UNIVERSIDAD ESTATAL PENINSULA DE SANTA ELENA
INSTITUTO DE INVESTIGACIÓN CIENTÍFICA
Y DESARROLLO TECNOLÓGICO
INCYT - UPSE

Minería de datos con técnica cluster, caso de estudio: estudiantes de Universidad Estatal Península de Santa Elena.

Mariuxi Alexandra De La Cruz De La Cruz.
Dirección de Planeamiento.
Universidad Estatal Península de Santa Elena.
Calle Principal Vía Salinas-La Libertad.
La Libertad, Ecuador.
mdelacruz@upse.edu.ec

Resumen

Este trabajo fue realizado para la Universidad Estatal Península de Santa Elena, muestra el proceso de minería de datos aplicando la técnica cluster o conglomerado, con herramientas de software libre como: PostgreSQL, apropiada para el almacenamiento de los datos y R para el análisis estadístico y la generación de los algoritmos que implica la técnica. El análisis cluster fue implementado en datos sociales, económicos y académicos de los estudiantes inscritos en la Facultad de Sistemas y Telecomunicaciones de esta institución de educación superior. El proceso es detallado en tres secciones, la primera donde se describe los conceptos asociados a minería de datos, técnica cluster y las herramientas tecnológicas utilizadas para explotación de información. En la segunda sección se describe el proceso de aplicación la técnica cluster con el método particional y con el método jerárquico. En la sección tres se muestra los conglomerados obtenidos al aplicar la técnica en el grupo de datos. Finalmente se presentan las conclusiones y recomendaciones de este trabajo.

Palabras Claves: Minería de datos, cluster, algoritmo particional, algoritmo jerárquico, PostgreSQL, Lenguaje R**.**

Abstract

This project was done for Universidad Estatal Península de Santa Elena, the objective was to show the process data mining and cluster analysis application, using free software tools: PostgreSQL, suitable for storage data y R for statistical analysis and algorithms generation. The cluster analysis was implemented in the social, economic and academic data, of students at the Systems and Communications School . This process is described in three section distributed as follows: Section I, describes the concept associated with data mining, cluster analysis, tools for exploration of data. Section II, describes the process of cluster analysis with partitioned algorithm and using hierarchical algorithm on data set. Section III show the clusters after applying the technique in the academic, social and economic data set. Finally the conclusions and recommendations are mentioned.

Keywords: Data mining, cluster analysis, partitional algorithm, hierarchical algorithm, PostgreSQL, Languaje R.

1. Introducción.

En la actualidad, toda entidad pública o privada cuenta con procesos automatizados, los mismos que generan gran cantidad de datos, existe el proceso denominado "minería de datos" o "explotación de conocimiento", que nos permite a través de técnicas y algoritmos obtener información relevante, para ayudar a tomar mejores decisiones a nivel gerencial.
En el presente trabajo se describe la técnica de minería de datos conocida como: "cluster" o "conglomerados", con dos de sus métodos de clasificación: el particional y el jerárquico, utilizando en todo el proceso, herramientas de software libre.
La técnica expuesta será aplicada en datos académicos, sociales, y económicos de los estudiantes inscritos en la Facultad de Sistemas y Telecomunicaciones de la Universidad Estatal Península de Santa Elena, con la finalidad de identificar el perfil del estudiante universitario que está formándose en esta institución de educación superior.

1.1. Minería de datos.

La minería de datos, es el análisis de gran cantidad de datos para encontrar relaciones desconocidas y describir los datos en nuevas formas que son comprensibles para el tomador de decisiones. (Hand D.,2001) (5).

1.2. Fases de minería de datos.

Figura 1. Fases de minería de datos
Fuente: Hernández Orallo, José , Ma. José Ramírez Quintana, Cesar Ferri Ramírez. *Introducción a la Minería de datos.* España, 2004.(4)

Según la Figura 1, se tienen seis fases: 1. La etapa inicial que tiene que relación a la planificación del trabajo.
2. La comprensión del negocio, es la exploración de los datos identificando una necesidad en el mercado para ser explotada.
3. La extracción y limpieza de los datos, tiene mucha relación con calidad del dato, identificar en este paso si existen datos anómalos, atípicos para tratarlos apropiadamente.
4. La transformación y elección de variables, se encuentran las tareas de discretización, numerización, estandarización y normalización, se elige la adecuada dependiendo de la variable que formará parte del análisis.
5. La explotación de los datos involucra la elección de la técnica de minería de datos a utilizar.
6. Fase de interpretación y evaluación tiene relación con la descripción de los resultados obtenidos. (4)

1.3. Técnica cluster.

La técnica utilizada en el presente trabajo es "cluster", cuya finalidad esencial es revelar concentraciones en los datos (casos o variables), pudiéndose utilizar variables cuantitativas y cualitativas (Myatt y Jhonson, 2009).
Existen dos pasos que implican la técnica cluster:

1. Calcular las distancias entre los datos (matriz de distancias).
2. Elegir el algoritmo de clasificación.

Distancia entre los datos.
Enfocándonos a describir el coeficiente a utilizar para el cálculo de las distancias entre observaciones que vienen de variables mixtas, el ***Coeficiente de Gower***, es descrito por Myatt y Jhonn:
El coeficiente Gower es calculado con la Ec.(1) para dos observaciones p y q, sobre i variables:

$$d(p,q) = \sqrt{\frac{\sum_{i=1}^{n} w_i d_i^2}{\sum_{i=1}^{n} w_i}} \qquad (1)$$

donde w_i es el peso para la i-ésima variable toma el valor de uno cuando ambos valores son conocidos, en otro caso es cero.
El valor d_i^2 es el cuadrado de la distancia entre el i-ésimo valor de las dos observaciones (pi y qi). (6)

Algoritmo de clasificación.
Los métodos utilizados para la agrupación se dividen en jerárquicos y no jerárquicos, los primeros además tienen las técnicas aglomerativas y divisivas, y se destacan los algoritmos: enlace simple, enlace completo, enlace en la media; mientras que los métodos no jerárquicos también llamados particionales tienen técnicas como: k-medias, k-modas, k-medoides.

1.4. Herramientas tecnológicas.

Para el almacenamiento de los datos se utiliza el gestor de base de datos PostgreSQL, software de código abierto, de acceso libre. (7)
Para el análisis y aplicación de los algoritmos del análisis cluster, se utiliza el lenguaje R, software estadístico de acceso libre. (3)

2. Metodología.

2.1 Sistemas informáticos usados.

Las matrices iniciales donde se aplica el análisis cluster, fueron proporcionadas en formato Excel, para el almacenamiento y transformación de estos datos se recurrió a la herramienta PostgreSQL y al lenguaje estadístico R para la ejecución de los algoritmos que implica la técnica cluster.

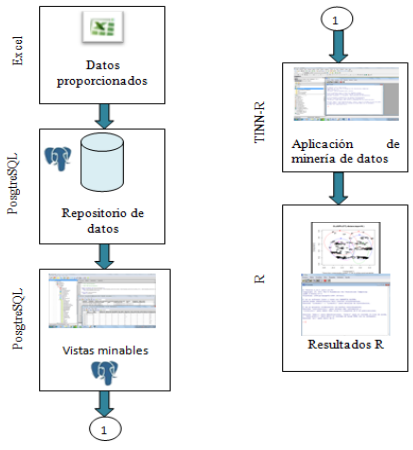

Figura 2. Proceso desde recopilación hasta resultados de minería en cada herramienta tecnológica.

En la Figura 2, están todos los pasos generales ejecutados en las dos herramientas utilizadas: PostgreSQL, y R; las tareas ejecutadas en el entorno de Postgresql son enunciadas:

- ✓ Carga de los datos proporcionados en formato Excel (creación de estructura de tablas y llenado de datos)
- ✓ Funciones que permitieron transformar y seleccionar los datos
- ✓ Integración de datos con la creación de vistas.
- ✓ Generación de la vista minable (tabla final donde se incluía todos los campos con quienes se trabajaría el análisis cluster). (8)

El lenguaje "R" posee un editor de texto de acceso gratuito denominado "TINN-R; se trabajó en los dos entornos para:

- ✓ Creación de instrucciones para imputación de datos.
- ✓ Creación de instrucciones para estandarización de datos.
- ✓ Creación de instrucciones donde se aplicaba el análisis cluster.
- ✓ Creación de instrucciones para gráfico de resultados. (8)

2.2 Técnica cluster con método particional.

La Unidad de Producción de la Escuela de Informática (UPEI), es la unidad responsable de los sistemas que automatizan los procesos académicos de la Universidad Península de Santa Elena (UPSE), nos proporcionó los datos históricos para proceder a realizar el análisis.

2.2.1 Recopilación de datos.

Creación de la estructura de la tabla con las notas de estudiantes.

```
CREATE TABLE "PerfilFSE"."Notas_Facistel_tot"
(
  id_ord varchar(255),
  carrera varchar(255),
  periodo varchar(255),
  sistema_estudio varchar(255),
  nivel integer,
  estudiante varchar(255),
  matricula varchar(255),
  materia varchar(255),
  promedio double precision
)
WITH (
  OIDS=FALSE
);
ALTER TABLE "PerfilFSE"."Notas_Facistel_tot"
  OWNER TO postgres;
```

Instrucción para cargar los datos de la tabla al gestor de base de datos

```
copy "PerfilFSE"."Notas_Facistel_tot" from 'c:\Notas_facistel_tot.csv' with delimiter ',' null as ''
```

Figura 3. Instrucciones para crear estructuras de tablas y cargar datos en el gestor de base de datos.

La matriz proporcionada contenía campos con datos generales y académicos del estudiante obtenidos en el proceso de matriculación, esta matriz se encontraba en formato Excel fueron transformados a archivos con formato .csv para ser cargados en el gestor de base de datos PostgreSQL, las instrucciones utilizadas son las mostradas en la Figura 3.

2.2.2 Transformación, limpieza y selección de datos. En este paso se establece la creación de atributos, en el entorno de postgreSQL, en este caso se adicionaron atributos y se crearon "vistas" (tablas virtuales donde se almacena consultas), estas vistas se convertirían luego en las "vistas minables" (tablas que formarán parte del análisis).

En la Tabla 1 se expone los atributos creados y la descripción indica los criterios considerados para su creación.

Tabla 1. Campos creados y su descripción.

Atributos	Descripción
cantidad_materias_aprobadas	Cantidad de materias cuyo promedio es >=70 por cada estudiante
cantidad_materias_reprobadas	Cantidad de materias cuyo promedio es< 70 por cada estudiante
Promedio	Promedio de calificación en la universidad por estudiante
fecha_inicio	Primera fecha en que se registró el estudiante
fecha_fin	Última fecha en que se registró el estudiante
tiempo_univ	fecha_fin - fecha_inicio (diferencia entre las dos fechas)
Indice_estudio	Cantidad de materias aprobadas/tiempo_univ

En la Tabla 2, se presenta los atributos que forman parte de la vista denominada "minable", donde se trabajó solo con registros que tengan datos completos en todos sus campos.

En esta fase se ejecuta también la transformación de todos los valores de las variables a datos numéricos, en este caso se utilizó la creación de variables **"dummy**" o ficticias que pasarían a ser binomiales.

Finalmente se procede a realizar la **normalización de las variables** cuantitativas, en un rango entre 0 y 1, para evitar el sesgo al calcular las distancias entre los datos.

Tabla 2. Campos de "vista minable".

Nombre de variable	Descripción
Masculino	1.- Masculino
	0.- Femenino
Fima	1. Si / 0. No
Informática	1. Si / 0. No
Electrónica	1. Si / 0. No
Comercio y administración	1. Si / 0. No
Quibio	1. Si / 0. No
Otras	1. Si / 0. No
Fiscal	1. Si / 0. No
Particular	1. Si / 0. No
Otro	1. Si / 0. No
Trabaja	1. Si / 0. No
Edad	Edad
Califi_cole	Calificación promedio con el que se graduó en el colegio
Prome_calif	Promedio de calificación en la universidad
Mate_reprob	Materias reprobadas en toda su estadía en la universidad
Tiempo_univ	Tiempo que tiene en la universidad
Indice_mat	Promedio de materias aprobadas por período académico

2.2.3 Extracción del conocimiento. Al aplicar la técnica cluster uno de los pasos iniciales es el cálculo de la **matriz de distancias**, para este procedimiento se utiliza el lenguaje R.

En la Figura 4, se encuentran las instrucciones ejecutadas en R para calcular las distancias entre los datos.

```
library(cluster)
distanciaperfil<-
daisy(dataintegradausar,metric="gower")
agddatos<-agnes(distanciaperfil, method = "average")
```

Figura 4. Instrucciones para calcular distancia entre los datos.

Luego se procede a elegir él método de clasificación, en este caso se utiliza ***la técnica particional***, Partitioning Around Medoids (PAM), donde inicialmente se debe elegir la cantidad de grupos en los que se clasificará las observaciones.

Para seleccionar la cantidad de grupos, se utilizó el criterio **de "siluetas"** que representa la distancia promedio entre la observación de un grupo y la observación de otro grupo, mientras mayor sea el coeficiente de silueta, mejor será la estructura de los grupos.

```
library(cluster) # necesaria para correr PAM
asw <- numeric(20)
## Note that "k=1" won't work!
for (k in 2:20)
asw[k] <- pam(distanciaperfil,k) $silinfo $avg.width
k.best <- which.max(asw)
cat("Numero Optimo de Grupos (por Silueta):", k.best, "\n")
plot(1:20, asw, type= "h", main = "Evaluacion de Grupos con pam()",
xlab= "k (No. Grupos)", ylab = "Ancho Medio de Silueta")
axis(1, k.best, paste("mejor",k.best,sep="\n"), col = "red", col.axis = "red")
```

Figura 5. Algoritmo para definir la cantidad de conglomerados óptimos utilizando PAM.
Fuente: Bernardis, Reeb, Bramardi. "Agrupamiento de pozos de petróleo en base de datos de perforación". Argentina, 2009. (2)

.

En la Figura 5, se encuentra las funciones utilizadas para aplicar el método de clasificación y las instrucciones utilizadas para identificar la cantidad de grupos, sentencias que fueron ejecutadas en Lenguaje R.

2.2.3 Validación de la técnica.

Para el proceso de validación se utiliza la definición de silueta, la misma que identifica la disimilitud(desigualdad) entre los grupos, si el coeficiente de la silueta entre grupos es próximo a uno significa que un objeto i está bien clasificado en el grupo, mientras que el coeficiente se aleje de 1 entonces se considera que no hay una apropiada clasificación, la interpretación para el coeficiente de la silueta se presenta en la Tabla 4. (1)

Tabla 3. Interpretación subjetiva del coeficiente de silueta (SC), definido como el ancho de Silueta Promedio Máximo para todo el conjunto de datos.

Coeficiente de silueta	Interpretación propuesta
0,71 - 1,00	Estructura Sólida
0,51 - 0,70	Estructura Razonable
0,26 - 0,50	Estructura débil, puede ser artificial, probar métodos adicionales
<=0,25	No hay estructura

Fuente: Ayala Gallejo, Guillermo. Análisis de datos con R, para Ingeniería Informática , 2008.

2.3 Técnica cluster con método jerárquico.

Debido a la falta de dato existente sobre aspectos sociales, y económicos del estudiante que complementan el aspecto académico, se propuso mejorar la ficha socioeconómica adicionando nuevos campos, y plantear el ingreso de los datos en línea.

2.3.1 Recopilación, transformación y limpieza de datos.

Tabla 4. "Vista minable" con variables académicas, sociales, y económicas.

Nombre de Variables	Característica
Calificación	Calificación del estudiante al graduarse de bachiller
Edad	Edad del estudiante actualmente
M_reprobadas	Materias reprobadas en el período de estudios
Tiempo_univ	Tiempo que se encuentra en la universidad
Promedio	Promedio de calificaciones en la universidad
Medtoting	Promedio del ingreso familiar
Indice_materias	Promedio por año, de materias aprobadas
Masculino	1.- Masculino
	0.- Femenino
Fima	1. Si / 0. No
Informática	1. Si / 0. No
Administración de sistemas	1. Si / 0. No
Electrónica	1. Si / 0. No
Técnico Industrial	1. Si / 0. No
Contabilidad	1. Si / 0. No
Ciencias Administrativas	1. Si / 0. No
Quibio	1. Si / 0. No
Otras	1. Si / 0. No
Fiscal	1. Si / 0. No
Fisco-Misional	1. Si / 0. No
Particular	1. Si / 0. No
Municipal	1. Si / 0. No
Casado	1. Si / 0. No
Unión Libre	1. Si / 0. No
Divorciado	1. Si / 0. No
Separado	1. Si / 0. No
Viudo	1. Si / 0. No
Soltero	1. Si / 0. No

La propuesta fue aceptado por las autoridades académicas y a partir del mes de julio del 2011, en la UPSE, empezó el proceso de llenado obligatorio de la ficha accediendo al link desde la página web de la universidad.
Con el 47% de los datos actualizados hasta octubre del 2011, se procedió a ejecutar cada uno de los pasos que indica la técnica cluster utilizando como método de agrupación, el jerárquico, se detalla la aplicación de la técnica en los datos respectivos.
En la tarea de recopilación se volvió a cargar los datos desde Excel a postgreSQL, en la transformación se crearon las variables **"dummy"** para ejecutar la tarea de numerización, se integraron todos las variables procediendo a tener la "vista minable", con los atributos que se describen en la Tabla 3.
2.3.2 Extracción del conocimiento.- Para la fase de explotación de los datos se utilizó el análisis cluster pero considerando ***la técnica de clasificación jerárquica (AGNES)***, se empezó calculando la matriz de distancias.

```
Técnica AGNES y UPGMA
agddatos<-agnes(distanciaperfil, method = "average")
cofagddatos<- cophenetic(agddatos)
cor(distanciaperfil, cofagddatos)
```

Figura 6. Instrucciones del método jerárquico.

Las instrucciones utilizadas en R para aplicar la técnicas son las que se presentan en la Figura 6 y representan al uso del método UPGMA(Unweighted Pair Group Method with Arithmatic Mean).
En la Figura 7, están las instrucciones ejecutadas en R, para observar el dendograma(gráfico que permite observar los conlgomerados encontrados), en el mismo se pudieron identificar seis grupos posibles.

```
hclddatos<- (as.hclust(agddatos))
ramas <-cutree(hclddatos, k = 6)
datosfinales<-data.frame(dataintegradaprimera,ramas)
pltree(agddatos,xlab="EstudiantesFacistel",ylab="Altura"
,cex=0.6, main="DENDOGRAMA PARA CLUSTER DE
ESTUDIANTES")
inforcorte<-rect.hclust(agddatos, k=6, border="blue")
```

Figura 7. **Instrucciones del dendograma.**

2.3.3 Validación de la técnica.
En este segundo análisis se utilizó el método jerárquico aglomerativo, y según el criterio de coeficientes de siluetas, se obtuvo un coeficiente de 0,27, por lo tanto tiene una estructura débil considerando como valores base, los que se observan en la Tabla 4. (1)

3. Resultados.

La finalidad del análisis cluster es dividir un conjunto de objetos y clasificarlos en grupos tal que las características de los entes ubicados en un mismo grupo sean similares entre sí. (8).
En este trabajo se aplicó el mencionado análisis para obtener el perfil de los estudiantes de la Facultad de Sistemas y Telecomunicaciones, las matrices fueron proporcionados por la unidad de producción de informática, ente responsable del almacenamiento de estos datos.
El análisis cluster posee varios métodos de clasificación, se han utilizado dos: el particional y el jerárquico aglomerativo, siendo el segundo el que nos proporcionó una mejor interpretación en los grupos formados.
La primera ejecución de la técnica cluster fue con el método de clasificación particional, aplicado en atributos académicos y en ciertos atributos que contenían datos generales de los estudiantes, en este primer proceso existía un alto porcentaje de dato faltante, se identificaron 11 grupos, sin embargo al validar el modelo aplicado presentó un coeficiente que indicaba una estructura de clasificación débil.
La segunda ejecución de la técnica cluster fue utilizando el método jerárquico aglomerativo con más atributos que se recopilaron de los estudiantes (sociales, económicos y académicos), el método de clasificación es considerado más apropiado dado que los conglomerados se van formando paso a paso y automáticamente, no así el método particional donde con anterioridad se debe indicar la cantidad de grupos en los que se clasifican los datos.
En la segunda ejecución, se obtuvo un coeficiente de silueta, que indicaba una estructura débil, sin embargo la cantidad de grupos identificados fueron 6.
La metodología para aplicar el análisis cluster y las instrucciones utilizadas están claramente identificadas en este trabajo, permitiendo que exista posibilidad de volver a aplicar en conjuntos de datos distintos, por lo tanto buscando una mejor interpretación de los grupos que pudieran formase, se realizó una tercera ejecución de la técnica cluster con el método de clasificación jerárquico en los datos de los

estudiantes pero separándolos por sistemas de estudio: Semestral y Anual.
A pesar que en esta última ejecución, al validar la técnica se obtuvo un coeficiente de silueta de 0,26, tanto para aquellos individuos que estaban en sistema anual como para aquellos que estaban en sistema semestral, valor que indica una estructura débil de agrupación, sin embargo al ejecutar por separado, según el sistema de estudio, los grupos o conglomerados formados pueden describirse mejor que en las anteriores ejecuciones.
Para describir cada uno de los conglomerados formados, se analiza los atributos de cada grupo, a través del cálculo de frecuencias, diagramas de barras o dispersión.

3.1 Cluster o conglomerados identificados.

Al aplicar el análisis cluster en los estudiantes del ***sistema de estudio anual*** se obtuvieron siete grupos, tal como se observa en el dendograma presentado en la Figura 8.

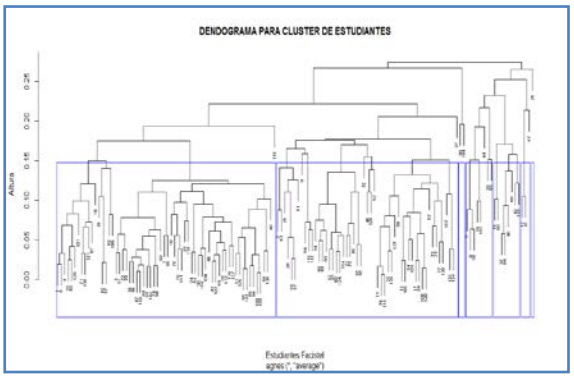

Figura 8. Dendograma obtenido para estudiantes del sistema de estudio Anual.

Los estudiantes del ***sistema de estudio semestral*** fueron clasificados en cinco conglomerados, utilizando el análisis cluster, con método de clasificación jerárquico aglomerativo, tal como se observa en el dendograma de la Figura 9.

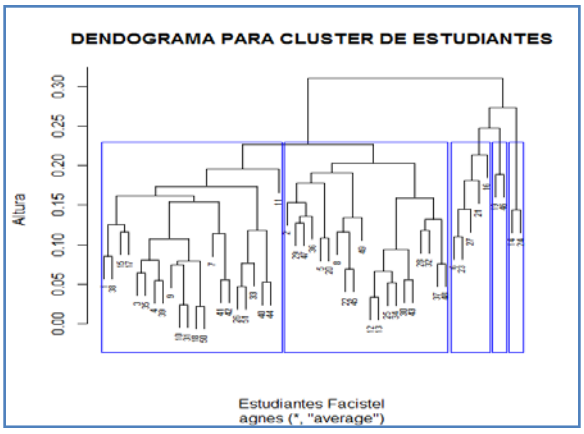

Figura 9. Dendograma obtenido para estudiantes del sistema de estudio Semestral.

3.2 Descripción del conocimiento.

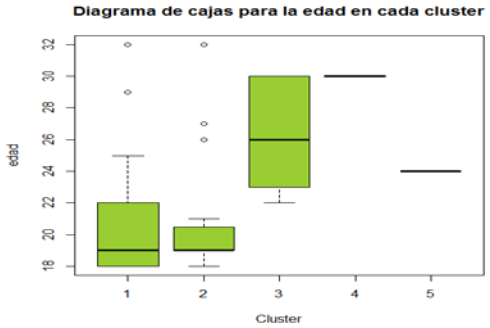

Figura 10. Gráficos de dispersión (Boxplot) para la variable edad en cada grupo identificado.

Para cada cluster formado se realizaron gráficos de dispersión (boxplot), cuando se trataba de variables cuantitativas y para variables cualitativas se utilizaron los diagramas de barras, gráficamente permite determinar el comportamiento de las variables en cada uno de los grupos; en la Figura 10 se encuentra un boxplot para los grupos formados en el sistema semestral para la variable edad.
Considerando sólo el comportamiento de la variable edad, se observa que el grupo 1 y 2 están formados por estudiantes jóvenes, en el grupo 3 en cambio la edad mínima es 22 y la máxima 30, el grupo 4 y 5 tiene pocos datos, el grupo 4 tiene gente más adulta que el grupo 5.

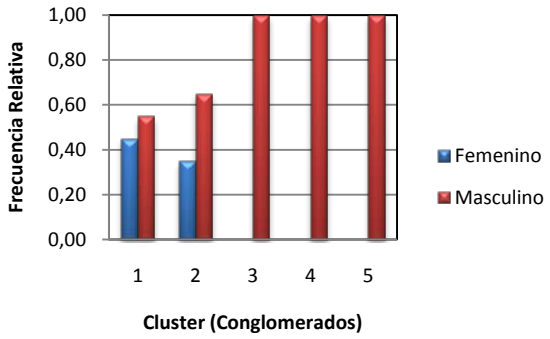

Figura. 11. Distribución de frecuencias del sexo de los estudiantes en sistema de estudio semestral por cada conglomerado.

En la Figura 11, observamos diagramas de barras para mostrar en forma gráfica el comportamiento de la variable "sexo" en cada cluster identificado, en los estudiantes de sistema semestral
Se observa que el grupo 1 y 2 están formados por estudiantes de sexo masculino y femenino, donde la mayor proporción son masculinos, el grupo 3, 4 y 5 en cambio son todos de género masculino.
Todas las variables participantes del análisis fueron analizadas de la misma manera, para obtener conclusiones sobre el comportamiento de cada variable en cada cluster formado, tanto en estudiantes con sistema de estudio anual, como en estudiantes con sistema de estudio semestral, los grupos obtenidos para los estudiantes de cada sistema de estudio se describen.

3.3 Conglomerados en el sistema de estudio anual.

Para esta interpretación la variable "indice_materias" se discretizó a una variable nominal tal como se indica en la tabla 5.

Tabla 5. Codificación de la variable "indice_materias" para sistema de estudio anual.

Indice_materias	Codificación	Categoría
[0-1]	1	Pésimo
(1-3]	2	Regular
(3-5]	4	Bueno
(5-6]	5	Muy bueno

Grupo 1: (46% de casos): Jóvenes tanto hombres como mujeres, solteros sin responsabilidades, la mayoría con el mínimo puntaje promedio para pasar el año, según la codificación del índice de materia se pueden considerar como "Buenos", con "Alto" recurso tecnológico a su disposición.
Grupo 2: (38% de casos).- Jóvenes más hombres que mujeres, estudiantes, solteros y sin responsabilidades, considerando el índice materias aprobadas pueden catalogarse como: "Buenos", con "aceptable" recurso tecnológico disponible, provenientes de colegios fiscales.
Grupo 3: (1% de casos).- Grupo formado por mayor cantidad de hombres que mujeres, adultos, casados y con hijos de rendimiento "Bueno", y con "bajo" acceso a recurso tecnológico.
Grupo 4: (6% de casos).- Estudiantes adultos, con responsabilidad de padres y conyugal, de rendimiento "Regulares" con "alto" recurso tecnológico disponible.
Grupos 5: (6% de casos).- Mujeres con responsabilidades, jóvenes, de rendimiento "Regulares", pero con "alto" recurso tecnológico disponible.
Grupo 6: (2% de casos).- Según las características de este grupo el perfil se etiquetaría como: Mujeres jóvenes con hijos, de rendimiento "Muy Bueno", con "alto" recurso tecnológico disponible, en este grupo se destaca el esfuerzo de las chicas que a pesar de tener responsabilidades en su familia se esfuerzan por mantener un rendimiento aceptable.
Grupo 7(1%).- Las variables de este grupo indican el siguiente perfil: Hombres la mitad soltero y la mitad en unión libre con rendimiento "Regular" con aceptable recurso tecnológico disponible.

3.4 Conglomerados en el sistema de estudio semestral.

Para esta interpretación la variable "indice_materias" se discretizó a una variable nominal tal como se indica en la Tabla 6.

Tabla 6. Codificación de la variable "indice_materias" para sistema estudio semestral

Indice_materias	Codificación	Categoría
[0-1]	1	Pésimo
(1-2]	2	Malo
(2-4]	3	Regular
(4-7]	4	Bueno
(7-12]	5	Muy bueno

Grupo 1 (43% de casos): Para las características de este grupo se identifica el siguiente perfil: Hombres y mujeres solteros, sin responsabilidad, cuyas edades indican que

tienen poco tiempo de graduados de bachilleres no alcanzan el mínimo requerido para aprobación de materias y por ser cinco el promedio de materias aprobadas por año, se etiquetan de acuerdo a la tabla 6 como "Buenos", tienen además aceptable acceso a recursos tecnológicos.

Grupo 2 (39% de casos): Hay más hombres que mujeres todos solteros, la mayoría sin hijos con estudiantes jóvenes, que si llegan al mínimo requerido en el promedio de universidad, donde el 50% de estudiantes tiene nueve materias aprobadas por materia, lo que permitiría etiquetar a este grupo de estudiantes como "Muy Buenos" según lo definido en tabla 6, además con un porcentaje "aceptable" de acceso a recurso tecnológico.

Grupo 3 (39% de casos): Estudiantes hombres casados, con edades adultas, la mayoría con hijos, la cantidad de materias aprobadas por año es 7 consideradas como "Buenos" y con un promedio universitario que no llega al mínimo requerido apenas es de 65 puntos, un porcentaje "bajo" tiene acceso a recursos tecnológicos.

Grupo 4 (4% de casos): Estudiantes hombres casados con hijos, mayores de edad, cuyo promedio de materias aprobadas por año es 1 considerados por lo tanto como "pésimos" y con alto porcentaje de estudiantes con recursos tecnológicos disponibles.

Grupo 5 (4% de casos): Estudiantes hombres, casados con hijos, adultos, con un promedio por año de materias aprobadas de 7 que nos permite etiquetar a los estudiantes como "Buenos", pero que no llega al mínimo requerido en el promedio universitario, con poco acceso a recursos tecnológicos disponibles.

4. Conclusiones y recomendaciones

Conclusiones

1. El lenguaje R, software libre, tiene diversas librerías y funciones para ejecutar análisis estadísticos y presentación de resultados gráficos, de fácil uso, con recurso disponible en web, es un paquete que no se ha utilizado hasta la actualidad en la UPSE.
2. El análisis cluster fue aplicado en datos de los estudiantes de la facultad de Sistemas y Telecomunicaciones, los patrones obtenidos reflejan solo a los estudiantes de esta facultad.
3. En el primer análisis no se pudo obtener una descripción muy acertada del rendimiento del estudiante universitario y su relación con el aspecto socio-económico por falta del más del 10% de datos.
4. Una descripción mejor de los conglomerados identificados se obtiene cuando se separa a estudiantes del sistema académico anual, con los del sistema de estudio semestral
5. Según la descripción del perfil de cada grupo identificado se infiere que el rendimiento de los estudiantes puede verse afectado positivamente al poseer recursos como portátiles o computadoras de escritorio y también poseer internet dentro de su hogar, especialmente ésta relación se observa en aquellos grupos donde las personas tienen menos edad, son solteros y sin responsabilidades.
6. Existe además en la Facultad de Sistemas y Telecomunicaciones, un grupo de personas que tienen recursos como computadora de escritorio o portátiles y que además tienen acceso a internet desde su hogar sin embargo son etiquetados como "pésimo", en este grupo no influye mucho el tener recursos tecnológicos a su disposición en su rendimiento, dado que se identifica que son personas casadas y con hijos, se considera a estos criterios como razones que afectan negativamente en el rendimiento académico de este grupo.
7. El rendimiento del estudiante no sólo está relacionado al desempeño y motivación del profesor en el aula, deberían ser consideradas otras variables como las de este estudio; estos resultados nos indican que el estudiante y su rendimiento puede verse afectado por los recursos que posee para su desempeño no sólo en el aula sino fuera de ella, y por el ambiente social en el que vive y se desenvuelve.
8. El proceso mostrado en este trabajo puede utilizarse como modelo para la aplicación de la técnica cluster con dos de sus métodos de clasificación: particional y jerárquico, en cualquier grupo de datos, donde se necesite extraer el conocimiento.

Recomendaciones

1. Realizar la réplica del análisis del perfil socio-económico aplicando la técnica de cluster para cada una de las facultades de la universidad.
2. Diseñar la aplicación que muestre resultados en tiempo real del análisis planteado en este trabajo.
3. Investigar librerías adicionales en el lenguaje R para mejorar la aplicación de las técnicas estadísticas y la presentación de resultados, y fomentar el uso de éste lenguaje para trabajos estadísticos y en la parte académica.
4. Los directivos de la Facultad de Sistemas y Telecomunicaciones, deberían considerar las diferentes características del recurso humano que están formando, concientizar en los estudiantes, la importancia de su preparación universitaria.

5. Referencias.

[1] Ayala, G. 2008. Análisis de datos con R para Ingeniería Informática. Departamento de Estadístico e Investigación Operativa. Universidad de Valencia. Disponible en: http://www.uv.es/ayala/docencia/ad/ad09.pdf

[2] Bernardis, A., Reeb, P and Bramardi, S. 2009. Agrupamiento de Pozos de Petróleo en Base a Datos de Perforación. Libro de Resúmenes y trabajos completos. Disponible en: http://gab.org.ar/GAB2009/resumenesytrabajos/Resumenes_GAB2009.pdf.

[3] Free Software Foundation's GNU. The R Project for Statistical Computing. Consultado en mayo del 2011. Disponible en: http://www.r-project.org/.

[4] Hernández, J, Ramírez, Ma. José and Ferri,C. 2004. Introducción a la minería de datos. Madrid: Pearson Educación.

[5] Hand, D., Mannila H. y Padharic S. 2001 Principles of Data Minning, Massachusets: Institute of technology.

[6] Myatt Glen J., Myatt, W.. *A.* 2009. Practical Guide to Data Viualization, Advanced Data Mining Methods, and Applications. Hoboken, NJ, USA: Wiley. Disponible en: http://site.ebrary.com/lib/upse/.

[7] PostgreSQL Global Development Group. PostgreSQL. Consultado en mayo del 2011. Disponible en: http://www.postgresql.org/about/.

[8] De La Cruz, M. 2012. "Minería de datos para detección de patrones de los estudiantes inscritos en carreras técnicas, en la Universidad Estatal Península de Santa Elena", Tesis Magister, Escuela Superior Politécnica del Litoral.