado en Octubre del 2016.
[9] Escuela Profesional de Ingeniería de Sistemas. http://www.ucsm.edu.pe/ingenieria-de-sistemas/
[10] Universidad Católica de Santa María. http://www.ucsm.edu.pe/
[11] Página oficial de Pentaho, http://www.pentaho.com/, Revisado en Octubre del 2016.
[12] Wikipedia, CSV, https://es.wikipedia.org/wiki/CSV, Revisado en Octubre del 2016.
[13] Vitutor, Histograma, http://www.vitutor.com/estadistica/descriptiva/a_6.html, Revisado en Octubre del 2016.
[14] M. Gurrea, Análisis de componentes principales, Proyecto e-Math Financiado por la Secretaría de Estado de Educación y Universidades, http://www.uoc.edu/in3/emath/docs/Componentes_principales.pdf, Revisado en Octubre del 2016.
[15] J. Riquelme, R. Ruiz and K. Gilbert, "Minería de datos: Conceptos y Tendencias," *Revista Iberoamericana de Inteligencia Artificial*, vol. 10, no. 29, pp. 11-18, 2006.