ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE INGENIERÍA EN ELECTRICIDAD Y COMPUTACIÓN

TESIS DE GRADO

Minería de datos aplicada a la detección de patrones de los estudiantes inscritos, en las carreras técnicas, en la Universidad Estatal Península de Santa Elena (UPSE)

PREVIA A LA OBTENCIÓN DEL TÍTULO DE:

MASTER EN SISTEMAS DE INFORMACIÓN GERENCIAL

PRESENTADA POR:

MARIUXI ALEXANDRA DE LA CRUZ DE LA CRUZ

GUAYAQUIL - ECUADOR

AÑO 2012

ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

TESIS DE GRADO

Minería de datos aplicada a la detección de patrones de los estudiantes inscritos, en las carreras técnicas, en la Universidad Estatal Península de Santa Elena (UPSE).

Previa a la obtención del Título de:

MASTER EN SISTEMAS DE INFORMACIÓN GERENCIAL

Presentada por:

Mariuxi Alexandra De La Cruz De La Cruz.

GUAYAQUIL-ECUADOR
AÑO 2012

DEDICATORIA

A mis padres, por ser el eje fundamental en todos los años de mi vida y a mis hermanos por estar siempre presente con su apoyo incondicional

AGRADECIMIENTO

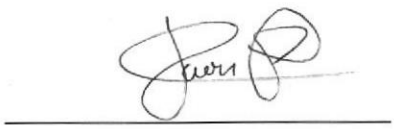
A Dios por darme vida para cumplir con uno más de mis propósitos

A mis padres y mis hermanos por estar siempre pendientes para darme su motivación, su apoyo y comprensión

A todas las personas amigos y compañeros que contribuyeron de alguna u otra manera para hacer posible este trabajo.

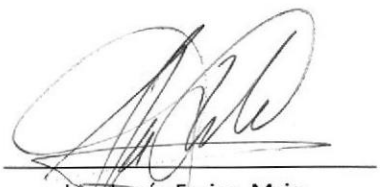
TRIBUNAL DE GRADO

Msc. Fabricio Echeverría
Miembro del tribunal



Msc. Juan García.
Miembro del Tribunal

Ing. Lenín Freire. Msig
Director de Tesis



Ing. Lenín Freire. Msig
Representante del Sub-Decano de la FIEC

RESUMEN

En el presente trabajo es aplicada la técnica de minería de datos conocida como: análisis cluster o de conglomerados, en datos sociales, económicos y académicos, de los estudiantes de la Universidad Estatal Península de Santa Elena, específicamente en aquellos inscritos en la Facultad de Sistemas y Telecomunicaciones, se buscará encontrar relación entre variables relacionadas al rendimiento académico y variables socio-económicas del estudiante.

Para garantizar la calidad del dato, proceso importante en minería, se planteó una propuesta de actualización de los datos sociales y económicos que hasta antes de este trabajo eran prácticamente inexistentes en esta institución de educación superior, proceso también descrito en los capítulos posteriores.

Para el almacenamiento de los datos es propuesto el uso del gestor de base de datos PostgreSql y para la aplicación de la técnica de minería, son propuestos algoritmos construidos en Lenguaje R (paquete estadístico de acceso libre).

Contenido

Índice de figuras

Índice de Tablas

INTRODUCCIÓN

En el presente trabajo se describe la aplicación del algoritmo de explotación de conocimiento conocido como "cluster", considerando datos históricos almacenados en los sistemas académicos y de bienestar universitario, que controla la Unidad de Producción de Informática de la Universidad Estatal Península de Santa Elena (UPSE).

En el capítulo I, se expone los conceptos teóricos de: dato, información y conocimiento; las fases de minería de datos, las características de la técnica cluster y además una breve descripción de las herramientas tecnológicas que se utilizan en el proyecto.

En el capítulo II, se describe el problema que llevó a proponer el tema, el objetivo general, los objetivos específicos, el alcance y la justificación del trabajo.

En el capítulo III, se explica la aplicación del proceso de minería, en los datos proporcionados, obteniendo un perfil académico de los alumnos inscritos desde el año 2001 hasta el 2011, en la facultad de Sistemas y Telecomunicaciones.

En el capítulo IV, se detalla las variables de la nueva ficha socio-económica, propuesta en este trabajo, que utilizará la UPSE (Universidad Estatal Península de Santa Elena), además la metodología implementada a partir del año 2011 para la recolección de los datos respectivos; en una muestra de datos de estudiantes de la Facultad de Sistemas y Telecomunicaciones, se explica la aplicación de las fases de la minería de datos para la obtención de un perfil académico, social y económico del estudiante.

Finalmente se detalla en el capítulo V, la interpretación, y validación de los resultados obtenidos en los capítulos III y IV, planteando sugerencias para mejorar la presentación de los resultados, además de las conclusiones y recomendaciones de este trabajo.

CAPÍTULO I

MARCO TEÓRICO

En este capítulo se expone definiciones técnicas sobre la diferencia entre dato, información y conocimiento, importancia de la explotación de los datos en la actualidad, y el aporte a la toma de decisiones en las empresas que manejan gran cantidad de datos.

Se realiza además, la descripción teórica de las fases de explotación de datos, especificando el algoritmo utilizado para la clasificación y agrupación de datos a partir de la técnica cluster y se enuncia el recurso tecnológico que se utilizará en las diferentes fases de minería de datos.

1.1 Dato, información y conocimiento

En la actualidad el progreso de la tecnología de información ha provocado que las instituciones públicas o privadas automaticen sus procesos, provocando el almacenamiento de una gran cantidad de datos, los mismos que manejados adecuadamente pueden generar información importante que se transformará finalmente en conocimiento, para la toma de decisiones en cualquier empresa, negocio e institución.

En las organizaciones donde la tecnología de información está presente, la información debería ser el resultado de la consolidación de los datos que se encuentran en los diversos sistemas de información.

Un sistema de información, es el conjunto de medios que permiten recolectar, clasificar, integrar, procesar, almacenar y difundir información interna y externa que la organización necesita para tomar decisiones en forma eficiente y eficaz (Meléndez par. 1).

Según Davenport y Prusak, las definiciones de dato, información, conocimiento e inteligencia son las siguientes:

> **Datos.-** son la mínima unidad semántica, y se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones. También se pueden ver como un conjunto discreto de valores, que no dicen nada sobre el por qué de las cosas y no son orientativos para la acción.
> Los datos describen únicamente una parte de lo que pasa en la realidad y no proporcionan juicios de valor o interpretaciones.

A pesar de todo, los datos son importantes para las organizaciones, ya que son la base para la creación de información.

Información.- Definida como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre. Los datos se pueden transforman en información añadiéndoles valor.

Conocimiento.- Se define como una mezcla de experiencia, valores, información y know-how que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción. Se origina y aplica en la mente de los conocedores. En las organizaciones con frecuencia no sólo se encuentra dentro de documentos o almacenes de datos, sino que también está en rutinas organizativas, procesos, prácticas, y normas.

Inteligencia.- Esta relacionada con la sabiduría que se obtiene cuando se ha procesado adecuadamente la información , tiene un conocimiento sustentando técnicamente. (30)

Figura 1. Pirámide del Conocimiento
Fuente: Hernández Orallo, José , Ma. José Ramírez Quintana, Cesar Ferri Ramírez. *Introducción a la Minería de datos* .España,2004

Según la fig. 1, la inteligencia se encuentra en la cúspide de la pirámide, y la base son los datos, normalmente se indica que los datos y la información, tienen relación con el proceso operacional de un negocio, empresa o institución, mientras que el conocimiento y la inteligencia, está relacionado con la aplicación de minería de datos.

1.2 Definición de minería de datos.

También llamada Data Mining (DM) es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Según Hand, 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 o dueño de la data. (2).

Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesables, implícito en las bases de datos.

1.3 Fases de la minería de datos

Las fases que involucra un proyecto de minería de datos son: planificación del proyecto, comprensión del negocio, filtrado de datos, selección de variables, transformación de los datos, extracción del conocimiento e interpretación y validación de los resultados tal como se puede apreciar gráficamente en la fig.2.

Figura 2. Ciclo de las Fases de Minería de Datos
Fuente: Hernández Orallo, José , Ma. José Ramírez Quintana, Cesar Ferri Ramírez. *Introducción a la Minería de datos.* España, 2004.

1.3.1 Comprensión del negocio y del problema que se quiere resolver.

La exploración de los datos empieza identificando una necesidad en el mercado que puede ser explotada.

1.3.2 Extracción y limpieza de datos

Este paso tiene mucha relación con trabajar con datos de calidad, para obtener una información confiable, los datos iniciales deben ser correctos, por lo tanto se debe en este paso analizar los datos con los que se cuenta, e identificar si existen datos anómalos o atípicos, teniendo que aplicar técnicas para tratarlos adecuadamente.

1.3.3 Transformación y selección de variables.

Dependiendo del algoritmo de minería de datos que se desea aplicar, se identifica las variables que deben considerarse para el modelo, analizar el trabajo con variables cualitativas, con variables cuantitativas, o con la combinación de ambos tipos de variables.

En este paso además se define la mejor transformación de las variables de acuerdo a las necesidades, generalmente se puede transformar variables cualitativas en variables nominales u ordinales.

1.3.4 Explotación de los datos.

Parte importante de la minería de datos, en este paso mediante una técnica se establece patrones de comportamiento observados en los valores de las variables del problema, pueden existir diversos modelos como: regresión, árboles de clasificación, cluster, redes neuronales, técnicas de inferencia estadística, aprendizaje bayesiano, inducción de reglas; se escoge el algoritmo dependiendo del análisis que desee ejecutarse con los datos.

1.3.5 Interpretación y validación.

Una vez que se ha realizado la extracción del conocimiento, el analista debe interpretar los resultados que determinan si es suficientemente fuerte el modelo para considerarlo

1.4.1.1 Medidas de disimilitud

Pérez César, expone la clasificación de Sneath y Sokal, sobre las cuatro medidas de similitud:

> ***Distancias:*** Se trata de las distintas medidas entre los puntos del espacio definido por los individuos. Se trata de medidas inversas de similitudes, es decir disimilitudes. El ejemplo más clásico es la distancia euclídea.
> ***Coeficientes de Asociación:*** se utilizan cuando trabajamos con datos cualitativos, aunque también se pueden aplicar a datos cuantitativos si se está dispuesto a sacrificar alguna información proporcionada por los individuos o las variables.
> ***Coeficientes angulares:*** se utilizan para medir la proporcionalidad e independencia entre los vectores que definen los individuos.
> ***Coeficientes de similitud probabilística***: miden la homogeneidad del sistema por particiones o subparticiones del conjunto de los individuos e incluye información estadística (610).

Existen según Myatt y Jhonn diversas metodologías para calcular las distancias entre las observaciones y dependiendo del tipo de variables con las que se trabajará, pueden ser: la distancia euclídea, manhattan, máxima distancia, jaccard, russell and rao, gower, mahalanobis, coeficiente de correlación, canberra (77-79).

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 fórmula (1) para dos observaciones p y q, sobre i variables:
>
> $$d(p,q) = \sqrt{\frac{\sum_{i=1}^{n} w_i d_i^2}{\sum_{i=1}^{n} w_i}} \quad (1)$$
>
> donde w_i es el peso para la i-ésima variable y 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 iésimo valor de las dos observaciones (pi y qi) donde la distancia se calcula con la fórmula. (2)

$$d_i = \frac{|p_i - q_i|}{R_i} \quad (2)$$

R_i : es el rango sobre todos los valores de la i-ésima variable. Para las variables categóricas d_i es cero, si el pi y qi son iguales, caso contrario es uno. (84)

1.4.1.2. Métodos utilizados para agrupación

Métodos Jerárquicos.- Un método jerárquico consiste en la construcción de diversos grupos o cluster sin conocer desde el inicio la cantidad de grupos que se formará, generalmente es presentado a través de un gráfico en forma de árbol denominado dendograma.

Los métodos jerárquicos se clasificarán en aglomerativos y divisivos; los primeros consideran tantos para grupos como individuos, se va fusionando sucesivamente hasta obtener los dos grupos más similares para llegar a una clasificación determinada, mientras que los divisivos parte inicialmente de un sólo grupo formado por todos los individuos y va sucesivamente identificando grupos más pequeños.

Hernández, Ramírez y Ferri indica que dependiendo de la forma de calcular las distancias de enlace entre los grupos se pueden distinguir tres algoritmos:

Enlace simple (single linkage).- donde se calcula la distancia entre todos los puntos de dos grupos y se toma como distancia entre grupos la menor.
Enlace completo (complete linkage).- igual que el anterior pero se toma la distancia entre grupos la mayor de todas
Enlace en la media (average linkage).- se toma como distancia la existente entre los representantes(centroides) de los grupos. (438)

válido, este puede ser replicado, siendo presentado a la gerencia y utilizado para análisis futuros.

1.4 Explotación de los datos.

1.4 .1 Minería de datos - Técnica Cluster

La minería de datos, es asociada con el análisis de los datos con la finalidad de obtener patrones en los datos que sean comprensibles para los tomadores de decisiones; para Hernández, Ramírez, y Ferri entre las tareas más importantes están:

> **Predictivas**, permiten estimar un valor a partir de otros, entre las que se puede mencionar la clasificación, la categorización, la regresión.
> **Descriptivas**, en este paso el objetivo no es predecir nuevos valores sino describir los datos existentes se menciona a: Agrupamiento (clustering), correlaciones, reglas de asociación, dependencias funcionales.
> Las tareas mencionadas requieren métodos, técnicas o algoritmos para resolverlas, como: las técnicas algebraicas y estadísticas, las técnicas bayesianas, técnicas basadas en árboles de decisión, algoritmo neuronales (137).

Una de las técnicas mencionadas y utilizadas en el presente trabajo es "cluster", que según lo expuesto por Myatt y Jhonson, tiene como finalidad esencial revelar concentraciones en los datos (casos o variables), pudiéndose utilizar variables cuantitativas y cualitativas (67).

Existen ciertos pasos que implica la técnica cluster:

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

Métodos no Jerárquicos.- También denominado particional se especifica al inicio el número de cluster en los que se dividirán las observaciones, método sensible a datos atípicos, el principal problema para su uso es elegir el adecuado número de cluster, y como alternativa se sugiere ejecutar varias veces el algoritmo con diferentes números de cluster, eligiendo el apropiado de acuerdo a la experiencia que tenga el analista sobre los datos.

Matt y Jonn indican que los algoritmos más usados para la aplicación de estos métodos son:

> **k-medias.-** algoritmo limitado para variables continuas, inicialmente se especifica la cantidad de grupos que se desea formar, se procede a calcular la media en cada grupo y luego todos las observaciones son reasignadas a los grupos de acuerdo a la distancia de cada observación con respecto a la media calculada, este proceso se realiza repetidamente hasta que las observaciones sean correctamente asignadas en cada cluster.
> **k-modas.-** Este algoritmo puede ser usado para variables categóricas y opera de forma similar que el algoritmo de la k-medias, la diferencia radica en no calcular la media de los datos sino calcular la moda de los datos.
> **K-medoides.-** No se calcula la media, se procede a elegir un valor representativo de cada grupo, cada una de las observaciones son reasignadas considerando que la distancia promedio del representante a cada uno de los otros objetos del grupos debe ser mínima. Por esta razón, el objeto representativo se llama medoide del grupo, también se los llama centrotipos. (98-102)

1.5 Tecnología de la Información

1.5.1 Herramientas para el almacenamiento y transformación de datos

PostgreSQL.- Es un gestor de base de datos de código abierto, ofrece control de concurrencia multi-versión, soporta casi toda la sintaxis SQL (incluyendo subconsultas, transacciones, tipos y funciones definidas por el usuario, contando también con un amplio conjunto de enlaces con lenguajes de programación (incluyendo C, C++, Java, perl, tcl y python), el software es de libre acceso y se encuentra disponible junto con una gran cantidad de recursos en la web www.postgresql.org.

1.5.2 Herramientas para la técnica de minería de datos

R.- Es un lenguaje de programación funcional, de uso público, tiene un modulo BASE, lo necesario para iniciar una sesión, el modulo BASE contiene las herramientas de programación, procedimientos estadísticos y gráficos frecuentemente utilizados.

El programa permite interactuar con el computador, pasos a paso, según los resultados que son mostrados en la consola, es decir que puede uno monitorear los procesos.

EL Programa R opera bajo diferentes plataformas: Windows, Linux, Unix y Mac, los recursos están disponibles para este software en www.r-project.org.

Se utilizará el software "R" que tiene opciones y módulos que permiten conectarse con PostgreSQL, y tiene capacidad de aplicar las técnicas de minería de datos.

Según lo expuesto en el trabajo realizado por Bernardis, Reeb y Bramard:

> El lenguaje R ofrece la librería "cluster" en R para operar el análisis de agrupamientos con el mismo nombre, pone a disposición cuatro programas de agrupamiento, dos particionales ("pam" y "fanny") y dos jerárquicos ("agnes" y "diana").
> El algoritmo "pam" hace agrupamientos alrededor de medoides y "fanny" hace agrupamientos "difusos" atribuyendo a cada objeto un grado de pertenencia a cada grupo.
> El algoritmo "agnes" hace agrupamientos jerárquicos aglomerativos y "diana", divisivos (82).

1.5.3 Herramientas para diseño de encuesta

LimeSurvey.- (anteriormente PHPSurveyor) Es una aplicación open source, es de libre acceso en la dirección www.limesurvey.org; aplicación de encuestas en línea, escrita en PHP y que utiliza bases de datos MySQL, PostgreSQL o MSSQL. Esta utilidad brinda la posibilidad a usuarios sin conocimientos de programación, desarrollo, publicación y recolección de respuestas de sus encuestas.

Las encuestas incluyen ramificación a partir de condiciones, plantillas y diseño personalizado usando un sistema de plantillas web, y provee utilidades básicas de análisis estadístico para el tratamiento de los resultados obtenidos. Los resultados pueden ser anónimos, separando los datos de los participantes de los datos que proporcionan, inclusive en encuestas controladas.

1.6 Aplicación de la minería de datos

Según Prabhu, la minería de datos actualmente se la está aplicando en diversos campos en el mundo entero debido al crecimiento acelerado de la información y a la necesidad de aprovechar este recurso, el mismo autor presenta una clasificación tomada de la IBM-2003, acerca de las industrias donde es utilizada y es reflejada en la figura 3.

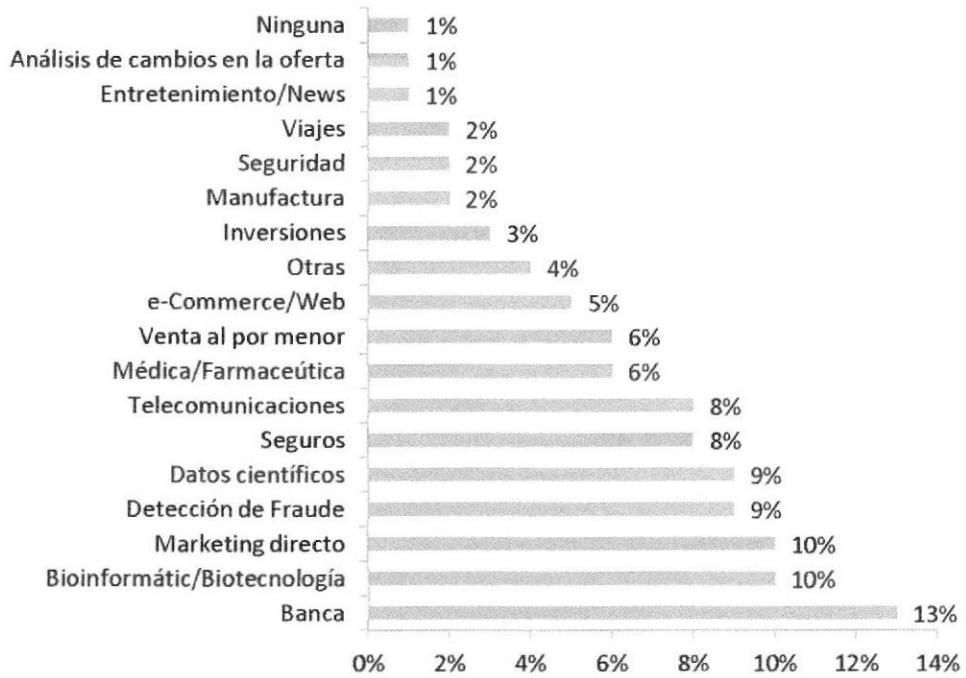


Figura 3 Aplicaciones industrial de la Minería de datos.
Fuente: IBM, clasificación IBM, 2003

El sector bancario es según esta clasificación el área donde más se utiliza la minería de datos, la misma es aplicada como técnica para predecir la reacción de los clientes, definir cuando los clientes presentan un alto riesgo de tener préstamos impagos, predecir clientes que pueden requerir los diferentes tipos de préstamos, detectar actividades fraudulentas en créditos y transacciones.

La bioinformática y biotecnología ocupa el segundo puesto, según esta clasificación y es un área de investigación en desarrollo donde se combina la biología y la tecnología de la información, las aplicaciones de minería de datos en estas áreas puede ser la predicción de estructura de varias proteínas, mapeos de la estructura del ADN.

El primer campo donde se usó la minería de datos es el área de mercadeo, que en esta clasificación se ubica en el tercer lugar, aplicando la minería de datos para identificar la segmentación de los clientes, decidir la promoción y ubicación de los productos dentro de una tienda y comparar con los resultados entre tiendas grupos de compradores de acuerdo a ciertas características.

1.6.1 Minería de datos en la educación superior

Según un artículo publicado por Luan J. de IBM Corporation, uno de los grandes retos a los que se enfrenta la educación superior hoy en día es el pronóstico de las trayectorias individuales de los estudiantes y de los antiguos alumnos. (1)

El artículo indica que las autoridades de las instituciones de educación superior deberían preocuparse de lo que puede suceder con el estudiante en el paso por la universidad, si será un estudiante de buen rendimiento, si necesitará apoyo académico o de orientación, si los estudiantes necesitarán de módulos extras por las falencias que traen. (2)

El trabajo de Luan expresa que se puede aplicar técnicas de minería también en las instituciones de educación superior para descubrir información de sus estudiantes inscritos, por ejemplo con la técnica de clasificación se podría obtener un análisis completo de las características de los estudiantes; o la estimación para predecir la probabilidad de una variedad de resultados, la persistencia, o el éxito de un estudiante en un determinado curso. (4)

Teniendo presente este artículo y analizando ejemplos de diversas organizaciones donde se utiliza la minería de datos para obtener resultados que mejoran la toma de decisiones, se plantea en el presente trabajo, la aplicación de esta técnica en las bases de datos que posee la institución de educación superior identificada como Universidad Península de Santa Elena (UPSE), para realizar el presente trabajo se cuenta con el apoyo de las autoridades de Dirección de Planeamiento, Dirección de Bienestar Universitario, Decanato de la Facultad de Sistemas y Telecomunicaciones, y Vicerrectorado Académico.

CAPÍTULO II

DESCRIPCIÓN DE LA SITUACIÓN ACTUAL

En este capítulo se encuentra la descripción de los sistemas informáticos que maneja la Universidad Estatal Península de Santa Elena, exponiendo el requerimiento de reportes estadísticos básicos, más aún se nota la ausencia de reportes estadísticos donde una de las prioridades principales sea la combinación de datos de los diversos sistemas existentes. Siendo éste motivo, la principal razón para el planteamiento de nuestro trabajo, se cita por lo tanto el objetivo general, los objetivos específicos, el alcance y la justificación del presente proyecto.

2.1 Descripción de la situación de la Universidad Estatal Península de Santa Elena y de los sistemas informáticos

El Plan Estratégico de desarrollo institucional, actualizado en el 2010, indica:

> La Universidad Península de Santa Elena (UPSE) es una institución pública de educación superior, que comenzó su funcionamiento con la promulgación de la ley el 22 de julio de 1998, las actividades académicas la empezó con un total de 716 alumnos
> La UPSE al 2010 estaba conformada por ocho facultades, 15 escuelas y 27 carreras; tiene dos modalidades de estudio: Modalidad Presencial, con clases de lunes a viernes en horario diurno y nocturno y la Modalidad Semi-

presencial, con clases de sábados y domingos en horario diurno; en este mismo año se matricularon 7145 estudiantes que representan el 13,65% de aumento con respecto al número de matriculados en el 2009 (3).

Como toda institución de educación superior, la UPSE necesitó automatizar los procesos académicos y administrativos, por lo tanto, en abril del 2005 fue creada la Unidad de producción de la Escuela de Informática (UPEI), cuyo objetivo es desarrollar y automatizar los procesos de la universidad.

Al año 2011, cuenta en la parte académica, con los siguientes sistemas desarrollados e implementados:

- Sistema de registro académico.
- Sistema de ingreso de calificaciones.
- Sistema de bienestar estudiantil.
- Sistema de control para ingreso de egresados y titulados.
- Sistema de control docente.

Sistema de registro académico

Permite el mantenimiento de los datos del estudiante, el registro de admisión, el registro del pre-universitario, el registro de matriculación.

Actualmente el mantenimiento de los datos del estudiante, especialmente de los nuevos, es responsabilidad de las asistentes administrativas de cada carrera, en este año el

registro al preuniversitario y la matriculación se realizaron en línea, los estudiantes se registraron directamente accediendo a la página web de la institución.

Sistema de ingreso de calificaciones

Presenta el módulo de generación de actas, ingreso de actas de calificaciones, mantenimiento y bloqueo de actas, el proceso del ingreso de calificaciones en línea se está efectuando desde julio del 2011.

Sistema del departamento de bienestar estudiantil

Su función es el ingreso y mantenimiento de la ficha de bienestar estudiantil, hasta el 2010, el mantenimiento y actualización de estos datos era responsabilidad de las asistentes administrativas de cada una de las carreras.

Sistema de control para ingreso de egresados y titulados

Presenta el registro de egresados, registro de temas de tesis y seminarios, registro de los graduados por proyectos de tesis y/o seminarios, promedios finales de los graduados, la actualización y mantenimiento de los datos que se requiere para alimentar este sistema es de las asistentes administrativas de las diferentes carreras.

Sistema de control docente

Mantenimiento de los datos del docente, distributivo de carga académica, planificación de horarios, control de asistencia, falta de docentes, y evaluación docente; la

actualización y mantenimiento de estos módulos están bajo responsabilidad de las asistentes administrativas o de los directores de cada carrera.

De los sistemas antes nombrados los dos primeros son actualizados automáticamente dado que la matriculación y el registro de calificaciones son ejecutados en línea, sin embargo los otros tres sistemas tienen un lento proceso de actualización y almacenamiento de los datos, debido a la falta de tiempo para los últimos procesos.

Los cinco sistemas tienen módulos donde presentan reportes estadísticos a nivel básico (obtener listas, conteos, sumatorias, promedios, porcentajes), en estos reportes se consideraron las peticiones de informes que generan las autoridades académicas, se observa entonces que una gran cantidad de datos que se posee en cada uno de los módulos no es aprovechada para ser convertida en información relevante que ayude a tomar decisiones sobre diversos aspectos tanto académicos como administrativos.

Si todos estos datos se analizaran de una forma más profunda utilizando análisis estadístico univariado, multivariado, algoritmos exploratorios o de predicción, se pueden obtener más información sobre los procesos y mejorar las decisiones de las autoridades académicas y administrativas.

En este trabajo se propone el uso de la minería de datos y la técnica de conglomerados o "cluster", para la aplicación de esta técnica se utilizará datos del área académica, de aspectos sociales y económicos de los estudiantes, se busca agrupar o clasificar a los estudiantes en grupos característicos, con la finalidad de describir su perfil.

El modelo de aplicación de la técnica cluster en un grupo de datos, será una propuesta para las autoridades académicas que ejemplificará la obtención no sólo de estadísticas básicas sino de información que permitan tomar mejores decisiones en el aspecto académico y sobre el rendimiento estudiantil.

El análisis inicial se realizará para dos carreras consideradas como técnicas en la UPSE, siendo estas: la carrera de Ingeniería en Sistemas y la carrera de Electrónica y Telecomunicaciones, ambas pertenecientes a la Facultad de Sistemas y Telecomunicaciones.

2.2 Alcance del proyecto.

Se plantea una mejora en la calidad y la pertinencia de los datos, especialmente de aquellos que recopila el departamento de bienestar estudiantil, mediante un análisis de la ficha que hasta el 2010 se utilizó, proponiendo una nueva ficha para los siguientes

años y finalmente actualizar la información socio-económica de los estudiantes que al presente año están inscritos en la UPSE.

Concientizar a las autoridades académicas sobre la importancia de poseer datos que sean pertinentes y actualizados para obtener reportes confiables a nivel gerencial.

Proponer el uso de modelos de minería de datos, utilizando para el manejo, tratamiento, y análisis de los datos herramientas tecnológicas de libre acceso, lograr con estas técnicas la caracterización del estudiante universitario, difundir el uso de estas herramientas para que las autoridades tomen mejores decisiones sobre aspectos académicos y administrativos, este proyecto será el ejemplo de muchas otras investigaciones que se pueden realizar a nivel de explotación de información en la Universidad Estatal Península de Santa Elena, y siendo el punto de referencia para proponer análisis similares para las instituciones públicas y privadas a nivel peninsular.

2.3 Objetivo general.

Identificar patrones sociales, económicos y académicos de los estudiantes inscritos, en las carreras técnicas, en la Universidad Estatal Península de Santa Elena (UPSE) utilizando técnicas de minería de datos.

2.4 Objetivos específicos.

- Analizar la calidad de los datos existentes en la base de datos académicos y en la base de datos del departamento de bienestar estudiantil.
- Mejorar la ficha de recolección de datos del departamento de bienestar estudiantil.
- Actualizar la base de datos del departamento de bienestar estudiantil, aplicando un censo estudiantil, que permita tener información confiable y oportuna.
- Analizar modelos descriptivos de minería de datos para encontrar patrones del estudiante en el proceso académico, utilizando los datos que se encuentran registrados en el sistema académico de esta institución, complementados con los datos del censo estudiantil.
- Concientizar a las autoridades académicas sobre la necesidad de tener datos completos y actualizados que puedan generar información, considerando la aplicación de herramientas relacionadas a la minería de datos.
- Poseer un proyecto base para la generación de investigaciones similares a nivel institucional y provincial.

2.5 Hipótesis.

El tratamiento adecuado de los datos y la explotación de los mismos utilizando técnicas de minería, permitirán obtener un perfil académico, social y económico del recurso

humano que está formando la UPSE, contribuyendo al planteamiento planificado de propuestas de mejoras en los servicios y en la formación que se brinda a los estudiantes.

2.6 Operacionalización de variables.

Variable dependiente: Construcción del perfil académico, social, y económico del estudiante inscrito en UPSE.

Variable independiente: Tratamiento y explotación de datos con el uso de técnicas de minería de datos.

2.7 Justificación.

En los sistemas informáticos que son manejados en la UPSE, existen los módulos para reportes estadísticos básicos (conteo, promedios, listas, porcentajes), para las autoridades académicas universitarias, no existen requerimientos de informes estadísticos donde se deba aplicar más que el análisis univariado básico.

Sin embargo muchas veces las autoridades del área académica, especialmente los representantes de direcciones de carreras técnicas se plantean ciertas interrogantes como:¿Qué factores podrían influir en el rendimiento de los estudiantes?, ¿Cuáles podrían ser las razones de la deserción existente en las carreras? o ¿Existe o no relación entre el uso de las herramientas tecnológicas y el rendimiento académico de los estudiantes? .

Consideramos que algunas de las respuestas a las interrogantes planteadas se generan analizando y enlazando los datos almacenados en el sistema de registro de notas, en el sistema de matriculación y en el sistema de bienestar universitario; en el presente trabajo se propone la aplicación del proceso de minería de datos como forma de explotación de información y la técnica de clasificación denominada "cluster o conglomerados", sobre los datos académicos, sociales y económicos de los estudiantes.

El sistema de matriculación y el de registro de calificaciones desde el 2010 son actualizados automáticamente, el primero por los estudiantes quienes se registran en línea y el segundo por los docentes, quienes también ingresan las calificaciones vía web, sin embargo al analizar el sistema de bienestar universitario, se observó que los datos almacenados no estaban actualizados, en la ficha socio-económica no existían campos donde se requieran los datos sociales y económicos de los estudiantes.

Frente a la necesidad de tener las mencionadas variables y los datos respectivos, proponemos una nueva ficha-socioeconómica cuyos datos serán ingresados por los estudiantes a través de la página web, para realizar este trabajo se tiene el respaldo y autorización del Vicerrectorado Académico, de las autoridades académicas de la Facultad de Sistemas y Telecomunicaciones, de Dirección de Planificación Universitaria y del Departamento de Bienestar Universitario, la propuesta inicial de aplicación del

proceso de minería de datos se lo ejecutará con datos de la Facultad de Sistemas y Telecomunicaciones.

El modelo de minería de datos podrá ser replicado, para obtener el perfil social, económico y académico de los estudiantes inscritos en las ocho facultades de la Universidad Estatal Península de Santa Elena.

CAPÍTULO III

APLICACIÓN DE MINERÍA CON DATOS HISTÓRICOS DE UPSE.

Se describe en este capítulo las herramientas utilizadas en las diferentes fases de minería de datos, la integración, el análisis de calidad de los datos y la aplicación de la técnica denominada análisis de conglomerados o cluster, que nos permitirá agrupar o clasificar a los estudiantes en grupos característicos, el análisis inicial y expuesto en este capítulo considera datos generales y académicos de los estudiantes desde el año 2001.

3.1 Descripción de las herramientas utilizadas

Los sistemas hasta ahora diseñados en UPSE son realizados en software comerciales, la minería de datos no es aplicada aún en la universidad.

La Ley Orgánica de Educación Superior (2010) en el artículo 32 indica sobre el uso de programas informáticos: "Las instituciones de educación superior obligatoriamente

incorporarán el uso de programas informáticos con software libre", debido a esta obligatoriedad este trabajo fue realizado en herramientas de acceso gratuito.

Para el almacenamiento de los datos se decidió por el programa de libre acceso, Postgresql, que tiene una interfaz amigable y cuyo lenguaje utilizado es el SQL (Lenguaje Estructurado de Consultas), cuenta con una gran cantidad de recurso disponible en la web.

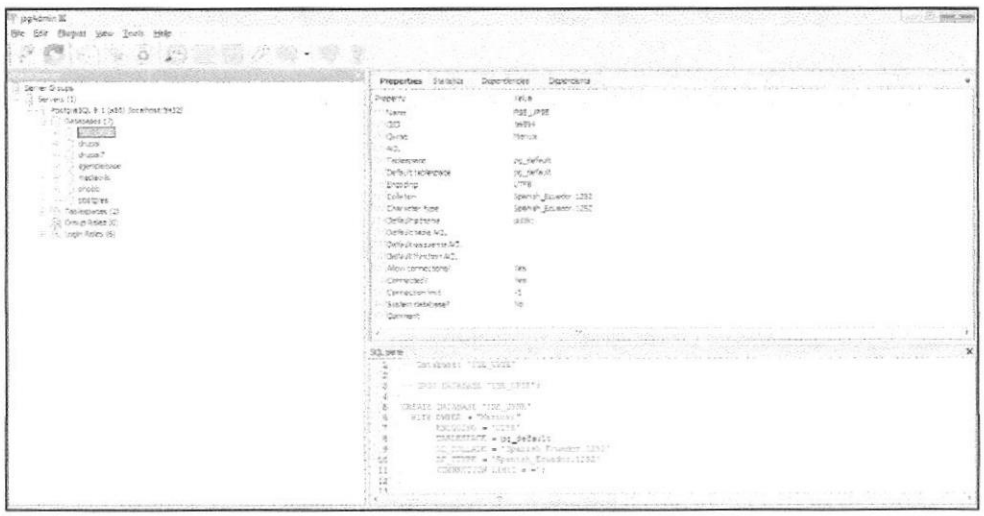


Figura 4. Entorno de Trabajo de PostgreSql-PgAdmin.

Para aplicar la minería de datos existen diversos sistemas y herramientas como: SPSS Clementine, Weka, Kepler, Darwin, DBMiner, Yale, DB2, Intelligent Miner, STATISTICA Data Miner, R, algunos de estas herramientas son software comerciales y otros de código libre.

Para este trabajo se eligió el programa denominado "R", por ser un software estadístico de código libre, muy funcional y que opera sobre diferentes plataformas; el lenguaje "R" poseía un editor de texto de acceso gratuito denominado "TINN-R", también fue utilizado para el manejo de instrucciones de R.

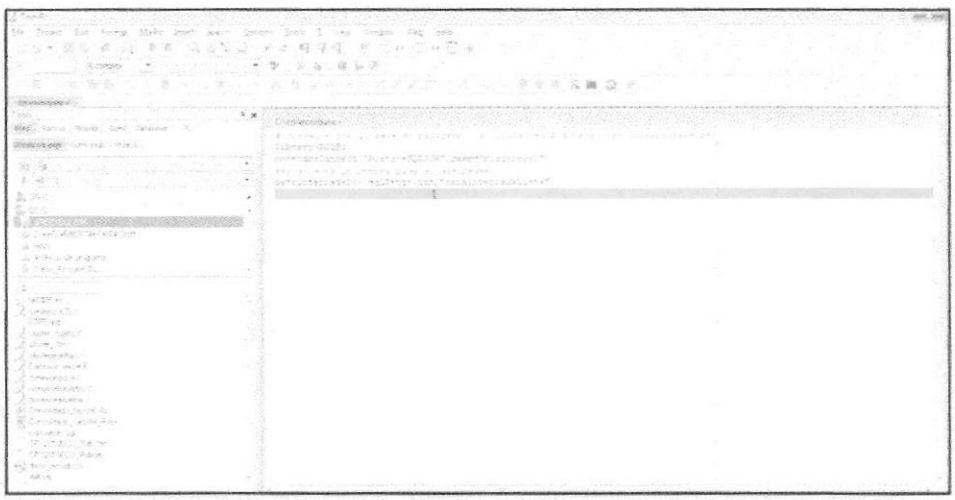


Figura 5. Entorno de Trabajo en Tinn-R

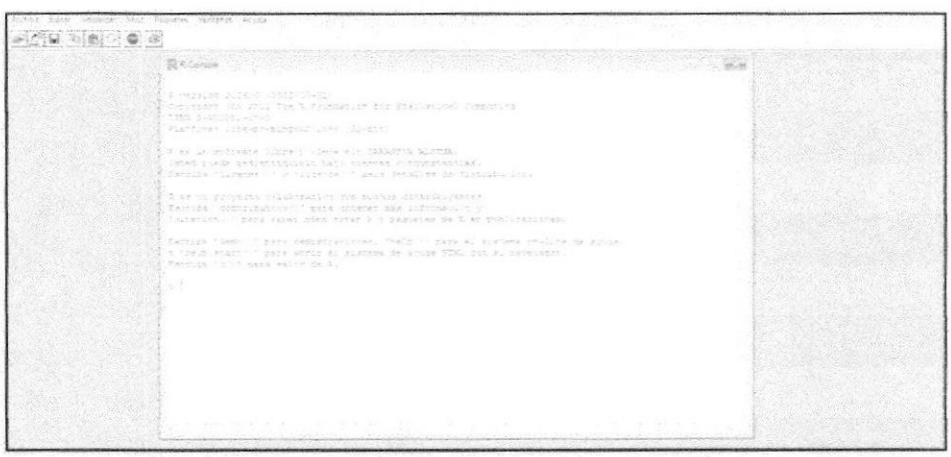


Figura 6. Entorno de trabajo de Projet R

3.2 Fases de la minería de datos

Las fases de minería de datos aplicadas:

- ❖ Preparación de los datos: Recopilación de datos.
- ❖ Transformación, limpieza de los datos, y selección.
- ❖ Explotación de los datos.
- ❖ Interpretación.

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), proporcionó los datos históricos para proceder a realizar el análisis estos datos que fueron entregados en formato Excel.

En el entorno de Postgresql, se ejecutaron los siguientes pasos:

- 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).

En el entorno de TINN-R fueron ejecutados los siguientes pasos:

- 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.

En el entorno R fueron ejecutadas todas las instrucciones que se diseñaron en el editor TINN-R.

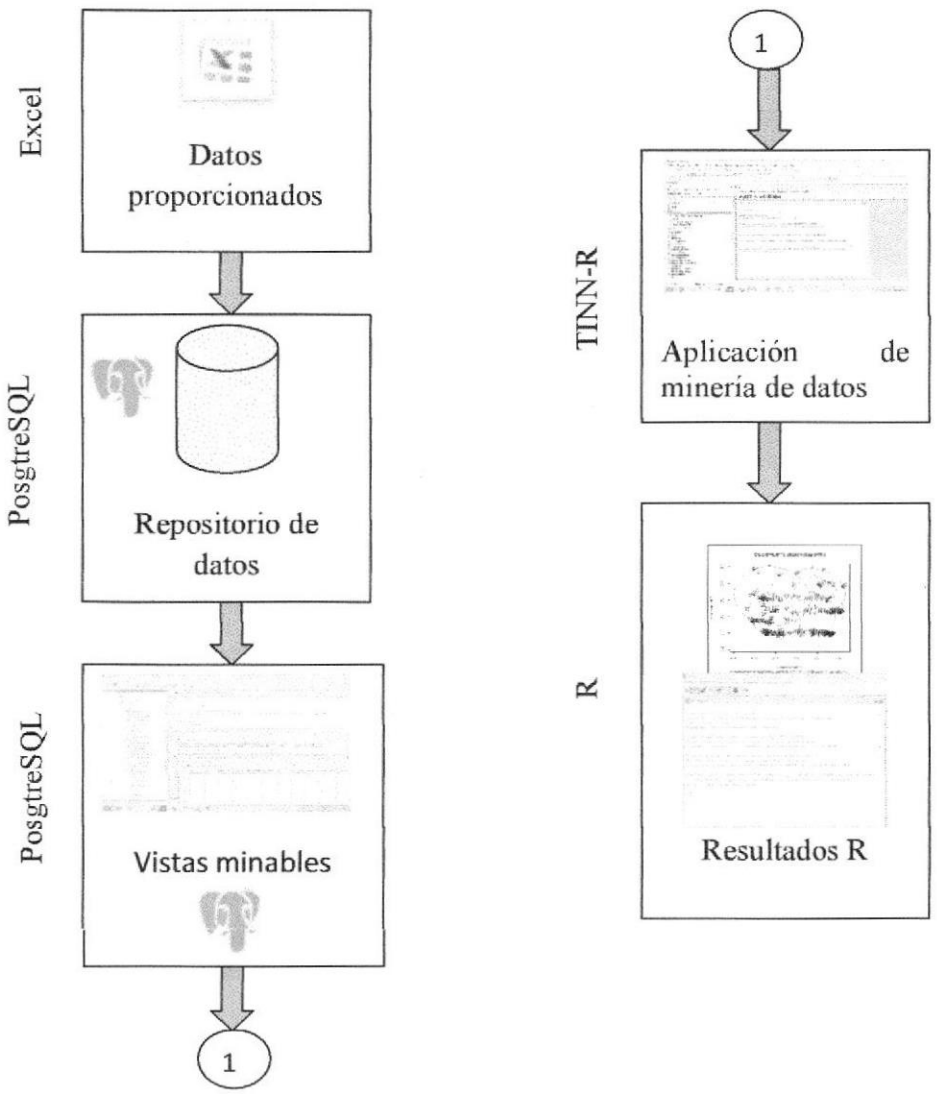


Figura 7. Proceso desde recopilación hasta resultados de minería en cada herramienta tecnológica

En la fig. 7, se indica los pasos generales que se siguió para aplicar la minería de datos en este trabajo, se explica a continuación en qué consiste cada fase.

3.2.1 Recopilación de los datos

La Unidad de Producción de la Escuela de Informática (UPEI), tiene cinco sistemas, para el análisis que se ejecutará solo fueron considerados los siguientes sistemas:

- Sistema de Bienestar Estudiantil.
- Sistema de Ingreso de Calificaciones.

Del primero se requería datos de identificación del estudiante y de su perfil socioeconómico, del segundo las calificaciones obtenidas desde su registro como estudiante universitario.

Los datos fueron proporcionados, por la UPEI, en formato Excel; se obtuvo acceso a la tabla ***"Notas_Facistel_UPSE"***, con campos que fueron obtenidos del ***Sistema de Ingreso de Calificaciones*** y son: carrera, período, sistema de estudio, nivel, estudiante, matrícula, materia, promedio; estos campos permitirían obtener información referente al promedio y rendimiento académico del estudiante inscrito desde el 2001.

Para el análisis son considerados los datos de estudiantes inscritos en las carreras de la Facultad de Sistemas, que fueron un total de 22.779 registros (se especifican la nota obtenida en cada materia en la cual se encontraban registrados los estudiantes), se

necesitó crear vistas donde se especificarían ciertos campos calculados que necesitaría este análisis.

La otra tabla que fue proporcionada fue ***"Datos_Generales"*** con los campos : matrícula, facultad, escuela, carrera, identificación, nombre, sexo, fec_nacimiento, lugar_nacimiento,nacionalidad, provincia_nacimiento, cantón_nacimiento, parroquia_nacimiento, provincia_reside, cantón_reside, parroquia_reside, barrio_reside, edad, libreta _militar, colegio, especialidad, año de graduación, calificación, domicilio, trabaja, teléfono, dirección, fecha de matriculación, sistema de estudio, modalidad, período, promedio.

Se tenía los datos antes mencionados de todos los estudiantes inscritos en la universidad desde el año 2001, sin embargo el análisis se lo realizaría solo para la Facultad de Sistemas y Telecomunicaciones, por lo tanto teníamos un total de ***5423 registros*** pertenecientes a los estudiantes de las dos carreras de esta facultad.

Los datos proporcionados en formato Excel fueron transformados a archivos con formato .csv para ser cargados en el gestor de base de datos PostgreSQL, previamente fueron creadas las estructuras de las dos tablas que se almacenaría.

En la fig.8, se muestra las instrucciones para la creación de estructura de una tabla ejemplo y para cargar la mencionada tabla en el gestor de base de datos PostgreSQL.

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

3.2.2 Transformación, limpieza de los datos

3.2.2.1 Creación de atributos

En un primer momento fueron proporcionadas solo dos tablas, aquella con los datos generales del estudiante, y aquella con las notas promedios obtenidas en cada materia, tal como fueron proporcionadas no se las podía utilizar directamente, sin embargo algunas de las variables de las tablas permitían obtener atributos derivados que daba más información del estudiante.

Las instrucciones para el tratamiento de las variables y de los datos fueron ejecutadas en PostgreSQL, se generaron las siguientes "vistas" para aumentar atributos:

Vista Materias Aprobadas

Con los siguientes campos tomados directamente de la tabla "Notas_Facistel_Upse"

- Matrícula: Campo que identifica al estudiante
- Estudiante: Campo donde se determina el nombre del estudiante
- Sistema de Estudiante: Campo donde se determina el sistema de estudio en el que se encuentra inscrito el estudiante.

Campo calculado considerando el campo "materia" y "promedios" de la tabla "Notas_Facistel_upse".

- Cantidad de materias aprobadas: cantidad de materias con promedios superiores o iguales a 70 puntos por cada estudiante.

Vista Materias Reprobadas

Con los siguientes campos tomados directamente de la tabla "Notas_Facistel_Upse"

- Matrícula: Campo que identifica al estudiante.
- Estudiante: Campo donde se determina el nombre del estudiante.
- Sistema de Estudiante: Campo donde se determina el sistema de estudio en el que se encuentra inscrito el estudiante.

Campo calculado considerando el campo "materia" y "promedios" de la tabla "Notas_Facistel_upse".

- Cantidad de materias reprobadas.- Cantidad de materias que tenían promedios inferiores a 70 puntos por cada estudiante.

Vista con el promedio por estudiante

Con los siguientes campos tomados directamente de la tabla "Notas_Facistel_Upse"

- Matrícula: Campo que identifica al estudiante.
- Estudiante: Campo donde se determina el nombre del estudiante.
- Sistema de Estudiante: Campo donde se determina el sistema de estudio en el que se encuentra inscrito el estudiante.

Campo calculado considerando el campo "materia" y "promedios" de la tabla "Notas_Facistel_upse"

Promedio de la calificación.- Promedio de notas obtenidas en todas las materias para cada uno de los estudiantes.

Creación de vistas materias aprobadas, materias reprobadas, promedio de cada estudiante.

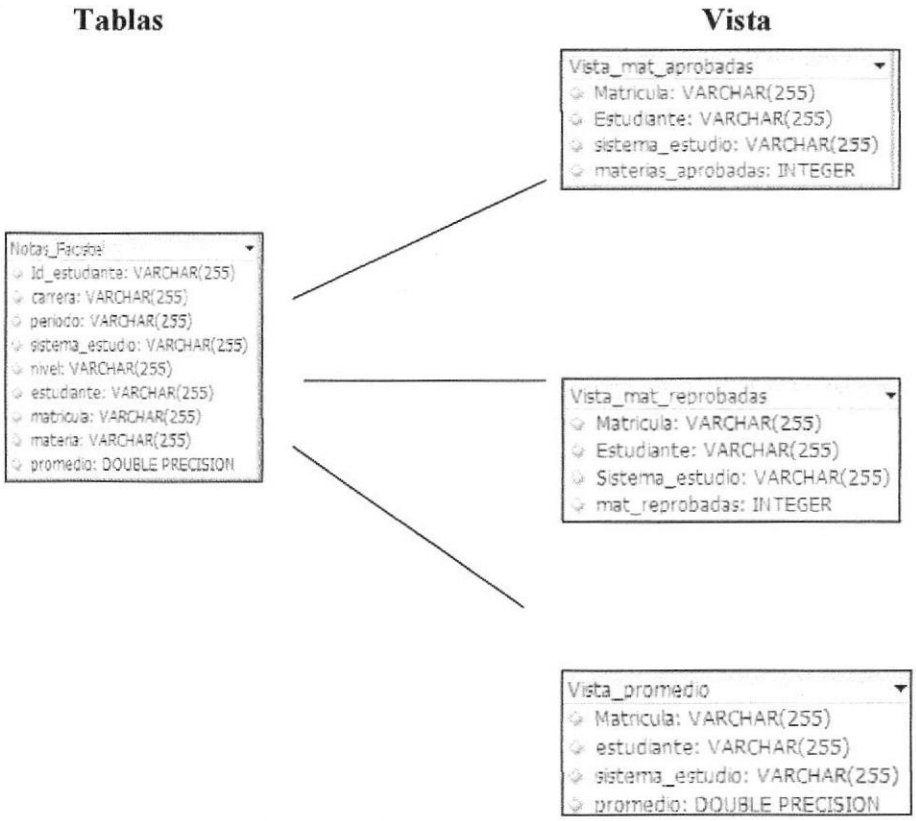


Figura 9. Vistas obtenidas con atributos derivados

Vista tiempo universidad

De la tabla ***"Upse_Dat_Estudi"***, y del campo fecha de matrícula fue determinada la primera fecha de matrícula de cada estudiante, y la última fecha de matrícula, la diferencia de estos dos campos permitió obtener el campo cantidad de tiempo en la universidad.

Vista datos generales.

De la tabla ***"Upse_Dat_Estudi"*** se creó la vista con campos donde se establecía datos históricos del estudiante: nombre, edad, sistema_estudio, sexo, trabaja, colegio, especialidad, calificación.

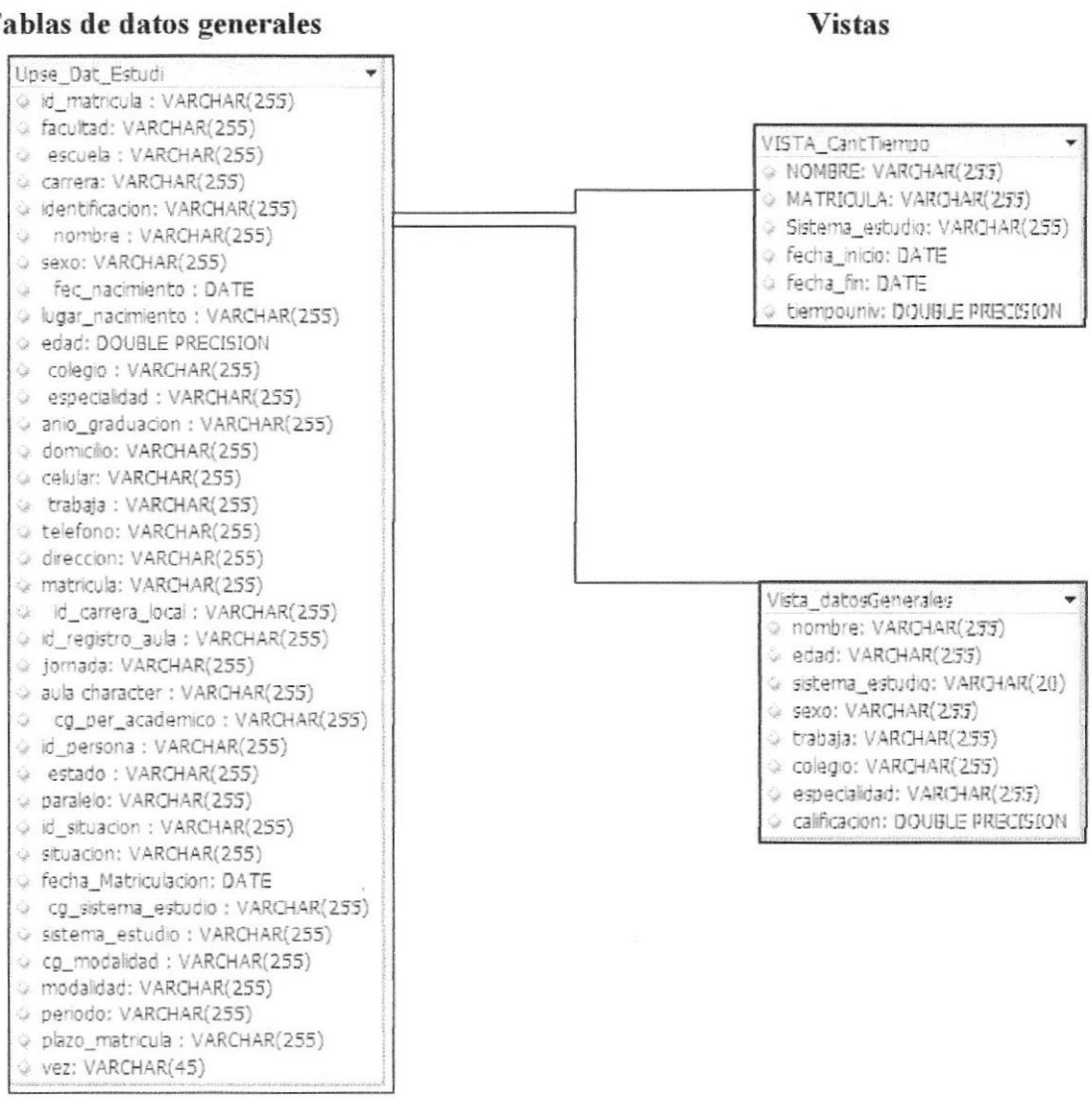


Figura 10. Vista que identifica datos generales de estudiantes.

Tabla 1. Vistas con atributos derivados

Tabla	Vistas	Atributos	Tipo de atributo	Descripción
Notas facistel Upse	Vista_mat_aprobada	Matrícula	Directo	El número de matrícula de cada estudiante
		Estudiante	Directo	Nombres y apellidos de los estudiantes
		Sistema_estudio	Directo	Sistema de estudio del estudiante
		cantidad_materias_aprobadas	Derivado	Cantidad de materias cuyo promedio es >=70 por cada estudiante
	Vista_mat_reprobada	Matrícula	Directo	El número de matrícula de cada estudiante
		Estudiante	Directo	Nombres y apellidos de los estudiantes
		Sistema_estudio	Directo	Sistema de estudio del estudiante
		cantidad_materias_reprobadas	Derivado	Cantidad de materias cuyo promedio es< 70 por cada estudiante
	Vista_promedio	Matrícula	Directo	El número de matrícula de cada estudiante
		Estudiante	Directo	Nombres y apellidos de los estudiantes
		Sistema_estudio	Directo	Sistema de estudio del estudiante
		Promedio	Derivado	Promedio de calificación en la universidad por estudiante

Tabla 2. Vistas y atributos generados desde la tabla "Upse_dat_estudi".

Tabla	Vistas	Atributos	Tipo de atributo	Descripción
Upse_Dat_Estudi"	Vista_cant_tiempo	Nombre	Directo	Nombres y apellidos de los estudiantes
		Matrícula	Directo	El número de matrícula de cada estudiante
		sistema_estudio	Directo	Sistema de estudio del estudiante
		fecha_inicio	Derivado	Primera fecha en que se registró el estudiante
		fecha_fin	Derivado	Última fecha en que se registró el estudiante
		tiempo_univ	Derivado	fecha_fin - fecha_inicio (diferencia entre las dos fechas)
Upse_Dat_Estudi"	Vista_datos_generales	Nombre	Directo	Nombres y apellidos de los estudiantes
		Edad	Directo	Edad de cada estudiante
		Sistema_estudio	Directo	Sistema de estudio del estudiante
		Sexo	Directo	Sexo del estudiante
		Trabaja	Directo	Define si el estudiante trabaja o no
		Colegio	Directo	Nombre del colegio donde se graduó el estudiante
		especialidad	Directo	Nombre de la especialización en al que se graduó el estudiante
		calificación	Directo	Calificación con la que se graduó de bachiller el estudiante

3.2.2.2 Integración de atributos.

Una vez creadas las vistas: vista_mat_aprobadas, vista_mat_reprobadas, vista_promedio,vista_cant_tiempo, se procedió a integrarlas con finalidad de agregar un último campo calculado (indice_materias), considerado importante para medir el rendimiento del estudiante universitario, el campo almacena el promedio de materias que aprueba el estudiante por año académico.

El campo indice_materias se obtiene con el cálculo: (mat_aprobadas / tiempouniv)

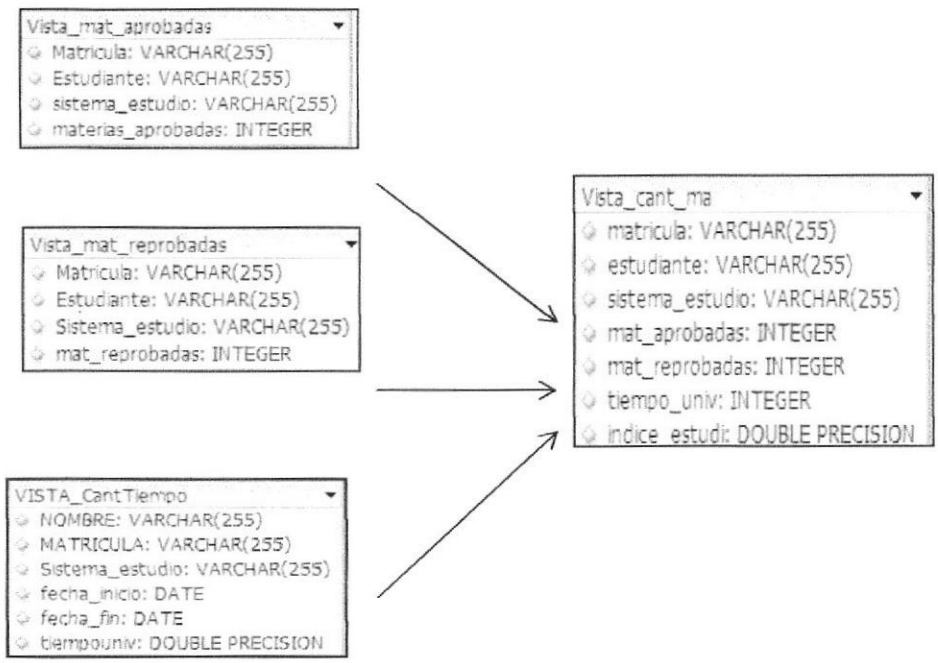


Figura 11. Vistas integradas.

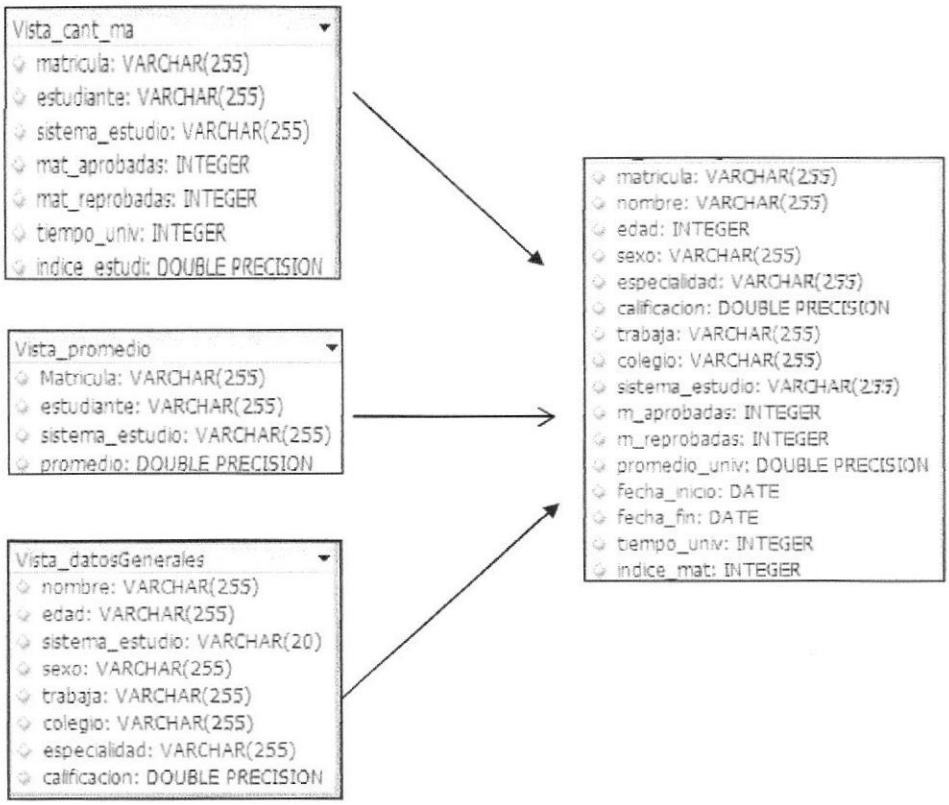


Figura 12. Vista final con los datos integrados

La vista: "dataprimeramarzo", es la que finalmente quedó, y será denominada "vista minable", tendrá los registros de los estudiantes inscritos en la Facultad de Sistemas y Telecomunicaciones que tiene bajo su responsabilidad dos carreras: Ingeniería en Sistemas e Ingeniería en Electrónica y Telecomunicaciones.

En la tabla 3. se presentan las variables que constituye la vista final denominada "vista minable" :

Tabla 3. Atributos de la "vista minable"

Nombre de atributo	*Tipo de atributo*	*Descripción*
Matrícula	Cualitativa	El número de matrícula de cada estudiante
Nombre	Cualitativa	Nombres y apellidos de los estudiantes
Edad	Cuantitativa	Edad de cada estudiante
Sexo	Cualitativa	Sexo del estudiante
Especialidad	Cualitativa	Nombre de la especialización en la que se graduó el estudiante
Calif_cole	Cuantitativa	Calificación promedio con el que se graduó en el colegio
Trabaja	Cualitativa	Define si el estudiante trabaja o no
Colegio	Cualitativa	Nombre del colegio donde se graduó el estudiante
Sistema_estudio	Cualitativa	Sistema de estudio del estudiante
Mat_aprobadas	Cuantitativa	Materias aprobadas en toda su estadía en la universidad
Mate_reprob	Cuantitativa	Materias reprobadas en toda su estadía en la universidad
Prome_calif	Cuantitativa	Promedio de calificación en la universidad
Fecha_inicio	Cualitativa	Primera fecha en que se registró el estudiante
Fecha_fin	Cualitativa	Última fecha en que se registró el estudiante
Iiempo_univ	Cuantitativa	Tiempo que tiene en la universidad
Indice_mat	Cuantitativa	Promedio de materias aprobadas por período académico

3.2.2.3 Numerización

Para el análisis cluster se necesita tener datos numéricos en todas las variables por lo tanto las variables categóricas deben ser convertidas a numéricas, se denomina a este proceso "numerización"

La numerización es un proceso donde se crea variables indicadores denominadas "dummy". Si una variable o atributo nominal tiene posibles valores (a1,a2..an) se crean n

variables numéricas con valores de 0 a 1 dependiendo si la variable nominal toma ese valor o no. Éste proceso es útil cuando se espera uno de los posibles valores de la variable sea significativo" (Hernández O, 2004)

Tabla 4. Atributos (variables) cualitativas de la "vista minable"

Nombre de atributos	Categoría
Sexo	Femenino / Masculino
Especialidad	Secretariado/Electrónica/Otros Qui-bio/ Comercio y Administración / Fima, Informática
Tipo de colegio	Fiscal / Fisco-Misional / Particular/ Municipal
Trabaja	Si /No

Tabla 5. Atributos (variables) cualitativas de la "vista minable"

Nombre de variable	Variables dummy	Tipo de variable	Significado
Sexo	Masculino	Binomial [0-1]	1.- Masculino 0.- Femenino
Especialidad	Fima	Binomial [0-1]	1. Si / 0. No
	Informática	Binomial [0-1]	1. Si / 0. No
	Electrónica	Binomial [0-1]	1. Si / 0. No
	Comercio y administración	Binomial [0-1]	1. Si / 0. No
	Quibio	Binomial [0-1]	1. Si / 0. No
	Otras	Binomial [0-1]	1. Si / 0. No
Tipo de colegio	Fiscal	Binomial [0-1]	1. Si / 0. No
	Particular	Binomial [0-1]	1. Si / 0. No
	Otro	Binomial [0-1]	1. Si / 0. No
Trabaja		Binomial [0-1]	1. Si / 0. No

En la tabla 4 apreciamos los atributos (variables) cualitativas y en la tabla 5 las variables "dummy" creadas en el proceso de numerización, a partir del atributo "colegio" que

constaba en la "vista minable" pudo identificarse el tipo de colegio, atributo cuyos registros fueron convertidos a datos numéricos

El proceso de numerización de los datos fue realizada con funciones diseñadas en PostrgeSQL

La tabla 6, observamos los atributos (variables) cuantitativas que no necesitaron ser numerizadas.

Tabla 6. Atributos (variables) cuantitativas de la "vista minable"

Nombre de variable	*Tipo de variables*	*Características*
Edad	Cuantitativa	Edad
Califi_cole	Cuantitativa	Calificación promedio con el que se graduó en el colegio
Prome_calif	Cuantitativa	Promedio de calificación en la universidad
Mate_reprob	Cuantitativa	Materias reprobadas en toda su estadía en la universidad
Tiempo_univ	Cuantitativa	Tiempo que tiene en la universidad
Indice_mat	Cuantitativa	Promedio de materias aprobadas por periodo académico

3.2.2.4 Valores faltantes

La "vista minable" una vez que pasó por el proceso de numerización fue analizada para determinar datos anómalos y datos faltantes.

Para el tratamiento de datos anómalos y faltantes se utilizó el editor de R, TINN-R y el lenguaje R.

Tabla 7. Datos faltantes en las variables cualitativas y cuantitativas

Datos	*Variables*	*Datos faltantes*	*Datos completos*	*Totales*
Porcentaje de registros por cada atributo (variable) cualitativa	*espe_fima*	39%	61%	100%
	espe_infor	39%	61%	100%
	espe_com_adminis	39%	61%	100%
	espe_quibio	39%	61%	100%
	espe_otro	39%	61%	100%
	fiscal	39%	61%	100%
	particular	39%	61%	100%
	otro	39%	61%	100%
Porcentaje de registros por cada atributo (variable) cuantitativa	*Edad*	11%	89%	100%
	calif_cole	41%	59%	100%

La tabla 7 indica, que la falta de datos en diez atributos (variables) tiene un porcentaje superior al 10%.

La falta de dato en atributos de la "vista minable" puede producir estimaciones incorrectas en los resultados, al aplicar la técnica cluster, las acciones que pueden ejecutar para el tratamiento de los valores faltantes son:

- Ignorar.- Algunos algoritmos de minería de datos son lo suficientemente robustos para aceptar la falta de datos, el método cluster, no es uno de ellos.

- Eliminar.- quitar el atributo donde se presenta la falta de dato, el número de variables reduciría demasiado y no se podría obtener características del rendimiento antes de entrar a la universidad.

- Filtrar la fila.- Se elimina toda la fila donde está el dato faltante, este paso produce un sesgo en los datos, los valores faltantes podrían ser importantes para el modelo.

- Remplazar el valor.- Se puede elegir remplazar el valor por la media en el caso de los datos numéricos y por la moda en el caso de los datos nominales, proceso en el cual se debe ser muy cuidadoso y también podría ser generador de texto.

Para el presente análisis inicialmente se ha decidido eliminar los registros que tengan la ausencia de datos, y finalmente la "vista minable" quedó con 564 registros que serán analizados aplicando la técnica cluster.

3.2.2.5 Normalización de rango.

En el análisis cluster se debe normalizar todos los datos de los diferentes atributos al mismo rango para evitar que las distancias entre las observaciones esté afectado por las unidades de los atributos, antes de aplicar la técnica de conglomerados, se procedió a

ejecutar el proceso de normalización más común, normalización lineal uniforme, cuyo modelo matemático es: (1)

$$x' = \frac{x-min}{max-min} \quad (1)$$

Se creó una función en el lenguaje R con la ecuación (1).

3.2.3 Extracción del conocimiento

El agrupamiento (clustering) consiste en encontrar grupos entre conjuntos de individuos, el concepto de distancia es importante debido a que individuos similares deben pertenecer al mismo grupo.

Los métodos de agrupación pueden ser jerárquicos y no jerárquicos, en el primer método existe la posibilidad de unir cluster pequeños en más grandes o dividir cluster grandes en más pequeños; en los métodos no jerárquicos, el conjunto de inviduos es particionado en una cantidad predefinida de cluster.

En el lenguaje R, existe una gama de paquetes estadísticos para análisis de grupos, en este trabajo se utilizó el paquete "cluster", donde se encuentran instrucciones que se aplican dependiendo del tipo de agrupamiento elegido.

Independiente del método elegido el análisis empieza con el cálculo de las distancias entre las observaciones, lo que genera un matriz cuadrada (nxn) denominada matriz de distancia de los datos.

3.2.3.1 Matriz de distancias

La "vista minable" contenía variables con características tanto cuantitativas como cualitativas, a pesar que existen diversas métricas para el cálculo de las distancias, se usa la métrica "Gower" considerada la más apropiada cuando hay combinación de datos continuos y discretos.

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

Figura 13. Instrucciones para calcular distancia entre los datos
Fuente: Paquete "cluster" de R

3.2.3.2 Programa de agrupamiento

Esta misma librería proporciona además ciertos programas de agrupamientos: dos particionales ("pam" y "fanny") y dos jerárquicos ("agnes" y "diana"); "pam" hace agrupamientos alrededor de medoides y "fanny" hace agrupamientos difusos atribuyendo a cada objeto un grado de pertenencia a cada grupo, "agnes" en cambio hace agrupamientos jerárquicos aglomerativos y "diana" divisivos.

Se utilizará el método no jerárquico particional denominado PAM, donde para obtener k grupos, el programa elige k objetos (llamados objetos representativos) del conjunto de datos.

Los grupos correspondientes se construyen asignando cada uno de los objetos restantes al objeto representativo más "cercano". Como no todos los objetos pueden ser buenos

"representantes", la clave es elegir objetos que, en cierta manera, queden en el "centro" de los grupos (medoide) , la distancia promedio del representante a cada uno de los otros objetos del grupos debe ser mínima, esta técnica se llama k-medoide.

3.2.3.3 Identificar la cantidad de grupos

El programa "PAM", presenta un análisis detallado de las características del agrupamiento por medio de un gráfico de "siluetas". Es conveniente hacer una estimación o evaluación del número de grupos óptimo. La propuesta se basa en encontrar el agrupamiento cuyo coeficiente de silueta promedio es máximo.

Función que permite elegir la mejor cantidad de conglomerados considerando el ancho de silueta de cada conglomerado formado

```
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 14. 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". Libro de resúmenes y trabajos completos. XIV Reunión científica del grupo argentino de biometría. Argentina, 2009

Cuando se aplicó la evaluación para definir el número de grupos más apropiados en el lenguaje de programación R, se obtuvo el siguiente resultado, con los datos que utilizamos para nuestro análisis.

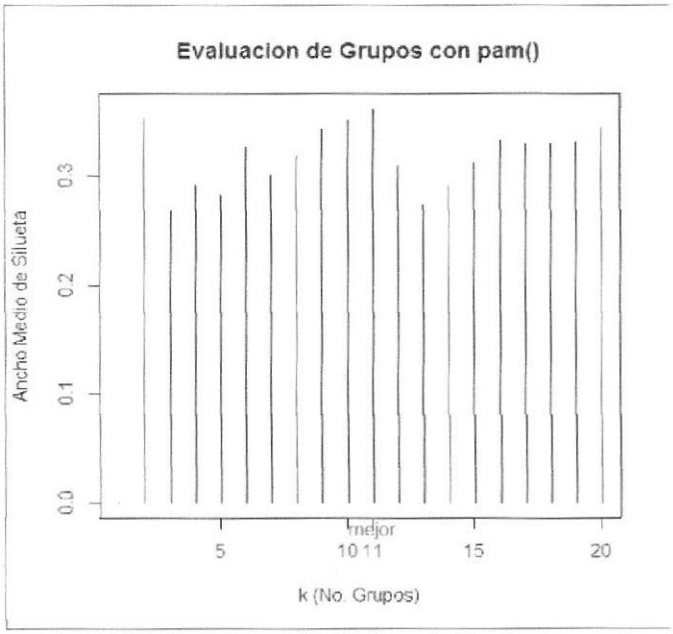


Figura 15. Resultado con la cantidad de grupos óptimos

Este resultado que se observa en la fig. 15, nos indica que según el ancho de la "silueta" la cantidad de grupos más apropiados para aplicar el método de agrupación ***es 11.***

3.2.3.3.1 Aplicación del algoritmo PAM (Agrupamiento alrededor de los medoides)

Una vez que se realizó la aplicación de algoritmo **PAM** para dos grupos, los resultados son presentados en la tabla 8:

Tabla 8. Análisis de la Fortaleza primer modelo (k=11)

Grupos	Tamaño	Máxima distancia	Promedio de distancias	Diámetros	Separación
Grupo 1	25	0,278	0,09	0,359	0,037
Grupo 2	125	0,16	0,06	0,265	0,037
Grupo 3	56	0,142	0,054	0,2	0,065
Grupo 4	41	0,225	0,085	0,345	0,044
Grupo 5	17	0,219	0,087	0,388	0,037
Grupo 6	34	0,189	0,11	0,296	0,023
Grupo 7	125	0,244	0,066	0,352	0,062
Grupo 8	40	0,191	0,061	0,268	0,054
Grupo 9	28	0,257	0,081	0,344	0,023
Grupo 10	30	0,224	0,092	0,338	0,044
Grupo 11	25	0,229	0,099	0,328	0,046

Las características mostradas en la tabla 8, en este análisis son:

- Tamaño del grupo: cantidad de elementos en el grupo

Disimilaridades:

- Máxima distancia y distancia promedio, entre objetos y medoide.
- Diámetro es la máxima disimilaridad entre dos objetos del mismo grupo.
- Separación es la disimilaridad mínima entre un objeto del grupo y una observación de otro grupo.

Considerando que máxima distancia, distancia promedio y diámetro son indicativos de la cohesión del grupo, analizando la tabla 8, en la columna "promedio de distancia", el

grupo con menor distancia promedio de las observaciones a su centrotipo es el 3 (0,054) es el grupo donde se observa los datos más concentrados, el grupo 6 en cambio presenta mayor distancia promedio entre sus datos y el centrotipo.

La tabla 8, indica en la columna "separación" que el grupo que presenta mayor separación de sus datos con respecto a los datos de otro grupo es también el grupo 3 (0,065).

Observación gráfica de los grupos formados.

La fig. 16 es una representación gráfica de los individuos agrupados en los 11 cluster, considerando dos componentes principales, se observa cada grupo o cluster, encerrado en una elipse y lineas entre los diferentes grupos que representa la distancia entre ellos.

Gráficamente el cluster 3 es el que presenta mayor concentración entre los individuos, existe mayor homogeneidad entre los mismos.

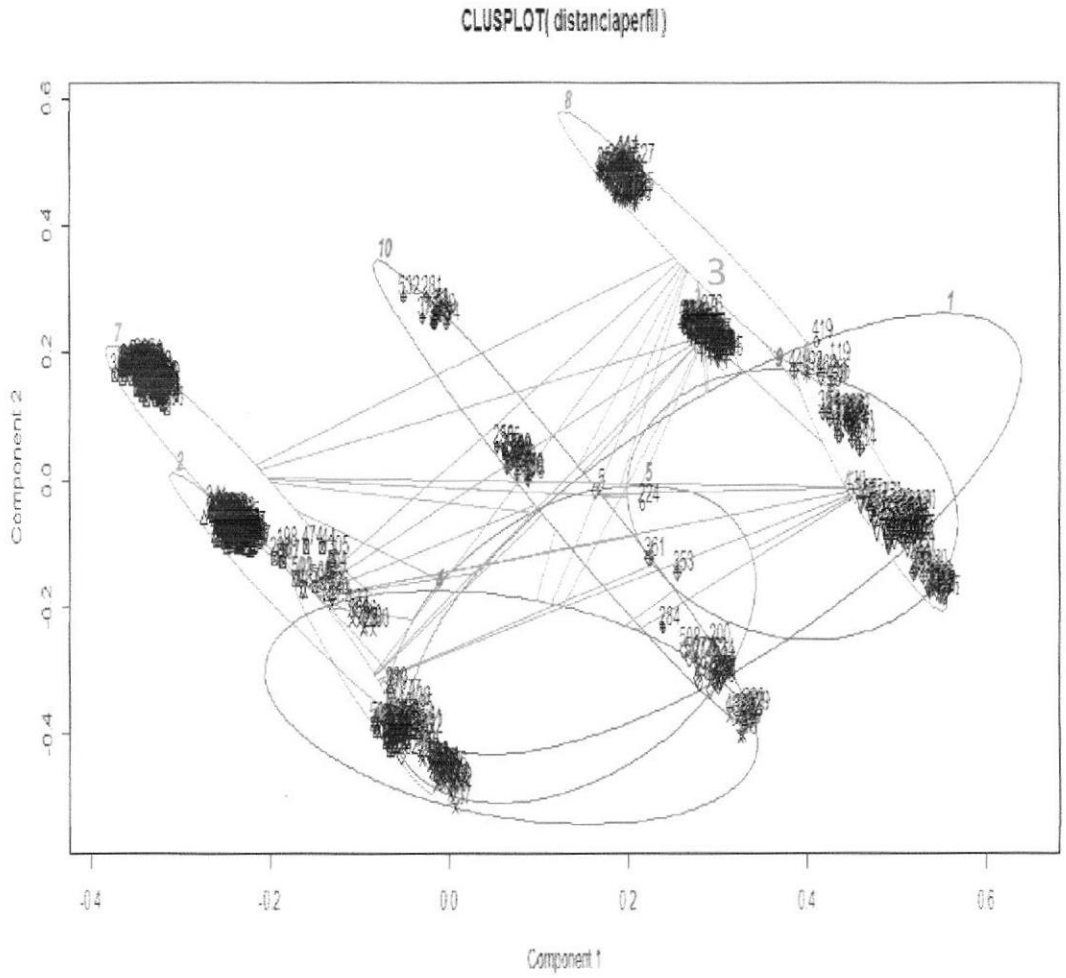


Figura 16 Gráfico de los cluster identificados

3.2.4 Interpretación

3.2.4.1 Análisis de las característica de cada grupo (K=11)

En este análisis se presentan las variables tanto cualitativas como cuantitativas, para las primeras se obtuvo la frecuencia relativa para cada una de sus categorías, mientras que para las variables cuantitativas se obtuvo el promedio.

Variable "Sexo"

Tabla 9. Distribución de frecuencias del sexo del estudiante por cada cluster

Variables		Sexo	
Categorías		Femenino	Masculino
Cluster	*1*	0,160	0,840
	10	0,333	0,667
	11	0,080	0,920
	2	0,000	1,000
	3	0,000	1,000
	4	0,180	0,820
	5	0,059	0,941
	6	0,029	0,971
	7	1,000	0,000
	8	1,000	0,000
	9	0,679	0,321

En la tabla 9, se observa que los 8 primeros grupos tienen un porcentaje alto de personas de sexo masculino, en los tres últimos grupos la presencia femenina es superior, los mismos resultados son presentados en la fig.17 con un diagrama de barras.

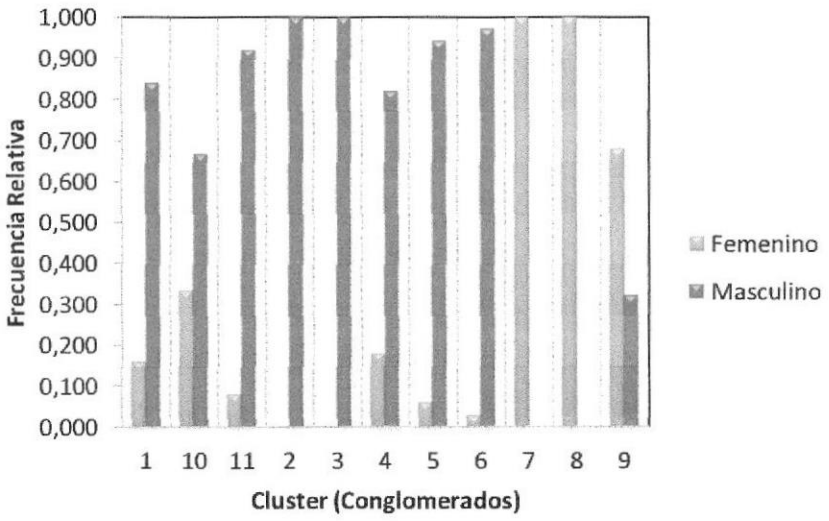


Figura 17. Distribución de frecuencias del sexo del estudiante en cada uno de los 11 conglomerados

Variables "Especialización en bachillerato"

Tabla 10. Distribución de frecuencias de la "especialización en el bachillerato" del estudiante por cada cluster.

Variables	*Categorías*	*Cluster*										
		1	*10*	*11*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Físico-matemático	**No**	1,000	0,933	1,000	1,000	1,000	0,000	1,000	0,794	1,000	1,000	0,036
	Si	0,000	0,067	0,000	0,000	0,000	1,000	0,000	0,206	0,000	0,000	0,964
Informática	**No**	1,000	0,133	1,000	0,048	0,000	1,000	1,000	1,000	0,088	0,050	1,000
	Si	0,000	0,867	0,000	0,952	1,000	0,000	0,000	0,000	0,912	0,950	0,000
Ciencias Administrativas	**No**	0,000	0,967	0,960	0,984	1,000	1,000	1,000	1,000	0,968	1,000	1,000
	Si	1,000	0,033	0,040	0,016	0,000	0,000	0,000	0,000	0,032	0,000	0,000
Electrónica	**No**	1,000	1,000	0,360	1,000	1,000	1,000	0,941	0,676	1,000	1,000	1,000
	Si	0,000	0,000	0,640	0,000	0,000	0,000	0,059	0,324	0,000	0,000	0,000
Quibio	**No**	1,000	1,000	0,760	1,000	1,000	1,000	1,000	0,941	0,976	1,000	0,976
	Si	0,000	0,000	0,240	0,000	0,000	0,000	0,000	0,059	0,024	0,000	0,024
Otras especializaciones	**No**	1,000	1,000	0,920	0,984	1,000	1,000	0,176	0,735	0,968	0,950	1,000
	Si	0,000	0,000	0,080	0,016	0,000	0,000	0,824	0,265	0,032	0,050	0,000

En la tabla 10 y en la fig. 18, se observa que el grupo 1 tiene personas con especialización en el colegio de "Ciencias Administrativas".

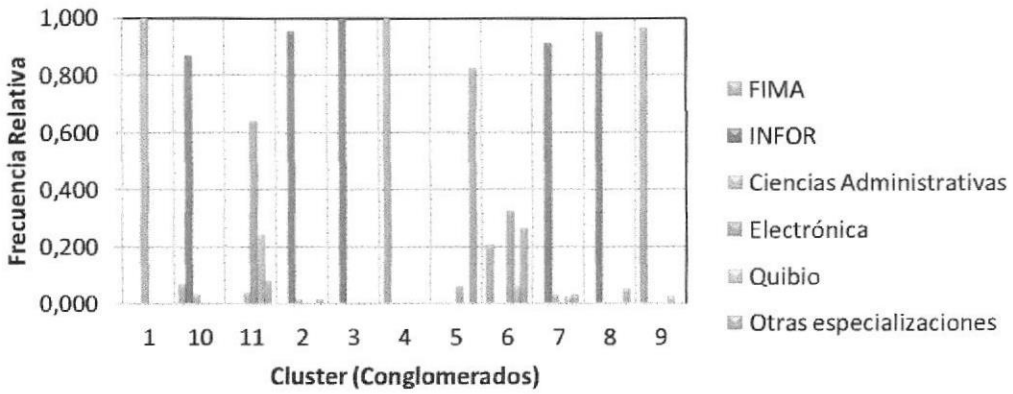


Figura 18. Distribución de frecuencias de la especialización del estudiante, en el colegio, para cada uno de los 11 conglomerados

En la fig. 18, la especialización "informática", se destaca en el grupo 10, en el 2, 3, 7, y 8; en el grupo 9 y 4 en cambio se destacan las personas con especialización "fima".

Variables "Tipo de colegio"

Tabla 11. Distribución de frecuencias del "tipo de colegio" del estudiante por cada cluster

Variables	*Categorías*	*Cluster*										
		1	*10*	*11*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Fiscales	**No**	0,20	1,00	1,00	1,00	0,00	1,00	1,00	0,00	1,00	0,00	0,07
	Si	0,80	0,00	0,00	0,00	1,00	0,00	0,00	1,00	0,00	1,00	0,93
Particular	**No**	0,88	1,00	0,00	0,00	1,00	0,15	0,88	1,00	0,00	1,00	1,00
	Si	0,12	0,00	1,00	1,00	0,00	0,85	0,12	0,00	1,00	0,00	0,00
Otro Tipo de Colegio (Municipal-fiscomisional)	**No**	0,92	0,00	1,00	1,00	1,00	0,86	0,12	1,00	1,00	1,00	0,93
	Si	0,08	1,00	0,00	0,00	0,00	0,15	0,88	0,00	0,00	0,00	0,07

Considerando el tipo de colegio del estudiante en la tabla 11, se observa que el 100% de estudiantes que vienen de colegios conocidos como particulares, se encuentran en el grupo 11, en el grupo 2, y en el grupo 7.

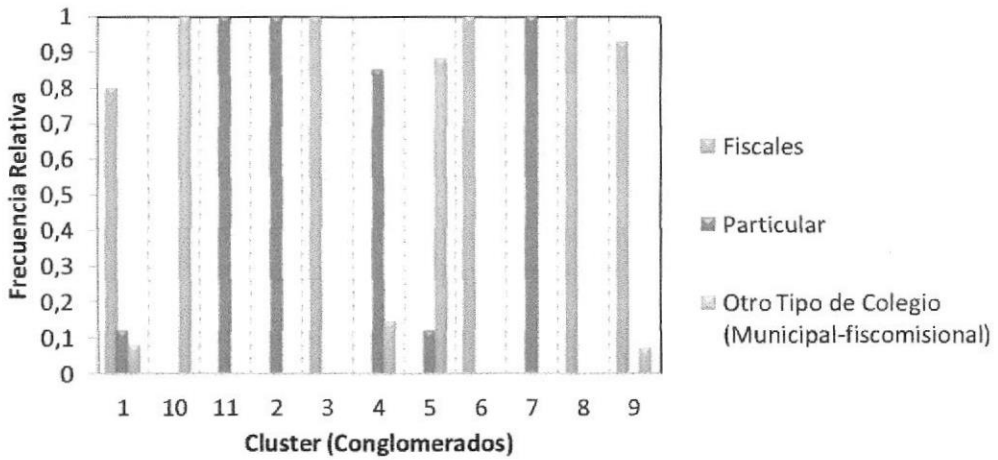


Figura 19. Distribución de frecuencias de la especialización del estudiante, en el colegio, para cada uno de los 11 conglomerados.

En la fig. 19 puede observarse los resultados del tipo de colegio para cada cluster representado a través de un diagrama de barras.

Tabla 12. Análisis descriptivo de las variables cuantitativas por grupo

Variables cuantitativas	*Cluster*										
	1	*10*	*11*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
calificación	17	18	17	18	18	18	17	17	18	18	18
edad	26	26	25	24	24	25	26	25	24	25	24
tiempouniv	2.7	3	3,8	3,8	3,2	3,4	2,8	2,5	3,3	3,5	3,6
m_reprobadas	6,4	6,7	8	8,1	7	5,8	4,8	6,4	7,4	7,4	7,2
prom_total	66	66	65	66	67	67	66	65	68	67	67
ind_materias	3,8	3,9	3	3,4	3,7	4,1	2,8	3,8	4,2	4,3	4,4

Edad

El promedio mínimo en la edad se encuentra en el grupo 9, 2, y 3 con 24 años, el mayor promedio de edad se lo encuentra en el grupo 1, 10, y 5 con 26 años de edad; en la tabla 12, se observa que en promedio el grupo que tiene más cantidad de materias reprobadas en todo los años universitarios es el grupo 11 y el grupo 2.

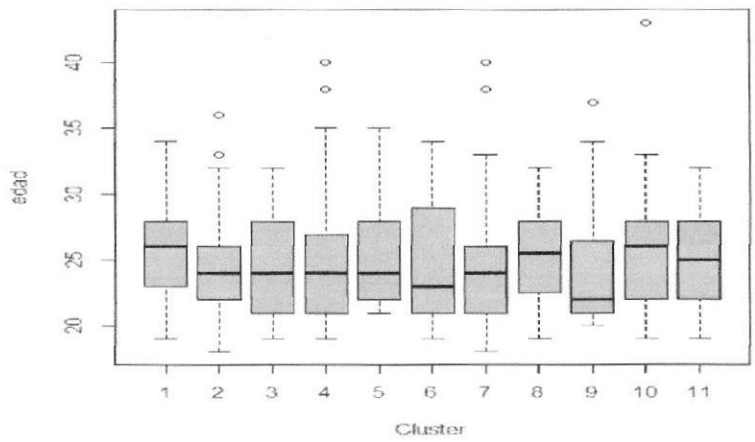


Figura 20. Diagramas de caja para la "edad" en cada uno de los 11 conglomerados

Se puede observar en la fig. 20, en el grupo 1, se encuentran el 75% de las personas con menos de 28 años de edad, el 75% de las personas del grupo 2 y el grupo 7 tiene menos de 26 años de edad.

Calificación de bachiller

En el grupo 3 se encuentran el 75% de estudiantes con edades inferiores a 29 años, y en el grupo 6 el 75% tiene edades inferiores a 30 años.

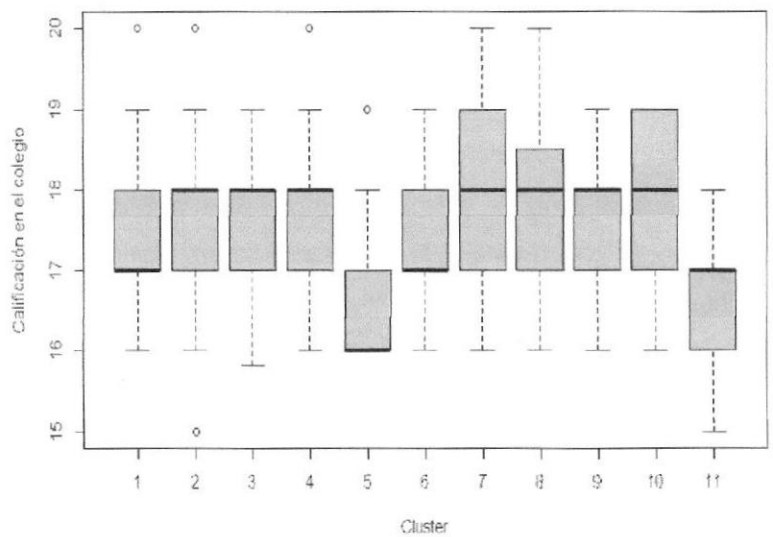


Figura 21. Diagramas de caja para "calificación de bachiller" en cada uno de los 11 conglomerados

Se destaca que el grupo 5 tiene calificaciones que están entre los 16 a los 18 puntos, el grupo 11 en cambio el mínimo puntos es 15 y el máximo 19. El grupo 7 y el grupo 10 tienen un 75% de estudiantes con calificaciones inferiores a 19 puntos.

Promedio de calificación en la universidad

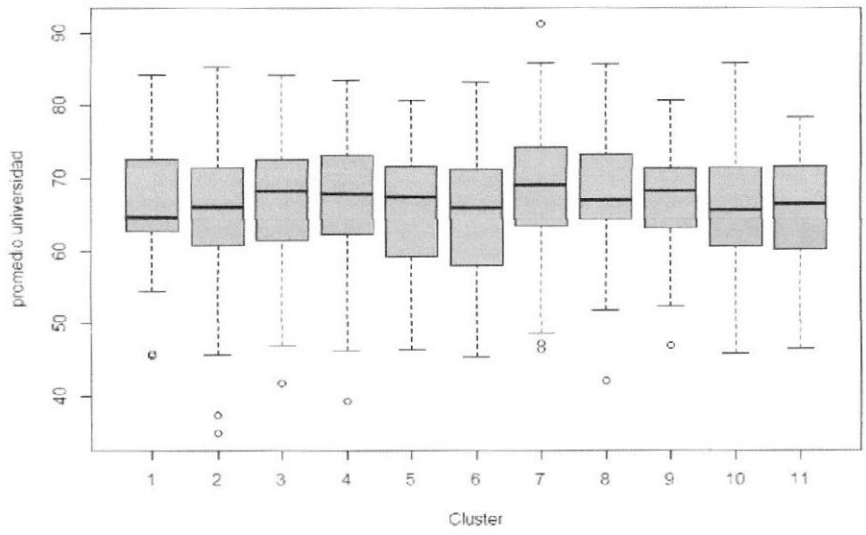


Figura 22. Diagramas de caja para "promedio de calificación en la universidad", en cada uno de los 11 conglomerados.

Ninguno de los onces grupos de estudiantes clasificados tienen un promedio en la universidad superior a 90 puntos; en el grupo 8 menos del 25% de estudiantes tienen notas inferiores a 65 puntos, y menos del 75% tiene notas inferiores a 71 puntos.

Índice de materias

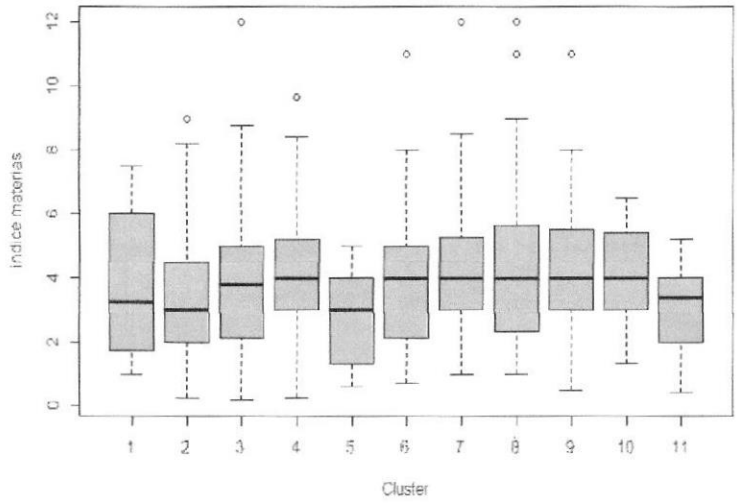


Figura 23. Diagramas de caja para "índice_materias", en cada uno de los 11 conglomerados

En el grupo 5 se encuentran los estudiantes con menos cantidad de materias aprobadas por año, el mínimo es una materia y lo máximo 5, en este grupo el 50% a pasado tres materias por período, en el grupo 1 el 75% de los estudiantes tienen menos de 6 materias aprobadas.

En el grupo 11 se encuentran estudiante con mínimo 1 materias aprobada y máximo cinco materias aprobadas.

Materias reprobadas

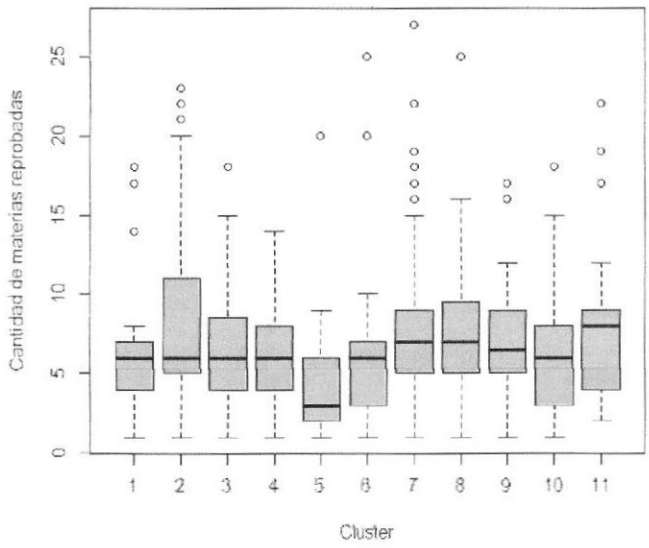


Figura 24. Diagramas de caja para "materias reprobadas", en cada uno de los 11 conglomerados

El grupo 5 son los estudiantes que tienen menos materias reprobadas en todos sus años de estudio, mínimo es 2, el 75% tiene menos 6 materias reprobadas; en el grupo 2 se encuentra un 75% de estudiantes que tienen menos de 11 materias reprobadas.

Descripción de todas las variables en cada uno de los cluster

Grupo 1: Estudiantes en mayor porcentaje **de género masculino** con especialización en ciencias administrativas, y un 80% de colegio fiscal, ***con 4 materias aprobadas por período, en promedio,*** con calificación en el colegio de 17, edad promedio de 26 años, con 6 materias reprobadas en promedio, con un promedio de calificación en la universidad de 66 puntos y con un promedio de estadía en la universidad de 3 años

Grupo 2: Estudiantes de **género masculino en un gran porcentaje (66,7%)** con especialización en informática (86.75), que tienen 4 materias aprobadas por período, en promedio, con 26 años de edad en promedio, con calificación promedio en la universidad de 66 puntos y en promedio 3 años de estadía en la universidad.

Grupo 3: Clasifica a estudiantes de género masculino, de colegios particulares, que tiene 3 materias aprobadas por período en promedio, un promedio universitario de 65 puntos y con promedio de graduación en el colegio de 17 puntos , 25 años de edad en promedio, y cuatro años en la universidad, cantidad de materias reprobadas 8.

Grupo 4: Estudiantes de género masculino, de colegios particulares, especialización informática, con un promedio de calificaciones en la universidad de 66 puntos, con un promedio de graduación en el colegio de 18 puntos, con 8 materias reprobadas y 3 materias por período aprobadas.

Grupo 5: Estudiantes de género masculino, de colegios fiscales, especialización informática, con un promedio de calificaciones en la universidad de 67 puntos, con un promedio de graduación en el colegio de 18 puntos, con 7 materias reprobadas y 4 materias por período aprobadas.

Grupo 6: Estudiantes de género masculino, de colegios fiscales, especialización FIMA, con un promedio de calificaciones en la universidad de 67 puntos, con un promedio de

graduación en el colegio de 18 puntos, con 6 materias reprobadas y 4 materias por período aprobadas.

Grupo 7: Estudiantes de género masculino, de colegios fiscales, especialización FIMA, con un promedio de calificaciones en la universidad de 67 puntos, con un promedio de graduación en el colegio de 17 puntos, con 5 materias reprobadas en 3 años y 3 materias por período aprobadas.

Grupo 8: Estudiantes de género masculino, de colegios fiscales , especialización Electrónica y otras, con un promedio de calificaciones en la universidad de 66 puntos, con un promedio de graduación en el colegio de 17 puntos, con 5 materias reprobadas en 3 años y 4 materias por período aprobadas.

Grupo 9: Estudiantes de género femenino, de colegios particulares, especialización informática, con un promedio de calificaciones en la universidad de 68 puntos, con un promedio de graduación en el colegio de 18 puntos, con 7 materias reprobadas en 3 años y 4 materias por período aprobadas.

Grupo 10: Estudiantes de género femenino, de colegios fiscales, especialización informática, con un promedio de calificaciones en la universidad de 67 puntos, con un promedio de graduación en el colegio de 18 puntos, con 7 materias reprobadas en 3 años y 4 materias por período aprobadas.

Grupo 11: Estudiantes de género femenino, de colegios fiscales, especialización FIMA, con un promedio de calificaciones en la universidad de 67 puntos, con un promedio de graduación en el colegio de 18 puntos, con 7 materias reprobadas en 4 años y 4 materias por período aprobadas.

CAPITULO IV

APLICACIÓN DE LAS FASES DE MINERÍA CON NUEVOS DATOS

Hasta el 2010, la universidad poseía el sistema de bienestar universitario que constaba de 47 variables, sin embargo para el análisis presentado en el capítulo III, sólo fueron proporcionados datos completos en cinco de estas variables, los datos restantes estaban incompletos, los datos proporcionados fueron analizados y tenían relación al aspecto académico histórico y actual del estudiante.

Se sugirió a partir de estos inconvenientes la actualización no sólo de los datos sino también de las variables que se encontraban en la ficha socio-económica, en este capítulo se describe las variable propuestas para la nueva ficha, y se establece el procedimiento para la recolección de los respectivos datos.

Finalmente se escogerá una muestra de estudiantes de la Facultad de Sistemas y Telecomunicaciones que hasta octubre del año 2011 ingresaron los datos respectivos en la nueva ficha socio-económica, datos que serán analizados aplicando la técnica de minería de datos denominada análisis de conglomerados ("cluster").

4.1 Propuesta de la nueva ficha socio - económica

4.1.1 Descripción de las variables presentes en la nueva ficha socio-económica

Analizando la ficha de bienestar estudiantil vigente hasta el año 2010, se consideró apropiado actualizarla, reorganizando las secciones, adicionando ciertos campos en las diferentes secciones. La nueva ficha la denominaríamos "Ficha actual Bienestar" y constaría de siete secciones:

Sección 1: Identificación Universitaria: Identificar a cada uno de los estudiantes en su respectiva carrera, modalidad, nivel, paralelo, además de requerir si el estudiante antes se ha inscrito en otra carrera de esta u otra universidad, los campos de esta sección son: *Apellidos y Nombres, Carrera, Modalidad, Nivel, Paralelo, Sistema de estudio, Año de ingreso, Forma de ingreso a la universidad, Inscripción en otra carrera.*

Sección 2: Datos personales: Requerir información personal del estudiante, importante en toda institución de educación superior para conocer a su principal cliente y poder ubicarlo, dentro de esta sección se solicita al estudiante información sobre algún contacto en el caso de alguna emergencia que podría ocurrir dentro de la institución de educación superior: *Cédula de identidad, Lugar y fecha de nacimiento, Nacionalidad, Teléfono Convencional, Teléfono Celular, Correo electrónico, Edad, Sexo, Tipo de sangre, Estado civil, Nombre del esposo, Tiene hijos, Lugar de residencia Habitual: Provincia, Cantón, Parroquia, y Dirección exacta, Nombre de algún contacto de*

emergencia, Parentesco con el contacto de emergencia, Dirección y Teléfono del contacto de emergencia.

Sección 3: Información educativa: Conocer información sobre la historia educativa del estudiante, colegio, especialidad, tipo de colegio, año de graduación, ubicación del colegio y calificación que el estudiante obtuvo al egresar como bachiller en el nivel medio*: Colegio de Procedencia, Especialidad, Tipo de colegio, Año de graduación, Calificación promedio de graduación, Ubicación del colegio: Provincia, cantón, parroquia, barrio/comuna, dirección del colegio.*

Sección 4: Información económica - laboral: Esta se divide en varias sub- secciones y se solicita considerando la dependencia económica del estudiante.

4.1 Grupo Familiar.- En la primera se requiere que el estudiante defina de quien depende económicamente, que brinde un detalle de su grupo familiar, requiriendo para cada habitante de su hogar su estado civil, su edad, el nivel educativo y la ocupación. Se solicitará además la cantidad de miembros de la familiar que recibe bono del estado y la cantidad de miembros que dependen del jefe de hogar.

4.2 Padres.- En esta parte se requiere la situación laboral tanto del padre como de la madre, en el caso que de ellos dependa económicamente, se realiza esta solicitud de información con cierto detalle de su nivel de instrucción, la profesión, la empresa donde

trabaja, la categoría ocupacional, el cargo en la empresa, la dirección de su trabajo, y su ingreso promedio mensual.

4.3 Otros familiares Este aspecto requiere la situación laboral del familiar de quien depende económicamente el estudiante, con cierto detalle de su nivel de instrucción, la profesión, la empresa donde trabaja, la categoría ocupacional, el cargo en la empresa, la dirección de su trabajo, y su ingreso promedio mensual.

4.4 Del estudiante: Se pregunta a todos los estudiantes si en los actuales momentos se encuentra trabajando, de ser así, se requiere el nombre de la empresa, la dirección, el cargo, la categoría ocupacional, el ingreso promedio mensual, y del estudiante se necesita conocer el tiempo que lleva trabajando y las horas que dedica al trabajo diariamente.

Sección 5: Condiciones habitacionales: Se considera importante conocer sobre la vivienda del estudiante, el tipo de construcción, los servicios básicos que posee y la condición económica de la vivienda, esta sección consta de los siguientes campos: *Clase de vivienda, Tipo de construcción, Piso, Cantidad de divisiones, Servicios que posee, Condiciones económicas de la vivienda.*

Sección 6: Información adicional: Se requiere información del estudiante sobre familiares que han emigrado, la forma de transportarse a la universidad, las actividades que practica, el tiempo que dedica al internet, entre otros campos que también nos

indicaran ciertas características del estudiante: *Estudiante posee discapacidad, Tipo de discapacidad (sólo en caso de tenerla), Familiares que han emigrado, Forma de transportarse a la universidad, Actividades que practica, Navegación en internet por trabajo / estudios, Navegación en internet por diversión, Tenencia de computadora personal, Tenencia de computadora portátil, Acceso al servicio de internet.*

Sección 7: Financiamiento Educativo: Se pregunta al estudiante sobre el tipo de financiamiento que tiene para sus estudios: *Tipo de financiamiento del estudiante, Beca recibida por el estudiante.*

4.2 Implementación de la ficha e ingreso de datos.

La "ficha actual Bienestar" fue propuesta en noviembre del 2010 y después de un proceso de prueba, considerando una recolección piloto, fue revisada, y validada por Vicerrectorado Académico y Decanos de las ocho facultades de la Universidad Estatal Península de Santa Elena, siendo finalmente aprobada en Junio del 2011, éste instrumento quedó como insumo para la universidad y específicamente para el Departamento de Bienestar Estudiantil.

Entre marzo y mayo del 2011, fue diseñada una aplicación con los campos principales de la ficha, esta aplicación fue realizada en el software LimeSurvey, la misma que permite construir de una forma sencilla encuestas, y tiene la posibilidad de presentar

resultados estadísticos básicos, esta encuesta fue diseñada para ambiente web y se accedía a la misma a través de un link ubicado en la página web de la universidad.

Una vez diseñada la aplicación, esta sería propuesta al estudiante como un link en la página web de la universidad, en el mes de julio hasta diciembre del 2011, la población de estudiantes universitarios de la UPSE, sería censado con la "Ficha actual Bienestar", colocando como condicionante que los estudiantes no podrían observar sus notas sin antes llenar en línea la ficha socioeconómica.

A partir del siguiente período académico la ficha solo sería de llenado obligatorio para los estudiantes nuevos, que se registren en la Universidad Península de Santa Elena (UPSE).

Hasta octubre del 2011, 3400 estudiantes han ingresado sus datos en la ficha socio-económica, la cantidad de estudiantes matriculado al 2011 son 7200, el 47% de la cantidad de alumnos inscritos en la UPSE ingresaron sus datos.

4.3 Aplicación de minería con nuevos datos de la ficha socio-económica

4.3.1 Recopilación de los datos

Los datos fueron proporcionados por UPEI, en formato Excel se realizó el proceso de transformación para ser cargados en el gestor de base de datos PostgreSQL, la tabla estructurada y cargada es: "FSE_nueva".

4.3.2 Transformación, limpieza de los datos

4.3.2.1 Creación y selección de atributos

Vista FSE

En esta vista fue creado el atributo "ingreso_familiar" como campo calculado, donde se suma todos los ingresos de las personas de quienes el estudiante había mencionado su dependencia económica.

Las diferentes variables que constan en esta vista son mencionadas en la tabla 13.

Descripción de las variables

Las variables de la ficha socioeconómica que se eligió para el análisis son:

Tabla 13. Variables elegidas de la nueva ficha socio-económica

Nombre de variable	Categoría
Sexo	Variable binomial considerando dos opciones(1.-Femenino, 2.-Masculino)
Edad	Variable cuantitativa, el entrevistado indica la cantidad de años cumplidos
Especialidad	El entrevistado indica la especialización con la que se graduó de bachiller: Secretariado/Electrónica/Otros Qui-bio/ Comercio y Administración / Fima, Informática
Tipo de colegio	El entrevistado indica el tipo de colegio del que proviene: Fiscal / Fisco-Misional / Particular/ Municipal
Estado Civil	Variable cualitativa con las siguientes opciones: Casado / Unión Libre / Divorciado / Separado / Viudo / Soltero
Tiene Hijos	Variable Binomial (Si /No)
Computadora de escritorio	Variable Binomial (Si /No)
Computadora portátil	Variable Binomial (Si /No)
Acceso de Internet en casa	Variable Binomial (Si /No)
Calificación de graduación	La calificación con la que terminó el bachillerato
Ingreso promedio en el hogar	Variable calculada obtenida de la suma de los ingresos familiares de los miembros del hogar de quienes depende económicamente el entrevistado

Fuente: Autor

Fueron utilizadas las mismas vistas que se obtuvieron para el análisis de los datos propuestos en el capítulo 3:

Vista_mat_aprobada, con el campo calculado de la cantidad de materias aprobadas

Vista_mat_reprobada, con el campo calculado de la cantidad de materias reprobadas.

Vista_promedio, con el campo calculado con el promedio de calificaciones en la universidad.

Vista_cant_tiempo: con el campo calculado cantidad de años en la universidad, considerando la primera fecha de inscripción y la última fecha de inscripción.

4.3.2.2 Integración de atributos

Se integraron las vistas creadas a partir de la tabla con notas y las vistas creadas considerando la tabla con datos de la nueva ficha socioeconómica.

La data integrada tiene 190 registros considerando sólo a los estudiantes de la Facultad de Sistemas y Telecomunicaciones que hasta Octubre del 2011, llenaron la ficha socioeconómica, esto representa una muestra del 46% de los estudiantes matriculados al 2011.

Vistas integradas

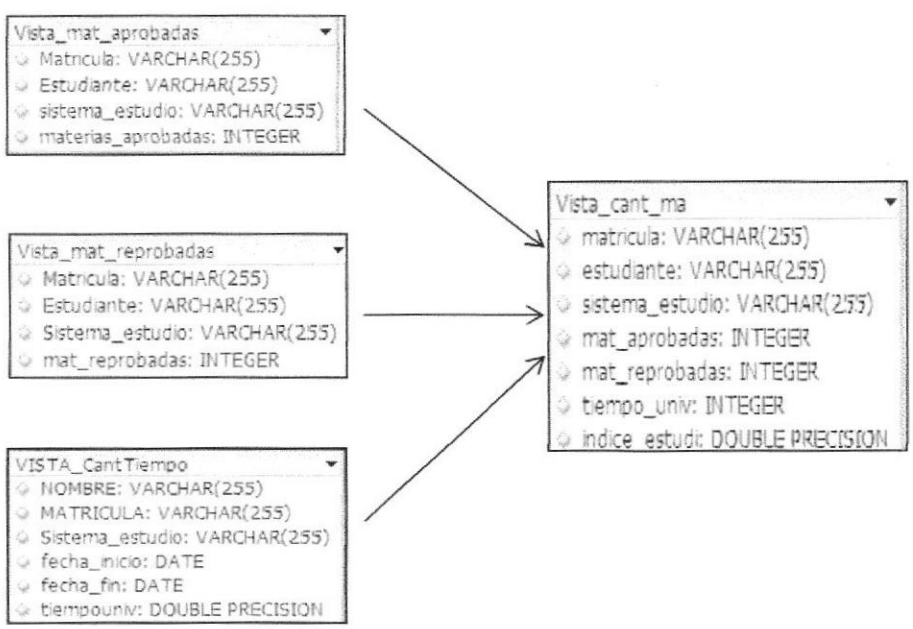


Figura 25. Integración de vistas creadas con tabla "Notas_facistel"

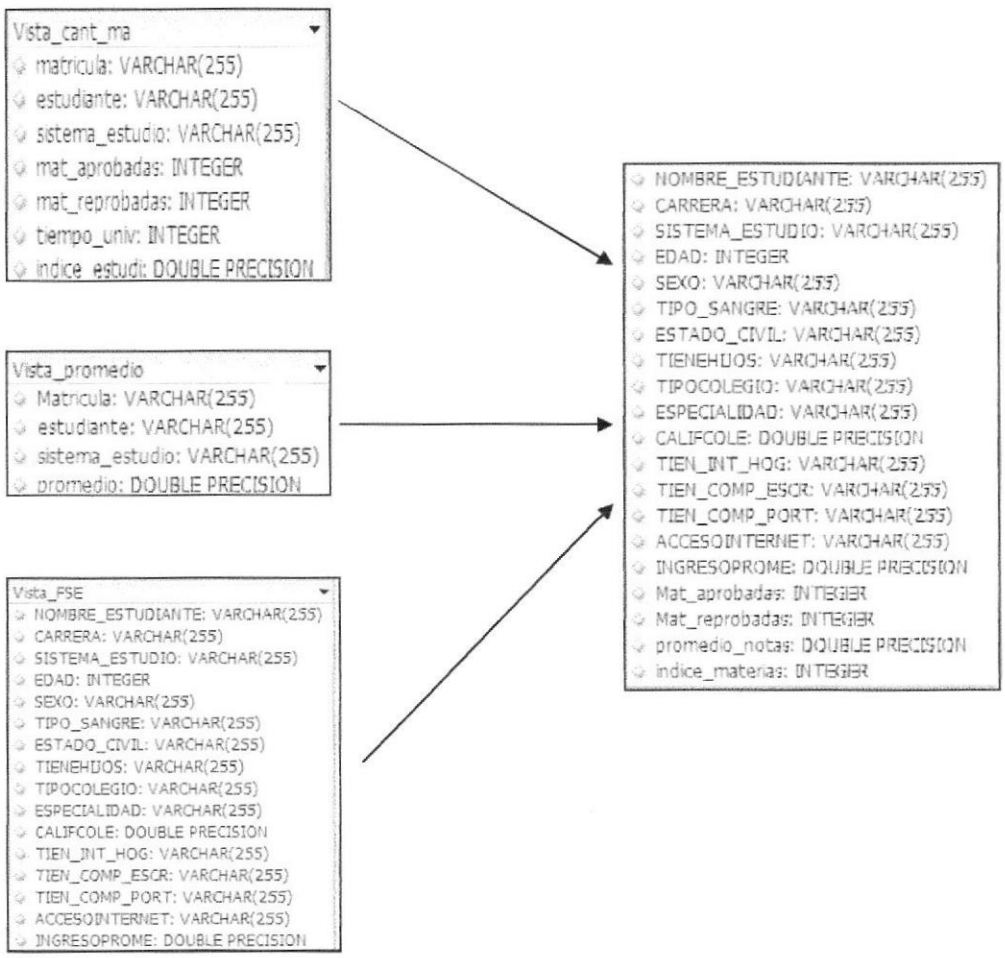


Figura 26. Vista integrada para segundo análisis

La vista: "data_integrada _nueva" es la que finalmente quedó, los registros que serán analizados en este trabajo serán los pertenecientes a la Facultad de Sistemas y Telecomunicaciones que tiene bajo su responsabilidad dos carreras: Ingeniería en Sistemas e Ingeniería en Electrónica y Telecomunicaciones.

4.3.2.3 Numerización y Selección

La matriz de datos a la que denominaremos "vista minable", consta de variables cualitativas y cuantitativas, a continuación presentadas:

Tabla 14. Variables cuantitativas en la data integrada.

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

Fuente: Autor.

Tabla 15. Variables cualitativas que se encuentran en la data integrada.

Nombre de variable	Categoría
Sexo	Femenino / Masculino
Especialidad	Secretariado/Electrónica/Otros Qui-bio/ Comercio y Administración / Fima, Informática
Tipo de colegio	Fiscal / Fisco-Misional / Particular/ Municipal
Estado Civil	Casado / Unión Libre / Divorciado / Separado / Viudo / Soltero
Tiene Hijos	Si /No
Computadora de escritorio	Si /No
Computadora portátil	Si /No
Acceso de Internet en casa	Si /No

Fuente: Autor

Uno de los requerimientos del análisis cluster es contar con una vista minable con datos solo numéricos, por lo tanto en este trabajo donde las variables cualitativas poseen datos categóricos, se procedió a transformar las variables apoyándonos en la generación de variables denominadas "dummy", este procedimiento consiste en transformar las variables cualitativas a valores 0 ó 1, que indica la presencia o ausencia de una categoría específica, se transformará las variables en binomiales, la transformación de dichas variables son mostradas en la Tabla 16.

Tabla 16. Transformación de las variables cualitativas que se encuentran en la data integrada.

Nombre de variable	Variables dummy	Tipo de variable	Significado
Sexo	Masculino	Binomial [0-1]	1.- Masculino 0.- Femenino
Especialidad	Fima	Binomial [0-1]	1. Si / 0. No
	Informática	Binomial [0-1]	1. Si / 0. No
	Administración de sistemas	Binomial [0-1]	1. Si / 0. No
	Electrónica	Binomial [0-1]	1. Si / 0. No
	Técnico Industrial	Binomial [0-1]	1. Si / 0. No
	Contabilidad	Binomial [0-1]	1. Si / 0. No
	Ciencias Administrativas	Binomial [0-1]	1. Si / 0. No
	Quibio	Binomial [0-1]	1. Si / 0. No
	Otras	Binomial [0-1]	1. Si / 0. No
Tipo de colegio	Fiscal	Binomial [0-1]	1. Si / 0. No
	Fisco-Misional	Binomial [0-1]	1. Si / 0. No
	Particular	Binomial [0-1]	1. Si / 0. No
	Municipal	Binomial [0-1]	1. Si / 0. No
Estado Civil	Casado	Binomial [0-1]	1. Si / 0. No
	Unión Libre	Binomial [0-1]	1. Si / 0. No
	Divorciado	Binomial [0-1]	1. Si / 0. No
	Separado	Binomial [0-1]	1. Si / 0. No
	Viudo	Binomial [0-1]	1. Si / 0. No
	Soltero	Binomial [0-1]	1. Si / 0. No

Fuente: Autor

El proceso de transformación de los datos a variables "dummy" se realizó en PostgreSQL.

4.3.2.4 Imputación de los valores faltantes

A partir de este paso, el trabajo se utilizará con el editor de texto del Lenguaje R (Tinn-R), en este software estadístico sería ejecutado el análisis cluster.

Utilizando una conexión ODBC y la librería de R destinada para aquello se procedió a realizar el enlace entre PostgreSQL y R; las instrucciones utilizadas para este paso fueron:

```
library(RODBC)
con=odbcConnect("PostgreSQL35W",case="postgresql")
dataintegrada1<- sqlFetch(con,"dataintegrada_prueba1")
```

Figura 27. Instrucciones que vinculan PostgreSQL y R

Una vez que tenemos los datos en R, se procede a analizar la existencia de valores faltantes en la matriz de datos, e identificándose a la variable: "ingreso promedio mensual de la familia", como la única que poseía valores nulos o faltantes, consideramos que eran datos que seguro se resistían a contestar, el porcentaje de valor faltante en esta variable fue el 23%, se procedió por lo tanto remplazar los valores considerando el valor promedio de la variable.

```
promingreso<- mean(dataintegrada1$total_ing_fam,na.rm=TRUE)
dataintegrada1$total_ing_fam<-
replace(dataintegrada1$total_ing_fam,is.na(dataintegrada1$total_ing_fam),promingreso)
```

Figura 28. Instrucciones para remplazar los valores faltantes

4.3.2.5. Normalización de los datos

La matriz de datos poseía tantos valores continuos como discretos para normalizar los mismos tal que la distancia entre las observaciones no estuviera afectada por los rangos de los datos se procedió a estandarizar todos los valores de la matriz de datos, especialmente los cuantitativos en valores que se encontrarán entre 0 y 1.

```
estandarconrango <- function(archivo, columna)
{
x<- archivo[,columna]
xminimo<- min(x,na.rm=T)
xmaximo<-max(x,na.rm=T)
archivo[,columna]<-(x-xminimo)/((xmaximo-xminimo))
return(archivo[,columna])
}
```

Figura 29. Función para estandarizar los datos cuantitativos en valores entre 0 y 1.

Instrucciones que ejecutan la estandarización en las variables cuantitativas

```
edad1<-estandarconrango(dataintegrada1,3)
calif<-estandarconrango(dataintegrada1,23)
ingfam<-estandarconrango(dataintegrada1,34)
mrepro<-estandarconrango(dataintegrada1,35)
promtot<-estandarconrango(dataintegrada1,36)
tiempouniv<-estandarconrango(dataintegrada1,37)
ind_mat<-estandarconrango(dataintegrada1,38)
```

Figura 30. Instrucciones que ejecutan la estandarización en las variables cuantitativas

4.3.3. Extracción del conocimiento

Se procede a aplicar la técnica de minería de datos denominada análisis "cluster" o de "conglomerados", utilizando la herramienta "R", los pasos a seguir fueron:

- Calculo de las distancias entre las variables (medida de disimilaridades).
- Aplicación del método de agrupamiento.
- Identificar los grupos y realizar la interpretación.

4.3.3.1 Matriz de distancia y programa de agrupamiento

El análisis "cluster" generalmente requiere la matriz de distancias o disimilaridades, y para el cálculo de esta matriz son utilizados diferentes métodos dependiendo del tipo de variable (continuas o discretas), en este trabajo la matriz de datos estaba compuesta de datos continuos y discretos por lo tanto se utilizará la distancia Gower, éste método permite calcular la distancia entre las observaciones, considerando la combinación de variables continuas y discretas.

En lenguaje R el algoritmo que permite utilizar la métrica Gower es denominado "Daisy"

```
library(cluster)
distanciaperfil<-daisy(dataintegradausar,metric="gower")
```

Figura 31. Instrucciones para determinar la matriz de distancias

El análisis cluster se realiza utilizando los métodos jerárquicos (exploratorios) y los métodos no jerárquicos (la cantidad de cluster son definidos en forma anticipada).

Para aplicar los métodos jerárquicos en R existe el paquete "cluster" y tiene a disposición el algoritmo "agnes" el mismo que permite realizar la agrupación jerárquica aglomerativa.

El algoritmo "agnes" incluye métodos de agregación distintos que son: Método del promedio de grupos apareados no ponderados (average o UPGMA), Vínculo simple (Single linkage), Vínculo completo (Complete linkage), Método de Ward, Método del promedio ponderado, en este trabajo se comprobó el uso de los cinco métodos y se evalúo la calidad del dendograma, por medio de la correlación cofenética.

Las instrucciones para aplicar cada uno de los métodos y el cálculo de la respectiva correlación es:

Método UPGMA

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

Método Single linkage

```
agddatos<-agnes(distanciaperfil, method = "single")
cofagddatos<- cophenetic(agddatos)
cor(distanciaperfil, cofagddatos)
```

Método de Ward

```
agddatos<-agnes(distanciaperfil, method = "ward")
cofagddatos<- cophenetic(agddatos)
cor(distanciaperfil, cofagddatos)
```

Sigue...

Continua....

Método Complete linkage

```
agddatos<-agnes(distanciaperfil, method = "complete")
cofagddatos<- cophenetic(agddatos)
cor(distanciaperfil, cofagddatos)
```

Método del promedio ponderado

```
agddatos<-agnes(distanciaperfil, method = "weighted")
cofagddatos<- cophenetic(agddatos)
cor(distanciaperfil, cofagddatos)
```

Figura 32. Instrucciones donde se aplica el algoritmo agnes

Los resultados de la correlación cofenética son:

Tabla 17. Comparación de los métodos aglomerativos con sus coeficientes de correlación respectivos.

Métodos	Coeficiente de correlación
Método UPGMA	0,674
Método Single linkage	0.57
Método de Ward	0,579
Método Complete linkage	0,5153
Método del promedio ponderado	0,667

El coeficiente de correlación puede estar entre 0 y 1 mientras más cercano al 1 se indica que existe una mejor estructura jerárquica de las observaciones, considerando los resultados, el método que tiene una más alta correlación es el de AVERAGE o UPGMA.

Por lo tanto será el método utilizado para identificar los clusters o conglomerados

Para elegir la cantidad de cluster, procedemos a ejecutar el método Average considerando diferentes cantidades de cluster y obtuvimos lo siguiente:

Tabla 18. Comparación de la cantidad de casos dependiendo de la cantidad de cluster.

Grupos	2	3	4	5	6	7	8
	Casos	Casos	Casos	Casos	Casos	Casos	Casos
1	172	75	75	75	75	75	75
2	18	97	93	93	93	93	93
3		18	4	4	4	4	4
4			18	18	5	5	5
5				13	10	9	8
6					3	1	1
7						3	1
8							3

Se observa que cuando ya se tiene más de 7 cluster hay grupos sólo con un elemento, consideramos que era apropiado entonces considerar 6 conglomerados.

Los resultados del método jerárquico aglomerativo serán presentados gráficamente a través de un dendograma, las instrucciones en R para obtenerlo son:

```
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 33. Instrucciones para obtener el dendograma con la identificación de los grupos

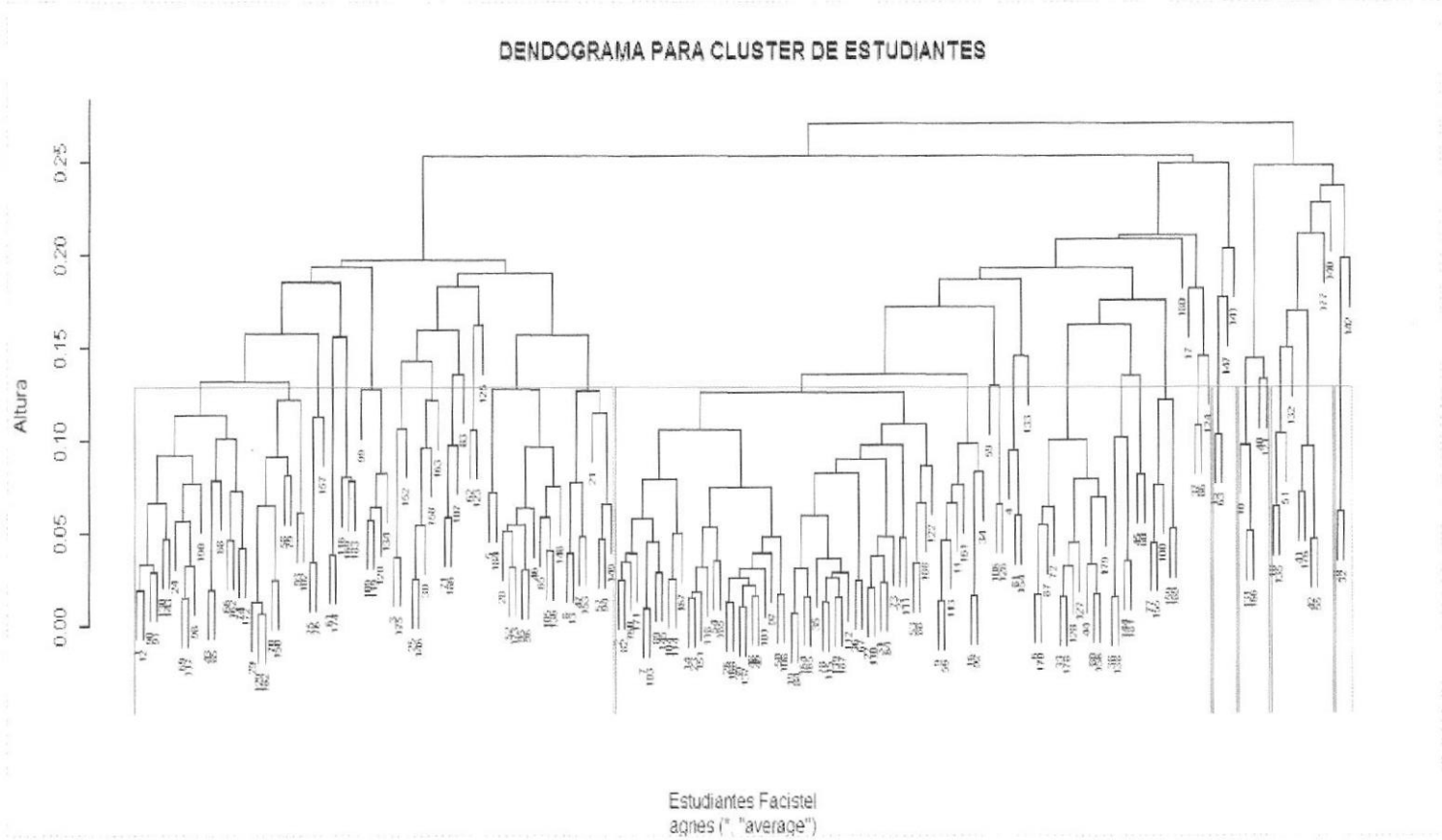


Figura 34. Dendograma obtenido con el método jerárquico

4.3.4. Interpretación de los resultados considerando todos los datos.

En la fig. 34, observamos que existe más casos en el primero y segundo grupo, los cuatro grupos restantes tienen pocas observaciones, para describir mejor las variables en cada uno de los grupos se presentan los siguientes resultados.

Los resultados serán presentados en forma tabular y gráfica para las variables cuantitativas y para las variables cualitativas.

Análisis de las variables cuantitativas

Tabla 19. Análisis del promedio en las variables cuantitativas en cada cluster.

Variables	*Cluster*					
	1	2	3	4	5	6
Calicole	18	18	18	16	18	18
Edad	22	22	30	23	27	26
Ingreso	581	449	668	456	502	320
Tiempouniv	3.3	4.1	7.6	4.0	3.0	1.3
m-reprobadas	7.3	8.6	8.4	9.8	6.1	6.0
Promuniv	65	65	62	62	68	64
Indicmat	5.1	4.4	2.1	4.8	5.0	8.2

En la tabla 19, se encuentran los promedios de las variables cuantitativas en cada uno de los seis conglomerados identificados; en promedio el grupo 3 es el que mayor ingreso ($668), en dólares tiene y también el grupo con mayor edad en promedio (30 años); en promedio el grupo que tiene más materias reprobadas (10) en los años de estudio es el 4, y éste mismo grupo tiene la menor calificación en promedio obtenida en el bachillerato, el grupo con mayor promedio en la universidad (68) es el grupo 5; y el grupo 6 es el que tiene más materias aprobadas por año lectivo.

Edad.

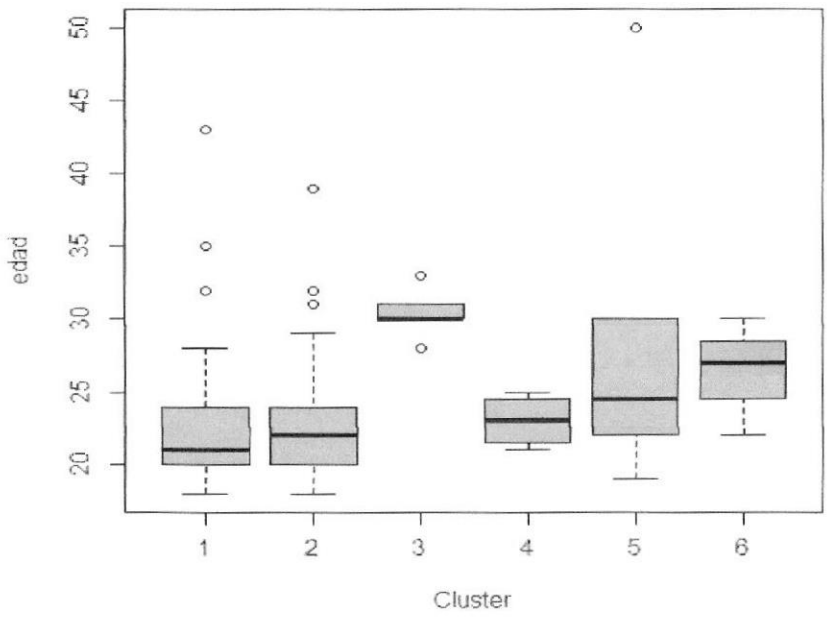


Figura 35. Diagrama de caja para la "edad" en cada conglomerado

Considerando la variable edad los dos primeros grupos se parecen mucho, el tercer grupo en cambio está constituido por personas que tienen la mayor edad, oscilan por los 30 años, el quinto grupo se destaca por tener personas que tienen máximo 20 años de edad, existe en éste grupo un valor que se podría denominar anómalo y es una persona con 50 años de edad siendo la máxima edad encontrada en todos los grupos.

En los dos primeros grupos la concentración del 75% de los datos están en personas entre 20 a 24 años de edad, no así en los cuatro grupo restantes donde están personas que tienen más de 21 años de edad.

Ingreso familiar.

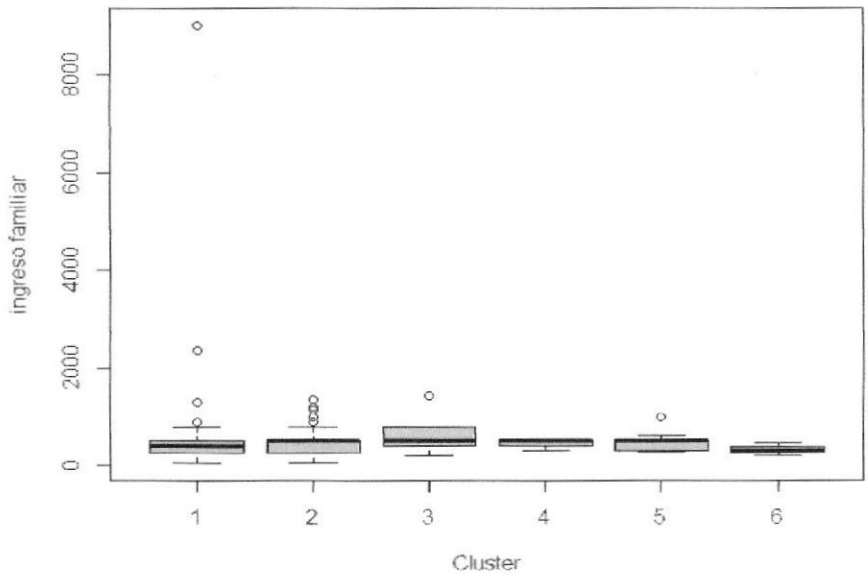


Figura 36. Diagrama de caja para la ingreso familiar en cada conglomerado.

Los datos permiten observar que existe similitud en la concentración y dispersión de las observaciones en cada uno de los grupos, el primer grupo presenta una característica particular donde el ingreso familiar mensual es superior a 8000 dólares.

Calificación del colegio.

El grupo 3, se destaca por tener personas cuyo promedio al salir del colegio fue de 16 puntos, en el grupo 1 y 3 entre el 25% al 75% están personas cuyas calificaciones con las que se graduaron en el colegio oscilan entre 17 y 19 puntos, estos resultados pueden observarse a través de un diagrama de cajas en la fig. 37.

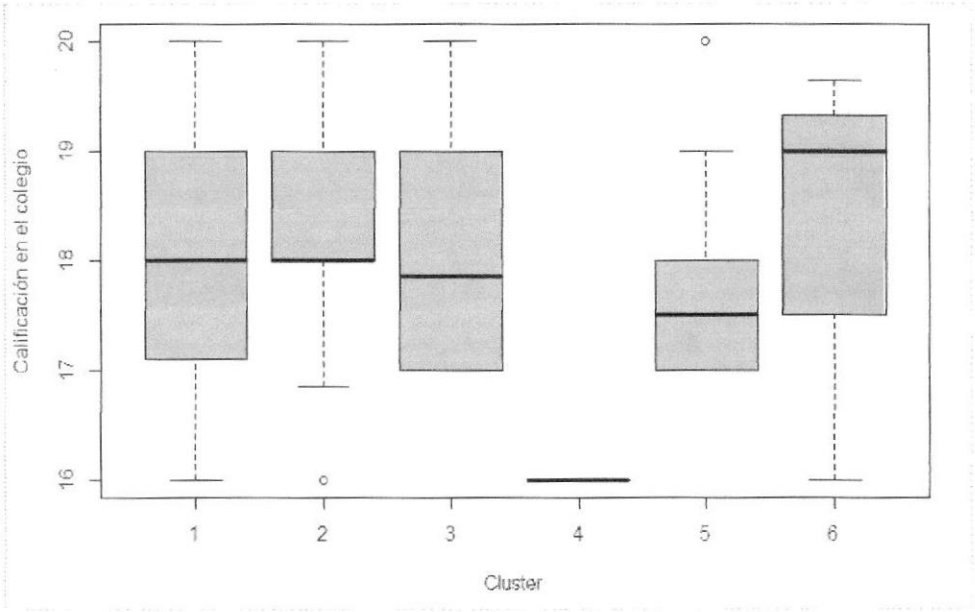
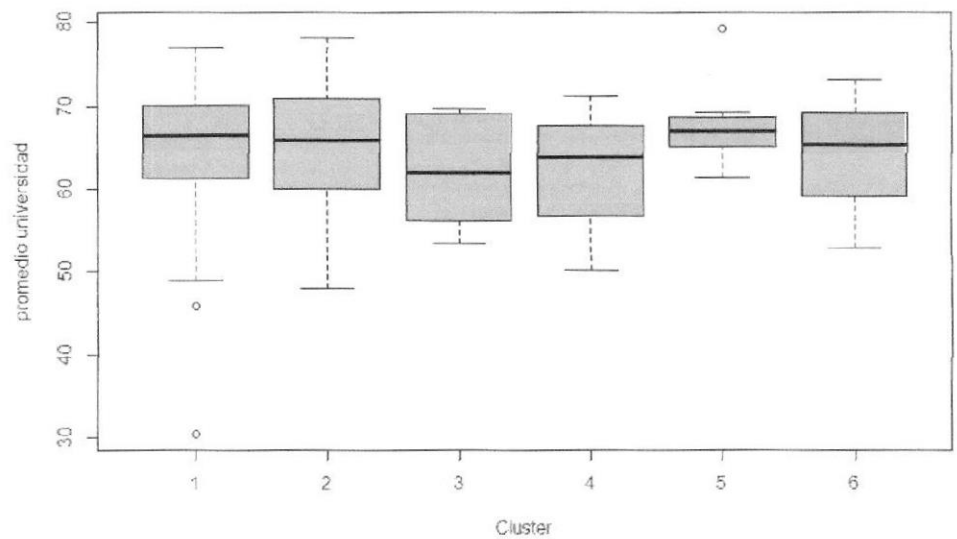


Figura 37. Diagrama de caja para "calificación del colegio" en cada conglomerado

Promedio en la universidad



Al analizar el promedio en la universidad los dos primeros grupos tienen promedios entre 60 y 70 puntos, en este rango se encuentra la mayor concentración de los datos, el

grupo 5 es aquel que tiene la concentración de sus datos en promedios entre 65 a 68 puntos.

Índice de materias

En esta variable se presenta la cantidad de materias aprobadas en promedio por cada año en la universidad.

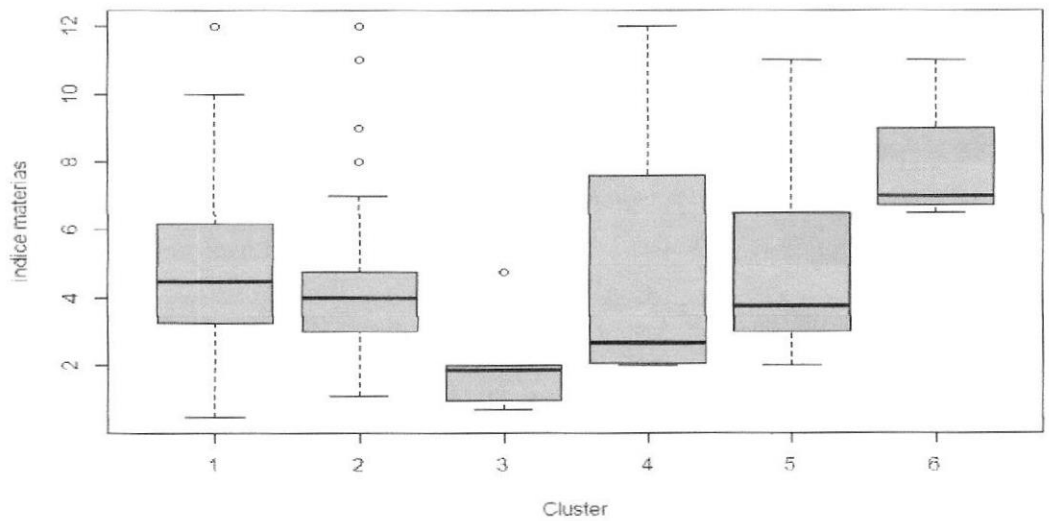


Figura 39. Diagrama de caja para "índice_materias" en cada conglomerado

En el grupo 1 se encuentra la concentración en observaciones cuya cantidad de materias aprobadas en promedio por año cursado está en los rangos de 3 a 6 materias, existiendo una dispersión hasta 10 materias, aquellos datos seguramente son de personas que tienen el período semestral, el grupo 6 constan de datos que provienen de períodos semestrales dado que la cantidad de materias aprobadas supera las seis materias.

Tiempo en la universidad

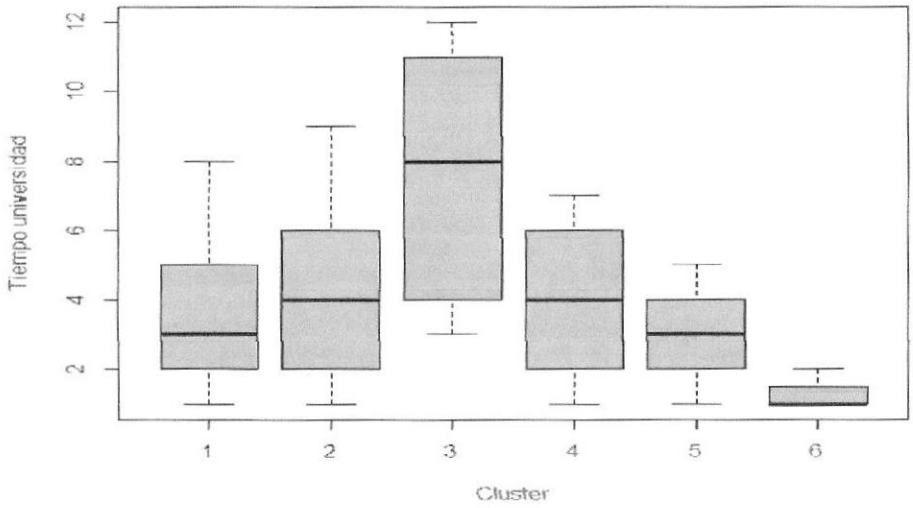


Figura 40. Diagrama de caja para "tiempo_univ" en cada conglomerado

El grupo 3 se destaca por ser el que tiene más tiempo en la universidad, cuando lo normal es hasta seis años, son personas que estuvieron inscritos, se retiraron y se inscribieron nuevamente, el grupo 6 en cambio son los que menos tiempo tienen en la universidad.

Materias reprobadas.

El grupo 5 y 6 son los que tienen menos materias reprobada mientras que el grupo 4 presenta una gran dispersión en las observaciones y entre el 25 al 75% de las personas tienen entre 6 a 13 materias reprobadas en todos sus años de estudio.

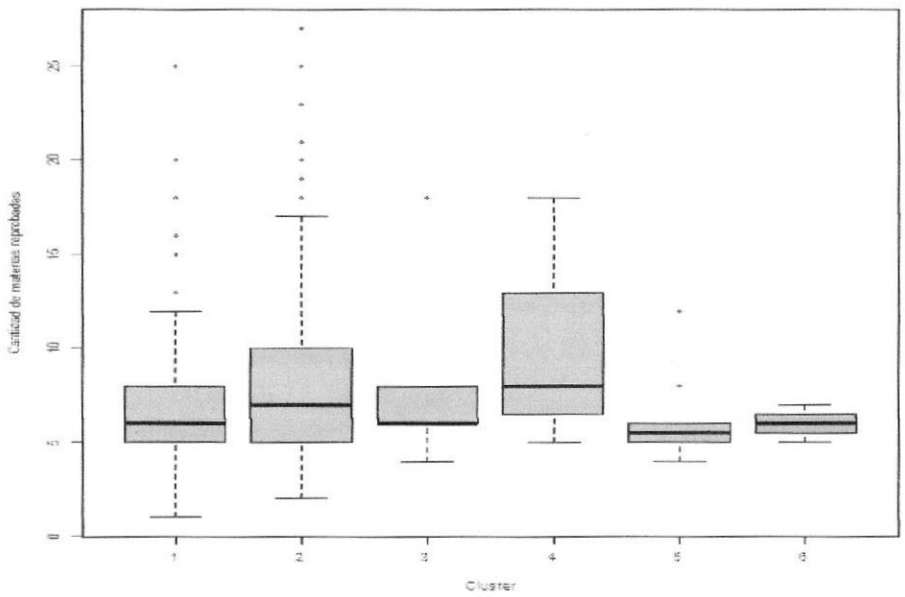


Figura 41. Diagrama de caja para "materias_reprobadas" en cada conglomerado

Se observa diferencia entre el grupo 1 y 2, en el segundo la dispersión es mayor.

Descripción de variables cualitativas

Sexo

Tabla 20. Distribución de frecuencias del sexo del estudiante en cada uno de los cluster.

Variables	*Categorías*	*Cluster*					
		1	*2*	*3*	*4*	*5*	*6*
Sexo	Femenino	0,31	0,44	0,2	0	0,1	0,33
	Masculino	0,69	0,56	0,8	1	0,9	0,67

Sólo un grupo el cuatro no tiene presencia del género femenino, la presencia de las mujeres se observa en mayor proporción en el grupo 2, sin embargo en todos los grupos la presencia masculina es mayor.

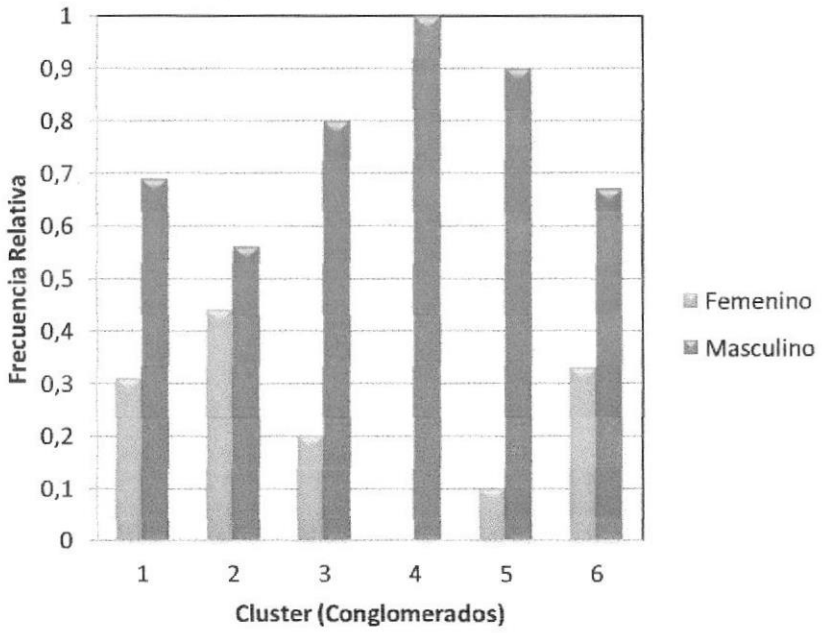


Figura 42. Distribución de sexo del estudiante en cada conglomerado

Variable estado civil y tiene hijos

Tabla 21. Distribución de frecuencias del "estado civil" del estudiante, y "tiene hijos", en cada uno de los cluster.

Variables	*Categorías*	*Cluster*					
		1	*2*	*3*	*4*	*5*	*6*
Casado	No	0,96	0,94	0,40	0,50	0,30	0,33
	SI	0,04	0,06	0,60	0,50	0,70	0,67
Unión Libre	No	1,00	1,00	0,60	0,75	0,90	0,67
	SI	0,00	0,00	0,40	0,25	0,10	0,33
Soltero	No	0,05	0,06	1,00	0,75	0,80	1,00
	SI	0,95	0,94	0,00	0,25	0,20	0,00
Tiene hijos	No	0,95	0,88	0,00	0,00	0,10	0,00
	SI	0,05	0,12	1,00	1,00	0,90	1,00

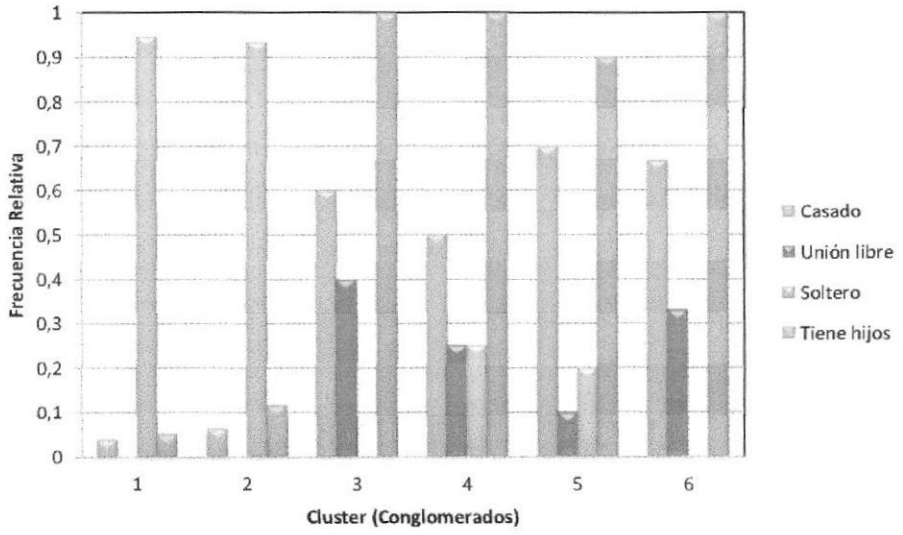


Figura 43. Distribución de "estado civil" del estudiante, y "tiene hijos", en cada conglomerado

El grupo 1 y 2 se destacan por tener a estudiantes solteros, el grupo 3 en cambio se encuentras las personas casadas, en unión libre y con hijos, todas las personas de los grupos 3, 4, y 6 tienen hijos.

No hay ninguna persona en el grupo 3 y 6 que sea soltera, el grupo 4 y 5 en cambio existe un pequeño porcentaje de personas que es soltera.

Variable especialidad

Tabla 22. Distribución de frecuencias del "especialización", en cada uno de los cluster.

Variables	*Categorías*	*Cluster*					
		1	*2*	*3*	*4*	*5*	*6*
FIMA	No	0,73	0,85	1,00	0,75	0,90	1,00
	SI	0,27	0,15	0,00	0,25	0,10	0,00
Informática	No	0,60	0,30	0,40	0,75	0,10	1,00
	SI	0,40	0,70	0,60	0,25	0,90	0,00
Administración de Sistemas	No	0,93	0,94	1,00	1,00	1,00	1,00
	SI	0,07	0,06	0,00	0,00	0,00	0,00
Electrónica	No	0,89	0,96	1,00	1,00	1,00	0,33
	SI	0,11	0,04	0,00	0,00	0,00	0,67
Técnico Industrial	No	0,97	1,00	1,00	1,00	1,00	1,00
	SI	0,03	0,00	0,00	0,00	0,00	0,00
Contabilidad	No	0,97	1,00	0,80	1,00	1,00	0,67
	SI	0,03	0,00	0,20	0,00	0,00	0,33
Ciencias administrativas	No	0,95	0,98	1,00	0,75	1,00	1,00
	SI	0,05	0,02	0,00	0,25	0,00	0,00
Quibio	No	0,99	1,00	1,00	0,75	1,00	1,00
	SI	0,01	0,00	0,00	0,25	0,00	0,00
Otras	No	0,96	0,99	0,80	1,00	1,00	1,00
	SI	0,04	0,01	0,20	0,00	0,00	0,00

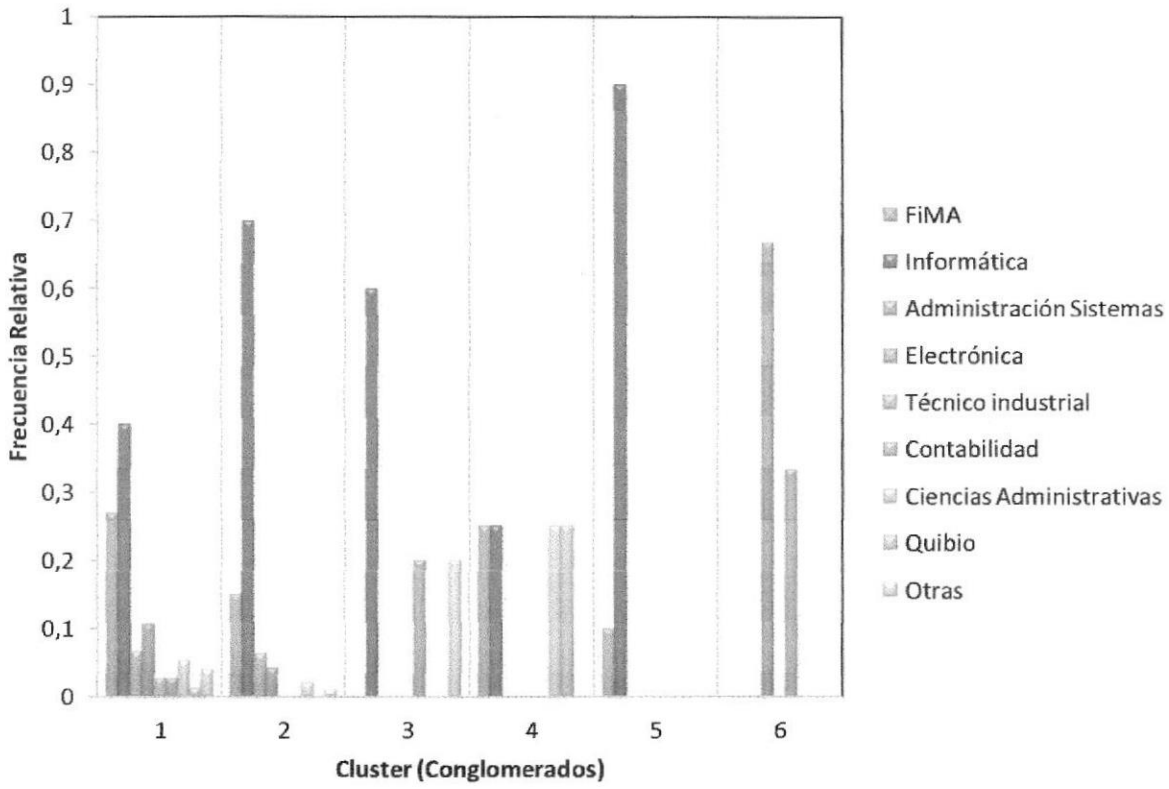


Figura 44. Distribución de "especialización en el colegio", del estudiante en cada conglomerado

El grupo 1,2,3,4,5 tiene personas con especialidad en "informática", existe una alta proporción en el grupo 5, mientras que en el grupo 6 se encuentras personas que estudio "electrónica" o " ciencias administrativas".

El grupo 4 tiene el 25% de las personas que provienen de especialidad "informática", el mismo porcentaje de personas con especialización en FIMA, en QUIBIO y en ciencias administrativas.

Variable tipo de colegio

Tabla 23. Distribución de frecuencias de "tipo de colegio", en cada uno de los cluster.

Variables	*Categorías*	*Cluster*					
		1	*2*	*3*	*4*	*5*	*6*
Fiscal	No	0,00	1,00	0,00	1,00	0,60	0,00
	SI	1,00	0,00	1,00	0,00	0,40	1,00
Particular	No	1,00	0,00	1,00	0,00	0,50	1,00
	SI	0,00	1,00	0,00	1,00	0,50	0,00
Municipal	No	1,00	1,00	1,00	1,00	0,90	1,00
	SI	0,00	0,00	0,00	0,00	0,10	0,00

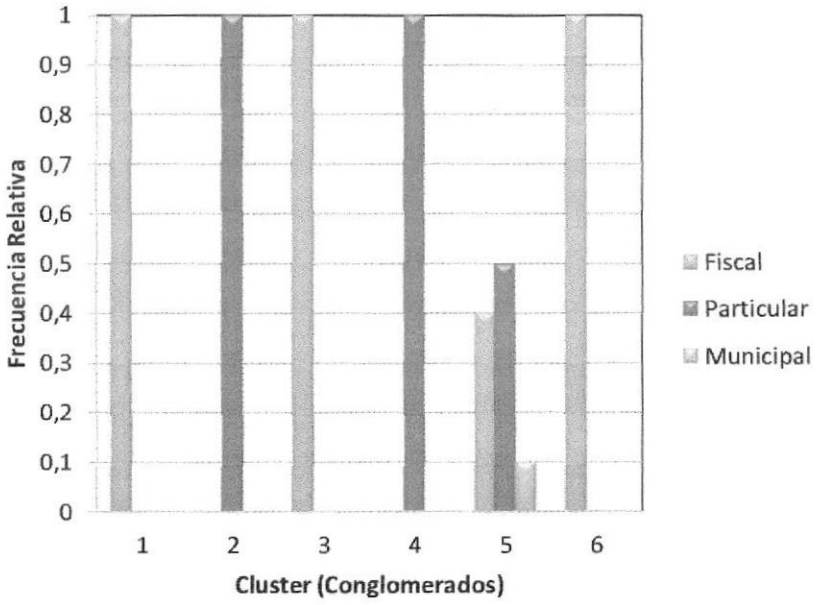


Figura 45. Distribución de "tipo de colegio", del estudiante en cada conglomerado

Solo el grupo 5 tiene personas provenientes de colegios fiscales, particulares y municipales, el grupo 2 y 4 están conformados por personas que provienen de colegios que son denominados particulares.

Computador de escritorio, computador portátil, posee internet

Tabla 24. Distribución de frecuencias de "computadora de escritorio", "computador portátil", "posee internet", en cada uno de los cluster.

Variables	*Categorías*	*Cluster*					
		1	*2*	*3*	*4*	*5*	*6*
Computadora de escritorio	No	0,35	0,22	0,00	0,75	0,60	1,00
	SI	0,65	0,78	1,00	0,25	0,40	0,00
Computadora portátil	No	0,63	0,55	0,00	0,25	0,90	0,33
	SI	0,37	0,45	1,00	0,75	0,10	0,67
Posee internet en casa	No	0,65	0,37	0,00	0,00	0,90	1,00
	SI	0,35	0,63	1,00	1,00	0,10	0,00

Todas las personas del grupo 3 tienen los tres recursos que mencionamos, el grupo 1 en un 60% tiene computadora de escritorio, y 35% de personas poseen computadora portátil y además acceso a internet, el grupo 4 se destaca porque son pocas las personas allí que tienen computadora de escritorio, mientras que portátiles y acceso al internet si tienen, en el grupo 6 en cambio sólo tienen portátiles como herramienta tecnológica.

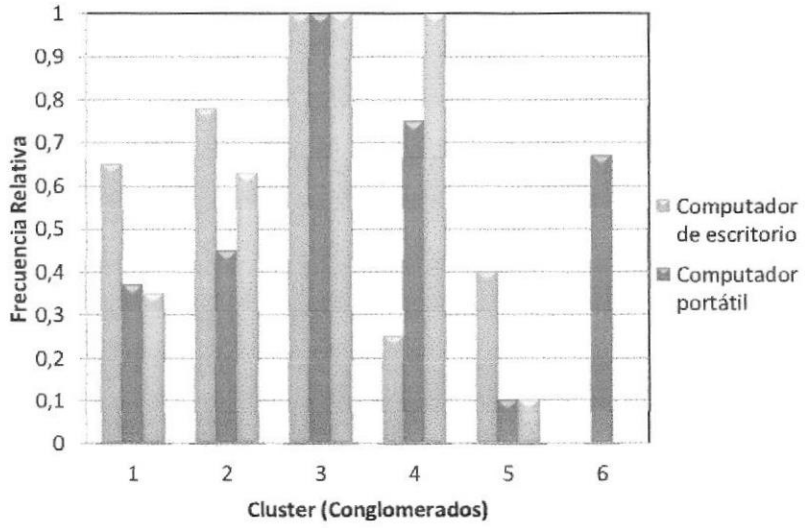


Figura 46. Distribución de "Recursos tecnológicos" utilizados por el estudiante en cada conglomerado.

Descripción de las características por cada cluster o conglomerado

Grupo 1: Son estudiantes hombres en su mayoría (70%), con aproximadamente **22 años de edad**, un promedio universitario **de 65 puntos** y un promedio en el colegio de **18 puntos,** cuyo ingreso promedio en **dólares en su hogar es ($ 581),** donde mayoritariamente son hombres de las **especializaciones fima o informática**, provenientes en su mayoría de ***colegios fiscales, son solteros,*** sin hijos, un 65% ***también tiene computadora de escritorio, sólo un 35% posee una computadora portátil***, la mayor concentración de estudiantes está los que tienen entre 3 y 6 materias aprobadas por período académico.

Grupo 2: En este grupo se clasifican los estudiantes ***de género masculino con especializaciones fima en el 15% de los casos e informática en el 70%,*** son además personas provenientes ***de colegios particulares, solteros,*** que por el momento ***no tiene hijos, posee computadora de escritorio (79%), mientras que la portátil es usada por el 45% de los estudiantes y el 61% tiene acceso a internet desde el hogar*** , la mayor concentración de los estudiantes se encuentra entre los que aprobaron ***entre 3 y 5 materias***, por el promedio de materias que aprueba cada período académico, tiene un promedio en la universidad ***de 65 puntos y un promedio de ingreso familiar mensual de 449 dólares.***

Grupo 3: Estudiantes hombres (80%), de informática, de contabilidad y de otras especializaciones, provenientes de colegios "fiscales" que son "casados" y "tienen hijos", tiene el servicio de internet en su hogar, tiene computadora portátil, y escritorio, sin embargo tienen en promedio dos materias aprobadas por periodo académico y un promedio de calificaciones en la universidad de 62 puntos ", con un promedio de calificaciones de 64 puntos y el promedio de edad de 30 años.

Grupo 4: Estudiantes hombres, de especialización en el colegio Fima o Informática provenientes de colegios particulares, son casados y tienen hijos, poseen internet como servicio adicional en el hogar, el 75% cuenta con computadora portátil, y sólo un 25% tiene computadora de escritorio, en las materias aprobadas en promedio durante sus años académicos, la concentración de las observaciones en ésta variable se encuentra entre 2 y 8 materias aprobadas, con el promedio de puntuación universitaria de (62.25), estos tienen un ingreso familiar de 456 dólares en promedio mensual.

Grupo 5.- En este grupo el 90% son estudiantes hombres y el 10% mujeres, que tienen especialización en bachillerato bien sea fima, informática o materias relacionadas a la electrónica, estudiantes casados en un 70%, y con hijos, provenientes de colegios particulares en un 50%, el restante proveniente de colegios fiscales y municipales, cuya calificación promedio es de 68 puntos en la universidad, la cantidad de materias que aprueban en promedio por período académico están entre 3 y 7 materias, además sólo un 40% tiene computadora de escritorio, y sólo un 10% tiene tanto computadora portátil

como interne, la mayor concentración de datos se establece para los que tienen entre 2 y cuatro años en la universidad.

Grupos 6 (13%) Aquí se han clasificado alumnos de género masculino (67%) graduados en colegios fiscales que son casados(67%), con hijos, calificación promedio en la universidad de 64 puntos, con un ingreso familiar mensual promedio de 320 dólares, con seis materias reprobadas en todo su tiempo en la universidad, el tiempo en la universidad apenas está entre 1 y 2 años, que NO poseen computadora de escritorio, pero si poseen portátiles (67%) y que no cuentan con servicio de internet en el hogar sin embargo el 75% de ellos tienen entre 6 y 10 materias aprobadas por período.

4.3.5. Análisis cluster para datos de estudiantes de período anual y de período semestral por separado.

Para mejorar la descripción de los grupos y considerando que en la matriz de datos que se utiliza para el análisis, se encuentran estudiantes tanto de período anual como de período semestral se dividirá estos dos grupos y se identificará cluster en cada uno de ellos.

El procedimiento utilizado tanto en los estudiantes que siguen un período anual como aquellos que tiene un sistema semestral, es el mismo que se mostró con anterioridad y las instrucciones son ejecutadas en lenguaje R.

Pasos ejecutados para matriz de datos de estudiantes con período anual

1.- Calcular la matriz de distancias utilizando el algoritmo daisy con la métrica "Gower"

2.- Aplicar el algoritmo "agnes" en la matriz de distancias utilizando los cinco métodos con los que puede trabajar.

3.- Realizar la comparación de los diferentes métodos y considerando el coeficiente de correlación definir el método que muestra una mejor estructura de los conglomerados.

Al realizar la comparación se obtuvo que el algoritmo agnes y el método "UPGMA" ó "average" es el que tiene una alta correlación cofenética.

Tabla 25. Comparación de los métodos jerárquicos y el coeficiente de correlación para los datos de estudiantes de sistema anual

Métodos	Coeficiente de correlación
Método UPGMA	0,73
Método Single Linkage	0,6179
Método de Ward	0,5894
Método Complete Linkage	0,6306
Método del promedio ponderado	0,7132

4.- Gráficamente se muestra el dendograma que genera el algoritmo "agnes" y el método "UPGMA"

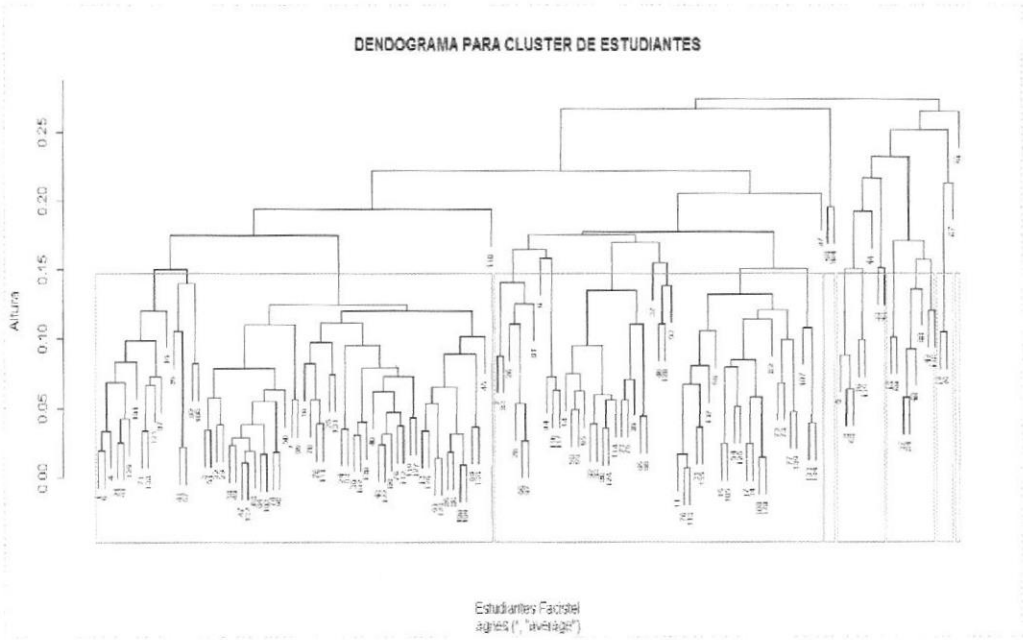


Figura 47. Dendograma obtenido al aplicar el método jerárquico

Observando la fig. 47, pueden identificarse algunos conglomerados, sin embargo para definir la cantidad de conglomerados apropiados se ejecutó el algoritmo "agnes" y el método "UPGMA" considerando diferentes cantidades de cluster.

Tabla 26. Comparación de los diferentes casos para cada número de conglomerados.

antidad de casos dependiendo de los grupos formados						
Grupos	2	3	4	5	6	7
	Casos	Casos	Casos	Casos	Casos	Casos
1	117	117	117	117	117	64
2	22	2	2	2	2	53
3		20	20	16	8	2
4			1	3	8	8
5				1	3	8
6					1	3
7						1

Al agrupar en siete conglomerados las observaciones hay un cluster que se formo sólo con una observación sin embargo al tener siete conglomerados se apreciará un mejor análisis sobre todo del cluster que tiene 117 elementos y que subdividió cuando la cantidad de cluster elegidos fue siete.

La división de los cluster se puede apreciar gráficamente en la fig. 48, donde se observa que el grupo 1 y 2 están más cercanos entre sí, son los que más casos tienen, pero también tienen mucha dispersión entre las observaciones dentro del grupo.

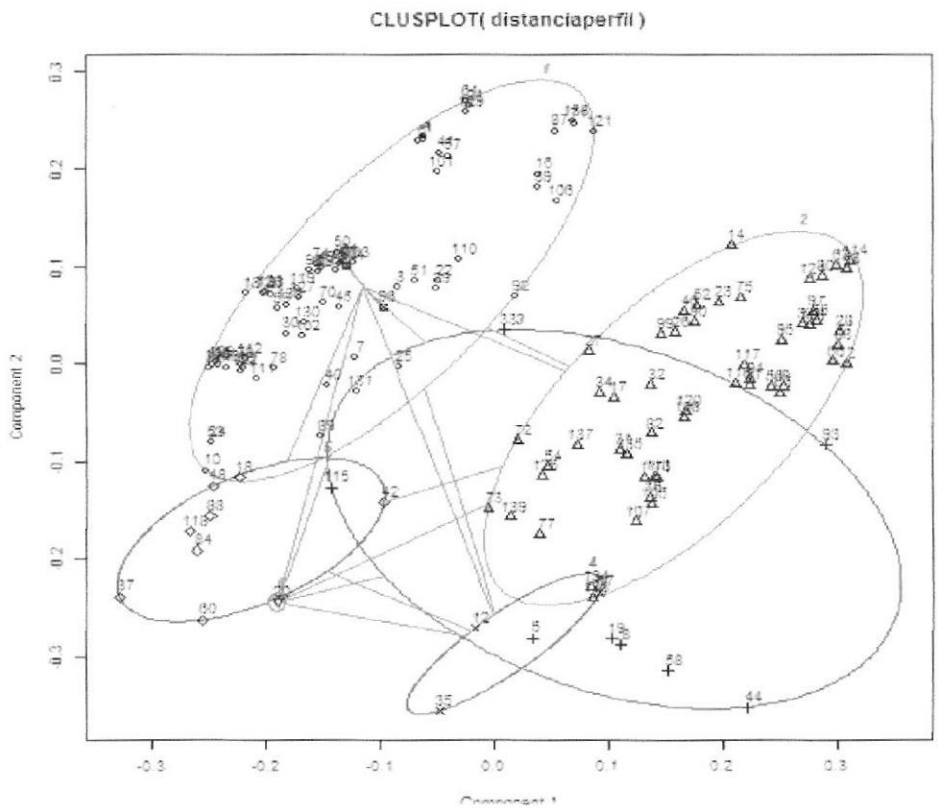


Figura 48. Diagrama de dispersión entre los conglomerados

5.- Finalmente se analiza tantos las características cuantitativas y cualitativas de cada grupo, los resultados se presenta en forma tabular y para tener una mejor visualización del comportamiento de las variables cuantitativas, se diseñaron diagramas de cajas, mientras que para las variables cualitativas se diseñaron los diagramas de las distribuciones de frecuencia.

Tabla 27. Análisis de las variables cuantitativas en los datos de estudiantes de sistema de estudio anual, por cada conglomerado.

Variables	Cluster						
	1	2	3	4	5	6	7
Calificación colegio	18	18	18	18	18	20	16
Edad	23	23	30	31	24	20	22
Ingreso	456	436	383	713	444	1000	376
Tiempo universidad	4.7	4.1	3.6	5.0	6.6	2.0	4.0
M_reprobadas	8.8	8.3	6.5	6.7	15.5	5.0	8.0
Promedio en la universidad	65	64	71	63	65	68	57
Indice _ materias	3.8	4.2	4.4	2.9	2.9	6.5	2.6

En el sistema de estudio anual, se detectaron siete grupos o cluster, en la tabla 28, se indican los promedios obtenidos de cada variable cuantitativa analizada, el grupo 3 y 4 se destacan por tener el promedio mayor en la variable edad, en el grupo 3 el promedio de edad es de 30 años, mientras que en el grupo 4 el promedio es de 31 años de edad.

En el grupo 4, se encuentran estudiantes que tiene un promedio de ingreso familiar de 713 dólares, siendo el mayor ingreso en comparación con los otros grupos, el grupo 1 en cambio tiene como promedio 9 materias reprobadas, es el grupo con mayor cantidad de

materias reprobadas, el grupo 6 se destaca por tener en promedio más cantidad de materias aprobadas (7) por año académico, mientras que el grupo 3 se destaca por tener el más alto promedio en calificación de la universidad.

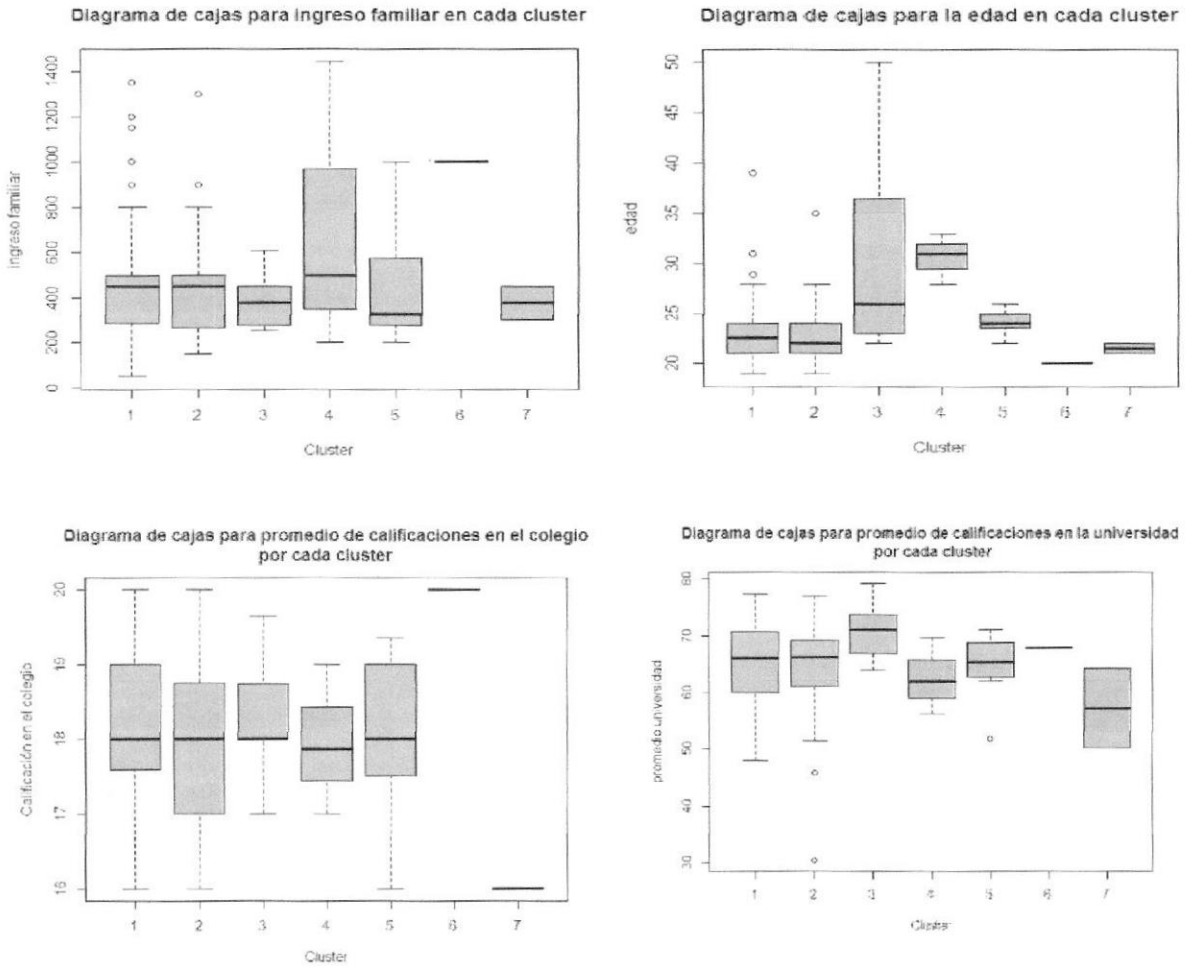


Figura 49. Diagramas de cajas para edad, ingreso familiar, calificación en el colegio, promedio de calificación en la universidad, para cada conglomerados. (Sistema de Estudio Anual)

La fig. 49 tiene los diagramas de cajas de algunas de las variables, puede observarse que el grupo cuatro tiene un 75% de estudiantes con ingresos inferiores a $1000 dólares, es el grupo que más dispersión presenta en la mencionada variable

El grupo 7 en cambio se destaca por ser el que tiene menores promedios de calificaciones en la universidad.

En la fig. 50, se observa que en el grupo 1 y 2 el 25% de los estudiantes tienen tres materias en promedio por año académico aprobadas, el grupo 4 y 5 se destaca porque el 75% de sus estudiantes tiene menos de 3 materias aprobadas por año lectivo.

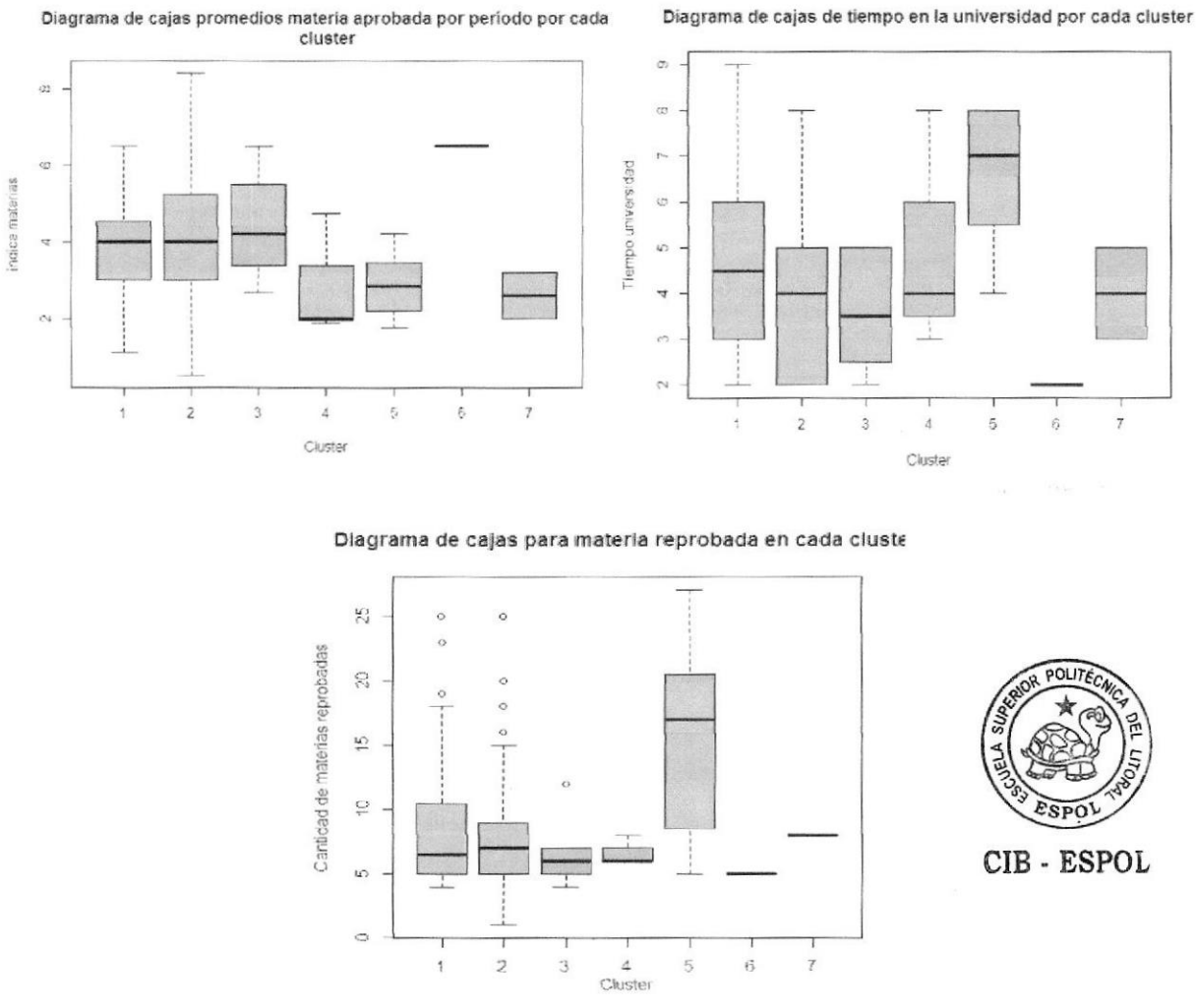


Figura 50. Diagramas de cajas para materia aprobada, ingreso familiar, tiempo en la universidad, materia reprobada, para cada conglomerados. (Sistema de Estudio Anual)

Sexo.

Tabla 28. Distribución de frecuencias del "sexo" del estudiante del sistema de estudio anual por conglomerado

Variable	Categoría	Cluster						
		1	2	3	4	5	6	7
Sexo	Femenino	0,38	0,30	0,38	0,33	0,62	1,00	0,00
	Masculino	0,62	0,70	0,62	0,67	0,38	0,00	1,00

En la tabla 28 y en la fig.51, se observa que el sexo femenino sólo predomina en el grupo 6, y el sexo masculino en el grupo 7, los cinco grupos restantes tienen hombres y mujeres, sin embargo la proporción de hombres es superior a la proporción de mujeres en el grupo 1 al 5.

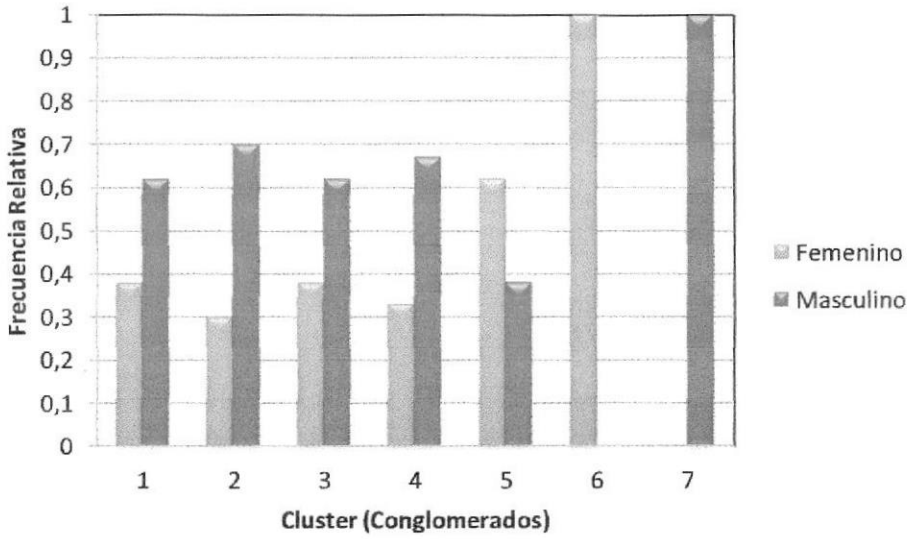


Figura 51. Distribución de frecuencias del sexo del estudiante en sistema anual, para cada conglomerado.

Estado civil.

Tabla 29. Distribución de frecuencias del "estado civil" en el estudiante del sistema de estudio anual por conglomerado.

Variable	Categoría	Cluster						
		1	2	3	4	5	6	7
Casado	**No**	1,00	1,00	0,12	0,67	0,12	1,00	1,00
	Si	0,00	0,00	0,88	0,33	0,88	0,00	0,00
Unión libre	**No**	1,00	1,00	1,00	0,33	1,00	0,00	0,50
	Si	0,00	0,00	0,00	0,67	0,00	1,00	0,50
Soltero	**No**	0,00	0,00	1,00	1,00	0,88	1,00	0,50
	Si	1,00	1,00	0,00	0,00	0,12	0,00	0,50
Tiene hijos	**No**	0,92	0,96	0,25	0,00	0,00	0,00	0,00
	Si	0,08	0,04	0,75	1,00	1,00	1,00	1,00

En la tabla 29 y en la fig. 52 se observa que los dos primeros grupos se encuentran formados por estudiantes "solteros", el grupo 3 y 5 son los que tienen un alto porcentaje de personas "casadas".

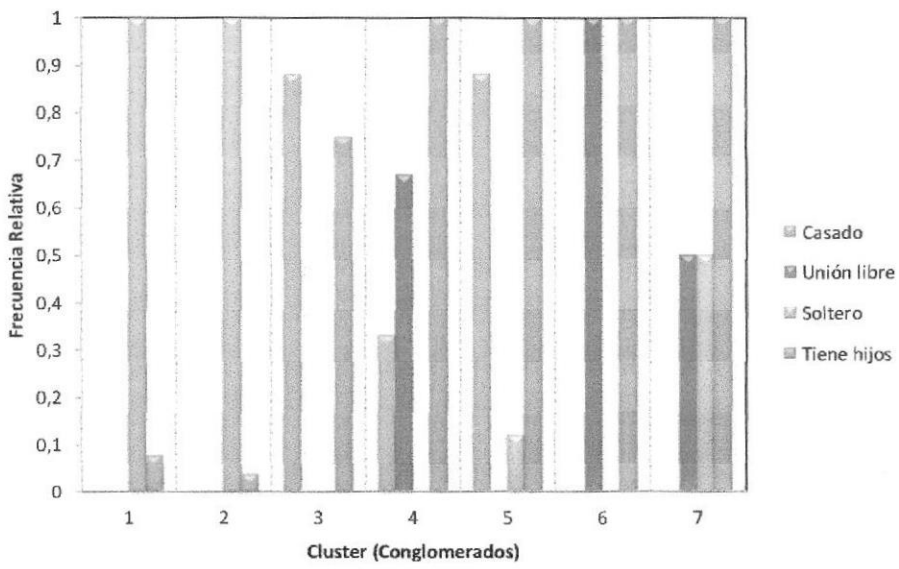


Figura 52. Distribución de frecuencias de "estado civil" del estudiante en sistema anual, para cada conglomerado.

Especialización.

Tabla 30. Distribución de frecuencias de "Especialización en el bachillerato" del estudiante en sistema de estudio anual por conglomerado.

Variable	Categoría	Cluster						
		1	2	3	4	5	6	7
Fima	**No**	0,81	0,74	0,75	1,00	1,00	1,00	0,50
	Si	0,19	0,26	0,25	0,00	0,00	0,00	0,50
Informática	**No**	0,28	0,57	0,38	0,33	0,12	0,00	1,00
	Si	0,72	0,43	0,62	0,67	0,88	1,00	0,00
Administración de sistemas	**No**	1,00	0,94	1,00	1,00	0,88	1,00	1,00
	Si	0,00	0,06	0,00	0,00	0,12	0,00	0,00
Electrónica	**No**	0,95	0,91	1,00	1,00	1,00	1,00	1,00
	Si	0,05	0,09	0,00	0,00	0,00	0,00	0,00
Técnico industrial	**No**	1,00	0,98	1,00	1,00	1,00	1,00	1,00
	Si	0,00	0,02	0,00	0,00	0,00	0,00	0,00
Contabilidad	**No**	1,00	0,98	0,88	1,00	1,00	1,00	1,00
	Si	0,00	0,02	0,13	0,00	0,00	0,00	0,00
Ciencias administrativas	**No**	0,95	0,94	1,00	1,00	1,00	1,00	1,00
	Si	0,02	0,06	0,00	0,00	0,00	0,00	0,00
Quibio	**No**	1,00	0,98	1,00	1,00	1,00	1,00	0,50
	Si	0,00	0,02	0,00	0,00	0,00	0,00	0,50
Otras	**No**	0,98	0,96	1,00	0,67	1,00	1,00	1,00
	Si	0,02	0,04	0,00	0,33	0,00	0,00	0,00

En el grupo 1, los estudiantes provenientes de especialización "informática" son los que existe en mayor porcentaje (72%), seguidos por los que tienen especialización "fima", en el grupo 5 , el mayor porcentaje son los que tienen especialización "informática" (88%), seguidos por los que tienen como especialización "Administración de sistemas".

El grupo 7 en cambio tiene la mitad de estudiantes de especialización "Quibio" y el otro porcentaje de especialización "Fi-ma"

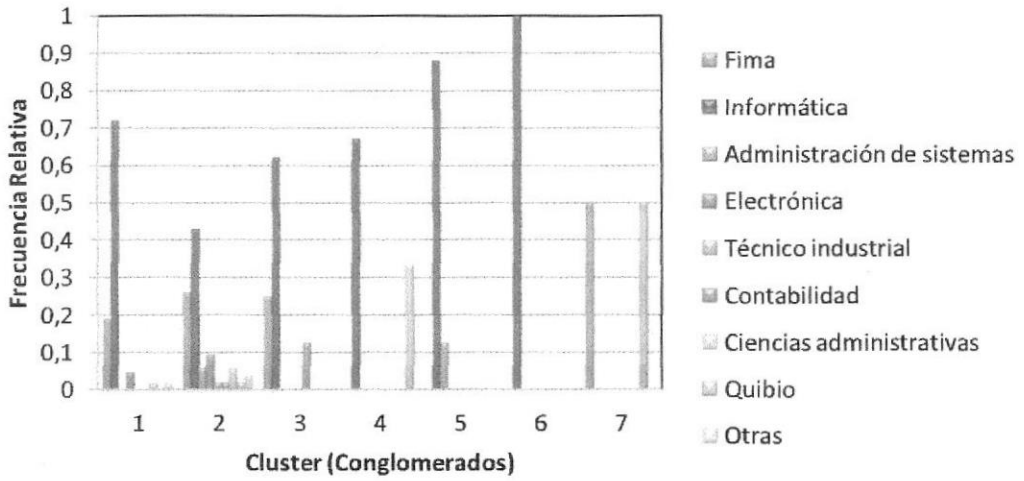


Figura 53. Distribución de frecuencias de "especialización en el colegio" del estudiante en sistema anual, para cada conglomerado.

En la fig. 53, se aprecia que el grupo 2 es el que tiene un porcentaje de estudiantes de cada una de las especializaciones: Fima, informática, administración de sistemas, electrónica, técnico industrial, contabilidad, ciencias administrativas.

Tipo de colegio

Tabla 31. Distribución de frecuencias de "tipo de colegio" en el estudiante del sistema de estudio anual, por conglomerado.

Variable	Categoría	Cluster						
		1	2	3	4	5	6	7
Fiscal	**No**	1,00	0,02	0,25	0,00	1,00	1,00	1,00
	Si	0,00	0,98	0,75	1,00	0,00	0,00	0,00
Particular	**No**	0,00	1,00	0,75	1,00	0,00	0,00	0,00
	Si	1,00	0,00	0,25	0,00	1,00	1,00	1,00
Municipal	**No**	1,00	0,98	1,00	1,00	1,00	1,00	1,00
	Si	0,00	0,02	0,00	0,00	0,00	0,00	0,00

El grupo 2, existen personas que vienen de colegios "fiscales" y un 2% de colegios "municipales", el grupo 1, 5, 6, y 7 están constituidos por estudiantes de colegios

"particulares", estos resultados se pueden apreciar gráficamente a través de diagramas de barras en la fig. 54.

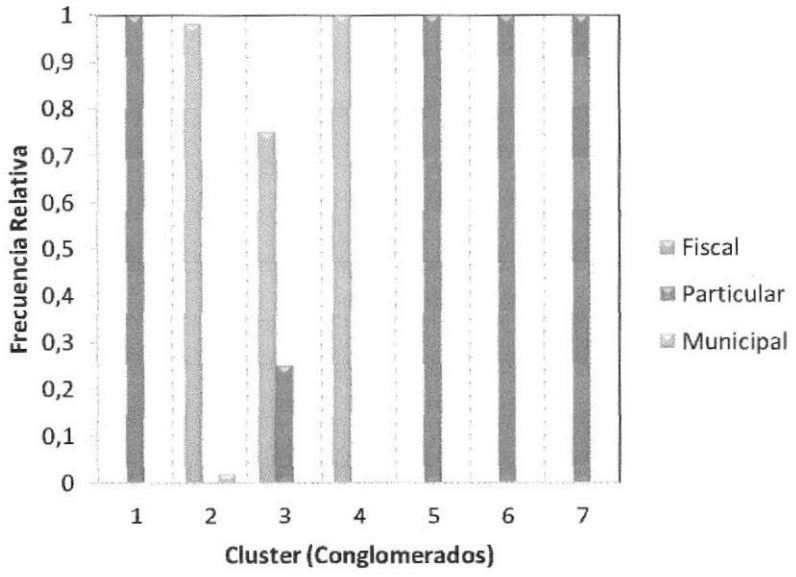


Figura 54. Distribución de frecuencias de "tipo de colegio" del estudiante en sistema anual, para cada conglomerado.

Recursos tecnológicos utilizados

Tabla 32. Distribución de frecuencias de "recursos tecnológicos" utilizados por el estudiante en sistema de estudio anual por conglomerado.

Variable	Categoría	Cluster						
		1	2	3	4	5	6	7
Computadora de escritorio	**No**	0,20	0,34	0,25	0,00	0,25	1,00	0,50
	Si	0,80	0,68	0,75	1,00	0,75	0,00	0,50
Computadora portátil	**No**	0,53	0,60	1,00	0,00	0,38	0,00	0,00
	Si	0,47	0,40	0,00	1,00	0,62	1,00	1,00
Internet	**No**	0,30	0,66	0,88	0,00	0,38	1,00	0,00
	Si	0,70	0,34	0,12	1,00	0,62	0,00	1,00

En la tabla 32. se aprecia que en el grupo 1 existe un 80% de estudiantes que tiene computadora de escritorio, un 70% tiene acceso a internet en el hogar y sólo un 47% tiene computadora portátil.

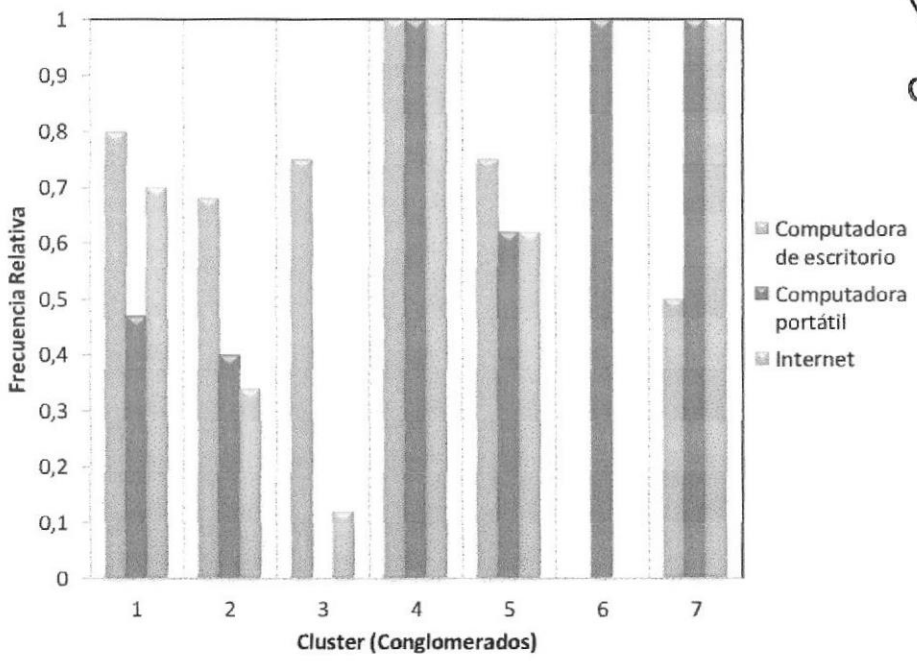


Figura 55. Distribución de frecuencias de "recursos tecnológicos" del estudiante en sistema anual, para cada conglomerado.

El grupo 4 es el que tiene en un 100% los tres recursos tecnológicos analizados, computadora portátil, de escritorio y acceso a internet desde el hogar

6. Interpretación de los datos, finalmente se identifica las características predominantes en cada grupo conformado

Descripción de las variables en cada uno de los conglomerados para estudiantes con sistema de estudio anual.

Grupo 1 (46% de casos). El 62% ***son hombres frente al 38% que son mujeres, todos solteros*** y en un 92% sin hijos, , en éste grupo se encuentran el 75% de jóvenes que tienen edades entre ***21 a 24 años***, un 72% se graduó en la especialización de informática, todos en ***colegios "particulares",*** el 75% de los jóvenes de este grupo se graduó de bachiller con calificación que están entre **17,5 a 19** puntos, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que en el 75% de las observaciones ***llegan a 70 puntos***, siendo el máximo puntaje en este grupo ***de 79 puntos***, la cantidad de materias que aprueban el 50% de los estudiantes por período académico ***es 4*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es **de 6**, con un ingreso familiar en promedio de 456 dólares, alrededor del 80% tiene computadora de escritorio, el 47% tiene computadora portátil y un 70% tiene acceso a internet desde su hogar, el 75% tiene entre 3 a 6 años en la universidad.

Grupo 2(38% de casos). ***El 70% son hombres frente al 30% que son mujeres***, todos ***solteros*** y en un 96% sin hijos, en éste grupo se encuentran el 75% de jóvenes que tienen edades ***entre 21 a 24 años,*** un 42% se graduó en la especialización de informática, y un 28% tiene como especialización "FIMA", el 98% ***provenientes de colegios "fiscales",*** el 75% de los jóvenes de este grupo se graduó de bachiller con calificación que están entre

17a 19 puntos, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 75% de las observaciones ***llegan a 69 puntos,*** la cantidad de materias que aprueban el 50% de los estudiantes por periodo académico ***es 4*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es de ***6*** , con un ingreso familiar en promedio de 436 dólares, alrededor del 70% tiene computadora de escritorio, el 40% tiene computadora portátil y un 38% tiene acceso a internet desde su hogar, el 75% tienen cinco años en la universidad.

Grupo 3(1% de casos). ***El 62% son hombres, el 88% casados***, un 75% dijo tener hijos, en éste grupo se encuentran el 75% de jóvenes que tienen edades entre ***24 a 35 años***, un 62% se graduó en la especialización de informática, y un 23% tiene como especialización "Fima", el 75% provenientes de ***colegios "fiscales***", el 50% de los jóvenes de este grupo se graduó de bachiller con calificación de ***18*** puntos, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones llegan a ***70 puntos,*** la cantidad de materias que aprueban el 50% de los estudiantes por periodo académico es ***4*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es **6,** con un ingreso familiar en promedio de 383 dólares, el 78% tiene computadora de escritorio y un 10% solamente tiene acceso a internet desde su hogar.

Grupo 4(6% de casos). El ***67% son hombres y el 33% mujeres , 33% son casados, y el 67% se encuentra en unión libre, todos dijeron tener hijos,*** en éste grupo se encuentran personas que tienen entre 29 a 31 ***años de edad***, un 67% se graduó en la especialización de informática, el 100% provenientes ***de colegios "particulares",*** el 50% de los jóvenes de este grupo se graduó de bachiller con calificación ***de 18 puntos***, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones **llegan a 60 puntos**, la cantidad de materias que aprueban el 50% de los estudiantes por período ***académico es 2*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es ***de 6***, con un ingreso familiar en promedio de 713 dólares, ***el 100% tiene computadora portátil, computadora de escritorio y acceso a internet desde sus hogares.***

Grupo 5(6% de casos). ***El 62% son mujeres, el 88% casados***, todos dijeron tener hijos, en éste grupo se encuentran personas que tienen ***24 años de edad,*** un 88% se graduó en la especialización de informática, y el otro porcentaje de especializaciones relacionadas a la administración de sistemas, el 100% provenientes de colegios "particulares", el 50% de los jóvenes de este grupo se graduó de bachiller con calificación de ***18 puntos,*** acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones llegan a 65 puntos, la cantidad de materias que aprueban el 50% de los estudiantes por período académico ***es 3*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es ***de 16,*** con un

ingreso familiar en promedio de 414 dólares, el 75% tiene computadora de escritorio y el 62% tiene computadora portátil, y tiene acceso a internet desde sus hogares.

Grupo 6(2% de casos). ***El 100% son nujeres el 100% casados***, todos dijeron tener hijos, en éste grupo se encuentran personas que tienen ***20 años de edad,*** un 100% se graduó en la especialización de informática, el 100% provenientes de colegios "particulares", el 50% de los jóvenes de este grupo se graduó de bachiller con calificación de ***17 puntos,*** acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que llegan a 70 puntos, la cantidad de materias que aprueban por período académico ***es 6*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad es ***de 5,*** con un ingreso familiar en promedio de 1000 dólares, el 100% tiene computadora portátil y tienen 2 años en la universidad.

Grupo 7(1% de casos).. ***El 100% son hombres, el 50% en unión libre y el mismo porcentaje solteros***, todos dijeron tener hijos, en éste grupo se encuentran personas que tienen ***21 años de edad,*** un 50% se graduó en la especialización fima, y el otro porcentaje en especialización quibio, el 100% provenientes de colegios "particulares", el 50% de los jóvenes de este grupo se graduó de bachiller con calificación de ***16 puntos,*** acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones llegan a 56 puntos, la cantidad de materias que aprueban el 50% de los estudiantes por período académico ***es 3*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es ***7,*** con un ingreso familiar en promedio de 376 dólares, el

50% tiene computadora de escritorio y el 100% tiene acceso a internet desde sus hogares, un mismo porcentaje posee computador portátil, el 50% de este grupo tienen 4 años en la universidad.

Análisis cluster para matriz de datos de los estudiantes con sistema de estudio semestral.

Pasos generales:

1.- Calcular la matriz de distancias utilizando el algoritmo daisy con la métrica "Gower".

2.- Aplicar el algoritmo "agnes" en la matriz de distancias utilizando los cinco métodos con los que puede trabajar.

3.- Realizar la comparación de los diferentes métodos y considerando el coeficiente de correlación definir el método que muestra una mejor estructura de los conglomerados.

Tabla 33. Comparación de los métodos jerárquicos y el coeficiente de correlación para los datos de estudiantes de sistema semestral

Métodos	Coeficiente de correlación
Método UPGMA	0,7583
Método Single linkage	0,7184
Método de Ward	0,6969
Método Complete linkage	0,7267
Método del promedio ponderado	0,7561

Al realizar la comparación se obtuvo que el algoritmo agnes y el método "UPGMA" ó "average" es el que tiene una alta correlación cofenética (0,758); los resultados pueden ser apreciados en la tabla 33.

4.- Gráficamente se muestra el dendograma que genera el algoritmo "agnes" y el método "UPGMA"

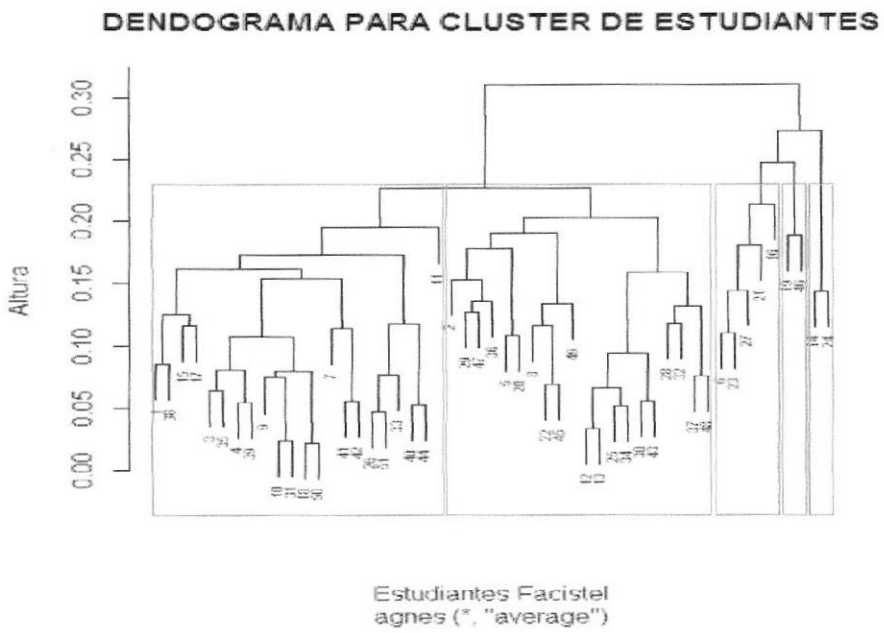


Figura 56. Dendograma al aplicar el método jerárquico en los estudiante con sistema de estudio semestral.

Establecen cinco grupos que se identifican claramente en el dendograma:

Grupo 1: 22 observaciones

Grupo 2: 20 observaciones

Grupo 3: 5 observaciones

Grupo 4: 2 observaciones

Grupo 5: 2 observaciones

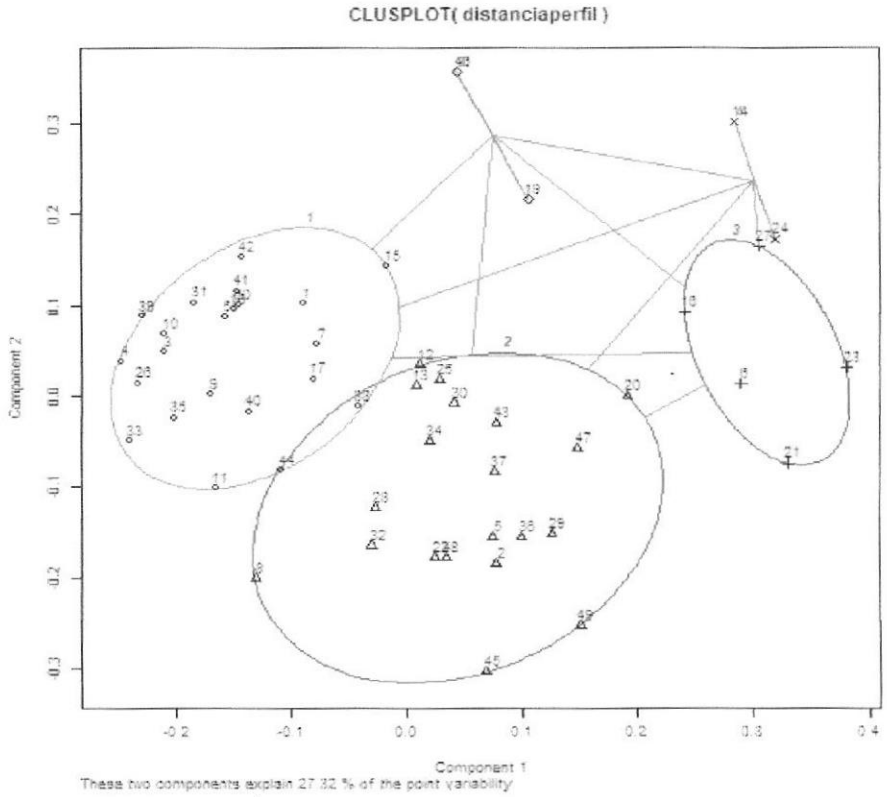


Figura 57. Diagrama de dispersión entre los diferentes conglomerados

5.- Analizar tantos las características cuantitativas y cualitativas de cada grupo, los resultados se presenta en forma tabular y en forma gráfica

Tabla 34. Análisis de las variables cuantitativas en los datos de estudiantes de sistema de estudio semestral, por cada conglomerado.

Cluster	*1*	*2*	*3*	*4*	*5*
califcole	18	18	17	18	17
edad	21	21	26	30	24
ingreso	444	952	431	600	458
m_repro	5.7	5.2	4.8	11.0	4.5
promuniv	64	66	64	61	70
indi_mat	6.59	8.10	7.00	0.86	7.00

En la tabla 34, están los cinco cluster identificados y el promedio de cada una de las variables cuantitativas analizadas.

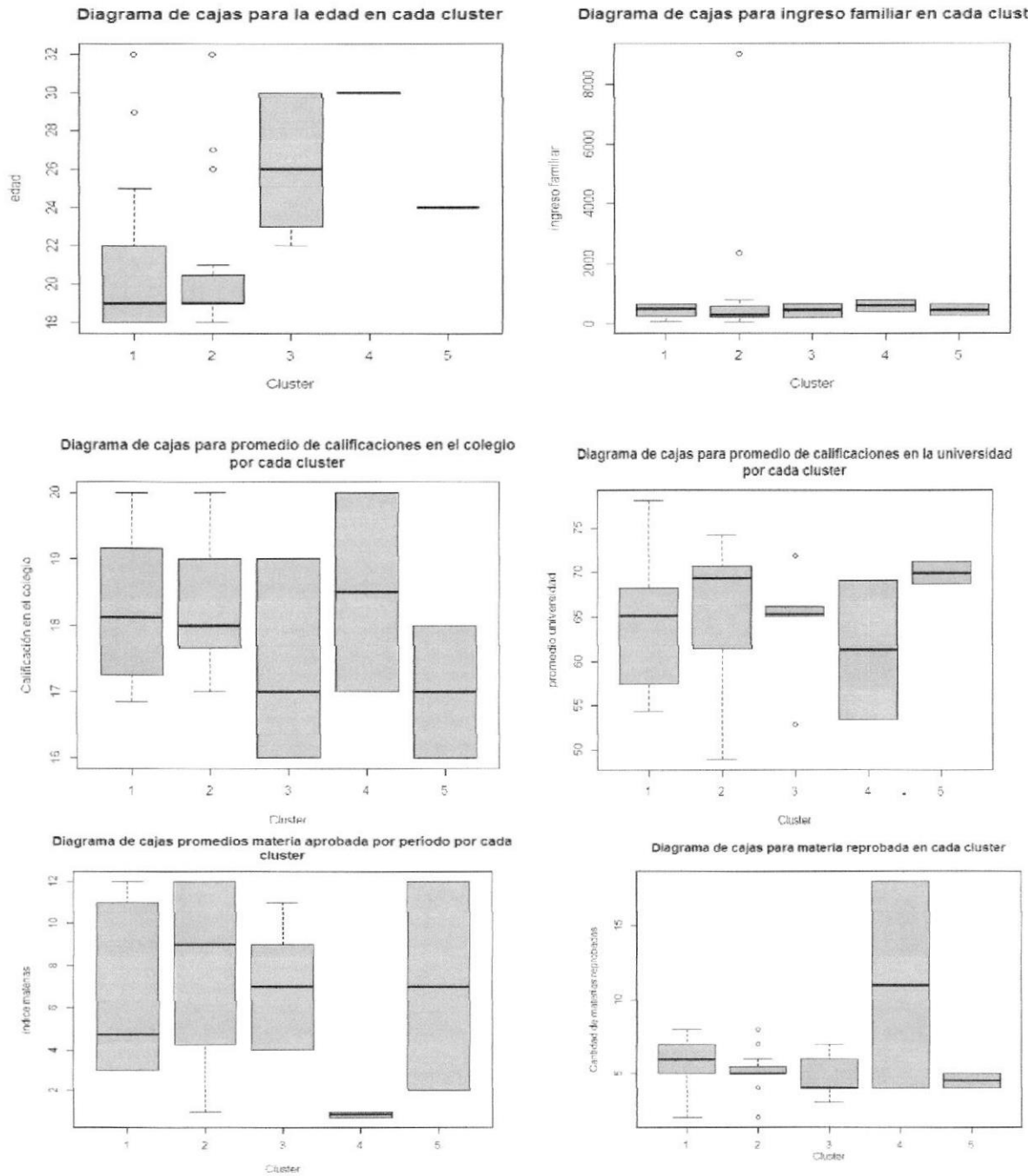


Figura 58. Diagrama de cajas para edad, ingreso familiar, calificación en el colegio, promedio de calificación en la universidad, índice de materias aprobadas, materia reprobadas", en los estudiantes en sistema de estudio semestral, por cada conglomerado.

En la fig. 63, se encuentran gráficamente diagramas de cajas para cada una de las variables cuantitativas, el grupo 2 se destaca por aquellos estudiantes cuyos ingresos familiares en promedio son superiores en comparación con los otros grupos, en el mismo grupo se encuentran los estudiante con mayor cantidad de materias aprobadas al año, considerando que el sistema de estudio de estas personas es semestral, ellos tienen la posibilidad de elegir 6 materias por cada semestre, el máximo número de materias en las que se inscriben es 12, el grupo 2 se destaca además por tener el 75% de sus estudiantes un promedio en la universidad inferior a 70 puntos, con un máximo de 74 puntos.

Totalmente contrario al grupo 2 es el 4, son aquellos estudiantes cuya edad promedio es de 30 años, es el grupo que en promedio tiene una mayor cantidad de materias reprobadas y el que tiene un índice de materias aprobadas mínimo, además considerando el promedio de la universidad, el 50% de los estudiantes tienen promedios inferiores a 60 puntos, el máximo puntaje es 70.

Descripción de variables cualitativas

Sexo

Tabla 35. Distribución del sexo en los estudiantes del sistema de estudio semestral

Variables	Categorías	1	2	3	4	5
Sexo	Femenino	0,45	0,35	0,00	0,00	0,00
	Masculino	0,55	0,65	1,00	1,00	1,00

En los grupos 3, 4 y 5 sólo se encuentran personas de sexo masculino, el grupo 1 y 2 tiene personas de sexo masculino y femenino, predominado en ambos grupos los hombres.

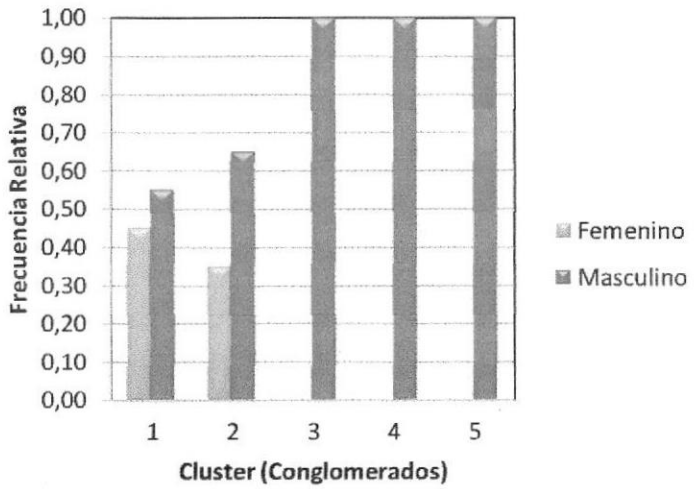


Figura 59. Distribución de frecuencias del sexo de los estudiantes en sistema de estudio semestral por cada conglomerado

Estado civil.

Tabla 36. Distribución del "estado civil" en los estudiantes del sistema de estudio semestral.

Variables	Categorías	Cluster				
		1	2	3	4	5
Casado	No	1,00	1,00	0,20	0,00	0,00
	Si	0,00	0,00	0,80	1,00	1,00
Unión libre	No	1,00	1,00	0,80	1,00	1,00
	Si	0,00	0,00	0,20	0,00	0,00
Soltero	No	0,00	0,00	1,00	1,00	1,00
	Si	1,00	1,00	0,00	0,00	0,00
Tiene hijos	No	0,96	0,96	0,40	0,00	0,00
	Si	0,05	0,05	0,60	1,00	1,00

Según la tabla 36 y la fig. 59 donde se muestran los resultados en cada uno de los grupos del estado civil del estudiante, el grupo 1 y 2 son los grupos formados por los estudiantes solteros, mientras que los tres grupos restantes son casados o se encuentran en "unión libre".

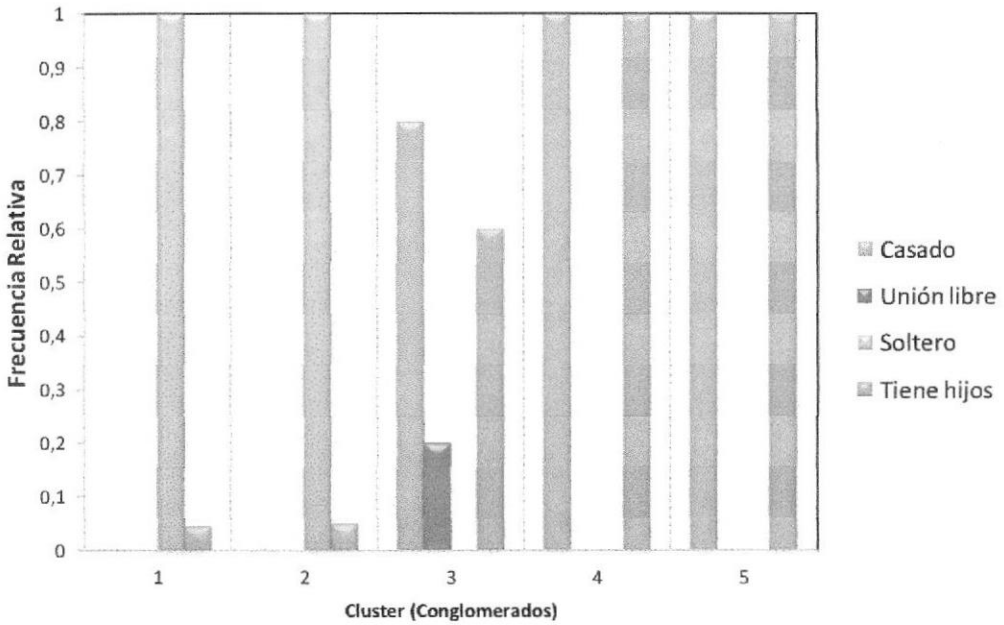


Figura 60. Distribución de frecuencias de "estado civil", de los estudiantes en sistema de estudio semestral por cada conglomerado.

Según estos resultados en el grupo 3, 4 y 5 se encuentran estudiantes que tienen bajo su responsabilidad los hijos.

Tipo de colegio

Tabla 37. Distribución del "tipo de colegio" en los estudiantes del sistema de estudio semestral.

Variables	Categorías	1	2	3	4	5
Fiscal	No	1,00	0,05	0,00	0,00	1,00
	Si	0,00	0,95	1,00	1,00	0,00
Particular	No	0,00	0,95	1,00	1,00	0,00
	Si	1,00	0,05	0,00	0,00	1,00

Grupo 1 y grupo 5 son estudiantes provenientes de colegios "particulares", mientras que el 2, el 3 y el 4 son estudiantes que se graduaron en colegios denominados "fiscales".

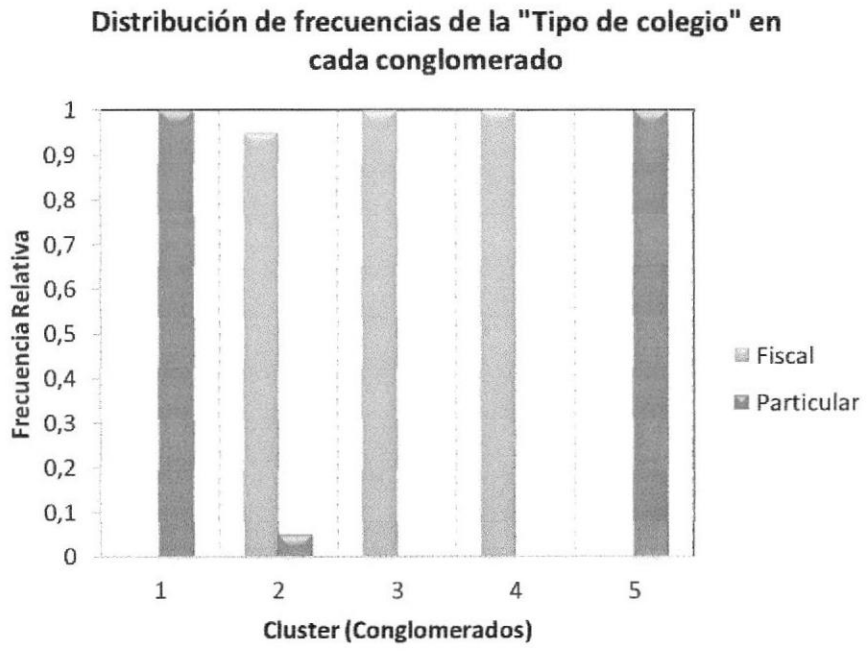


Figura 61. Distribución de frecuencias de "tipo de colegio", de los estudiantes en sistema de estudio semestral por cada conglomerado.

Tipo de colegio

Tabla 38. Distribución del "especialización" en los estudiantes del sistema de estudio semestral.

Variables	Categorías	Cluster				
		1	2	3	4	5
Fima	No	0,96	0,70	1,00	1,00	1,00
	Si	0,05	0,30	0,00	0,00	0,00
Informática	No	0,36	0,65	0,80	0,50	0,50
	Si	0,64	0,35	0,20	0,50	0,50
Administración de sistemas	No	0,77	0,90	1,00	1,00	1,00
	Si	0,23	0,10	0,00	0,00	0,00
Electrónica	No	0,96	0,90	0,40	1,00	1,00
	Si	0,05	0,10	0,60	0,00	0,00
Técnico industrial	No	1,00	0,95	1,00	1,00	1,00
	Si	0,00	0,05	0,00	0,00	0,00
Contabilidad	No	1,00	0,95	1,00	0,50	1,00
	Si	0,00	0,05	0,00	0,50	0,00
Ciencias administrativas	No	0,96	0,95	1,00	1,00	0,50
	Si	0,05	0,05	0,00	0,00	0,50
Otras	No	1,00	1,00	0,80	1,00	1,00
	Si	0,00	0,00	0,20	0,00	0,00

Analizando los resultados de la especialización del estudiante que estudia en sistema semestral, el grupo 1 tiene personas de "informática", "administración de sistemas". "electrónica" y "ciencias administrativas", el mayor porcentaje se encuentra en "informática".

El grupo 2 tiene personas de todas las especializaciones enunciada pero con mayor porcentaje en la especialización "informática" y "fima".

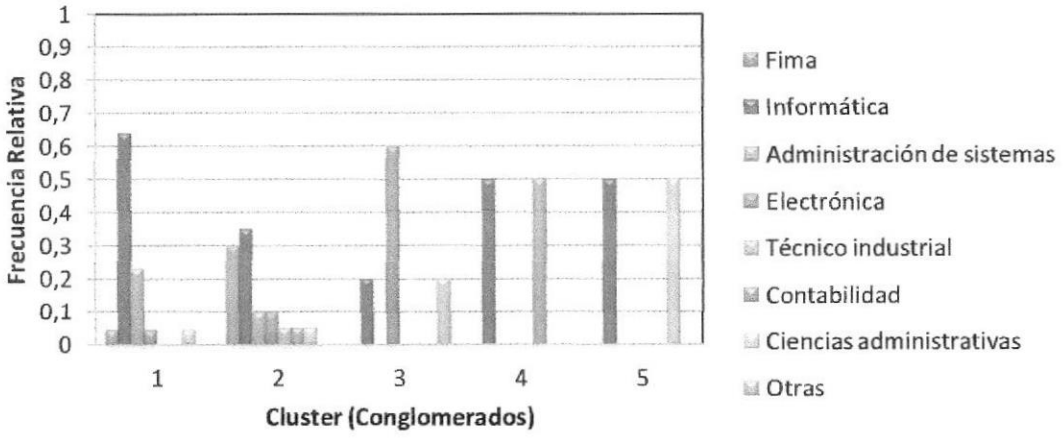


Figura 62. Distribución de frecuencias de "especialización en el colegio", de los estudiantes en sistema de estudio semestral por cada conglomerado.

El grupo 3 en cambio se destaca por tener personas graduadas en el bachillerato, con especialización en "electrónica".

Recursos tecnológicos

Tabla 39. Distribución de los "recursos tecnológicos" utilizados por los estudiantes del sistema de estudio semestral.

Variables	Categorías	Cluster				
		1	2	3	4	5
Computadora de escritorio	**No**	0,32	0,45	0,80	0,00	0,50
	Si	0,68	0,55	0,20	1,00	0,50
Computadora portátil	**No**	0,68	0,60	0,60	0,00	1,00
	Si	0,32	0,40	0,40	1,00	0,00
Internet	**No**	0,55	0,60	1,00	0,00	0,50
	Si	0,45	0,40	0,00	1,00	0,50

El 68% de los estudiantes del grupo 1 tiene computadora de escritorio, hay un 32% que posee computador portátil, mientras que sólo un 45% tiene acceso a internet; un poco más de la mitad de estudiantes del grupo 2 tiene computadora de escritorio, el 40% tiene computadora portátil, y el 40% tiene acceso a internet; el grupo 4 se destaca por tener los tres recursos analizados, computadora de escritorio, portátil, y acceso a internet

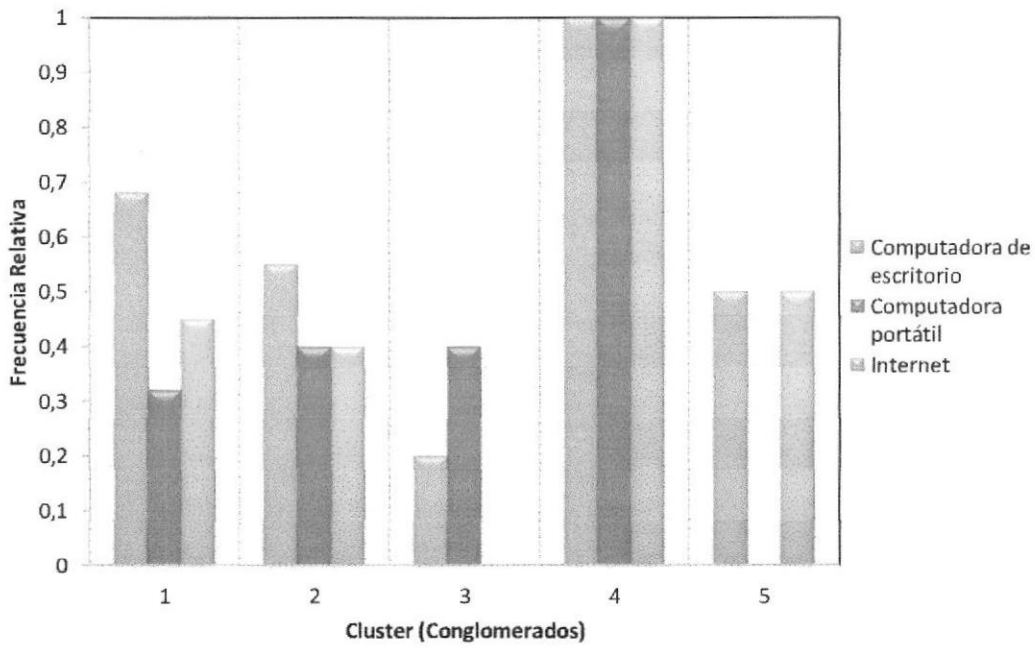


Figura 63. Distribución de frecuencias de "recursos tecnológicos", usados por los estudiantes en sistema de estudio semestral, por cada conglomerado.

6. Interpretación de los datos.

Descripción de las variables en cada uno de los grupos identificados sistema semestral

Grupo 1 (43% de casos). El 55% ***son hombres frente al 45% que son mujeres, todos solteros*** y en un 95% sin hijos, , en éste grupo se encuentran el 75% de jóvenes que tienen edades entre ***18 a 22 años***, un 65% se graduó en la especialización de

informática, todos en ***colegios "particulares"***, el 75% de los jóvenes de este grupo se graduó de bachiller con calificación que están entre **17 a 19** puntos, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que en el 75% de las observaciones ***llegan a 68 puntos***, siendo el máximo puntaje en este grupo de 78 puntos, la cantidad de materias que aprueban el 50% de los estudiantes por período académico ***es 5*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es **de 6**, con un ingreso familiar en promedio de 444 dólares, alrededor del 70% tiene computadora de escritorio, el 30% tiene computadora portátil y un 45% tiene acceso a internet desde su hogar.

Grupo 2(39% de casos). ***El 65% son hombres frente al 35% que son mujeres***, todos ***solteros*** y en un 95% sin hijos, en éste grupo se encuentran el 75% de jóvenes que tienen edades ***entre 19 a 20 años,*** un 35% se graduó en la especialización de informática, y un 30% tiene como especialización "FIMA", el 95% ***provenientes de colegios "fiscales",*** el 95% de los jóvenes de este grupo se graduó de bachiller con calificación que están entre 17.5 a 19 puntos, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones ***llegan a 70 puntos,*** la cantidad de materias que aprueban el 50% de los estudiantes por período académico ***es 9*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es de ***5*** , con un ingreso familiar en promedio de 952 dólares, alrededor del 58% tiene computadora de escritorio, el 40% tiene computadora portátil y un 40% tiene acceso a internet desde su hogar.

Grupo 3(10% de casos).- ***El 100% son hombres, el 80% casados*** y el porcentaje restante en unión libre, un 60% dijo tener hijos, en éste grupo se encuentran el 75% de jóvenes que tienen edades entre ***23 a 30 años***, un 20% se graduó en la especialización de informática, y un 60% tiene como especialización "Electrónica", el 100% provenientes de ***colegios "fiscales***", el 50% de los jóvenes de este grupo se graduó de bachiller con calificación de ***17*** puntos, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones llegan a ***65 puntos,*** la cantidad de materias que aprueban el 50% de los estudiantes por período académico es ***7*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es de ***4,*** con un ingreso familiar en promedio de 431 dólares, el 40% tiene computadora portátil y un 20% solamente tiene acceso a internet desde su hogar.

Grupo 4(4% de casos). - El ***100% son hombres, el 100% casados, todos dijeron tener hijos,*** en éste grupo se encuentran personas que tienen ***30 años de edad***, un 50% se graduó en la especialización de informática, y el otro porcentaje de especializaciones contables, el 100% provenientes ***de colegios "fiscales",*** el 50% de los jóvenes de este grupo se graduó de bachiller con calificación ***de 18,5 puntos***, acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones **llegan a 60 puntos**, la cantidad de materias que aprueban el 50% de los estudiantes por período ***académico es 1*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es

de 11, con un ingreso familiar en promedio de 600 dólares, ***el 100% tiene computadora portátil, computadora de escritorio y acceso a internet desde sus hogares.***

Grupo 5(4% de casos).- ***El 100% son hombres, el 100% casados***, todos dijeron tener hijos, en éste grupo se encuentran personas que tienen ***24 años de edad,*** un 50% se graduó en la especialización de informática, y el otro porcentaje de especializaciones relacionadas a las ciencias administrativas, el 100% provenientes de colegios "particulares", el 50% de los jóvenes de este grupo se graduó de bachiller con calificación de ***17 puntos,*** acerca del rendimiento en la universidad considerando su promedio de calificaciones destacamos que 50% de las observaciones llegan a 69 puntos, la cantidad de materias que aprueban el 50% de los estudiantes por periodo académico ***es 7*** y la cantidad de materias reprobadas en todo el tiempo que cursa la universidad de un 50% de estudiantes pertenecientes a este grupo es ***de 4,*** con un ingreso familiar en promedio de 458 dólares, el 50% tiene computadora de escritorio y un mismo porcentaje tiene acceso a internet desde sus hogares.

CAPÍTULO V .

VALIDACIÓN DEL ALGORITMO DE MINERÍA.

Luego de analizar los datos que anualmente genera la Universidad Estatal Península de Santa Elena, especialmente los del Sistema Académico y el Sistema de Bienestar Estudiantil, de aplicar las fases de explotación de conocimiento, en particular la técnica "cluster" , se llegó a obtener perfiles característicos de los estudiantes inscritos en la Facultad de Sistemas y Telecomunicaciones, facultad que sería el ejemplo piloto para replicar el mismo análisis en las otras facultades.

En este capítulo se presenta un análisis de los resultados obtenidos luego de aplicar las fases de minería de datos en los datos históricos y en los obtenidos a partir de la aplicación de la ficha socio-económica, se enuncia el proceso generar que se puede seguir para la aplicación de esta técnica en los datos de estudiantes de otras facultades.

5.1 Comprobación de los resultados generados por la técnica cluster

Como se mencionó inicialmente al aplicar la técnica cluster o de conglomerados permite obtener grupo de datos homogéneos, se debe cómo último paso validar la clasificación de las observaciones en cada uno de los grupos identificados.

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 dado que la disimilaridad con los objetos de su propio grupo es mucho menor que la disimilaridad con los objetos del grupo más próximo, 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 40.

Tabla 40. 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

En el primer análisis cluster ejecutado y presentado en el capítulo 3, con los datos académicos históricos y actuales del estudiante, existía más del 10% de datos faltantes, por lo tanto fueron eliminados dichos registros en éste primera aplicación del algoritmo cluster se utilizó la técnica particional donde inicialmente se identificó la cantidad de grupos, que en este caso fue: 11.

El análisis de silueta nos indico que el mayor coeficiente (0,35) es con 11 conglomerados, sin embargo este coeficiente indica una estructura débil en la clasificación tal como se define en la tabla 40.

El segundo análisis realizado para una muestra de estudiantes con variables académicas, sociales y económicas se utilizó la técnica jerárquica aglomerativa, se realizó el análisis de coeficientes de siluetas donde obtuvimos un coeficiente de silueta de 0,27 cuando el número de grupos en los que se dividía era 6, también con estructura débil.

Se aplicó además el análisis cluster en datos que estudiantes con sistema de estudio anual y aquellos que tenían sistema de estudio semestral, el coeficiente de silueta para aquellos que tenían sistema anual fue de 0,26 indicando una estructura débil, un coeficiente de silueta igual fue obtenido para el grupo de datos de sistema semestral.

A pesar que la estructura es débil al analizar los grupos por sistema de estudio, las características de cada grupo pudieron ser interpretadas como se enuncian a continuación

Sistema de Estudio Anual.

Para esta interpretación la variable "indice_materias" se discretizó a una variable nominal tal como se indica en la tabla 41.

Tabla 41. Codificación de la variable "indice_materias" para sistema de estudio anual

Indice_materias (cantidad de materias aprobadas por período académico)	Codificación	Categoría
[0-1]	1	Pésimo
(1-3]	2	Regulares
(3-5]	4	Bueno
(5-6]	5	Muy bueno

Fuente: Autor

Teniendo las consideraciones anteriores el análisis de cada grupo queda de la siguiente manera:

- **Grupo 1(46% de casos)**:Se encuentran los estudiantes tantos hombres como mujeres, cuyas edades están entre 21 a 24 años, que tienen desde 3 a 6 años en la universidad, son solteros, no tienen responsabilidades como padres, un gran porcentaje tiene computadora de escritorio y acceso a internet y pocos tienen computadora portátil, el promedio del 75% de estudiantes es de 70 puntos y el 50% de ellos han aprobado 4 materias por año académico. (***Jóvenes 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, con edades entre 21 a 24 años, el 50% con un promedio de 4 materias aprobadas por año académico, el 75% de los estudiantes tienen calificaciones promedio que no llegan al mínimo requerido para pasar una materia, la mayoría tiene computadora de escritorio, menos de la mitad portátiles y acceso a internet desde el hogar: ***estudiantes solteros y sin responsabilidades, "Buenos", con aceptable recurso tecnológico disponible, provenientes de colegios fiscales.***

- **Grupo 3 (1% de casos).-** Estudiantes más hombres que mujeres, casados y con hijos, que tienen entre 24 a 35 años de edad, el 50% de los estudiantes tiene en promedio de calificación en la universidad el mínimo requerido para la aprobación, la mitad de estudiantes de este grupo tienen 4 materias aprobadas por periodo, la mayoría tiene sólo computadora de escritorio y pocos tienen acceso a internet, se pueden considerar como ***Adultos, casados y con hijos de rendimiento "Buenos", y con bajo acceso a recurso tecnológico.***

- **Grupo 4: (6%).-** estudiantes con edades entre 29 a 31 años de edad, casados o en unión libre, el 50% de este grupo tiene un promedio de 60 puntos, en promedio la cantidad de materias aprobadas es 2, el 100% de estas personas

tienen computadora de escritorio, portátil y acceso al internet, por lo tanto pueden ser considerados como: ***adultos***, ***con responsabilidad de padres y conyugal, de rendimiento "Regulares" con alto recurso tecnológico disponible.***

- **Grupos 5 (6%):** En este grupo la mayoría son mujeres, son casadas, y con hijos, en promedio tiene 24 años de edad, en promedio el 50% tiene 3 materias aprobadas por año académico, además en un gran porcentaje tiene computadora de escritorio, portátil y acceso a internet desde el hogar, se puede considerar ***"Mujeres con responsabilidades, jóvenes, de rendimiento "Regulares", pero con alto recurso tecnológico disponible"***

- **Grupo 6(2%).-** Según las características de este grupo el perfil se etiquetaría como ***Mujeres jóvenes con responsabilidades, 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.***

Sistema de Estudio Semestral.

Para esta interpretación la variable "indice_materias" se discretizó a una variable nominal tal como se indica en la tabla 42.

Tabla 42. Codificación de la variable "indice_materias" para sistema estudio semestral

Indice_materias (cantidad de materias aprobadas por periodo académico)	Codificación	Categoría
[0-1]	1	Pésimo
(1-2]	2	Malo
(2-4]	3	Regular
(4-7]	4	Bueno
(7-12]	5	Muy bueno

Fuente: Autor

Grupo 1 (43%): 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 el promedio de materias aprobadas por año (5), se etiquetan de acuerdo a la tabla 42 como "Buenos", tienen además aceptable acceso a recursos tecnológicos disponibles.

Grupo 2 (39%): El siguiente perfil se describe para este grupo: 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 42, además con "aceptable" acceso a recurso tecnológico.

Grupo 3 (39%): 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, tienen bajo acceso a recursos tecnológicos.

Grupo 4 (4%): 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 altos recursos tecnológicos disponibles.

Grupo 5 (4%): 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.

Según estos perfiles podemos indicar que en la Facultad de Sistemas y Telecomunicaciones , los estudiantes tienen en general promedios inferiores a lo mínimo requerido (70 puntos), son estudiantes que podrían considerarse como "Buenos", por el promedio de materias que aprueban cada periodo, pero un tanto "regulares" por el promedio de calificación alcanzado en los años de estudios universitarios.

Hay grupos que se destacan por ser personas jóvenes con pocos años de graduados de bachillerato, que tienen ciertas comodidades económicas y acceso a ciertas herramientas tecnológicas, sin embargo son estudiantes considerados "Buenos", otros en cambio también son "Buenos" pero si poseen el acceso a las herramientas necesarias para su buen desempeño.

Además existe un grupo pequeño de personas que si bien tienen comodidades su rendimiento está afectado porque tienen otras responsabilidades como estar casado o tener hijos.

Existe entonces ciertas variables ajenas a la parte universitaria que debería considerarse para tomar decisiones, son "Buenos" pero no tienen acceso a herramientas necesarias para su desenvolvimiento por lo tanto pudiera implementarse un laboratorio donde se diera acceso a los estudiantes con el cobro de cierta tasa apropiada al estudiante.

Hay ciertos estudiantes que son considerados "regulares" o "pésimos" pero que su rendimiento se ve afectado por las responsabilidades que ya posee, se debería implementar un programa de orientación a los estudiantes que recién empiezan para brindarles motivación sobre su superación profesional.

Según este análisis el promedio de graduación del colegio no es un indicador para el buen rendimiento en la universidad, hay aspectos económicos, sociales y acceso a recursos tecnológicos que afectan el rendimiento estudiantil.

5.2 Formas de mejoras en el proceso planteado

Un futuro proyecto sería aplicar la técnica de minería de datos en cada una de las facultades existentes en la Universidad Estatal Península de Santa Elena

El proceso estándar que debería utilizarse es el propuesto en la fig. 64:

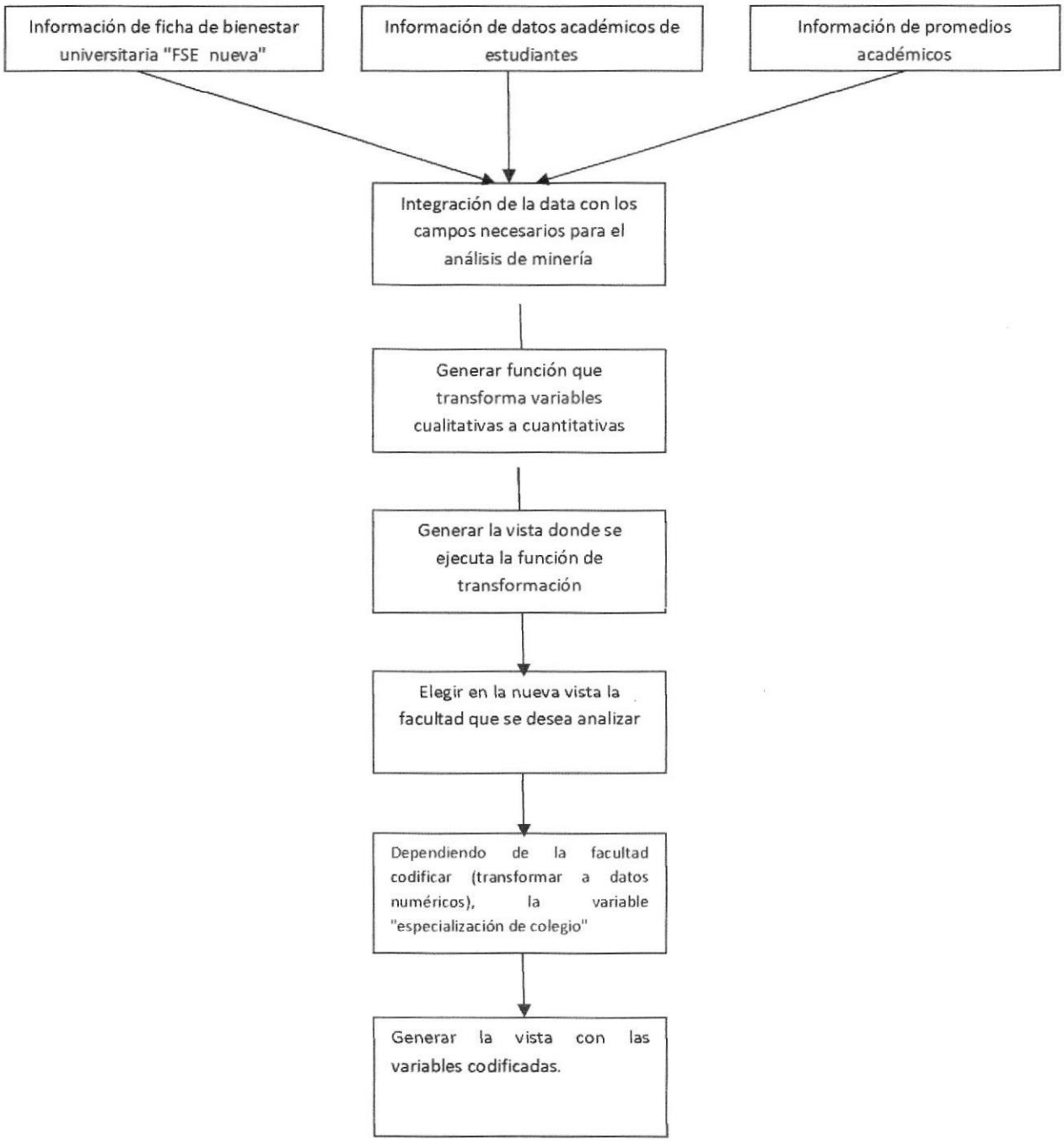


Figura 64. Proceso general que se debe ejecutar en PostgreSql

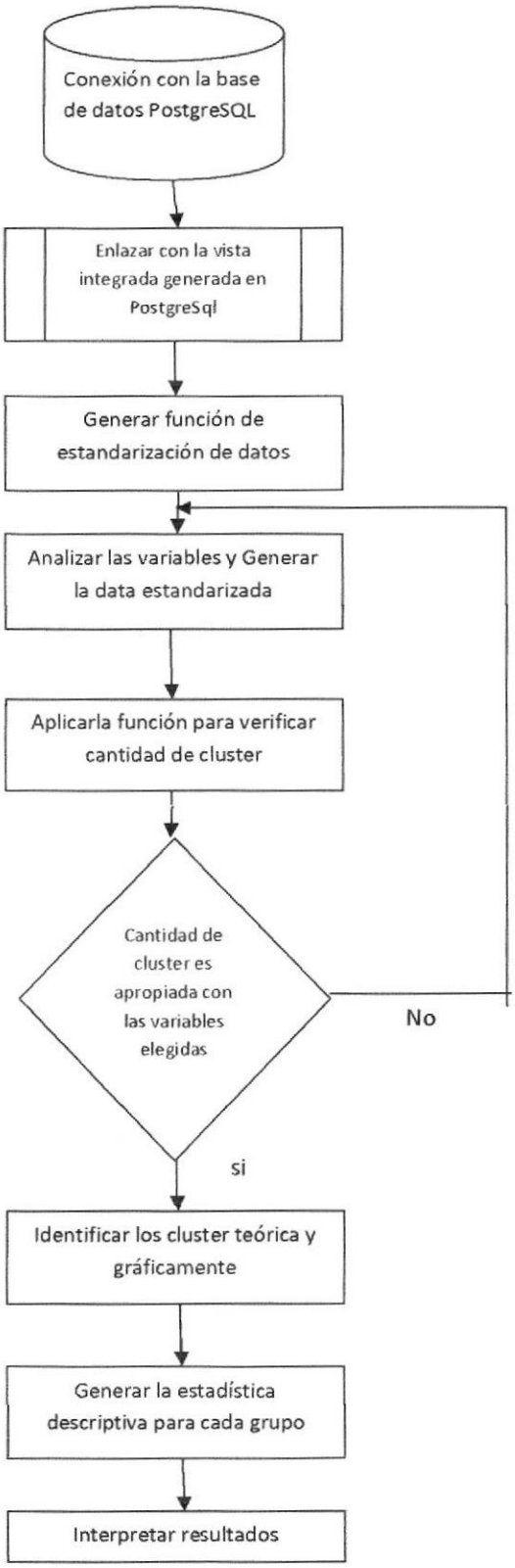


Figura 65. Procesos a ejecutar en R

Los pasos indicados en la figura 64 y 65 representan el proceso general que debe ejecutarse tanto en Postgresql (lugar donde se encuentran las tablas, funciones y vistas), como en R (aplicación para realizar análisis estadísticos), estas tareas a cumplir son repetitivas, si se desea realizar el análisis del resto de facultades existentes.

Aplicar otro de los algoritmos de minería denominado "árboles de clasificación" como modelo no sólo descriptivo sino más bien generar un modelo predictivo, con la finalidad de establecer una metodología para clasificar a un nuevo estudiante en algún grupo específico.

Un futuro proyecto es el diseño de una plataforma en ambiente web que muestre los resultados obtenidos en este trabajo, integrar las herramientas utilizadas y lograr la integración de la base de datos operacional con la base donde se almacenaran los registros que posteriormente formarán parte del proceso de minería de datos, mostrando resultados a los usuarios finales que requieran tomar decisiones.

Conclusiones

El trabajo realizado permite concluir sobre la importancia de tener datos de calidad, el uso de software libre para aplicar el análisis cluster, esta técnica fue aplicada en los primero en datos académicos de los estudiantes, en el primer momento no se posee variables relacionadas al aspecto social y económico; en un segundo momento ya se contaba con datos tanto académicos, sociales y económicos, finalmente fue aplicada la técnica de minería, en la última matriz de datos pero separando los estudiantes de sistema anual y sistema semestral.

- El lenguaje R tiene diversas librerías y funciones para ejecutar análisis estadísticos y presentación de resultados gráficos, es un software libre, de fácil uso, con recurso disponible en web, es un paquete que no se ha utilizado hasta la actualidad en la UPSE.

- Para el primer análisis, existía un porcentaje alto (más del 10%), de dato faltante específicamente en variables como: tipo de colegio, especialidades, y promedio de notas en el colegio, datos que me permitían analizar el rendimiento histórico del estudiante.

- 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 de dato.

- En la segunda aplicación de la técnica cluster se logró clasificar a los estudiantes en seis grupos, la técnica utilizada fue la jerárquica aglomerativa, para obtener una mejor descripción del perfil del estudiante fueron separados según el sistema de estudio: anual y semestral.

- El perfil de estudiantes del sistema académico anual los dividió en siete grupo 1: Jóvenes 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 recurso tecnológico a su disposición, de colegios particulares, el grupo 2, son estudiantes solteros y sin responsabilidades, "Buenos", con aceptable recurso tecnológico disponible, provenientes de colegios fiscales, el grupo 3: Adultos, casados y con hijos de rendimiento "Buenos", y con bajo acceso a recurso tecnológico, el grupo 4: adultos, con responsabilidad de padres y conyugal, de rendimiento "Regulares" con alto recurso tecnológico disponible, el grupo 5 son "Mujeres con responsabilidades, jóvenes, de rendimiento "malo", pero con alto recurso tecnológico disponible", el grupo 6, son mujeres jóvenes con responsabilidades, de rendimiento "Muy Bueno", con alto recurso tecnológico disponible, y el grupo 7: Hombres la mitad

soltero y la mitad en unión libre con rendimiento "Regular" con "aceptable" recurso tecnológico disponible.

- En el sistema de estudio semestral, los estudiantes fueron clasificados en cinco grupos, los dos primeros grupos, tienen la más alta proporción de casos; el primer grupo se puede etiquetar como "Buenos", y con "Aceptable" acceso a recursos tecnológicos, hombres y mujeres solteros y sin responsabilidades; el grupo 2 pueden ser etiquetadas como "Muy Buenos" jóvenes solteros y sin hijos, con aceptable recurso tecnológico disponible, el grupo 3 son adultos, casados, y con hijos, se considera como rendimiento "Buenos" con bajo recurso tecnológico disponible; el grupo 4, hombres casados, mayores de edad, con hijos, considerados "pésimos" en rendimiento con altos recursos tecnológicos disponibles; y el grupo 5 son hombres adultos, casados con hijos, de rendimiento "Buenos" y con poco recurso tecnológico.

- Según las últimas dos conclusiones 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.

- Existe además en esta Facultad 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", para este grupo no influye mucho el tener recursos tecnológicos a su disposición, en su rendimiento, sin embargo analizando otras características del mismo grupo 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.

- Según los resultados se infiere que poseer o no recursos tecnológicos como computadora de escritorio, portátil o acceso a internet desde el hogar puede influir positivamente en el rendimiento del estudiante, sin embargo existen otras características sociales como estar casado o tener responsabilidad de padres que también afectan pero en forma negativa en el rendimiento estudiantil, principalmente cuando son más jóvenes.

- En muchas ocasiones en esta institución de educación superior se ha mencionado que el rendimiento estudiantil está directamente relacionado con el desempeño y motivación del profesor en el aula, sin considerar otras variables como las que en este estudio se están analizando, 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.

Recomendaciones

- Se debe establecer un plan de análisis de la calidad del dato que se está almacenando en los diferentes sistemas que maneja la universidad, con la finalidad que los análisis estadísticos posteriores sean confiables, no presenten anomalías y permitan obtener resultados que ayuden adecuadamente a la toma de decisiones.

- Realizar la réplica del análisis del perfil socio-económico aplicando la técnica de cluster para cada una de las facultades de la universidad.

- Diseñar la aplicación que muestre resultados en tiempo real del análisis planteado en este trabajo.

- 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.

- Los directivos de la Facultad de Sistemas y Telecomunicaciones, deberían considerar las diferentes características del recurso humano que están formando, y planificar en conjunto con el departamento de bienestar estudiantil, charlas de orientación vocacional y de motivación para los estudiantes de la Facultad con la finalidad de concientizar en los estudiantes, la importancia de su preparación universitaria.

- Es importante la implementación de laboratorios acordes a las necesidades de los estudiantes, considerando que existen chicos que no poseen los recursos adecuados sin embargo se preocupan por mantener un rendimiento aceptable, y sobre todo para aquellos estudiantes que no tienen recursos y que tienen un rendimiento que los etiqueta como "regulares".

Referencias bibliográficas.

Ayala Gallejo, Guillermo. "Análisis de datos con R para Ingeniería Informática". Departamento de Estadístico e Investigación Operativa. Universidad de Valencia. Mayo 2008. Octubre 2011. <http://www.uv.es/ayala/docencia/ad/ad09.pdf>

Bernardis Alfredo, Pablo Reeb, y Sergio Bramardi. "Agrupamiento de Pozos de Petróleo en Base a Datos de Perforación". Libro de Resúmenes y trabajos completos. Septiembre 2009. Octubre 2011. <http://gab.org.ar/GAB2009/resumenesytrabajos/Resumenes_GAB2009.pdf>.

Berry Michael, Linoff Gordon S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Abril 2004. Hoboken. USA: Wiley. Octubre 2011. <http://site.ebrary.com/lib/upse/>

Diferencia entre dato, información y conocimiento. 2001. 15 octubre 2011. <http://www.gestiondelconocimiento.com/conceptos_diferenciaentredato.htm>.

Free Software Foundation's GNU. The R Project for Statistical Computing. 2011. noviembre 2011. <http://www.r-project.org/>

Hernández José, Ma. José Ramírez y César Ferri. Introducción a la minería de datos. Madrid: Pearson Educación, 2004.

Had David, Mannila H. y Padharic S. Principles of Data Minning, Massachusets: Institute of technology,2001.

Luan Jing. "Aplicaciones de minería de datos en la educación superior". IBM Corporation.2010.Marzo 2011. <ftp://ftp.software.ibm.com/common/ssi/ecm/es/imw14303eses/IMW14303ESES.PDF>

Meléndez M. Sistema de información, conocimiento y toma de decisiones. Julio del 2001. 17 Noviembre 2011. <http://www.sappiens.com/castellano/articulos.nsf/Comunicaci%C3%B3n/Sistemas_de_informaci%C3%B3n,_conocimiento_y_toma_de_decisiones/64A6FA79630082CC41256A9A00328FB0!opendocument>

Myatt Glen Jhonson, Wayne P. Myatt. A Practical Guide to Data Viualization, Advanced Data Mining Methods, and Applications. Marzo 2009. Hoboken, NJ, USA: Wiley. Noviembre del 2011 <http://site.ebrary.com/lib/upse/>

Pérez César, Daniel Santín. Minería de Datos. Técnicas y Herramientas. 2007. Madrid: Thomson Ediciones Paraninfo. <http://books.google.com.ec/books?id=wz-D_8uPFCEC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false >.

PostgreSQL Global Development Group. "PostgreSQL". 2011. Noviembre 2011. <http://www.postgresql.org/about/>

Prabhu S., N Venkatesan . Data mining and warehousing. 2007. Delhi: New Age International. Diciembre del 2011. <http://site.ebrary.com/lib/upse/Doc?id=10323316&ppg=1>

Schmitz Carsten. LimeSurvey. 20 diciembre 2011. <http://www.limesurvey.org/>

UPSE. Plan Estratégico Institucional de la Universidad Estatal Península de Santa Elena. Diciembre del 2010. Santa Elena.

Vallejo Raúl. Manual de escritura académica. Guía para estudiantes y maestros. Quito: Corporación Editora Nacional, 2006