Análisis de cociente triádico para procesos de admisión de estudiantes en la fundación Insutec mediante minería de datos

Triadic ratio analysis for student admission processes in the foundation Insutec through data mining

John Petearson Anzola Anzola*

Fecha de recepción: octubre 1 de 2010
Fecha de aceptación: noviembre 9 de 2010

Resumen

Actualmente, el proceso de admisión en las instituciones de educación superior a nivel nacional e internacional se ha convertido en un factor determinante. La selección de un programa profesional a través de las diferentes ofertas y demandas de formación representa, a menudo, un problema en el momento de elegir una carrera profesional. Así mismo, dentro del proceso de aprendizaje interfieren múltiples variables externas que pueden conducir a la no culminación de una carrera profesional.

El cociente triádico es una herramienta de caracterización de individuos cuyo fin es analizar las variables internas del ser, las cuales —al combinarse con las externas— modelan un ambiente incierto y especulativo frente al comportamiento de los seres humanos en el proceso de selección y admisión de una carrera profesional. Es justo en este punto donde converge la minería de datos, un conjunto de tecnologías que permite descubrir e inferir conocimiento a partir de diferentes técnicas.

Palabras clave: Cociente Mental Tríadico, minería de datos, caracterización, agrupación de datos, Weka.

---

* Aspirante a Magister en Ciencias de la Información y Comunicación de la Universidad Distrital Francisco José de Caldas. Docente investigador de la Fundación Universitaria Los Libertadores. Miembro del grupo de investigación Ideas de la Universidad Distrital. Correo electrónico: jpanzolaa@correo.udistrital.edu.co

Abstract

Currently, the process of admission to higher education institutions nationally and internationally, has become a key factor. Selecting a professional program through the various offers and demands of training, is often a problem when selecting a career, interfering in the process of forming multiple external variables that can lead to non-completion of a career training in Colombia.

The striated ratio is a tool for characterization of individuals, which analyzes the internal variables of being, which combined with external variables, model an environment of uncertainty and speculation against the behavior of individuals in the process of selection and admission of a career. It is at this point that converges data mining, tool in which we can discover and infer knowledge from different techniques.

Keywords: Mental Ratio Triadic, Data Mining, characterization, Data Association, Weka.

Introducción

Felder y Silverman [1] clasifican a los estudiantes según su forma de aprendizaje y de acuerdo a un conjunto de pares dicotómicos: observando y escuchando, pensando y analizando, reflexionando y actuando, razonando lógica e intuitivamente, memorizando y visualizando, construyendo analogías y modelos de asociación que, por lo general, son matemáticos. Por su parte, los estudiantes adoptan métodos de enseñanza que se encuentran definidos por el estilo y la formación de los docentes; es decir, algunos de ellos leen; otros, demuestran; unos discuten o se centran en principios y en leyes universales; muchos otros, en aplicaciones; algunos enfatizan en la memorización y otros, en la comprensión, etc. Entonces, lo que cada estudiante aprenda en una clase dependerá de la habilidad innata que posea, de su preparación previa y, además, de la compatibilidad entre su estilo de aprendizaje y el de la enseñanza del docente [2].

Como profesor de la Facultad de Ingeniería de la Fundación de Educación Superior (Insutec), consideré de gran utilidad determinar

las características del perfil de aprendizaje de los estudiantes; en función de ellas pueden adecuarse las estrategias de enseñanza. Bajo esta premisa y empleando el test propuesto por Felder y Soloman [3], fueron encuestados ochenta y nueve estudiantes de la misma facultad. Así mismo, con el fin de descubrir el conocimiento implícito en las respuestas, se siguió el proceso de descubrimiento de conocimiento en bases de datos (KDD-*Knowledge-discovery in databases*) y, a través de técnicas de minería de datos, pudo caracterizarse un estudiante en el proceso de admisión.

Objetivos del análisis

Un paso previo al análisis lo constituyó la fuente de información (base de datos), en la cual se establecieron los siguientes objetivos: extracción de búsqueda de relaciones, identificación de modelos subyacentes en los datos y comprensión del dominio de los mismos. Lo anterior con el fin de establecer una idea clara sobre la caracterización de un estudiante en el proceso de admisión. Así mismo, antes de comenzar con el análisis fue elegida la plataforma de software para aprendizaje automático y minería de datos *Weka* (*Waikato EnvironmentforKnowledgeAnalysis*), la cual está escrita en Java y fue desarrollada en la Universidad de Waikato [4].

Además, el uso de Weka en el proceso de análisis de datos (KDD) permitirá dirigir la búsqueda y el refinamiento de la información mediante una interpretación adecuada de los resultados generados. Cabe resaltar que los análisis efectuados no <<emergen>> de los datos, sino que deben ser considerados e interpretados con detenimiento como primer paso del estudio.

El objetivo principal de este artículo es relacionar los resultados obtenidos en las pruebas con las características o los perfiles de los estudiantes en el proceso de admisión (si bien la descripción e información de la base de datos disponible no es muy amplia y probablemente habrá que adaptarse y sujetarse al contenido con el que se cuenta). Por su parte, las siguientes son algunas premisas que se pueden plantear y responder como objetivos del análisis:

- ¿Qué características cerebrales tienen los estudiantes de cada programa?
- ¿Existen grupos de estudiantes, no conocidos de antemano, con características similares?
- ¿Existen diferencias significativas en los resultados obtenidos según las preguntas A0, A1, A2 y A3?
- ¿La elección de una carrera profesional depende del entorno o de qué variable?
- ¿Es posible predecir la selección de un programa mediante alguna variable previamente conocida?
- ¿Cuáles son las relaciones más significativas entre variables?

Como se observará más adelante, en los resultados obtenidos pueden encontrarse relaciones triviales, conocidas previamente o, incluso, no hallarse ninguna significativa; situación que sería relevante para el proceso de interpretación. A continuación dos ejemplos potenciales: **a)** determinar, después de un análisis exhaustivo, que la clase social no condiciona la escogencia de una carrera profesional; y **b)** que la prueba de ingreso debe considerarse homogénea sin importar su estratificación social (resultados como los anteriores podrían interpretarse como conclusiones válidas). Por otra parte, este análisis tiene un enfoque introductorio e ilustrativo que permite acercarse a las técnicas disponibles y a su misma manipulación desde la herramienta Weka, hecho que deja abierto —para el investigador— el estudio del dominio de datos a resultados y conclusiones más elaborados.

CT - Revelador del cociente triádico

El Test Revelador del Cociente Mental Tríadico (Rcmt) fue diseñado y validado por Waldemar de Gregory para diagnosticar las manifestaciones del cerebro triuno (triádico). Éste, a su vez, se caracteriza por poseer ciertas manifestaciones, en el comportamiento proporcional o desproporcional, que inciden en el desempeño educativo y social del estudiante —identificadas mediante el uso del Rcmt. Así mismo, la aplicación de este test puede incidir conscientemente en el desarrollo de las operaciones, habilidades y facultades mentales (en especial las relacionadas con el pensar: *crear-imaginar-sentir* y *concretar-actuar*) [5].

Tabla 1. Manifestaciones del cerebro triuno.

CEREBRO		
Izquierdo	**Central**	**Derecho**
Verbal, numérico, analítico, lógico descompositor, racional, abstracto, alerta, vigilante, articulador, crítico, investigador, visual y lineal.	Instintivo, vegetativo, motor, concreto, administrador, regulador, trabajador, profesional, negociante, apropiador, planificador, económico, político, mercader y ecosistémico.	Proverbial, magnético, intuitivo, sintético, reintegrador, holístico, emocional, sensorial, espacial, espontáneo, relajado, libre, asociativo, artístico, contemplativo, sonoro y no lineal.

Manifestaciones del cerebro triuno

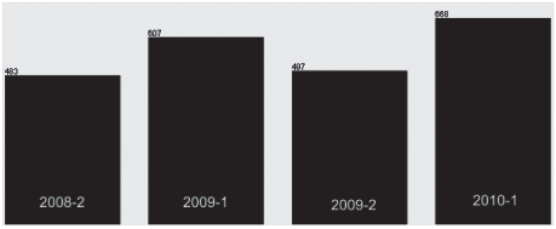
Fuente de datos

La principal fuente de datos utilizada en esta investigación la constituyeron los estudiantes que presentaron los exámenes de admisión en los periodos académicos comprendidos desde 2008-II hasta 2010-I. En total fueron 2.255 jóvenes, de ambas jornadas (diurna y nocturna), cuyas edades oscilan entre los 16 y los 27 años. Así mismo, la mayor parte de estos estudiantes se encuentran becados y pertenecen a estratos socioeconómicos bajos [6]. El conjunto de datos consta de 26 atributos y un total de 2.255 registros.

Análisis estadístico de los datos en Weka

La figura 1 ilustra el análisis de la clase-periodo, la cual contiene cuatro atributos que se destacan gráficamente:

Figura 1. Clase-periodo académico.



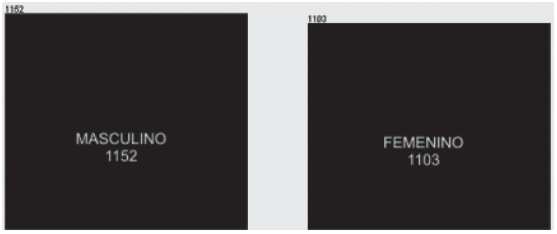
Por su parte, la siguiente tabla contiene la cantidad de estudiantes por periodo académico:

Tabla 2. Estudiantes por periodo académico.

Periodo	Número de estudiantes matriculados
2008-II	483
2009-I	607
2009-II	497
2010-I	668

Entre tanto, la figura 2 ilustra el análisis de la clase-sexo:

Figura 2. Clase-sexo.



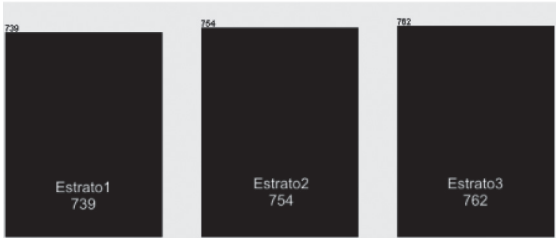
En seguida es expuesta la cantidad de estudiantes durante los cuatro periodos de observación:

Tabla 3. Cantidad de hombres y mujeres durante los cuatro periodos.

Sexo	Cantidad
Masculino	1152
Femenino	1103

La figura 3, que expone el análisis de la clase-estrato, contiene tres atributos:

Figura 3. Clase-estrato



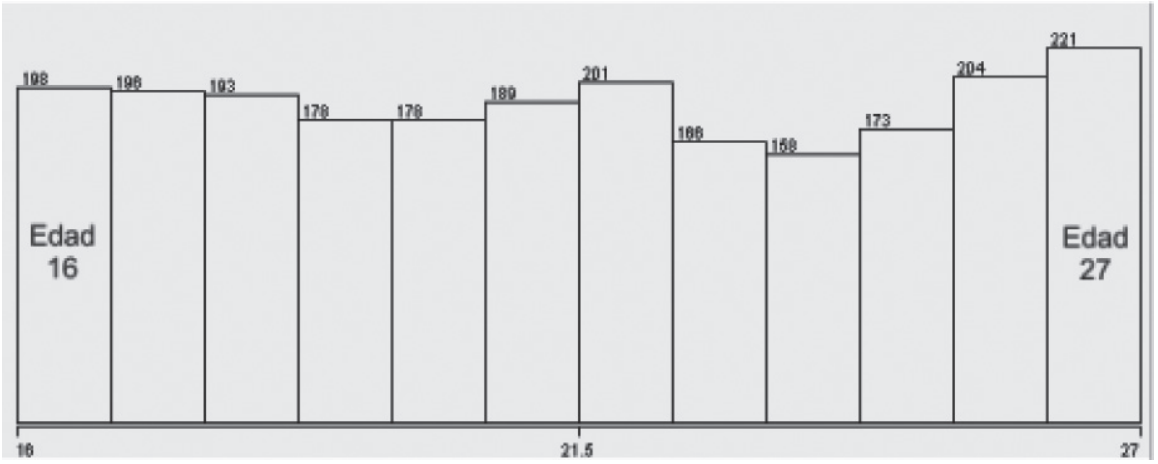
Así mismo, la siguiente tabla evidencia la cantidad total de estudiantes, durante los cuatro periodos, acorde a su estrato socioeconómico:

Tabla 4. Estratos de la totalidad de la población de muestra.

Estrato	Cantidad
1	739
2	754
3	762

A continuación se ilustra el análisis de la clase-edad, el cual se caracteriza por ser una variable de tipo entero:

Figura 4. Clase-edad.



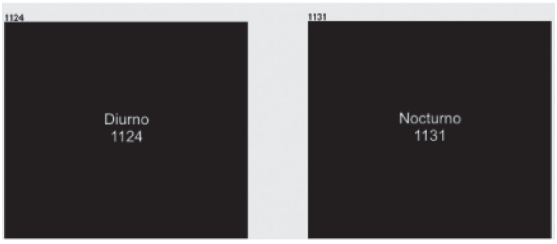
En la siguiente tabla pueden observarse los datos estadísticos de la clase-edad, destacando que fueron tomados de los cuatro periodos académicos:

Tabla 5. Datos estadísticos de la clase-edad.

Edad	Valor
Mínimo	16
Máximo	27
Promedio	21.514
Desviación Estándar	3.539

La figura 5 ilustra el análisis de la clase-jornada. Esta es una variable categórica:

Figura 5. Clase-Jornada.



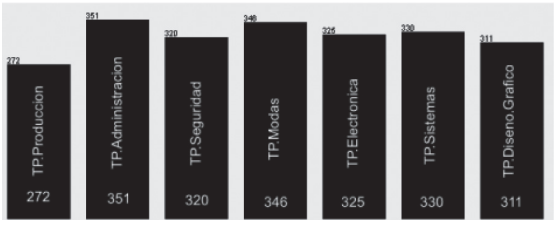
La tabla 6 contiene los datos estadísticos de la clase jornada:

Tabla 6. Cantidad de estudiantes por jornada.

Nº	Jornada	Cantidad
1	Diurna	1124
2	Nocturna	1131

En la siguiente figura se ilustra el análisis de la clase-programa, una variable categórica cuya distribución se observa a continuación:

Figura 6. Clase-programa.



Los datos estadísticos de la clase-programa conforman la tabla 7. Cabe resaltar que se tomaron de los cuatro periodos académicos:

Tabla 7. Cantidad de estudiantes por programa.

Nº	Programa	Cantidad
1	Producción	272
2	Administración	352
3	Seguridad	320
4	Modas	346
5	Electrónica	325
6	Sistemas	330
7	Diseño gráfico	311

En la siguiente figura se ilustra el análisis de la clase A0 correspondiente a la pregunta: ¿se encuentra trabajando actualmente?

Figura 7. Clase A0.

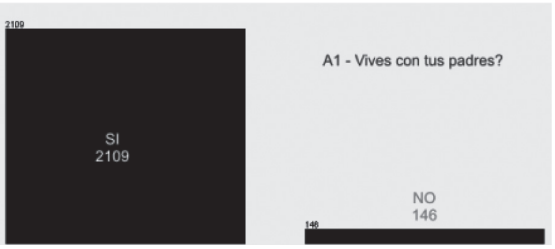
En seguida, los datos estadísticos del cuestionamiento A0:

Tabla 8. Datos totales de la pregunta A0.

Nº	A0	Cantidad
1	SÍ	487
2	NO	1768

Figura 8: clase A1, ¿vive con sus padres?

Figura 8. Clase A1.



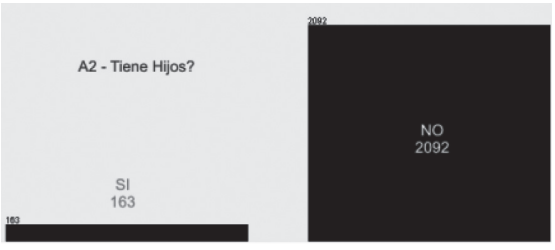
Datos estadísticos de la clase A1:

Tabla 9. Datos totales de la pregunta A1.

Nº	A1	Cantidad
1	SÍ	2109
2	NO	146

En la siguiente figura se ilustra el análisis de la clase A2, el cual corresponde a la pregunta: ¿tiene hijos?

Figura 9. Clase A2.

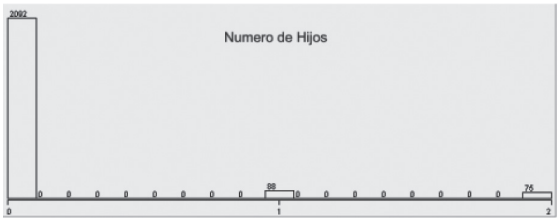


Datos estadísticos de la clase A2:

Tabla 10. Datos totales de la pregunta A2.

Nº	A2	Cantidad
1	SÍ	163
2	NO	2092

Figura 10. Clase A3.



La figura 10 ilustra la pregunta de la clase A3: ¿cuántos hijos tiene? La tabla 11 expone la totalidad de los datos de la clase A3:

Tabla 11. Datos totales de la pregunta A3.

Estadísticas	Cantidad
Mínimo	0
Máximo	2
Promedio	0.106
Des. Estándar	0.401

Aplicación de filtros

Weka contiene filtros integrados que permiten realizar manipulaciones sobre los datos en dos niveles: atributos e instancias. Las operaciones de filtrado pueden aplicarse <<en cascada>>, de manera que cada filtro tome como entrada el conjunto de datos resultante de uno anterior y guarde los resultados

de aplicar filtros en nuevos ficheros —que también serán de tipo Attribute-Relation File FormatARFF— para manipulaciones posteriores [7].

Filtros de atributos

A continuación se indica —entre todas las posibilidades implementadas la utilización de filtros para eliminar atributos, discretizar atributos numéricos y añadir nuevos atributos con expresiones según por la frecuencia con la que se realizan estas operaciones.

Al aplicar el filtro *Remove* se eliminan los atributos correspondientes a *id*, *epistemología*, *act_cientifica*, *clasificaciones*, *comunicación*, *administración*, *planeación*, *pro_empresa*, *imp_sobrevivencia*, *espiritualidad*, *estado_alfa*, *creatividad* y *afectividad*. Por tal motivo, sólo se trabajarán los atributos: *p_academico*, *sexo*, *estrato*, *edad*, *jornada*, *programa*, *A0*, *A1*, *A2*, *A3*, *cerebro_izq*, *cerebro_der* y *cerebro_cen*.

Otro filtro aplicado —antes de implementar algunos algoritmos que se detallarán posteriormente— es *Discretize*, el cual transforma los atributos numéricos seleccionados en atributos simbólicos con una serie de etiquetas que resultan tras dividir la amplitud total del atributo en intervalos. Por ejemplo, una vez empleado este filtro, si se dividen los intervalos del <<atribute>> *cerebro_der* en cuatro de igual frecuencia, se obtendrán los rangos delimitados por (29.5, 32.5, 35.5), destacando cómo el 31.2 % de los estudiantes desarrolla el cerebro izquierdo.

Visualización

La herramienta de visualización de Weka se utiliza para representar gráficas 2D que relacionan pares de atributos. La figura 11 representa el rango del atributo *cerebro_der* con la totalidad de los estudiantes que presentaron la prueba de admisión en el lapso 2008-II — 2010-I.

Figura 11. Filtro de discretización (4 Bins).

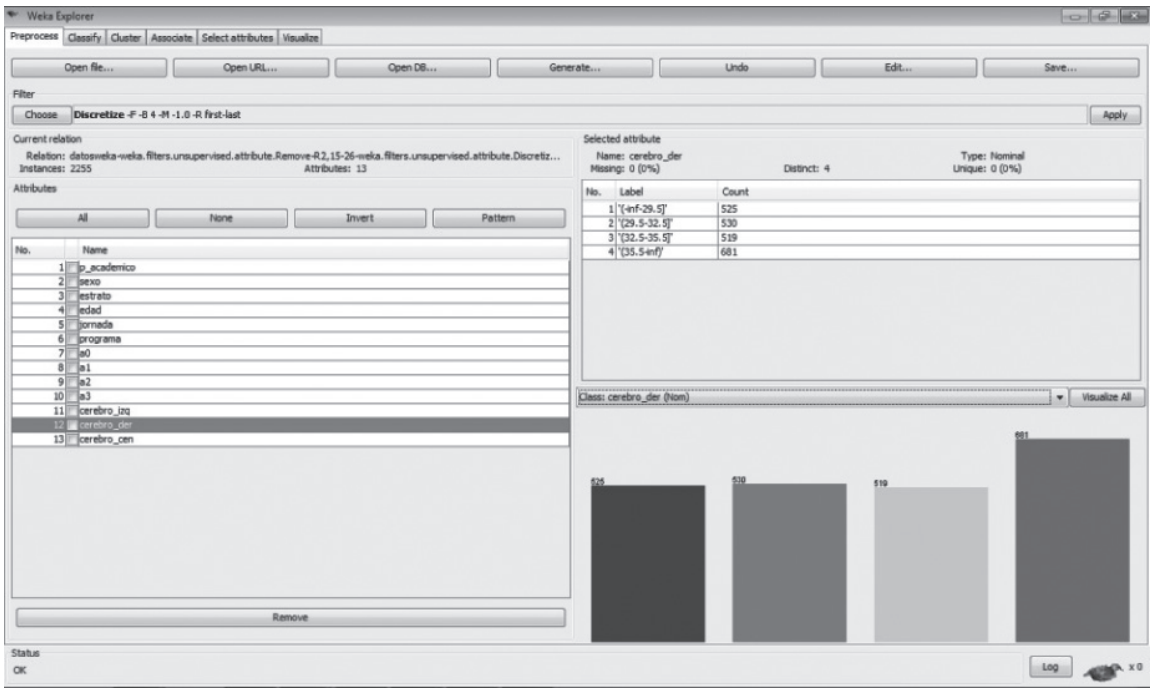
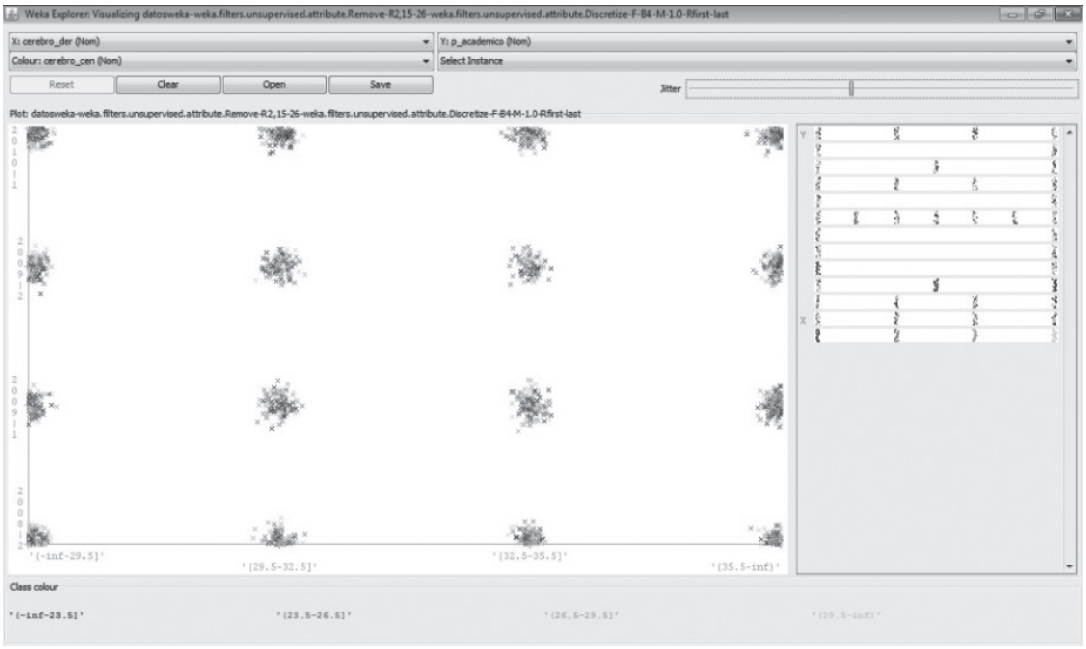


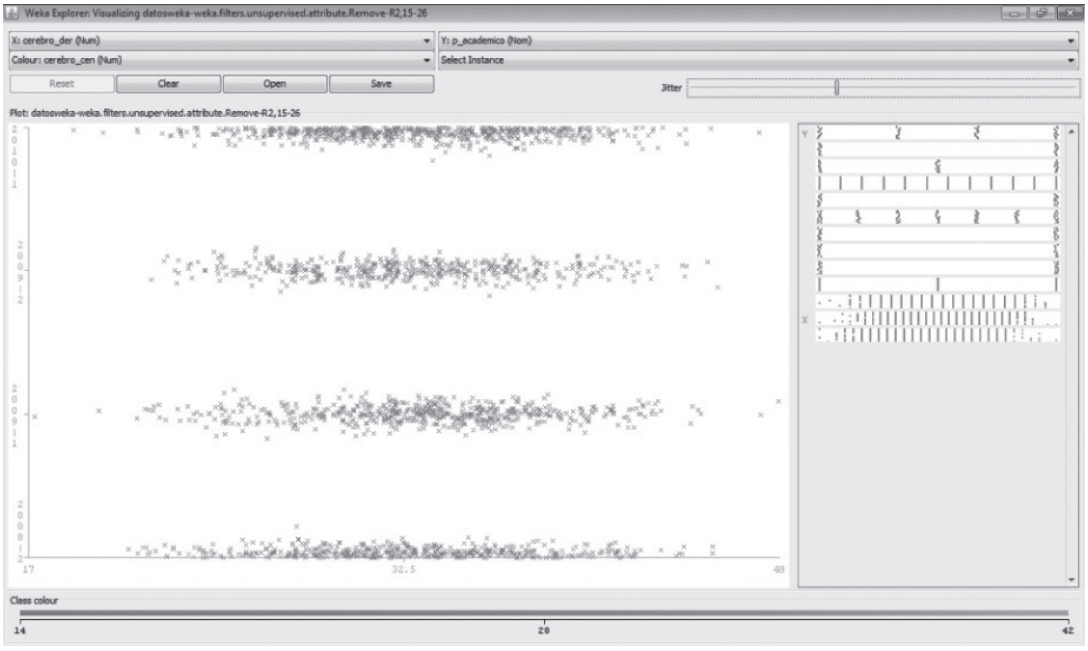
Figura 12. *cerebro_der Vs p_academico.*



En la figura 12 se observa la tendencia de los estudiantes que desarrollaron el cerebro derecho durante los cuatro periodos académicos:

A continuación, en la figura 13, se compara la tendencia de los estudiantes que desarrollaron el cerebro derecho —durante los cuatro periodos académicos— pero sin la aplicación del filtro de discretización.

Figura 13. *Cerebro_der Vs p_academico (sin discretizar).*



Luego de comparar la figura 12 con la 13 puede apreciarse la forma en que los datos están dispersos por cada periodo académico (teniendo una media en 32); mientras que en la figura 12 persiste la misma tendencia pero con una dispersión menos, facilitando, así mismo, la caracterización de los estudiantes por periodo académico.

A continuación se observarán algunas tendencias de los atributos sin discretizar:

La figura 14 expone la distribución de la edad frente al atributo del cerebro izquierdo. Al interior del gráfico se destaca la pregunta A1: ¿vive con sus padres? Los que respondieron afirmativamente son expuestos en color azul; los que contestaron "no" están en rojo.

Figura 14. *edad Vs cerebro_izq.*

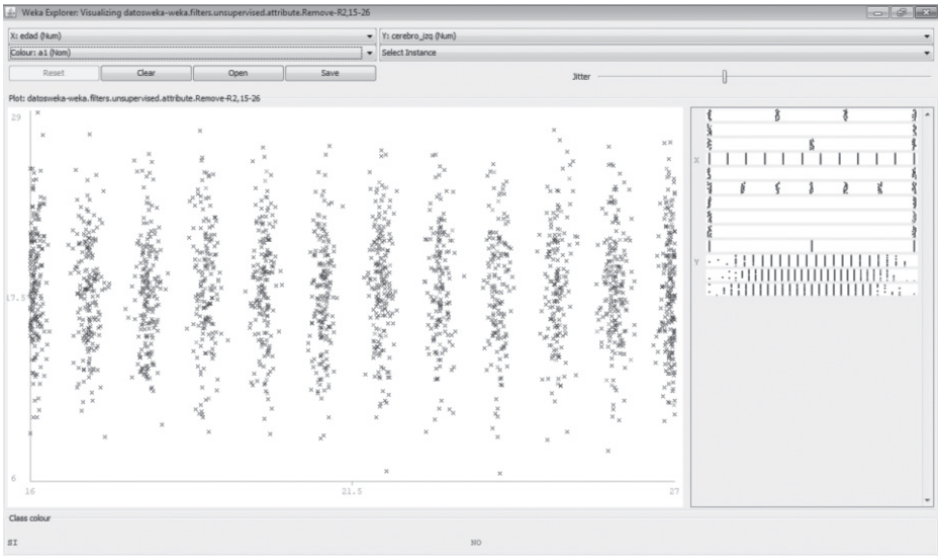
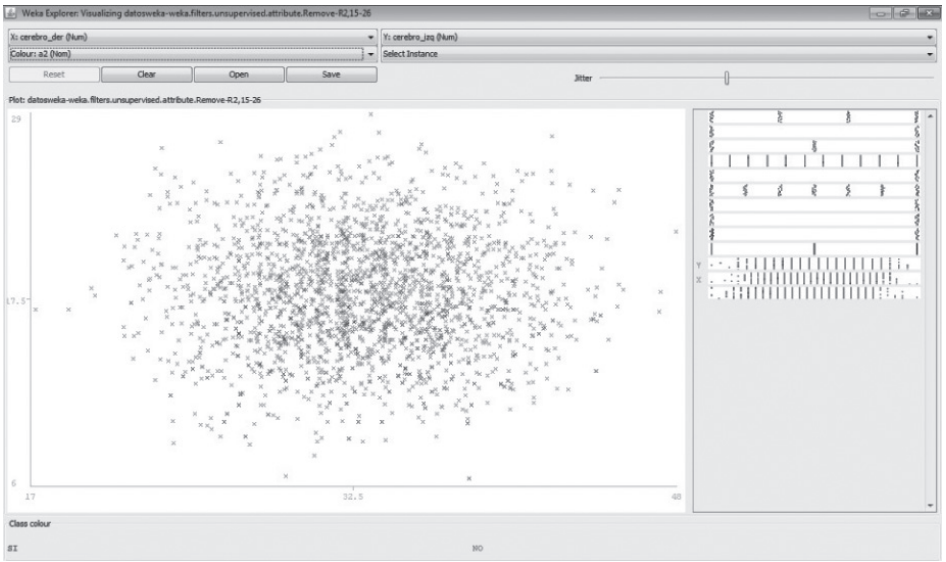


Figura 15. *cerebro_izq Vs cerebro_der.*



En la figura 15 se compara la caracterización de los individuos que desarrollaron el cerebro izquierdo y el derecho. Así mismo, en su interior están distribuidas, en color rojo y azul, las personas que tienen hijos.

En la figura 16 son comparadas la caracterizaciones de los individuos —por programa académico y cerebro derecho— con la distribución interna de la pregunta A0: ¿vive con sus padres?. Sin embargo, fue añadida la característica que se discretizó en tres grupos (manteniendo la frecuencia). La mayor parte de los individuos se encuentran en la parte central.

Otra etapa de filtrado consiste en la normalización, proceso en el cual se dejan todos los términos numéricos en un rango de 0 a 1. Entonces, el rango de valores se transforma en un intervalo determinado (normalmente [0,1]) y está dado por [8]:

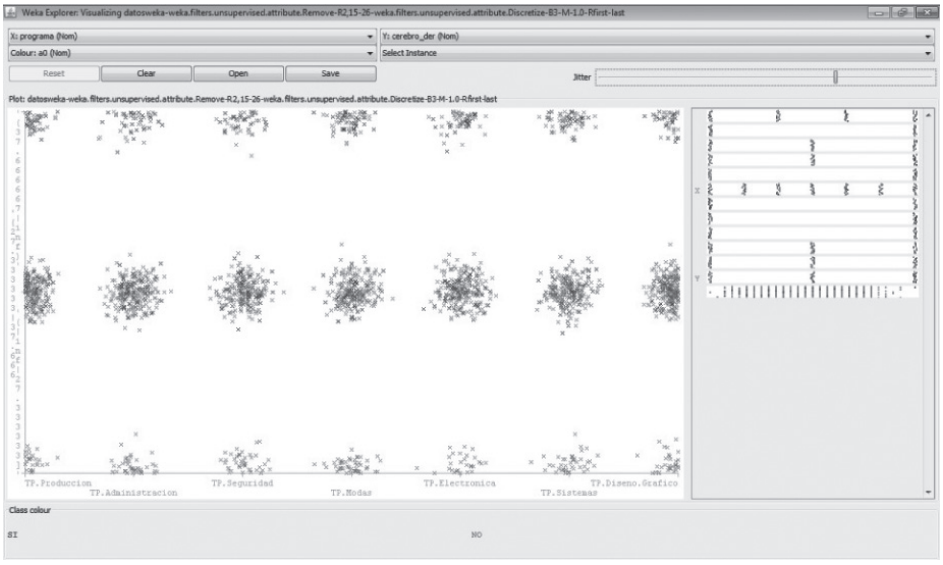
$$valor_{Normal} = \frac{valor - valor_{min}}{valor_{max} - valor_{min}}$$

Entre tanto, la normalización es necesaria si se van a aplicar algoritmos de aprendizaje basados en distancias para que todos los atributos estén en el mismo rango. Con las variables normalizadas pueden estimarse —visualmente— el estado de las variables y su distribución. Además, con el fin de aplicar este filtro de transformación para usar los datos experimentales, su lectura se interpreta sectorizada y en porcentaje (de esta forma se normalizan para eliminar los efectos de las fuentes de sesgo). Si la normalización se lleva a cabo correctamente, el proceso no alterará el contenido de los datos; simplemente corregirá las desviaciones que surgan durante la de recolección de información.

Agrupamiento

Los algoritmos de agrupamiento buscan grupos de instancias con características similares según un criterio de comparación entre valores de atributos de las instancias definidos en los algoritmos.

Figura 16. *P_academico Vs cerebro_der* (discretizado con blins en 3).



Agrupamiento numérico

Algoritmo K-Medias::k-means es un método de análisis de conglomerados que apunta a la partición de *n* observaciones en *k* grupos; en la que cada observación pertenece al grupo más cercano con la media. Es similar a la expectativa de algoritmo de optimización de mezclas de gaussianas porque ambos intentan encontrar los centros de las agrupaciones naturales en los datos [9].

Dado un conjunto de observaciones $(x_1, x_2, \ldots x_n)$, donde cada una es un d-vector real dimensión, a continuación, k-means tiene por objeto dividir las *n* observaciones en series *K* ($K < N$) $S = (S_1, S_2, S_3, \ldots, S_K)$ con el fin de reducir al mínimo la suma de los cuadrados:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$$

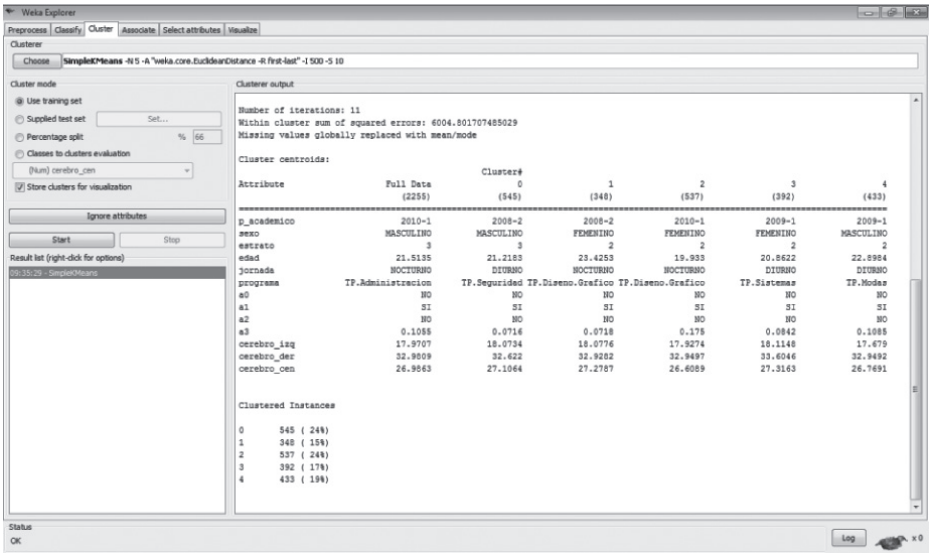
Donde μ_i es el punto medio en S_i.

Existen dos versiones del algoritmo k-medias. La primera es similar al Algoritmo EM y se basa en dos pasos iterativos: primero, reasigna todos los puntos a sus centros y demás cercanos; y segundo, vuelve a calcular los centroides de los nuevos grupos creados en el anterior. El proceso continúa hasta alcanzar un criterio de parada (por ejemplo, que no se realicen nuevas reasignaciones).

Esta versión se conoce como Algoritmo de Forgy [10]. Por su parte, la segunda [11] reasigna los puntos basándose en un análisis más detallado de los efectos causados sobre la función del objetivo al mover un punto de su *cluster* a otro nuevo. Si el traslado es positivo, se realiza, y en caso contrario, se queda como está. A diferencia de los anteriores algoritmos (*COBWEB* y *EM*, *k-medias*) éste necesita la especificación previa del número de *clusters* que se desean obtener. La implementación utilizada del Algoritmo de Forgy en el trabajo presente también es la ofrecida por *Weka* y se aplicará a los siguientes datos:

Se va a comprobar si el atributo *p_academico* divide naturalmente a los alumnos en grupos

Figura 17. Algoritmo SimpleKMeans, con numCluster = 5.



similares, para lo cual se seleccionará el algoritmo *SimpleKMeans* con un número de *clusters* igual a cinco. A continuación, aparecerán los cinco grupos de ejemplos más similares y sus centroides (promedios para atributos numéricos y valores más repetidos en cada grupo para atributos simbólicos). Por ejemplo, para tres de los cinco *clusters*, el programa más repetido es *TP.Administracion* y, para los otros dos, son *TP.Seguridad* y *TP.Diseno*. En todos los *clusters*, el periodo de mayor afluencia de estudiantes fue el 2010-I. El número de instancias agrupadas en cada *cluster* se observa en la figura 17.

La implementación de este algoritmo de agrupamiento permitió clasificar los programas más relevantes y sirvió como "filtro" puesto que la primera instancia que se obtuvo en los cinco grupos fue:

- No seleccionó el periodo académico 2009-II.
- Los periodos académicos en los que se agruparon mayor cantidad de hombres fueron 2008-II y 2009-I.
- El estrato socioeconómico 1 no fue relevante en la agrupación.
- El grupo con mayor edad fue el 1, que perteneció al periodo académico 2008-II. Su promedio fue de 23 años.

En la siguiente figura se describen los resultados obtenidos por el algoritmo *SimpleKMeans* en Weka.

En la figura 19 se encuentra ausente el periodo académico 2009-II. Por este motivo sólo aparecen cuatro grupos.

El Algoritmo Esperanza-Maximización (EM) se utiliza en estadística para encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables. Además, alterna pasos de esperanza (E) donde se computa la <<esperanza de la verosimilitud>> mediante la inclusión de variables latentes como si fueran observables; y un paso de maximización (M), en el que se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada de E. Los parámetros que se encuentran en M se usan para comenzar el siguiente paso E. De esta forma, el proceso se repite [12].

Figura 18. Resultados del agrupamiento SimpleKMeans con k = 5.

	Cluster#				
Full Data	0	1	2	3	4
(2255)	(545)	(348)	(537)	(392)	(433)
2010-1	2008-2	2008-2	2010-1	2009-1	2009-1
MASCULINO	MASCULINO	FEMENINO	FEMENINO	FEMENINO	MASCULINO
3	3	2	2	2	2
21.5135	21.2183	23.4253	19.933	20.8622	22.8984
NOCTURNO	DIURNO	NOCTURNO	NOCTURNO	DIURNO	DIURNO
TP.Administracion	TP.Seguridad	TP.Diseno.Grafico	TP.Diseno.Grafico	TP.Sistemas	TP.Modas
NO	NO	NO	NO	NO	NO
SI	SI	SI	SI	SI	SI
NO	NO	NO	NO	NO	NO
0.1055	0.0716	0.0718	0.175	0.0842	0.1085
17.9707	18.0734	18.0776	17.9274	18.1148	17.679
32.9809	32.622	32.9282	32.9497	33.6046	32.9492
26.9863	27.1064	27.2787	26.6089	27.3163	26.7691

Figura 19. p_academico vs. periodo.

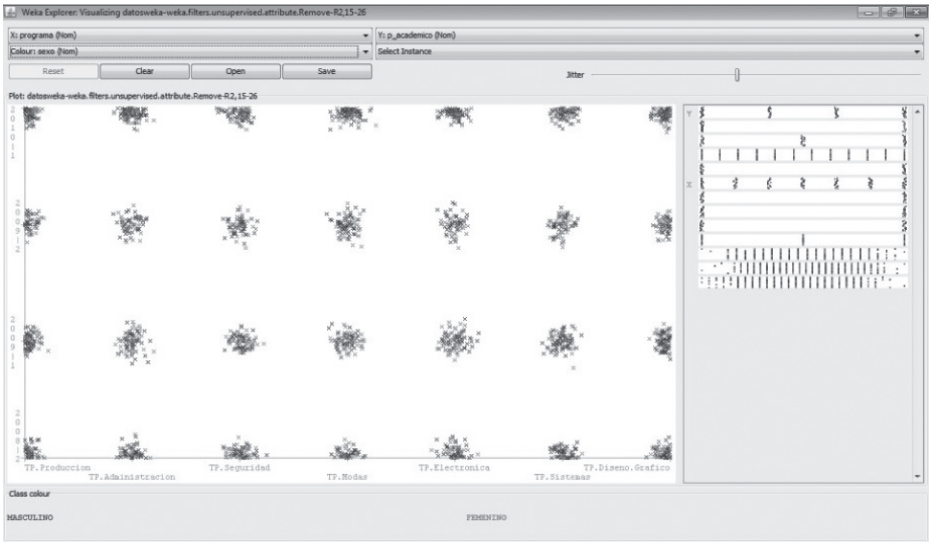
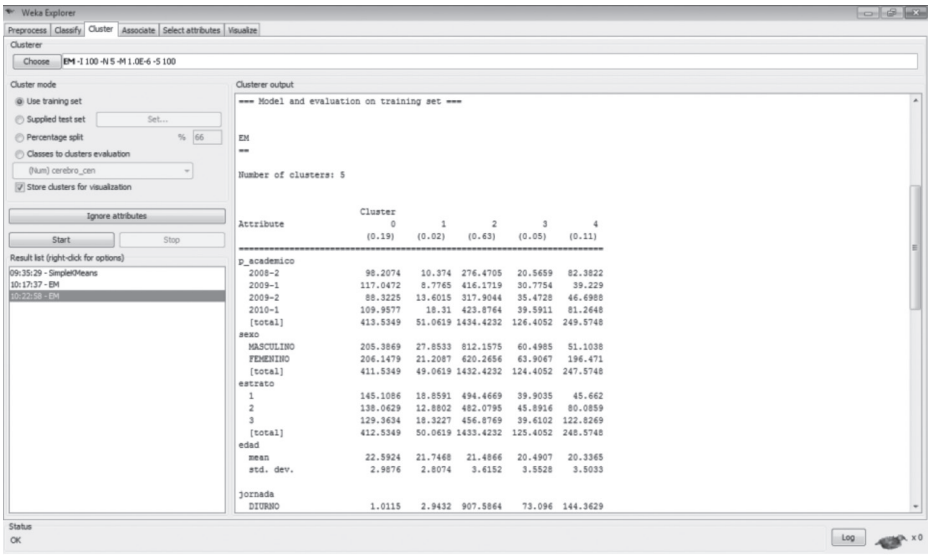


Figura 20. Algoritmo EM en Weka.



Implementando en Weka el algoritmo *EM, cin N = 5* se puede extraer e inferir la siguiente información, que se muestra en las figuras 20, 21 y 22:

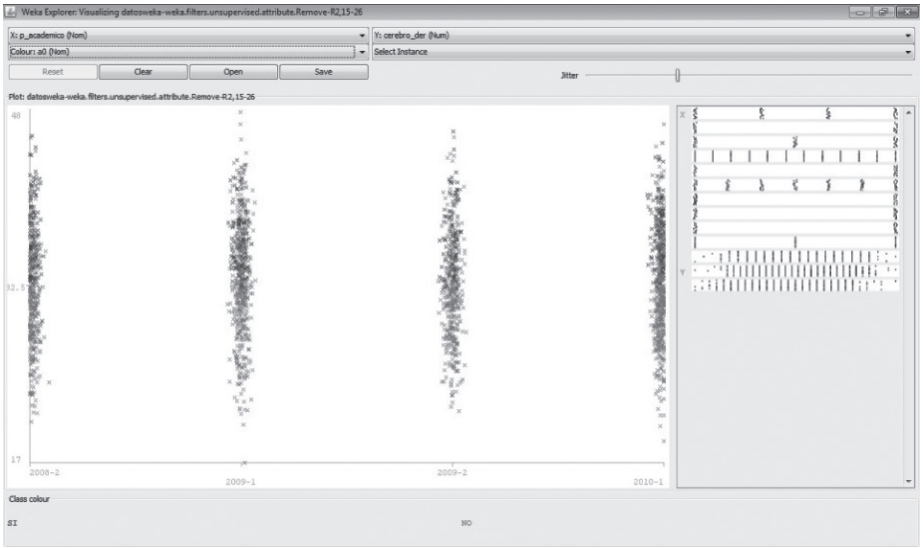
La figura 21 se destaca por el Algoritmo EM con N = 5 y una desviación estándar de $1x10^{-6}$ Entonces, se tiene:

- El grupo dos es el mayor. En él se encuentra concentrada el 72% de la información.
- Cuando los atributos son de tipo numérico, la información arrojada se centra en el promedio y en la desviación estándar. Una característica llamativa es que los cinco grupos tienen una predominancia del cerebro derecho; es decir, que la caracterización en promedio de toda la población está desarrollada en su hemisferio derecho.

Figura 21. Algoritmo EM (promedios).

```
a1
  SI                336.2892   38.3357 1380.4445  121.0987  237.8319
  NO                 75.2457   10.7262   51.9787    3.3065    9.7429
  [total]           411.5349   49.0619 1432.4232  124.4052  247.5748
a2
  SI                      1   41.9057         1  123.0943         1
  NO                410.5349    7.1563 1431.4232    1.3109  246.5748
  [total]           411.5349   49.0619 1432.4232  124.4052  247.5748
a3
  mean                    0     1.258         0    1.4607         0
  std. dev.          0.4012    0.6731    0.4012    0.5035    0.4012

cerebro_izq
  mean              18.1174   14.6887   18.2615   14.6778   18.3026
  std. dev.          3.2725    2.7019    3.4161    2.0517    3.5318

cerebro_der
  mean              36.8244   36.8597   32.0113   31.6163   32.1561
  std. dev.          2.9588    2.9199    4.4573    4.0579    4.8041

cerebro_cen
  mean              27.3838   21.3011   27.4761   20.4693   27.8077
  std. dev.          3.7343    2.8312    4.0242    1.8473    4.3011

Clustered Instances

0       447 ( 20%)
1        40 (  2%)
2      1621 ( 72%)
3       123 (  5%)
4        24 (  1%)
```

Figura 22. cerebro_der vs *p_academico.*



COBWEB [7] es un algoritmo de *clustering* jerárquico que se caracteriza por utilizar aprendizaje incremental; en otras palabras, realiza las agrupaciones <<instancia a instancia>>. Así mismo, durante la ejecución del algoritmo se forma un árbol (de clasificación) en el que las hojas representan los segmentos y el nodo-raíz engloba, por completo, el conjunto de datos de entrada.

En el inicio, el árbol consiste en un único nodo-raíz. Luego, las instancias se van añadiendo una a una y el árbol empieza a actualizarse en cada paso. La actualización, a su vez, consiste en encontrar el mejor sitio para incluir la nueva instancia, operación que puede requerir de la reestructuración de todo el árbol (incluyendo la generación de un nuevo nodo-anfitrión para la instancia y/o la fusión/partición de nodos existentes) o simplemente la inclusión de la instancia en un nodo existente. La clave para saber cómo y dónde se debe actualizar el árbol la proporciona una medida denominada *utilidad de categoría*, la cual valora la calidad general de una partición de instancias en un segmento.

Por su parte, la reestructuración que mayor utilidad de categoría proporciona es la adoptada en ese paso. El algoritmo, además, es muy sensible a otros dos parámetros:

Acuity: este parámetro es necesario ya que la utilidad de categoría se basa en una estimación de la media y la desviación estándar del valor de los atributos. Sin embargo, cuando se estima la desviación estándar del valor de un atributo para un nodo en particular, el resultado es cero (0) si dicho nodo sólo contiene una instancia. Así pues, el parámetro *acuity* representa la medida de error de un nodo con una sóla instancia; es decir, establece la varianza mínima de un atributo.

Cut-off: valor empleado para evitar el crecimiento desmesurado del número de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual.

En otras palabras, cuando el incremento de la utilidad de categoría no es suficiente al añadir un nuevo nodo, éste se corta conteniendo la instancia de otro ya existente.

Además, *COBWEB* pertenece a los métodos de aprendizaje conceptuales o basados en modelos. Esto significa que cada *cluster*, más que un ente formado por una colección de puntos, es considerado un modelo que puede describirse intrínsecamente. Al algoritmo *COBWEB* no hay que proporcionarle el número exacto de *clusters* deseados, pues en base a los parámetros mencionados encuentra el número óptimo. La implementación utilizada en este trabajo es la adoptada por el algoritmo implementado en *Weka*.

Aplicando el algoritmo de *COBWEB* en *Weka*, con los parámetros por defecto y solamente a los siguientes atributos:

- Sexo.
- Estrato.
- Programa.
- Cerebro_izq
- Cerebro_der
- Cerebro_cen

Entonces, el programa arroja los resultados mostrados en la figura 23 y en la visualización del árbol en la figura 24.

Figura 23. Algoritmo COWEB en Weka.

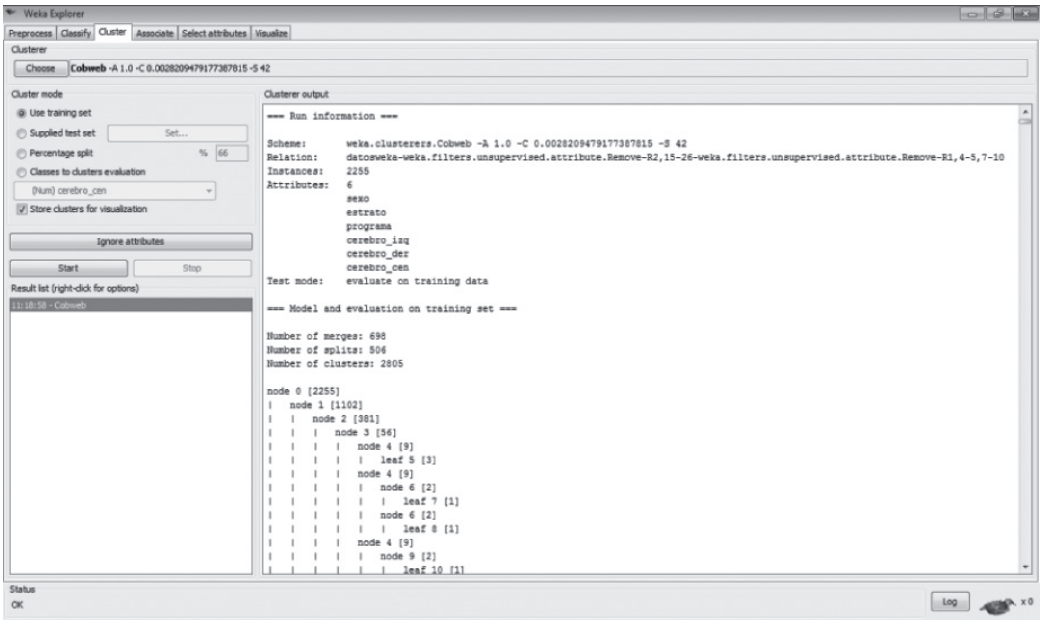
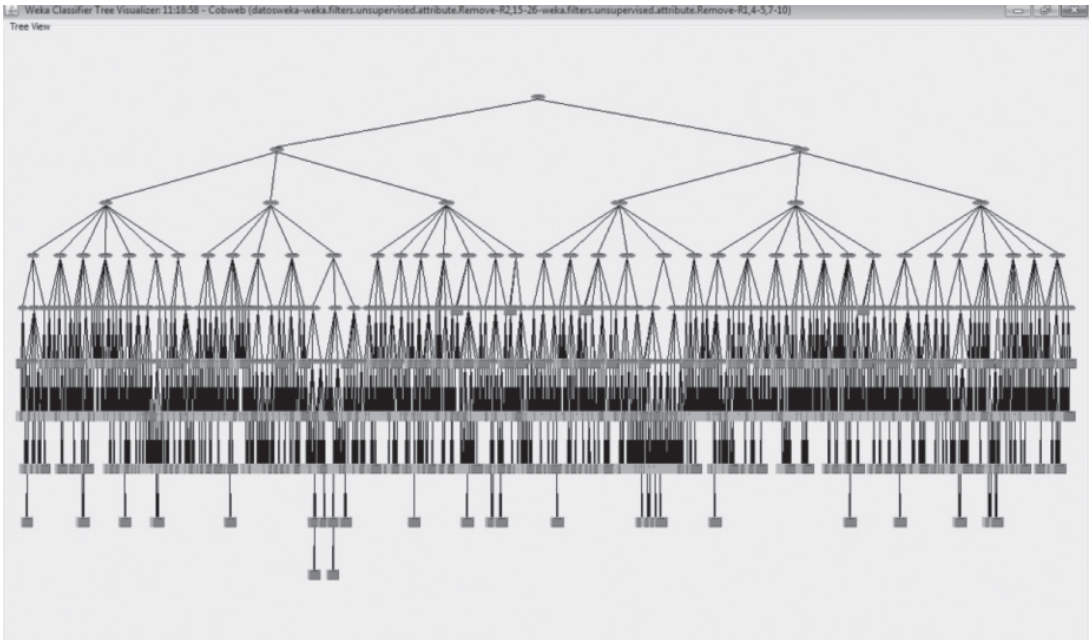


Figura 24. Árbol Algoritmo COWEB.



Conclusiones

La aplicación del proceso de KDD posibilitó determinar la existencia de un alto grado de homogeneidad en la población estudiantil promedio que integra la Fundación Insutec. Así mismo, mediante el análisis por *cluster*, se identificó el estilo dominante de la población: en el 72% predomina el uso del lóbulo derecho. Este resultado se comprobó aplicando el Algoritmo EM.

Por otra parte, en este trabajo también se plantearon lineamientos generales para adecuar las pruebas de admisión y selección del programa académico respecto al estilo dominante que caracteriza a la población de estudiantes. Cabe resaltar que dichos lineamientos deberán ser particularizados dentro del marco de cada programa académico.

El algoritmo *COBWEB*, perteneciente a la familia *clustering jerárquico*, no resultó adecuado para la segmentación de los datos trabajados en este artículo; tiende a agrupar a la mayoría en un sólo segmento, acción poco efectiva si se pretende la mejora en la estimación de la caracterización estudiantil. No obstante, *EM* y *k-medias* —algoritmos de particionado y recolocación— ofrecen mejores resultados que *COBWEB*, pues indican tareas de segmentación y clasificación en estudiantes, ya sea por programa académico, tendencia cerebral o caracterización socioeconómica.

Luego de comparar *EM* con *k-medias* pudieron apreciarse varias diferencias entre sus segmentos: *k-medias* agrupa en un sólo *cluster* los mismos programas académicos que *EM* divide en varios; es decir, *EM* realiza una división más específica que *k-medias*. Entonces *EM*, siendo un algoritmo que realiza *clustering probabilístico*, es más adecuado que *k-medias* para segmentar poblaciones de datos que se distribuyen normalmente. Lo anterior con el fin de mejorar la estimación de los promedios poblacionales de cada clase.

Referencias

[1] Felder, R.M. y Silverman, L.K. (1988), "Learning Styles and Teaching Styles in Engineering Education", en *Engr. Education*, vol. 78, núm. 7, pp. 674-681.

[2] Durán E. y Costaguta R. (2007), "Minería de Datos para Descubrir Estilos de Aprendizaje", en *Revista Iberoamericana de Educación*, VOL. NÚM. PP.

[3] Felder, F.M. y Soloman, B.A. (1991), "Index of Learning Styles".North Carolina State University.Disponibleon-line, [en línea], disponible en: <http://sn.umdnj.edu/studentsonly/cas/IndexofLearningStyles.pdf

[4] http://www.cs.waikato.ac.nz/ml weka/

[5] Waldemar, D.G. (1999), "En busca de una nueva noologia", en *Estudio pedagógico*, núm. 25, pp. 71-82.

[6] Fundación de Educación Superior. INSUTEC. (2010). Departamento de Bienestar Institucional. Bases de datos Admisiones.

[7] Alonso, C. (AÑO), "Waitako Environment for Knowledge Analysis. Introducción básica", [en línea], disponible en: http://www.infor.uva.es/~calonso/IAII/Aprendizaje/weka/IntroduccionWeka.pdf

[8] Merlino, H., *et. al.* (2005), "Un Método de Transformación de Datos Orientado al Uso de Explotación de Información". XI Congreso Argentino de Ciencias de la Computación, Buenos Aires, Argentina.

[9] Sáez Olivito, E., *et. al.* (1999), "*Caracterización estructural de explotaciones ovinas aragonesas mediante métodos estadísticos multivariantes*", SEOC XXII,

[10] Garre, M., *et. al.* (2007), "Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software", en *Revista Española de Innovación, Calidad e Ingeniería de Software*, vol. 3, núm. 1, pp. 1885–4486.

[11] Duda, R. Hart, P., "Pattern Classification and Scene Analysis", Wiley & Sons, 1973.

[12] Gómez García, J.; PalareaAlbaladejo, J. y Martín Fernández, J.A. (2006). "*Métodos de Inferencia Estadística con datos faltantes*", en *Estadística española*, vol. 48, pp. 240 -275.