ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE INGENIERÍA EN ELECTRICIDAD Y COMPUTACIÓN

TESIS DE GRADO

"SISTEMA ESTADÍSTICO INFERENCIAL APLICADO A LAS ENCUESTAS DEL CENACAD PARA FACILITAR LA TOMA DE DECISIONES"

Previa a la obtención del titulo de:

INGENIERO EN COMPUTACIÓN ESPECIALIZACIÓN SISTEMAS DE INFORMACIÓN
INGENIERO EN COMPUTACIÓN ESPECIALIZACIÓN SISTEMAS MULTIMEDIA

PRESENTADA POR:

GISSELLE MARÍA GUERRA DELGADO
JORGE GONZALO SÁNCHEZ VALAREZO

GUAYAQUIL - ECUADOR

2007

AGRADECIMIENTO

A todas las personas que de uno u otro modo colaboraron en la realización de este trabajo y especialmente al MSIG. Fabricio Echeverría Briones Director del Tema de Tesis, MSC. Carmen Vaca y al MSC. Guido Caicedo, vocales.

DEDICATORIA

A Dios

A nuestros padres

TRIBUNAL DE GRADO

PRESIDENTE

Ing. Holger Cevallos Ulloa

DIRECTOR DE TESIS

MSIG. Fabricio Echeverría

MIEMBROS PRINCIPALES

MSC. Carmen Vaca

MSC. Guido Caicedo

DECLARACIÓN EXPRESA

"La responsabilidad por los hechos, ideas y doctrinas expuestas en esta tesis, nos corresponden exclusivamente; y, el patrimonio intelectual de la misma, a la Escuela Superior Politécnica del Litoral"

(Reglamento de exámenes y títulos profesionales de la ESPOL)

Gisselle María Guerra Delgado

Jorge Gonzalo Sánchez Valarezo

RESUMEN

En la actualidad existen diversas maneras de interpretar los datos que se presentan en las evaluaciones realizadas por el CENACAD. Sin embargo, en la búsqueda por obtener datos fidedignos se ha tomado la iniciativa de fusionar dos ramas muy importantes, como lo son: la Estadística y la Minería de Datos para encontrar de esta manera patrones de conocimiento que sirvan para resolver inquietudes y a su vez tomar decisiones educativas.

La estadística inferencial permite evaluar datos de manera tal que se puedan obtener conclusiones que ayuden a beneficiar a quienes los están estudiando, en este caso los directivos interesados en verificar la calidad educacional de los profesores, materias que se dictan dentro de la ESPOL.

El presente proyecto de tesis está dirigido a implementar 4 modelos estadísticos inferenciales que permitan discriminar los datos atípicos y presentar las diversas formas de variaciones que puedan presentarse en las encuestas del CISE. Los modelos a ser estudiados serán:

- Análisis de Correspondencia
- Escalado Multidimensional
- Análisis Factorial
- Análisis de Conglomerados

En el capítulo 3 se hace referencia al **Análisis de Correspondencia que** es una técnica descriptiva, la cual presenta la frecuencia o aparición de 2 o más variables cualitativas que se encuentran en un conjunto de elementos, las cuales al inicio de la investigación parecen carecer de vinculación, pero que mediante este estudio se encuentran relacionadas. La proximidad entre los puntos representados está relacionada con el nivel de asociación entre las variables estudiadas la cual finalmente es presentada mediante biplots que presenta la información de manera gráfica

El capítulo 4 se refiere al **Escalado Multidimensional**, que permite entender la estructura de los elementos analizados, además de describirlos e interpretarlos. Con este método se puede analizar similitudes o diferencias existentes entre los elementos de un conjunto de variables en este caso se realizará el estudio de 3 casos.

En el capítulo 5 se tratará al **Análisis Factorial,** que permite reducir las dimensiones de un modelo obteniendo nuevas variables independientes, las cuales permiten prever el valor de otras variables dependientes existentes en el conjunto de elementos que se está evaluando.

Para finalizar en el capítulo 6 se estudiará el **Análisis de Conglomerados** que será implementado para agrupar o clasificar los elementos en grupos homogéneos en función de similitudes o similaridades. Este método particiona datos, construye jerarquías de los elementos de un conjunto por su similitud y realiza la clasificación de las variables en grupos.

Con la realización de este proyecto se busca la obtención de formas de conocimiento para el CENACAD, en los resultados de la evaluación. Además, de ofrecer nuevas opciones para la toma de decisiones de los directivos de la ESPOL.

No se desea competir con otras herramientas existentes en el mercado sino más bien poder utilizar medios locales para poder reducir los costos si se implementaran estos modelos en el sistema.

ÍNDICE GENERAL

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La estadística y los sistemas informáticos son utilizados en la actualidad como herramientas principales para la toma de decisiones en temas de gran relevancia. Desde la economía hasta la arquitectura, pasando por la física y la astronomía, el uso de sistemas estadísticos ha servido para definir conocimientos exactos y claros, además de conclusiones exactas y significativas.

Basados en los resultados obtenidos en los diversos estudios, el CENACAD ha implementado varios modelos estadísticos importantes para evaluar y decidir sobre los datos obtenidos y en este proyecto de tesis 4 de los más importantes métodos en la estadística inferencial han sido desarrollados.

En los capítulos siguientes se mostrarán resultados obtenidos a partir de los análisis realizados y se podrá verificar que con la información es posible tomar decisiones que ayuden a mejorar el proceso académico de la ESPOL

CAPÍTULO 1

1 PLANTEAMIENTO Y ANALISIS CONTEXTUAL

Este capítulo realiza una breve introducción a los objetivos y planteamientos por el cual se decidió realizar este proyecto de tesis, además de reseñar características y funcionamiento del CENACAD.

Serán revisados conceptos básicos de la Estadística Inferencial y se hará referencia teórica a los análisis que han sido implementados para el desarrollo de este Sistema Estadístico Inferencial.

1.1 Objetivos y Justificación del Proyecto de Tesis

La ESPOL tiene como misión brindar al estudiante una educación integral lo cual es sostenido en la siguiente cita:

"*Formar profesionales de excelencia, líderes emprendedores, con sólidos valores morales y éticos, que contribuyan al desarrollo del país, para mejorarlo en lo social, económico y político. Hacer investigación, transferencia de tecnología y extensión de calidad para servir a la sociedad*"[2]

Esta misión que impulsa a cada estudiante a formarse como profesional se da con la participación activa de los profesores que imparten sus conocimientos día a día, pero ¿cómo saber si un profesor está llevando su cátedra de manera eficiente para lograr estos objetivos propuestos?, ¿cómo saber si los alumnos se sienten cómodos con quien les dicta una materia en su unidad educativa? ¿Cómo estar seguros que la materia dictada está brindando al estudiante conocimientos para una vida profesional?

Para resolver estas dudas las ESPOL en conjunto con el CISE crearon lo que actualmente se conoce como el CENACAD (Censo

Académico) el cual permite evaluar puntos educacionales importantes que luego de ser evaluados e interpretados presentan el nivel en el que se encuentran las diferentes unidades académicas y profesores[1].

La información presentada en el CENACAD presenta a los evaluadores información estadística, pero no trabaja en la disipación de los datos que no deben de ser tomados como referenciales para una decisión, por lo que este proyecto de tesis pretende mostrar resultados que puedan ayudar a la correcta interpretación de lo contestado por los estudiantes en las distintas encuestas.

1.1.1 Toma de Decisiones

La tecnología ha avanzado en los últimos años y en el mundo actual es importante contar con soluciones informáticas que den la posibilidad de tomar decisiones relevantes sobre la información obtenida mediante distintos procesos, sean estos encuestas, cuestionarios, pruebas de conocimiento, etc.

Sobre los datos presentados en el CENACAD para los administradores es primordial encontrar información que les permita tomar decisiones académicas que influyan en el mejoramiento del

nivel educacional de la ESPOL donde su visión se proyecta a ser líder y referente de la Educación en América Latina[2].

Las decisiones que sean tomadas pueden ser preventivas o correctivas como por ejemplo el cambio de unidad académica de un profesor o el cambio de políticas manejadas por un profesor dentro de un paralelo e incluso la reestructuración de la malla académica de materias como ha sucedido en los últimos años.

Por lo tanto los análisis estadísticos inferenciales presentados les permitirán obtener ideas claras y firmes luego de un breve análisis. De esta manera habrá un criterio que podrá ayudarlos para tomar decisiones efectivas.

1.1.2 Administración y Evaluación

Evaluar el desempeño de cada uno de los catedráticos de la ESPOL permite encontrar las fortalezas y debilidades de estos. De este modo el nivel académico puede ser medido en resultados confiables, generados a partir de quienes receptan semestralmente la cátedra.

Con los resultados presentados se puede realizar proyecciones futuras, pues es posible evaluar el rendimiento general y así verificar que todos los procedimientos educativos estén siendo cumplidos.

La optimización de la educación que la ESPOL como centro superior ha adquirido a través de los años puede ser administrada y evaluada para que los niveles de calidad no se vean afectados por una administración sesgada por el incorrecto uso de los recursos (en este caso los profesores).

1.1.3 Supervisión educacional

La información obtenida cuando los métodos han sido desarrollados y analizados puede ser supervisada por los decanos y directores para verificar que los niveles de enseñanza de los docentes sean los apropiados para el aprendizaje funcional de los futuros profesionales y que éstos cumplan tanto con la misión y visión de institución.

Las unidades académicas (facultades) podrán tomar los resultados presentados en el sitio web y guiarse con estos resultados, evaluar periódicamente a sus docentes y exigirles que brinden un alto nivel académico.

Tomar como referencia la información presentada en este proyecto de tesis daría como resultado que los profesores como las unidades académicas alcancen las normas de calidad requeridas por la universidad y en caso de ser necesario se preocupen por mejorar los niveles tomando medidas correctivas.

1.1.4 Mejora en la calidad educacional

Con todo lo mencionado anteriormente es indudable que la ESPOL precautelará la calidad educacional además que optimizará y perfeccionará todos aquellos detalles que puedan mostrar debilidades y que puedan presentarse dentro de las evaluaciones de profesores o de la cátedra.

El firme compromiso de ser siempre una universidad de excelencia (visión), hará que el sistema desarrollado y las decisiones tomadas en base a los resultados de los análisis ayuden en el crecimiento profesional y educacional.

1.2 El CENACAD

1.2.1 Descripción y objetivos del CENACAD

El Censo Académico en Línea (CENACAD) fue desarrollado y puesto en producción en un proyecto conjunto del Vice-Rectorado General, el Centro de Investigaciones y Servicios Educativos (CISE) y el Centro de Investigación Científica y Tecnológica (CICYT)[1].

El objetivo primordial de este sistema en línea, no es sólo dejar a un lado las evaluaciones realizadas en papel que se realizaban hasta el 2004, sino dar inicio a una nueva etapa tecnológica en busca de resultados que permitan una evaluación inmediata de los docentes y al mismo tiempo que exista una retroalimentación entre los profesores, directivos y estudiantes.

Entre los objetivos generales que podemos mencionar del CENACAD podemos citar[1].

- Contribuir a mejorar la evaluación académica de los docentes de la ESPOL a través de la creación de un Sistema Automatizado de Censo Académico en Línea

- Minimizar los errores que pueden presentarse en los datos de las encuestas presentadas en línea.

De esa misma manera el CENACAD se ha planteado objetivos específicos entre los cuales se mencionan[1]:

- Reducir el tiempo de obtención de reportes para la evaluación a docentes.

- Evaluación del desempeño del docente.

- Garantizar que los resultados presentados en el sistema son de alta confiabilidad y que pueden ser usados para realizar el proceso de evaluación al docente.

- Utilizar técnicas estadísticas que permitan inferir de manera correcta en los resultados presentados en las encuestas.

1.2.2 Funciones y Servicios

La primera intención de funcionalidad del CENACAD era orientarse hacia la digitalización de las encuestas en papel, de esta manera se optimizaría el trabajo de evaluación y recopilación de datos y bajo

este procedimiento se disminuirían los costos en la compra e impresión de los formularios. Sin embargo, ante la necesidad de un mayor beneficio y luego de 2 años implementado, las funciones del sitio han incrementado considerablemente y se ha ajustado a nuevos requerimientos. En su mayoría, estos requerimientos son evaluativos y fueron descritos en las secciones anteriores del presente trabajo.

Al inicio la implementación del CENACAD estuvo enfocada a las evaluaciones de los docentes de las mismas. En la actualidad han sido integradas casi todas las unidades académicas existentes como los módulos de inglés del CELEX, el Prepolitécnico, los Sistemas de Gestión de Calidad de la ESPOL, el Índice de Satisfacción de Registro y la Evaluación docente a Nivel de colegios, todo esto con el fin de automatizar estos procesos de evaluación.

Es importante recalcar que el nuevo modelo del CENACAD permite realizar la evaluación digital a los docentes de cualquier organización que requiera tomar decisiones basados en los datos obtenidos. Por lo tanto el CENACAD no sólo tiene fines educacionales sino también organizacionales.

1.2.3 Estructura de los Datos

La estructura de los datos del CENACAD está representada por tablas de entidad, de relaciones, de dimensiones y de hechos.

Para la realización de este Sistema Estadístico las tablas relevantes para los análisis que serán mostrados en capítulos posteriores se presentan a continuación:

- Tabla Preguntas
- Tabla Respuestas
- Tabla Formularios
- Tabla Paralelos

Además, de estas tablas ha sido necesario crear ciertas tablas de hecho, que son usadas como datos de entrada para los análisis entre ellas, entre las cuales se tiene:

Tabla: reportes.promedio_grupo_par_encuesta

Contiene el promedio que cada paralelo ha obtenido en una encuesta específica.

Tabla: reportes.promedio_unidad_enc

Contiene el promedio global de toda una unidad académica en una encuesta específica.

Tabla: reportes.promedio_preg_grup_par_enc

Contiene el promedio que cada pregunta ha obtenido en un paralelo y una encuesta específicos.

1.2.4 Interfaz

El sitio web del CENACAD fue desarrollado e implementado en PHP y tiene 3 interfaces importantes:

La interfaz pública, la cual está disponible para cualquier persona. En ésta se pueden consultar los reportes de los diferentes paralelos, sus promedios generales, las respuestas de los alumnos, etc.

La interfaz del encuestador, la cual está disponible para los alumnos de la ESPOL y es donde se presenta el formulario de encuesta para que el alumno conteste las preguntas respecto al rendimiento del docente.

La interfaz del administrador, la cual es restringida. En ésta interfaz se crean las encuestas además que es posible observar los reportes más especializados, como las redes neuronales y la clusterización.

1.3 La Estadística Inferencial

La Estadística es una rama de las matemáticas encargada de reunir, organizar y analizar datos generalmente numéricos, ayuda a resolver problemas y además permite luego de realizados los cálculos tomar decisiones que puedan beneficiar al contexto que las estudia.

La estadística y los procedimientos que con ella pueden realizarse han permitido de manera efectiva describir con exactitud datos de casi todas las ramas del conocimiento entre ellas: economía, psicología, política, física, biología, química, medicina e informática y ha servido como herramientas útil para encontrarle relación a muchos de los datos estudiados por estas ciencias.

En la actualidad para un estadístico el trabajo va mas allá de reunir datos y calcularlos, debe de encargarse además de la difícil tarea de interpretar toda la información obtenida en los procesos estadísticos para que esta tenga un valor realmente importante.

La Estadística se encuentra dividida en dos grandes ramas, cada una con un propósito específico:

- La Estadística Inferencial
- La Estadística Descriptiva

Nuestro estudio está basado en la *Estadística Inferencial* por lo que ampliaremos el concepto de la misma en las siguientes secciones de este capítulo.

1.3.1 Marco Teórico

La *Estadística Inferencial* es una parte de la estadística que sólo trabaja con algunos de los datos de una población existente dentro de un grupo de elementos observados; es decir solo toma una muestra ***n*** de los ***N*** elementos existentes. Una vez que se obtiene este reducido grupo de datos la *estadística inferencial* trata de encontrar aspectos o propiedades relevantes para toda la población y basados en ellos tomar decisiones. Para obtener dichos resultados es necesario fundamentarse en como se selecciona la muestra, como realizar la inferencia de los datos y además la confianza que se puede tener en la información obtenida.

Cabe recalcar que para obtener datos fiables el nivel de conocimiento y comprensión de estadística, matemáticas y probabilidades debe de ser alto pues se debe recordar que los procedimientos están basados en pequeñas muestras las cuales pueden sufrir variación.

Con toda la información proporcionada es notorio que la estadística inferencial puede proveer de modelos importantes para estudiar un sinnúmero de datos multivariantes.

Métodos tales como Componentes Principales, Escalado Multidimensional, Análisis de Correspondencia, Análisis de Conglomerados, Análisis Factorial, Análisis Discriminante, entre otros brindan a los estudiosos grandes posibilidades de entender y predecir el comportamiento que los datos pueden tomar dada una condición.

Grandes análisis han sido desarrollados a través de la historia y mediante este proyecto de tesis serán implementados aquellos considerados primordiales y necesarios para el estudio de las variables que se presentan en el CENACAD.

Los modelos multivariantes a tratar son:

- Análisis de Correspondencia
- Escalado Multidimensional
- Análisis Factorial
- Análisis de Conglomerados

El ***Análisis de Correspondencia*** cuya traducción viene del francés *"Analyse des Correspondances"*, es una técnica descriptiva que fue desarrollada por el estadístico francés Jean Paul Benzecri en los años 60, con el objetivo de analizar, definir, describir e interpretar datos que presenten relación y a los cuales puedan dársele una interpretación conjunta [8].

El análisis de correspondencia captó la atención de Pearson, Guttman y Fisher, expertos estadísticos, quienes lo estudiaron y trataron de mejorarlo; sin embargo y ante la falta de herramientas informáticas que desarrollen operaciones matemáticas complejas en esa época tuvieron que dejar ese objetivo a un lado.

No fue sino hasta 1980 con el boom de la era informática y con el desarrollo y mejora de ciertos softwares estadísticos que el estudio

de este tipo de análisis tomó mayor importancia, pues estos programas se presentaban a los investigadores de manera mas amigable y práctica de tal forma que se lograba una fusión *investigador – software* que de esta manera consiguió mejores resultados. Uno de los aspectos técnicos que contribuyó más a este acercamiento fue el desarrollo del sistema operativo Windows, en el campo de las microcomputadoras y la eventual inserción de los programas estadísticos en este entorno, permitiendo así el manejo de dichos paquetes de datos de una manera mucho mas versátil y fácil de entender y desarrollar. Fue así que Crivisqui en el año 1993 citó:

“Los investigadores de hoy en día se encuentran ante un ‘nuevo mundo’ en el cual tienen ante sí, una forma diferente de acceder al dato, informatizado, descentralizado, interactivo y cuyas capacidades gráficas se han desarrollado rápidamente”[9].

Cabe señalar que el análisis de correspondencia como tal no explica claramente qué se está estudiando o investigando. Es el investigador el que le da sentido a los resultados de los datos, según la información inicial presentada por el software utilizado y el conocimiento adquirido con la experiencia.

El ***Escalado Multidimensional*** o EMD por sus siglas en inglés, es un método estadístico utilizado para descubrir similitudes o diferencias que puedan existir entre varias variables estudiadas mediante las distancias que sean halladas entre éstas. [4]

El origen del EMD podría ser atribuido a Adolfo Quatelec (1796 – 1874) astrónomo y estadístico belga quien fue el pionero en la aplicación de la probabilidad a las Ciencias Sociales. Sin embargo, el nacimiento de este método está unido a los estudios de psicología experimental en los años 50. Otros estadísticos que han sido de gran apoyo en el desarrollo de las investigaciones para mejorar este análisis han sido Torgensn, Shepard, Kruskal, Gower, entre otros.

El ***Análisis Factorial (AF)*** tiene su nacimiento dentro de 2 grandes e importantes ramas. Una de ellas la Psicología y la otra las Matemáticas. Para algunos estudiosos como Gomes Bezares (1985), Zaltman y Burger (1980) el Análisis Factorial se desarrolló gracias a la Psicología pues para ellos existían muchos aspectos desconocidos relacionados a la personalidad y a la inteligencia de los humanos que eran necesarios estudiar y que en los estudios básicos era casi imposible encontrar razones u obtener conclusiones, por lo tanto fue

necesario desarrollar un método en el cual en base a ciertos factores se pudiera detectar la presencia de esquemas del comportamiento para de ésta manera desarrollar amplias teorías que encontraran explicaciones al comportamiento humano en general[6].

Sin embargo, fue el psicólogo inglés Charles Spearman (1863 - 1945), quien inicialmente empezó el estudio del Análisis Factorial pues intentaba resolver la disyuntiva acerca de la inteligencia, pues para muchos de sus colegas ésta se generaba bajo un solo aspecto o característica; mientras que él quería probar que la inteligencia era desarrollada por varias habilidades específicas según el individuo. Con lo expuesto se puede concluir que realmente fueron los psicólogos quienes le dieron una aplicación al estudio de este método, pero que fueron los matemáticos quienes ofrecieron los primeros planteamientos y procedimientos de cómo resolver un problema con varios factores[18].

El Análisis Factorial, entonces es una técnica estadística multivariante la cual busca resumir una matriz de datos que contiene varias variables. Lo que se desea conseguir es encontrar factores que representen el modelo de correlaciones existentes entre todas las variables que han sido observadas.

El **Análisis de Clusterización (AC)** también llamado Análisis de Conglomerados, Taxonomía Numérica o Reconocimiento de Patrones fue usado por primera vez en el año de 1939 por Tryon es una técnica utilizada en la estadística para crear grupos, éstos pueden ser homogéneos (con características símiles) o heterogéneos (con características disímiles)[19].

El éxito de muchas investigaciones y estudios realizado por expertos se encuentra en encontrar patrones similares entre los grupos de personas, objetos, productos o incluso comportamientos analizados.

Básicamente el análisis de clusterización busca encontrar mediante una variable o criterio definido grupos que muestren en su interior que son iguales y que son externamente diferente a los otros grupos también existentes, con esto se podría decir que ésta técnica es exploratoria, pues estudia cada uno de los individuos de manera tal que encuentren agrupaciones naturales que definan a una colección de datos propuestos.

En este proyecto de tesis el análisis de conglomerados busca dar al investigador grupos de individuos del gran número de observaciones

encontradas en los cuestionarios tomados del CENACAD que sin este estudio pueden carecer de significado, pero con la técnica se puede explicar y dar conclusiones importantes.

1.3.2 Utilidad y Ventajas comparativas a otras herramientas estadísticas existentes

El *Sistema de Estadística Inferencial* desarrollado representa no sólo una herramienta útil para la toma de decisiones de directivos, presenta además varias ventajas que la hacen más integral y funcional que cualquier otra herramienta en el mercado. Entre las ventajas se pueden mencionar:

No existe costo por licencias

Está implementado en herramientas de código abierto, por lo cual no incide en costos de compras de licencias ni gastos de mantenimiento de versiones.

Integración al CENACAD.

Los módulos creados en este proyecto de tesis se integran de manera directa al sistema ya creado. Cada módulo puede ser revisado como una opción más de las que se presentan en el sitio Web.

Análisis Simples

Otras herramientas existentes hacen del proceso de obtención de datos y evaluación un procedimiento para usuarios expertos en el tema estadístico, con este sistema y sólo ingresando al sitio es posible obtener los resultados esperados con sólo hacer un clic en la opción que se desea evaluar.

Ayuda gráfica y explícita

Cada reporte presentado en este sistema estadístico muestra una ayuda práctica para que el docente o directivo sepa como debe de interpretar los datos.

1.4 Justificación del Sistema Estadístico

La decisión de desarrollar nuevos módulos para el CENACAD se dio ante la búsqueda por brindar información dedicada del desenvolvimiento general de profesores, materias y unidades académicas de la ESPOL.

El estudio de variables con análisis estadísticos conocidos como los que se verán en los capítulos 3, 4, 5 y 6, permitirá avisorar

comportamientos regulares de los elementos estudiados y se obtendrán valores indicativos que deberán ser analizados.

Utilizar la combinación de la estadística inferencial con la minería de datos, ha permitido que los módulos de este sistema estadístico interactúen, permitiendo resultados que ayudarán sin duda a cumplir con todos los puntos tratados en este capítulo.

CAPÍTULO 2

2 IMPLANTACIÓN EN EL CENACAD

En el presente capítulo se explica cuál fue el proceso a seguir para implantar este sistema en el CENACAD. Se especificará la arquitectura general diseñada para integrar los nuevos análisis al sistema ya existente.

Se describirá cómo se realizan los procesos estadísticos y cómo se almacenan en la base de datos. Adicionalmente se explica cómo se integraron los nuevos reportes a la interfaz Web del CENACAD.

2.1 Procesos en la base de datos

Los análisis estadísticos desarrollados, son procesos que se ejecutan diariamente en el servidor del CENACAD, de esta forma se mantienen los datos actualizados para las consultas y los reportes en la interfaz web.

En general todos los procesos trabajan de la siguiente manera (Figura 2.1):

- Se obtienen los datos a analizar de la base de datos del CENACAD.
- Se realiza el análisis o proceso estadístico requerido.
- Los resultados del análisis se almacenan de nuevo en la base de datos y a su vez en nuevas tablas de hecho.

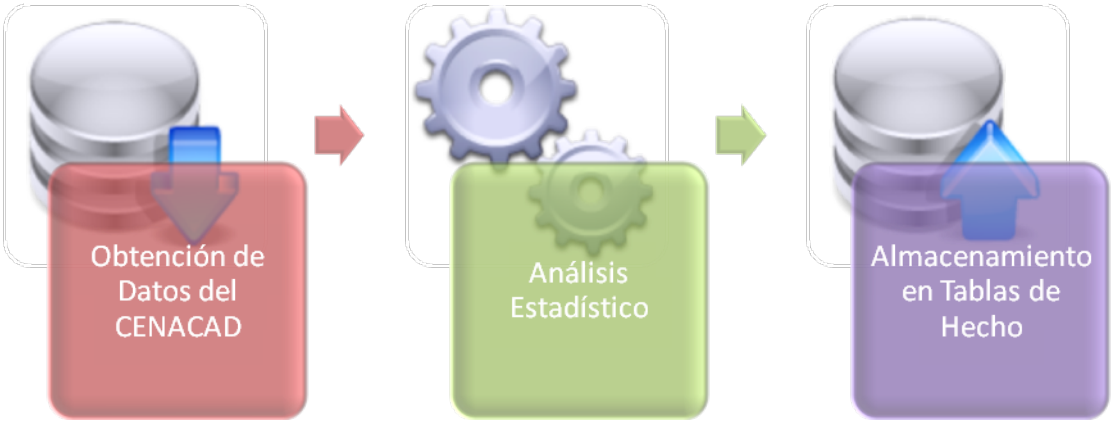


Figura 2.1 Arquitectura General de la Aplicación

Los procesos en detalle, junto con un flujo mucho más detallado por cada proceso, se muestran en el capítulo correspondiente a cada uno de los análisis.

Para la implementación de estos procesos, ha sido utilizado **Java** como el lenguaje de desarrollo. Entre las razones por las que esta plataforma de desarrollo fue elegida para la implementación de los análisis se encuentran:

Código abierto y multiplataforma: no está atado a licencias de ningún tipo, lo cual es requerido para la implantación en los servidores del CISE, los cuales tienen como sistema operativo a Linux.

Lenguaje de programación conocido: debido a los proyectos realizados en las diferentes materias durante la carrera de Ingeniería en Computación.

Eficiente: en cuanto al manejo de memoria y capacidad de procesamiento, lo cual es esencial, ya que los datos con los que trabajarán los algoritmos son grandes.

Diversidad de componentes: existen librerías ya implementadas bastante eficientes para el manejo de matrices y números complejos, las cuales son de gran utilidad para el desarrollo de este sistema.

Fácil Integración: el Cenacad tiene ciertos procesos que corren en Java, por lo tanto la integración es más sencilla.

Los algoritmos implementados, además de usar el JDK 1.4.2 de Sun, también usan dos librerías como apoyo, las cuales se detallan a continuación:

Jama: Paquete de Matrices Java

Esta librería fue desarrollada en conjunto por *The Math Works* (http://www.mathworks.com) y *the National Institute of Standards and Technology* (http://www.nist.gov). Provee un marco de trabajo eficiente para la construcción, manipulación y descomposición de matrices de números reales.

Esta librería fue elegida por su variedad de operadores sobre las matrices, y su gran capacidad para obtener los valores propios de una matriz, lo cual es crucial en los análisis estadísticos que se deben realizar.

Las clases y funciones que provee esta librería son usadas por las cuatro técnicas estadísticas implementadas en el presente trabajo de tesis.

A esta librería se le realizaron cambios para incrementar su fiabilidad al obtener los valores propios de ciertas matrices.

Jakarta Commons - Math

Esta librería fue desarrollada por *The Apache Software Foundation* (http://www.apache.org/), y es parte de un conjunto de librerías denominadas Jakarta-Commons, las cuales proveen una gran variedad de funcionalidades para aplicaciones Java.

Commons – Math (http://jakarta.apache.org/commons/math), provee diversas clases y funciones para el manejo de operaciones matemáticas y estadísticas complejas. En particular, lo que se usa de esta librería son las clases relacionadas con números complejos, pues en uno de los métodos (Análisis Factorial), es necesario trabajar con números con parte real e imaginaria.

2.2 Interfaz e Integración como Aplicación Web

La interfaz web para los reportes de los diferentes análisis estadísticos fue desarrollada en PHP4. Fue diseñada de esta manera porque el sitio web actual del CENACAD está implementado usando PHP.

Se buscó lograr una integración total del sistema estadístico con el CENACAD, por lo cual en las nuevas páginas se usó el mismo estilo de presentación, tipo de letra, colores, diseño existente.

El sistema estadístico además fue integrado al modelo en el que está desarrollado el CENACAD, es decir, usando la arquitectura MVC (Modelo, Vista y Controlador).

MVC es una arquitectura para el desarrollo web que divide a la aplicación en 3 capas: Modelo, Vista y Controlador [14].

- La capa de Modelo maneja la lógica del negocio
- La capa de Vista es la encargada de presentar la información al usuario

- La capa de Controlador es la interfaz entre las dos anteriores, es la que recibe las peticiones del usuario y se comunica con el Modelo para generar la Vista.

El CENACAD fue desarrollado usando un marco de trabajo (framework) que maneja la arquitectura MVC. Este framework se llama Mojavi[3] y provee las facilidades necesarias para separar adecuadamente los diferentes aspectos de la aplicación.

Todos los reportes desarrollados para el presente trabajo de tesis, se encuentran disponibles en línea en la sección de Administración del Sistema. Junto con cada reporte, se incluye además una página explicativa, a manera de ayuda en línea, para que al usuario se le haga mucho más sencillo entender e interpretar los reportes presentados por el sistema.

En este capítulo se trató acerca de los detalles de implantación, los cuales son comunes a todos los análisis y técnicas estadísticas implementadas. Se explicó la arquitectura tanto de los procesos como tal, corriendo en el servidor, como de l7os reportes mostrados en la interfaz web del CENACAD. Los cuatro siguientes capítulos

hablarán en detalle de cada uno de los análisis estadísticos implementados en el CENACAD.

CAPÍTULO 3

3 ANÁLISIS DE CORRESPONDENCIA

El presente capítulo explica la primera técnica estadística que fue aplicada a los datos de las encuestas del CENACAD: el Análisis de Correspondencia.

En primer lugar se expondrá el Marco Teórico del Análisis, a manera de introducción. Luego se explicará para qué sirve este Análisis y que se puede hacer con el mismo. Por último se presentarán los detalles del diseño e implementación de la aplicación y los reportes que se presentan en la interfaz web.

El Análisis de Correspondencia (AC), es usado generalmente para estudiar el comportamiento de variables cualitativas, las cuales al inicio de la investigación parecen carecer de vinculación pero que mediante este estudio se encuentran relacionadas, pues la proximidad entre los puntos representados está relacionado con el nivel de asociación entre las variables estudiadas la cual finalmente es presentada mediante biplots que nos presenta la información de manera gráfica [13]. Estos biplots muestran los dos conjuntos de datos, y la cercanía entre ellos representa su relación.

El análisis de correspondencia es concebido como una técnica estadística diseñada para analizar los siguientes puntos[8]

Tablas de Contingencia o Tabla Cruzada

Consiste en la presentación de dos variables que agrupan valores que han sido agrupados en categorías. (Consiste en el cruce de dos variables que agrupan a “individuos” en una serie de categorías.)

Tablas de Frecuencia

Consiste en una serie de atributos o características que representan a objetos o sujetos que vienen dadas por las columnas de la tabla. Las celdas contienen valores que muestran el grado de aceptación o

asociación de cada una de las columnas (objetos o sujetos estudiados). Los valores presentados pueden ser frecuencias absolutas o relativas.

Se puede tener una serie de atributos o características que corresponden a los objetos / sujetos que aparecen en columnas. Las celdas pueden expresar en términos absolutos o relativos, el grado de aceptación de cada uno de esos objetos o el nivel de asociación de cada característica a cada objeto.

Tablas de Valoración

A diferencia de las tablas de frecuencias este tipo de tabla no presenta los valores a ser analizados con frecuencias absolutas o relativas sino con puntaciones numéricas que han sido obtenidas para cada uno de los atributos a ser estudiados.

Los valores a ser estudiados no se presentan en frecuencias absolutas o relativas, sino en puntuaciones numéricas obtenidas para cada uno de los atributos.

Tablas Múltiples

Consisten en aquellas tablas que pueden presentar de 3 a más variables a ser estudiadas.

(Aquellas en las que se pueden tener tres o más entradas, estilos de vida, ambiente y atributos sociales.)

En esta tesis se estudiará el Análisis de Correspondencia Simple orientado al uso de tablas de contingencia de variables cualitativas.

Para finalizar con el *ACS* al referirnos al estudio con tablas de contingencia debemos puntualizar que esta técnica exploratoria utiliza valores positivos en los valores de sus elementos, pues lo que intenta descubrir es una asociación entre los elementos que están siendo estudiados, es decir; se trata de hallar una topología de la filas y una topología de las columnas que conforman la tabla para luego de esto fusionar ambas y encontrar la relación reflejada por los elementos.

3.1 Alcance de la Solución

Como se explicó en la introducción, el Análisis de Correspondencia permite representar tablas de contingencia, de manera que puedan ser fácilmente interpretadas.

El análisis de Correspondencia que será aplicado a las encuestas consta básicamente de dos partes: la construcción de las tablas de contingencia, y luego el análisis de las mismas.

Las tablas de contingencia se formarán de acuerdo a las respuestas que los estudiantes han dado a los cuestionarios de evaluación de profesores. Se considera pertinente hacer los análisis por paralelo, para en cada paralelo construir las tablas de contingencia necesarias y realizar el análisis respectivo.

3.2 Análisis de la Solución

Después de analizar los datos de las encuestas por paralelo, se concluyó que era necesario crear dos tablas de contingencia que representen a los datos:

Estudiantes vs Respuestas

En esta tabla, por cada estudiante se contarán las preguntas a las que haya contestado la alternativa 1, 2, 3, 4 o 5, construyendo de esta manera la tabla.

Este análisis pretende mostrar las tendencias en las puntuaciones otorgadas a las respuestas que fueron seleccionadas por cada uno de los estudiantes encuestados.

Bajo estas características se observará en el gráfico hacia cuales resultados los estudiantes suelen orientar sus respuestas. Así mismo, se podrá ver la preferencia u hostilidad que existe en ese grupo de estudiantes por las preguntas realizadas.

Preguntas vs Respuestas

Para construir esta tabla, por cada pregunta se contará el número de estudiantes que contestaron la alternativa 1, 2, 3, 4 o 5, obteniendo de esta manera los datos.

Este análisis busca mostrar las inclinaciones que tienen cada una de las preguntas que se presentan en una encuesta realizada por el

CENACAD a todos los estudiantes de un paralelo durante un determinado término.

De esta manera gráfica se podrá notar la similitud, afinidad, asociación o interacción entre las preguntas y las respuestas contestadas.

Se debe recordar que cada una de estas tablas contiene los datos por paralelo. Una vez que se han construido las tablas de contingencia, se procederá a la aplicación del análisis propiamente dicho, es decir, el algoritmo que procesa las tablas y trata de representarlas de la mejor manera en un gráfico bidimensional.

3.3 Diseño de la Aplicación

Para construir las tablas de contingencia (frecuencias) (primer paso), se recurre a un conteo exhaustivo en la tabla de respuestas del CENACAD, que contiene más de 8 millones de registros, lo que hace que este primer análisis tome bastante tiempo.

Una vez que se tienen las tablas de contingencia listas, se empieza a aplicar el algoritmo del ACS, el cual se explica en detalle en la Figura 3.1.

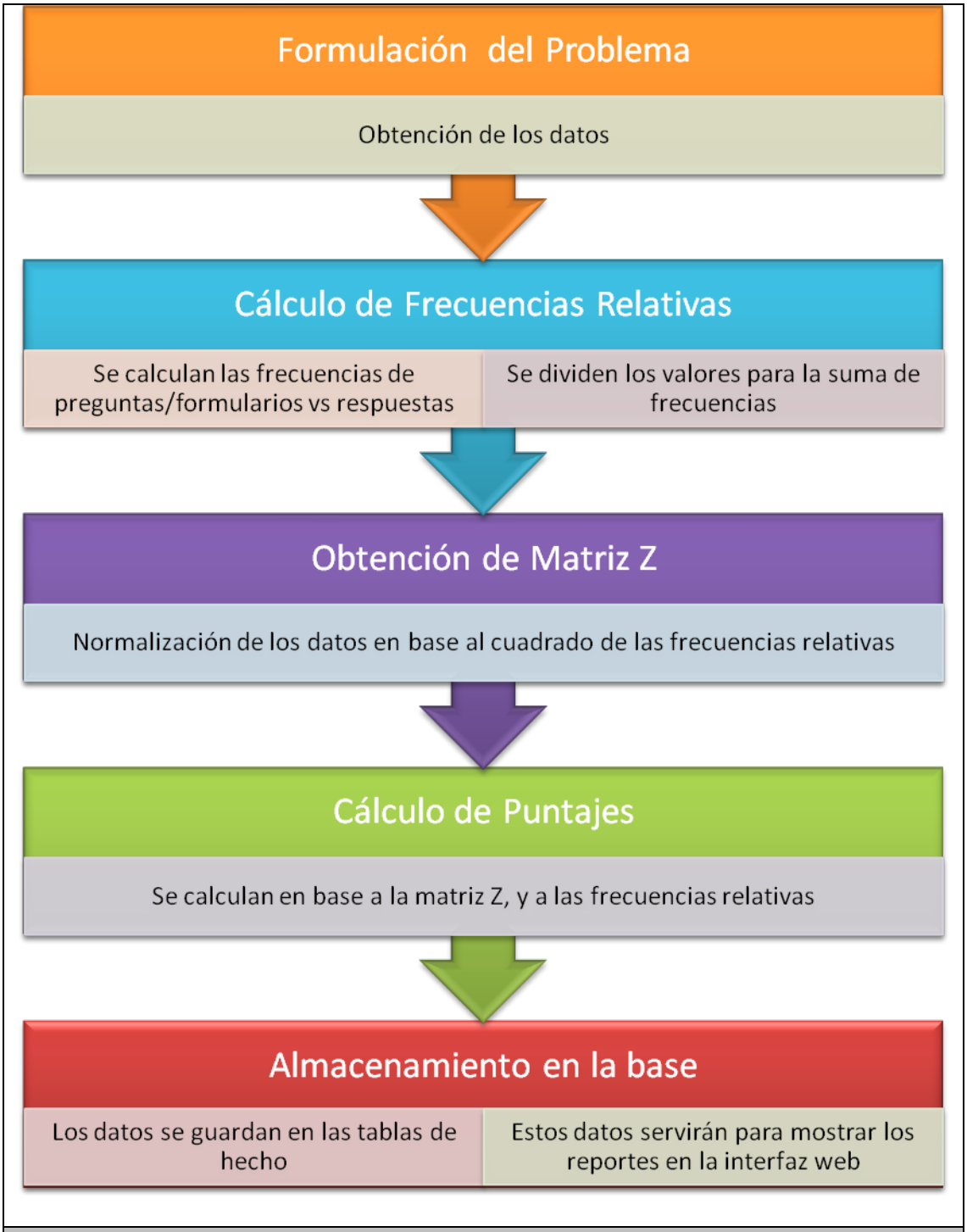


Figura 3.1 Diagrama de Flujo del Análisis de Correspondencia

A continuación se describen los pasos del análisis[5]:

1. Se calcula la tabla de frecuencias relativas F, dividiendo cada valor de la tabla de contingencia para el total de muestras.
2. Se calcula la tabla estandarizada Z, de frecuencias relativas de las mismas dimensiones de la tabla original (es decir F).
3. Se calculan los vectores propios ligados a valores propios mayores, pero distintos de uno.
4. Se calculan los puntajes de las variables sobre los vectores obtenidos en el paso anterior. Estos puntajes son los que se mostrarán en el gráfico bidimensional.

Para entender con facilidad el procedimiento que se realiza y por qué deben de ser tomados dichos valores se muestra continuación la explicación teórica de cada valor anteriormente descrito:

- Es necesario obtener la matriz de frecuencias relativas F relacionadas con las variables estudiadas. Para esto, se utilizará la matriz de contingencia que contiene las frecuencias absolutas obtenidas en las encuestas y a cada valor de esta matriz la dividiremos para *n*, donde *n* será el número total de elementos observados.

En este caso la frecuencia utilizada refleja el valor con que ha sido contestada cada una de las preguntas del cuestionario. Las respuestas contienen puntuaciones que van entre 1 y 5, siendo 1 el valor mínimo y 5 el valor máximo obtenido en cada respuesta.

- Es indispensable obtener la matriz estandarizada Z, ya que no es posible realizar el análisis con la matriz F pues esta matriz a pesar de contener los valores obtenidos en las encuestas incurre en errores de mala representación de los datos pues no refleja la estructura distinta de las filas o columnas.

  El proceso para obtener esta matriz es dividir cada celda de la tabla F por la raiz cuadrada de los totales de sus filas y columnas. Así tenemos:

$$z_{ij} = \frac{f_{ij}}{\sqrt{f_i \times f_j}}$$

- Una vez realizado el paso anterior, se deben obtener los valores y vectores propios de la matriz de menor dimensión entre el producto de la matriz Z por su transpuesta, ZZ', e inversa Z'Z. Encontrar los valores y vectores propios menores

busca elegir los valores que representen la menor variabilidad de los datos presentados en las matrices. De este modo el análisis es más exacto y proporciona resultados más fiables.

Cabe recalcar que si se obtiene los vectores propios del producto ZZ', es decir $A_{i,}$ los vectores del producto Z'Z se pueden obtener de la siguiente manera:

$$B_i = ZA_i$$

Una vez obtenidos estos valores se deben proceder a buscar los valores de las columnas y filas respectivas para la matriz final. Para ello la *I* filas de la matriz se presentarán como *I* puntos en el espacio R^h (donde *h=2* en nuestro caso), y cada coordenada vendrá dada por:

$$C_{f\,=}\, D_f^{\,-1/2}\, ZA_2$$

donde A_2 tiene en columnas los 2 vectores propios de Z'Z. La matriz D_f que se presenta no es más que la matriz de totales diagonales de las filas de la matriz estandarizada.

Las *J* columnas se representaran como *J* puntos en R^2 y las coordenadas de cada columna serán:

$$C_{c=}\, D_c^{\,-1/2}\, ZB_2$$

Así mismo la matriz D_c es la matriz de totales diagonales de las columnas de la matriz Z.

Una vez realizado todo el proceso anterior podremos obtener todos los valores de los coeficientes que serán mostrados en el biplot con los resultados finales del análisis de correspondencia.

3.4 Diseño e Interpretación del Reporte

En el reporte del Análisis de Correspondencia, se deben mostrar los resultados de los dos análisis, es decir, Estudiantes vs Respuestas y Preguntas vs Respuestas. De cada análisis se debe mostrar la tabla de frecuencias y el gráfico resultante del algoritmo. Por tanto, la página web del reporte contiene una estructura de cuatro pestañas, para poder mostrar los datos de una mejor manera.

La primera tabla corresponde a las frecuencias Estudiantes vs Respuestas, es decir, en la tabla se muestra cuántas veces cada estudiante respondió la alternativa 1, 2, 3, 4 o 5 en el cuestionario (Tabla 3.1). Como se puede observar, el total por cada fila es de 34, lo cual corresponde al número de preguntas que tiene el cuestionario evaluado.

Tabla Estudiantes - Respuestas | Gráfico Estudiantes - Respuestas | Tabla Preguntas - Respuestas | Gráfico Preguntas - Respuestas

Tabla de Contingencia: Estudiantes - Respuestas

Estudiante	Respuesta					
	1	2	3	4	5	Total
1	1	0	8	7	18	34
2	0	0	0	6	28	34
3	0	0	1	7	26	34
4	0	0	4	11	19	34
5	2	29	3	0	0	34
6	0	0	1	7	26	34
7	0	0	0	0	34	34
8	0	0	27	2	5	34
9	0	0	1	6	27	34
10	0	0	0	0	34	34
11	0	1	1	5	27	34
12	0	2	14	15	3	34
13	2	5	17	9	1	34
14	4	0	2	16	12	34
15	0	0	0	0	34	34
16	1	0	3	19	11	34
17	0	0	0	0	34	34
18	0	0	0	12	22	34
19	0	0	1	16	17	34
20	0	0	0	29	5	34
21	0	0	0	0	34	34
22	0	0	3	2	29	34
23	0	0	0	1	33	34
24	0	0	0	0	34	34
25	0	0	1	0	33	34
26	0	0	3	0	31	34
27	0	0	4	29	1	34
28	0	0	0	0	34	34
Total	10	37	94	199	612	952

Tabla 3.1 Análisis Estudiantes vs Respuestas

Datos de Frecuencia de los 28 estudiantes evaluadores de un paralelo

En la siguiente pestaña, se muestra el gráfico Estudiantes vs Respuestas. Los estudiantes se representan con puntos amarillos, y las respuestas, con cuadrados azules. Aquí se puede observar gráficamente las relaciones entre ellos. Cabe recalcar que las dos

dimensiones de los gráficos no tienen ningún significado, simplemente son resultados del análisis de frecuencia.

En la Figura 3.2, se presenta un ejemplo, donde se puede observar gran concentración de puntos amarillos (estudiantes) alrededor de la respuesta 5, lo cual sugiere que la mayoría de estudiantes tienen una percepción buena del paralelo, mientras que existe un estudiante que se encuentra muy alejado del resto (dato atípico), y muy cerca de la respuesta 2.

Figura 3.2 Análisis Estudiantes vs Respuestas.
Representación Grafica de los resultados mostrados en la Tabla anterior

La siguiente pestaña, es la tabla de frecuencias entre Preguntas y Respuestas (Tabla 3.2), es parecida a la primera tabla, pero en vez de estudiantes se muestran las preguntas. Esta tabla nos indica por cada pregunta cuántos estudiantes han contestado a la alternativa 1, 2, 3, 4 o 5. Como se puede observar, cada fila suma al total 28, que es el número de estudiantes evaluados en ese paralelo.

Tabla Estudiantes - Respuestas	Gráfico Estudiantes - Respuestas	Tabla Preguntas - Respuestas	Gráfico Preguntas - Respuestas

Tabla de Contingencia: Preguntas - Respuestas

No.	Pregunta	Respuesta					
		1	2	3	4	5	Total
1	Asistí regularmente a clases	1	0	6	0	21	28
2	Asistí puntualmente a clases	2	0	4	0	22	28
3	Permanezco en clase durante toda la sesión programada	1	0	2	0	25	28
4	Participé de forma activa en clases, solicitando aclarar dudas, respondiendo preguntas, aportando con ejemplos	2	0	7	0	19	28
5	En general juzgo que mi esfuerzo y dedicación fueron apropiados	0	0	5	0	23	28
6	Al inicio del término proporcionó y explicó a los estudiantes la programación y políticas del curso.	0	1	2	8	17	28
7	Refleja una adecuada preparación de sus clases.	0	1	1	8	18	28
8	Cumple con la programación propuesta al inicio del curso.	0	2	1	7	18	28
9	Considera los conocimientos previos de los estudiantes para el dictado de la clase.	0	1	3	5	19	28
10	Fomenta la utilización del texto guía y los textos de referencia.	0	1	4	7	16	28
11	Presenta los contenidos de la clase de una manera comprensible.	0	1	3	6	18	28
12	Enfatiza durante la clase los puntos principales de los temas que expone.	0	1	1	8	18	28
13	Utiliza tecnologías de información para reforzar los contenidos de las clases.	0	1	3	8	16	28
14	Presenta ejemplos prácticos sobre lo tratado en clase.	0	1	1	9	17	28
15	Promueve el razonamiento de los temas tratados.	0	1	1	9	17	28
16	Contesta en forma satisfactoria las preguntas formuladas en clase.	0	1	2	8	17	28
17	Asigna actividades que requieren investigación por parte de los estudiantes.	0	1	3	6	18	28
18	Organiza durante la clase actividades de autoaprendizaje.	0	2	2	7	17	28
19	Estimula en el estudiante el pensamiento creativo y admite pensamientos diferentes.	1	1	2	5	19	28
20	Desarrolla los contenidos de la materia con un ritmo apropiado.	0	1	3	6	18	28
21	Utiliza técnicas de trabajo en grupo durante la clase.	1	1	2	6	18	28
22	Exalta las buenas costumbres y conciencia social durante la clase.	0	1	2	10	15	28
23	Fomenta la participación activa de los estudiantes en clase.	0	2	2	11	13	28
24	Es respetuoso y cordial en el trato con los estudiantes.	0	2	3	4	19	28
25	Estimula en la clase el comportamiento ético de los estudiantes.	0	2	2	6	18	28
26	Realiza evaluaciones periódicas (deberes, lecciones, proyectos, pruebas, etc.)	0	2	4	11	11	28
27	Cumple con las políticas de evaluación señaladas al inicio del curso.	0	1	4	7	16	28
28	Formula claramente las preguntas en las evaluaciones escritas.	0	1	2	7	18	28
29	Los temas en las evaluaciones son representativo de lo enseñado.	0	1	3	5	19	28
30	Hace conocer los resultados de las evaluaciones periódicas en plazos oportunos a sus estudiantes.	0	1	2	5	20	28
31	Califica procedimientos y resultados en las evaluaciones de los temas de examen.	1	1	2	4	20	28
32	Asiste puntualmente a clases (llega y se retira dentro del tiempo reglamentario)	0	2	3	5	18	28
33	Asiste regularmente a clases (frecuencia)	1	2	2	4	19	28
34	De acuerdo con sus respuestas anteriores, usted evalúa el desempeño del profesor(a) como:	0	1	5	7	15	28
Total		10	37	94	199	612	952

Tabla 3.2 Análisis Preguntas vs Respuestas

Datos de Frecuencia de las 34 preguntas del formulario

La cuarta y última pestaña, muestra el gráfico de Preguntas vs Respuestas fruto del Análisis de Correspondencia (Figura 3.3). En este gráfico las preguntas se muestran con puntos amarillos, y muestra su relación con las respuestas. Los puntos tienen el mismo comportamiento que en el gráfico anterior. Es decir, si existen preguntas cercanas a la respuesta 5, quiere decir que esas preguntas obtuvieron valoración alta en el paralelo.

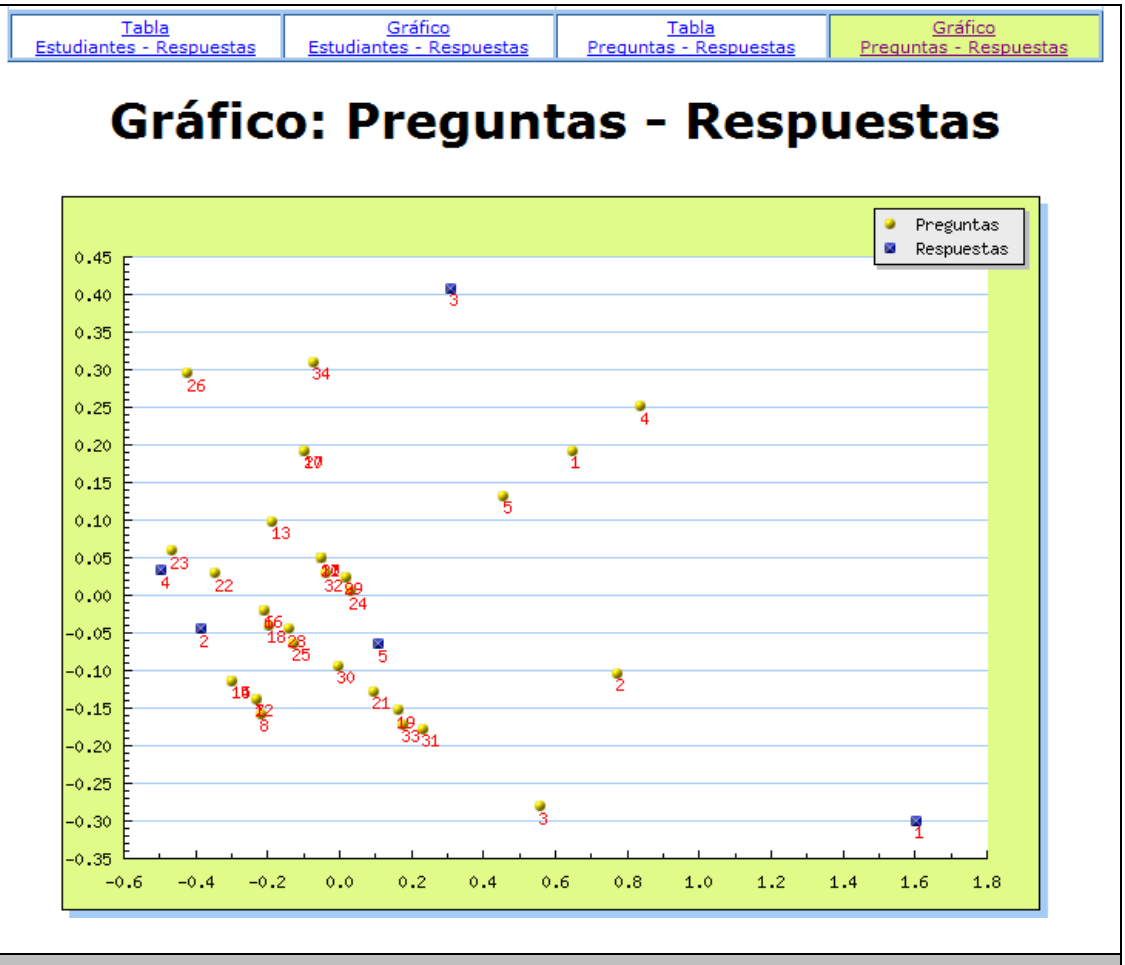


Figura 3.3 Análisis Preguntas vs Respuestas.
Representación Grafica de los resultados mostrados en la Tabla anterior

3.5 Plan de Pruebas

Para probar la validez del algoritmo desarrollado, los datos de algunos paralelos fueron extraídos, y luego analizados con el SPSS, usando la opción de Análisis de Correspondencia del mismo, y luego proyectando los datos resultantes en un gráfico.

En la Figura 3.4 se puede apreciar el gráfico resultante del Análisis de Correspondencia del SPSS en el paralelo con código 13130. En la Figura 3.5 en cambio, se puede apreciar el análisis del mismo paralelo, realizado con nuestro algoritmo.

Como se puede observar, los gráficos son iguales y demuestran que el proceso se implementó correctamente. Estas mismas pruebas fueron hechas con algunos paralelos, encontrando siempre coincidencias entre el proceso desarrollado por el SPSS y por nuestro algoritmo.

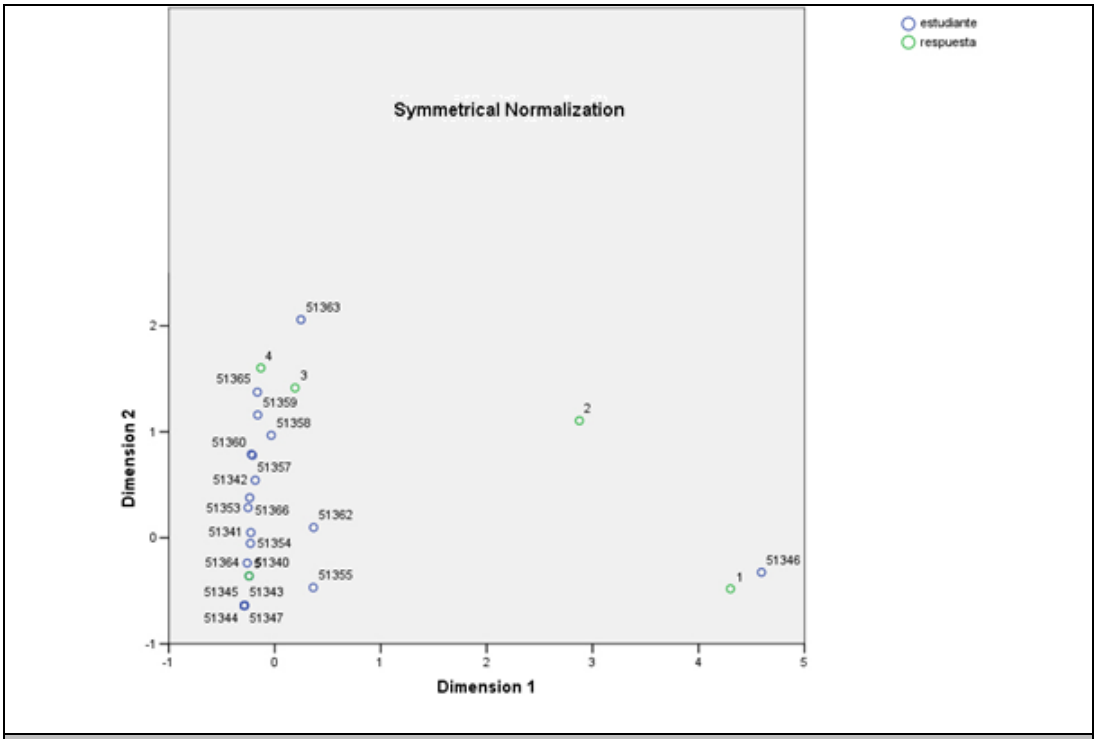


Figura 3.4 Análisis en el SPSS del paralelo 13130: Estudiantes vs Respuestas

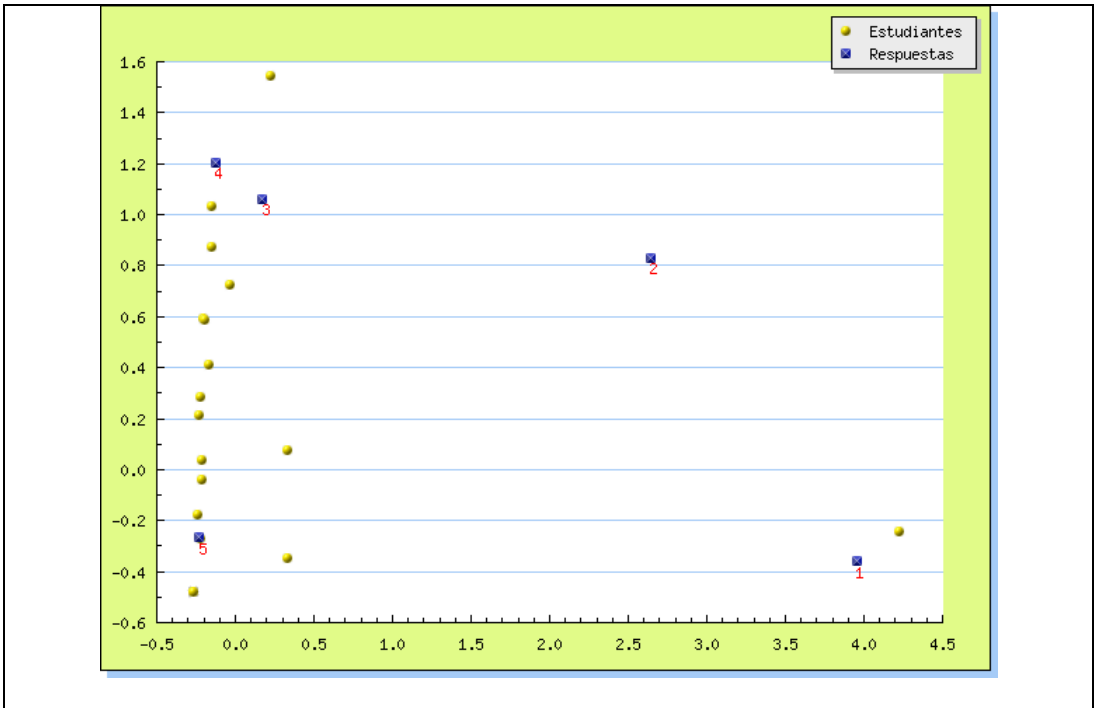


Figura 3.5 Análisis de nuestro sistema: Estudiantes vs Respuestas

Además, se realizaron pruebas de rendimiento, midiendo el tiempo que se tomaría el algoritmo en analizar los datos de todos los paralelos que se encuentran en la base de datos del CENACAD. La Figura 3.6 muestra los tiempos tomados en diferentes ejecuciones del algoritmo.

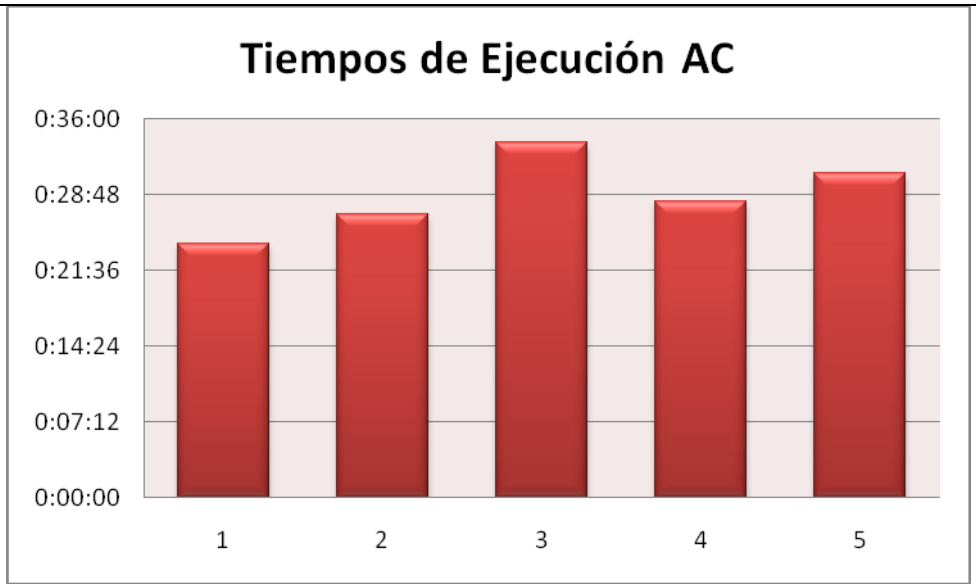


Figura 3.6 Tiempos de ejecución del Análisis de Correspondencia

Se puede observar que el tiempo promedio de ejecución de todo el algoritmo es de 28 minutos, lo cual demuestra eficiencia, teniendo en cuenta que sólo la primera vez se necesita hacer análisis sobre todos los datos, de ahí en adelante se hace sólo de los paralelos nuevos.

En este capítulo se ha detallado todo lo relacionado con el Análisis de Correspondencia, y como ha sido aplicado a las encuestas del CENACAD, realizando dos análisis simultáneos, Estudiantes vs Respuestas y Preguntas vs Respuestas.

El siguiente capítulo tratará acerca del Escalado Multidimensional, el cual permite medir distancias entre elementos, y representarlas gráficamente de manera comprensible.

CAPÍTULO 4

4 ESCALADO MULTIDIMENSIONAL

En el presente capítulo se hablará acerca de la técnica del Escalado Multidimensional, para analizar distancias entre diferentes elementos del mismo tipo.

Se detallará la aplicación de esta técnica estadística en las encuestas del CENACAD. Al igual que en el capítulo anterior, cada fase de su implementación (alcance, análisis, diseño, pruebas) será presentada y explicada en detalle.

El Escalado Multidimensional (EMD), está considerado como una de las herramientas estadísticas más importantes pues es posible determinar qué variables son importantes para la persona que desea tomar una decisión con respecto a los datos analizados.

Entre los objetivos del Escalado Multidimensional se pueden citar:

Encontrar un espacio de dimensiones pequeñas el cual pueda explicar las relaciones existentes entre los datos. De esta manera la solución obtenida será una combinación lineal de variables independientes, en donde cualquier persona es capaz de interpretar los resultados dados en la solución en términos de las variables expuestas al inicio.

Otro de los objetivos del Escalado Multidimensional es la de siempre encontrar una interpretación lógica y coherente para los datos, sin importar las circunstancias, ni de donde fueron obtenidos. Así se podrá observar luego de realizado el análisis correspondiente si existen datos o grupos atípicos entre los datos.

En el Escalado Multidimensional existen 2 métodos que pueden ser realizados el primero es el *escalado métrico* o también llamado de coordenadas principales y el segundo es el *escalado no métrico* [4].

Escalado Métrico

Es aquel que se realiza cuando la matriz es de similaridades. Este método utiliza las diferencias entre similitudes de los objetos estudiados.

Escalado No Métrico

En este análisis el punto de partida es una matriz de diferencias entre los objetos a estudiar que generalmente han sido obtenidos por algún tipo de procedimiento de ordenación o por algún tipo de consulta a expertos.

En el presente proyecto de tesis será estudiado y desarrollado el Escalado Multidimensional No Métrico.

4.1 Alcance de la Solución

Existen elementos dentro del modelo del CENACAD que son similares en naturaleza o que tienen algo en común, por ejemplo, todos los profesores que han dictado una misma materia. Variables

asociadas a estos datos, pueden ser comparados entre sí para obtener más información sobre su estructura, y tener una mejor visión sobre las distancias entre los mismos.

El Escalado Multidimensional sirve para medir o analizar distancias entre elementos del mismo tipo. Por lo tanto, se buscará identificar los elementos que pueden ser medibles y que se pueden beneficiar de este análisis.

Adicionalmente, es necesario identificar una medida de la distancia entre los elementos.

4.2 Análisis de la Solución

Después de realizar un estudio exhaustivo del modelo de datos del CENACAD, fueron obtenidos tres tipos de elementos que pueden ser usados para aplicar el Escalado Multidimensional.

A continuación, se explicará uno por uno estos elementos, y la forma en que es aplicado el Escalado Multidimensional en cada uno de ellos.

Escalado de Materias de un Profesor

En este análisis se mide la distancia que existe entre todos los cursos que ha dictado un profesor en particular. La medida de “distancia” a utilizar será el promedio que ha obtenido el profesor en cada curso que ha dictado.

El objetivo de este análisis es exponer todas las calificaciones que han sido obtenidas por el profesor durante toda su carrera como catedrático de la universidad.

Este análisis identificará fortalezas o debilidades del profesor para dictar una o más materias, pues se podrá observar en cuales materias obtiene mejor puntaje que en otras.

Escalado de Cursos de una Materia

En este análisis se mide la distancia que existe entre los diferentes cursos que se han dictado de una misma materia.

La medida de distancia en este caso, al igual que en el anterior, ha sido tomada como la diferencia de promedios que obtuvo un profesor en un paralelo.

En este análisis se trata de exponer todas las calificaciones que han recibido los cursos de una materia en particular, en todos los términos que ha sido dictada la misma, como ejemplo el puntaje obtenido por Algebra Lineal durante los últimos semestres.

Este análisis servirá para analizar qué profesores obtienen una buena calificación dictando una materia, y cuáles profesores obtienen calificaciones bajas dictando la misma materia. Eventualmente, servirá para decidir si un profesor debe seguir dictando una materia o debe de cambiar la manera de llevar su cátedra.

Escalado de Unidades Académicas por Encuestas

Este análisis permite medir la distancia que existe entre las diferentes Unidades Académicas por cada encuesta que se realiza. La medida de distancia, en este caso, es la diferencia de promedios que una Unidad Académica obtuvo en una encuesta en particular.

El presenta análisis expone a todas las Unidades Académicas que participaron en una encuesta (por ejemplo, “Encuesta de Materias Teóricas – II Término, 2005”) y el promedio que obtuvo cada una de ellas.

Este análisis ayudará a los tomadores de decisiones a identificar qué unidades han obtenido un nivel de calificación superior respecto a otras Unidades, en el mismo período de tiempo. Además muestra qué Unidades Académicas están por debajo del promedio.

4.3 Diseño de la Aplicación

El primer paso a seguir es el análisis de los datos, los cuales deben ser obtenidos de la fuente correcta. Por lo tanto, por cada análisis de los tres mencionados en la sección anterior, se hace una consulta a las tablas de hecho que tienen almacenados los promedios, y éstos son cargados en memoria.

El siguiente paso a cumplir es generar las distancias a partir de esos promedios, que no es más que la diferencia (en valor absoluto) de los mismos. Con la matriz de distancias, es posible iniciar el algoritmo, que a breves rasgos consiste en:

1. Construir la matriz *Q* de productos cruzados.
2. Obtener los valores propios de *Q*, y seleccionar los 2 mayores.
3. Obtener las coordenadas de los puntos que se mostraran en los gráficos respectivos.

A continuación se ampliará la explicación del método de manera más expresa para que pueda ser entendido con mayor facilidad[4]:

Para obtener la matriz *Q* definida como:

$$Q = -1/2 \; P * D * P$$

es necesario primero obtener la matriz *P* que se obtiene de la siguiente manera:

$$P = I - \frac{1}{2} \times 1 \times 1'$$

La matriz *P* no es más que una matriz de *datos centrados* que se define como resultado de restar a cada dato de su media. *P* será una matriz cuadrada y simétrica.

Luego de obtener *P*, se deberá obtener la matriz *D* que es la matriz de distancias existentes entre las variables que serán estudiadas y analizadas. Realizados estos procedimientos la matriz *Q* puede ser obtenida.

Se procederá a la obtención de los valores propios de la matriz *Q* obtenida en el literal anterior. Recordemos que los valores propios de una matriz de *n x n* son aquellos que no cambian la forma de una matriz si ésta se convierte en su transpuesta.

Los valores propios que serán tomados serán los 2 valores mayores encontrados en el procedimiento.

Para finalizar el análisis se deberán obtener los coeficientes de las filas y las columnas, de la siguiente manera:

$$C_c = C_f = V_i \ \lambda_i^{1/2}$$

Los coeficientes obtenidos se almacenarán en las tablas de hecho, y con esto se termina el algoritmo. El flujo completo del proceso se puede apreciar en la Figura 4.1.

Figura 4.1 Diagrama de Flujo del Escalado Multidimensional

4.4 Diseño e Interpretación del Reporte

A continuación, serán analizados los reportes de cada uno de los 3 análisis del Escalado Multidimensional. A pesar de ser similares en estructura los tres análisis, cada uno merece una explicación distinta acerca de su significado e interpretación.

Escalado de Materias de un Profesor

Este reporte es realizado por profesor. En primer lugar se presenta todos los cursos que el profesor ha dictado en una tabla, ordenados por promedio (Ver Tabla 4.1).

Reporte: **ESCALADO MULTIDIMENSIONAL: PROFESORES**

Explicación del reporte

Profesor: EDGAR EDUARDO, CERVANTES BERNABE

CURSOS QUE EL PROFESOR HA DICTADO

No.	Año	Término	Materia	Paralelo	Encuesta	Promedio
1	2006	1S	FUND.DE INGENIERIA	001	Encuesta de Materias Teóricas 2006 - 1S	**100.00**
2	2006	2S	FUND.DE INGENIERIA	001	Encuesta de Materias Teóricas 2006 - 2S	**100.00**
3	2006	2S	INGENIERIA I	001	Encuesta de Materias Teóricas 2006 - 2S	**99.85**
4	2007	1S	INGENIERIA II	001	Encuesta de Materias Teóricas 2007 - 1S	**99.08**
5	2004	2S	OCEANOGRAFIA DESCRIPTIVA	001	Encuesta de Materias teóricas 2004 - 2S	**98.25**
6	2006	1S	ING.PARA ACUACULTURA II	001	Encuesta de Materias Teóricas 2006 - 1S	**97.14**
7	2007	1S	ING. PARA ACUACULTURA I	001	Encuesta de Materias Teóricas 2007 - 1S	**97.11**
8	2005	2S	OCEANOGRAFIA DESCRIPTIVA	001	Encuesta de Materias Teóricas 2005 - 2S	**95.55**
9	2005	2S	MANEJO RECURSOS NATURALES	002	Encuesta de Materias Teóricas 2005 - 2S	**88.70**
10	2006	1S	ING. PARA ACUACULTURA I	001	Encuesta de Materias Teóricas 2006 - 1S	**86.94**
11	2005	1S	GEOGRAFÍA FÍSICA Y AMBIENTAL	003	Encuesta de Materias Teóricas 2005 - 1S	**86.20**
12	2004	2S	INGENIERIA I	001	Encuesta de Materias teóricas 2004 - 2S	**84.11**
13	2006	2S	ING.PARA ACUACULTURA II	001	Encuesta de Materias Teóricas 2006 - 2S	**81.16**
14	2005	2S	INGENIERIA I	001	Encuesta de Materias Teóricas 2005 - 2S	**80.37**
15	2005	1S	INGENIERIA II	001	Encuesta de Materias Teóricas 2005 - 1S	**78.00**
16	2005	1S	FUND.DE INGENIERIA	001	Encuesta de Materias Teóricas 2005 - 1S	**72.04**

Tabla 4.1 Tabla con todos los cursos que el profesor ha dictado en su carrera

Luego, en la Figura 4.2, se muestra el gráfico resultante fruto del Escalado Multidimensional. Generalmente los datos se colocan en forma de parábola. Mientras más a la derecha los puntos, significa que ese paralelo ha obtenido mayor puntaje.

Los ejes coordenados dividen el gráfico en cuatro cuadrantes, y por lo tanto, los puntos que se encuentren en el último cuadrante, son los cursos del profesor donde obtuvo menor calificación. Esto no necesariamente significa que sean malos, sólo que han obtenido una calificación baja con respecto a los otros paralelos que ha dictado dicho profesor.

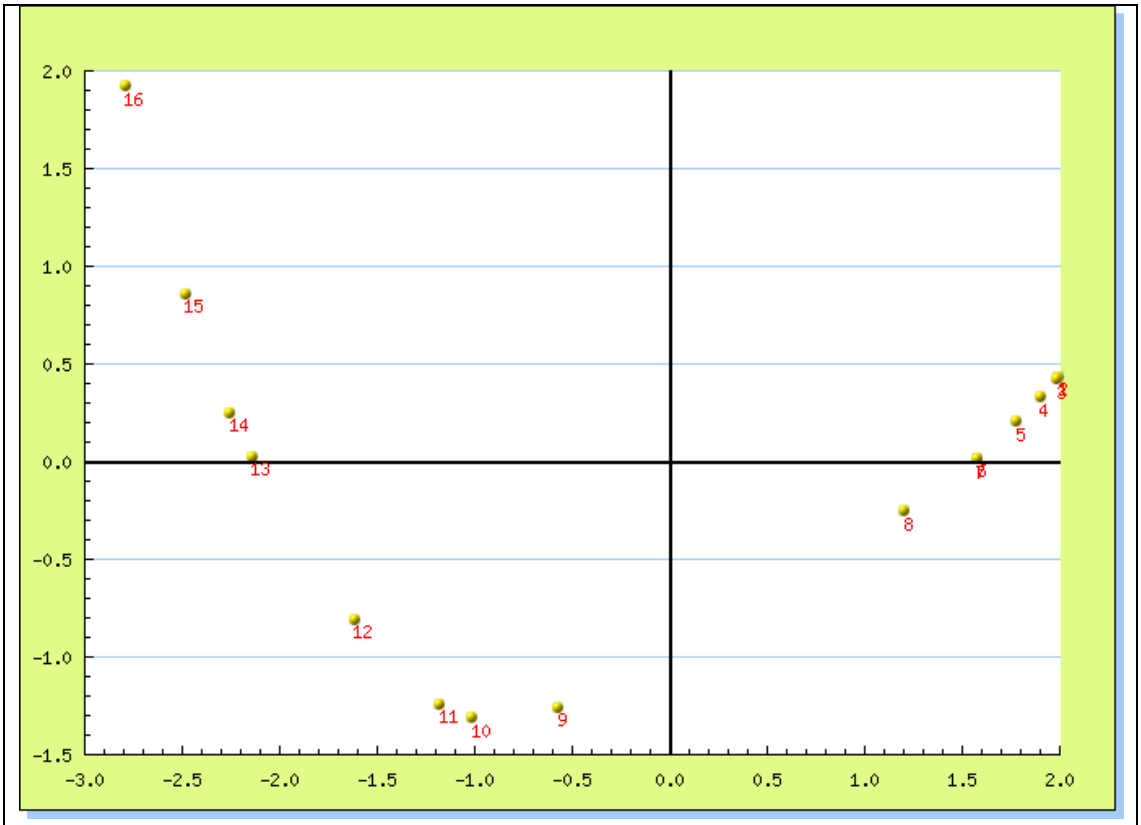


Figura 4.2 Escalado Multidimensional de todos los cursos dictados por un profesor

Escalado de Cursos de una Materia

Este reporte se hace por materia. En primer lugar son presentados todos los cursos que han sido dictados de esa materia, en todos los semestres, ordenados por promedio (Ver Tabla 4.2).

Reporte: **ESCALADO MULTIDIMENSIONAL: MATERIAS**

Explicación del reporte

Unidad: INSTITUTO DE CIENCIAS MATEMÁTICAS
Materia: CALCULO I (B)
Código: ICM00216

CURSOS DICTADOS DE LA MATERIA

No.	Año	Término	Profesor	Paralelo	Encuesta	Promedio
1	2004	2S	CARLOS MANUEL, ING. MARTIN BARREIRO	013	Encuesta de Materias teóricas 2004 - 2S	**97.41**
2	2004	2S	MARGARITA HELENA, ING. MARTINEZ JARA	002	Encuesta de Materias teóricas 2004 - 2S	**95.55**
3	2004	2S	SOVENY SORAYA, ING. SOLIS GARCIA	005	Encuesta de Materias teóricas 2004 - 2S	**95.47**
4	2004	3S	RAUL SEGUNDO TINGO SOLEDISPA	002	Encuesta de Materias Teóricas 2004 - 3S	**94.33**
5	2004	2S	JANET PATRICIA VALDIVIEZO	015	Encuesta de Materias teóricas 2004 - 2S	**93.42**
6	2004	2S	JORGE ROILANDI MEDINA SANCHO	003	Encuesta de Materias teóricas 2004 - 2S	**92.86**
7	2004	3S	JORGE ROILANDI MEDINA SANCHO	001	Encuesta de Materias Teóricas 2004 - 3S	**91.34**
8	2004	2S	MOISES, ING. VILLENA MUNOZ	008	Encuesta de Materias teóricas 2004 - 2S	**91.04**
9	2004	2S	SEGUNDO LEONARDO, ING. BARONA VALENCIA	010	Encuesta de Materias teóricas 2004 - 2S	**90.83**
10	2004	2S	RAUL SEGUNDO TINGO SOLEDISPA	011	Encuesta de Materias teóricas 2004 - 2S	**89.88**
11	2004	2S	ERWIN JOFFRE, ING. DELGADO BRAVO	006	Encuesta de Materias teóricas 2004 - 2S	**89.57**
12	2004	2S	ERWIN JOFFRE, ING. DELGADO BRAVO	007	Encuesta de Materias teóricas 2004 - 2S	**87.98**
13	2004	2S	PEDRO SENATORE, ING. RAMOS DE SANTIS	004	Encuesta de Materias teóricas 2004 - 2S	**85.18**
14	2004	2S	COLON MARIO, ING. CELLERI MUJICA	012	Encuesta de Materias teóricas 2004 - 2S	**84.91**
15	2004	2S	MOISES, ING. VILLENA MUNOZ	014	Encuesta de Materias teóricas 2004 - 2S	**84.03**
16	2004	3S	EDGAR JOHNI BUSTAMANTE ROMERO	003	Encuesta de Materias Teóricas 2004 - 3S	**82.99**
17	2004	2S	JANET PATRICIA VALDIVIEZO	001	Encuesta de Materias teóricas 2004 - 2S	**81.09**
18	2004	2S	SEGUNDO LEONARDO, ING. BARONA VALENCIA	009	Encuesta de Materias teóricas 2004 - 2S	**73.91**

Tabla 4.2 Tabla con todos los cursos dictados de alguna materia

A continuación, se muestra el gráfico resultante fruto del Escalado Multidimensional con esos datos. Al igual que en el anterior, los datos se ordenan de derecha a izquierda, y en forma de parábola. Nuevamente, los paralelos que se ubiquen en el último cuadrante, no necesariamente son malos, pero sí están distanciados del resto. (Figura 4.3)

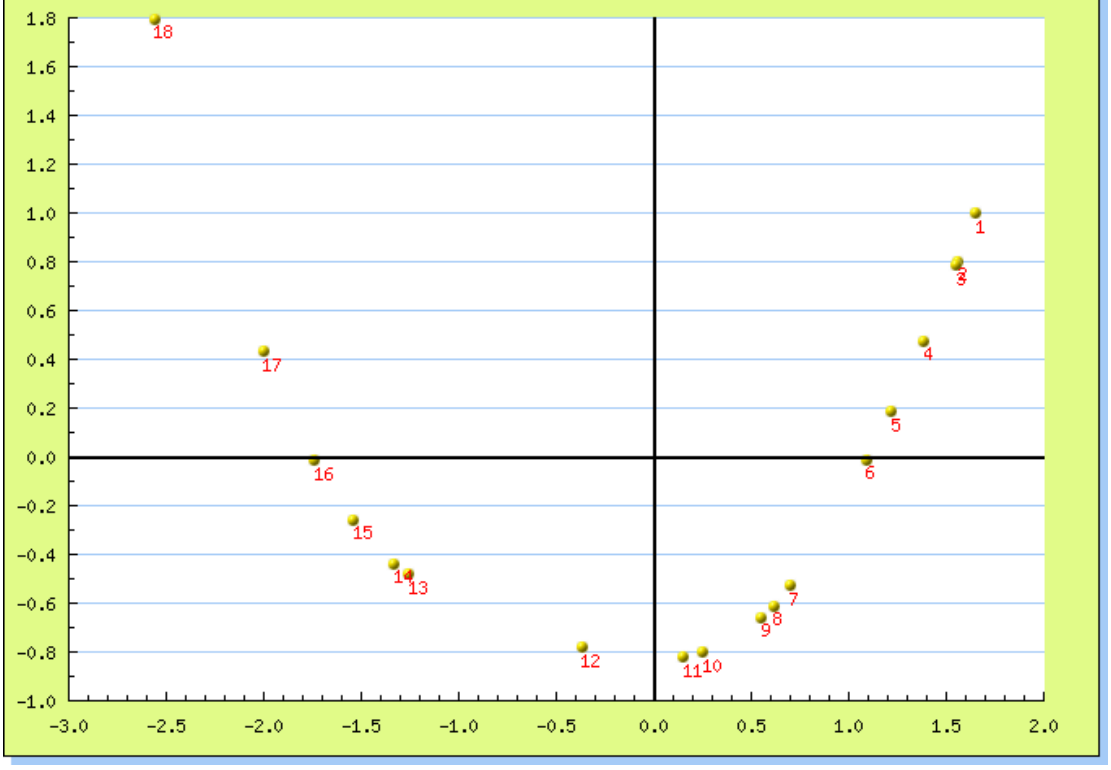


Figura 4.3 Escalado Multidimensional de todos los cursos dictados de una materia

Escalado de Unidades Académicas por Encuestas

Este reporte es realizado por encuesta. En primera instancia son presentadas todas las unidades que han realizado la encuesta en un término específico. Por ejemplo, en la Tabla 4.3 se muestran los promedios de todas las unidades que realizaron la Encuesta de Materias Teóricas el segundo término del 2006.

Reporte: **ESCALADO MULTIDIMENSIONAL: UNIDAD ACADÉMICA**

Explicación del reporte

Encuesta: Encuesta de Materias Teóricas 2006 - 2S
Año: 2006
Término: 2S

UNIDADES ACADÉMICAS

Unidad	Siglas Unidad	Nombre Unidad	Promedio Unidad
1	FIMCM	FAC. ING. MARÍTIMA Y CIENCIAS DEL MAR	**88.27**
2	CELEX	CENTRO DE ESTUDIOS DE LENGUAS EXTRANJERAS	**86.91**
3	PROTMEC	PROGRAMA DE TECNOLOGÍA MECÁNICA	**86.86**
4	ICQA	INSTITUTO DE CIENCIAS QUÍMICAS Y AMBIENTALES	**86.41**
5	EDCOM	ESCUELA DE DISEÑO Y COMUNICACIÓN VISUAL	**85.52**
6	ICM	INSTITUTO DE CIENCIAS MATEMÁTICAS	**85.43**
7	FIMCP	FAC.ING. EN MECÁNICA Y CC. DE LA PRODUCCIÓN	**84.78**
8	PROTCOM	PROGRAMA DE TECNOLOGÍA EN COMPUTACIÓN	**84.73**
9	ICHE	INST.DE CIENCIAS HUMANÍSTICAS Y ECONÓMICAS	**84.71**
10	PROTEL	PROGRAMA DE TEC. ELÉCTRICA Y ELECTRÓNICA	**84.52**
11	FIEC	FAC. ING. EN ELECTRICIDAD Y COMPUTACIÓN	**84.13**
12	FICT	FAC. ING. EN CIENCIAS DE LA TIERRA	**83.85**
13	PROTAL	PROGRAMA DE TECNOLOGÍA EN ALIMENTOS	**82.26**
14	ICF	INSTITUTO DE CIENCIAS FÍSICAS	**79.27**
15	PROTEP	PROGRAMA DE TECNOLOGÍA PESQUERA	**75.49**

Tabla 4.3 Tabla con todas las unidades que realizaron una encuesta

En la Figura 4.4 se muestra el gráfico del Escalado Multidimensional de los mismos datos. El gráfico tiene un comportamiento similar al de los dos análisis anteriores. En este gráfico se puede observar que las 4 primeras unidades están en el mejor cuadrante, mientras que las 2

últimas están en el último cuadrante, es decir, su promedio difiere significativamente del resto.

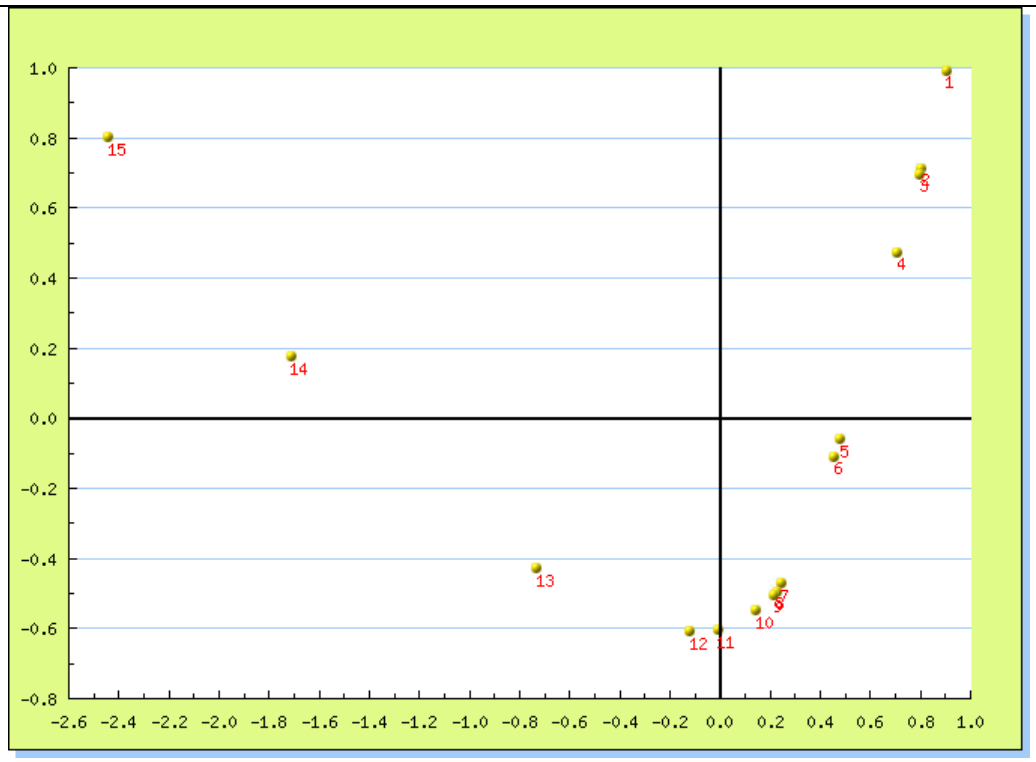


Figura 4.4 Escalado Multidimensional de las unidades que realizaron la encuesta

4.5 Plan de Pruebas

Para probar la validez del algoritmo desarrollado, se realizaron pruebas, donde fueron comparados los resultados de nuestro sistema con análisis hechos en Matlab 6.1 de la misma información, obteniendo los mismos datos de respuesta, probando así la

efectividad del método. Esto lo podemos comprobar en las Figuras 4.5 y 4.6.

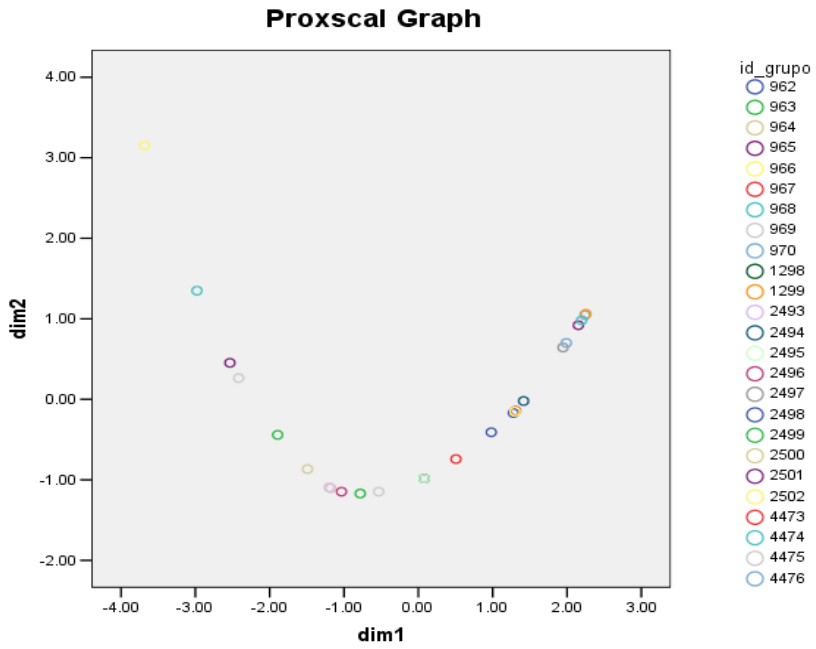


Figura 4.5 Escalado Multidimensional con el SPSS de los cursos de Cálculo I

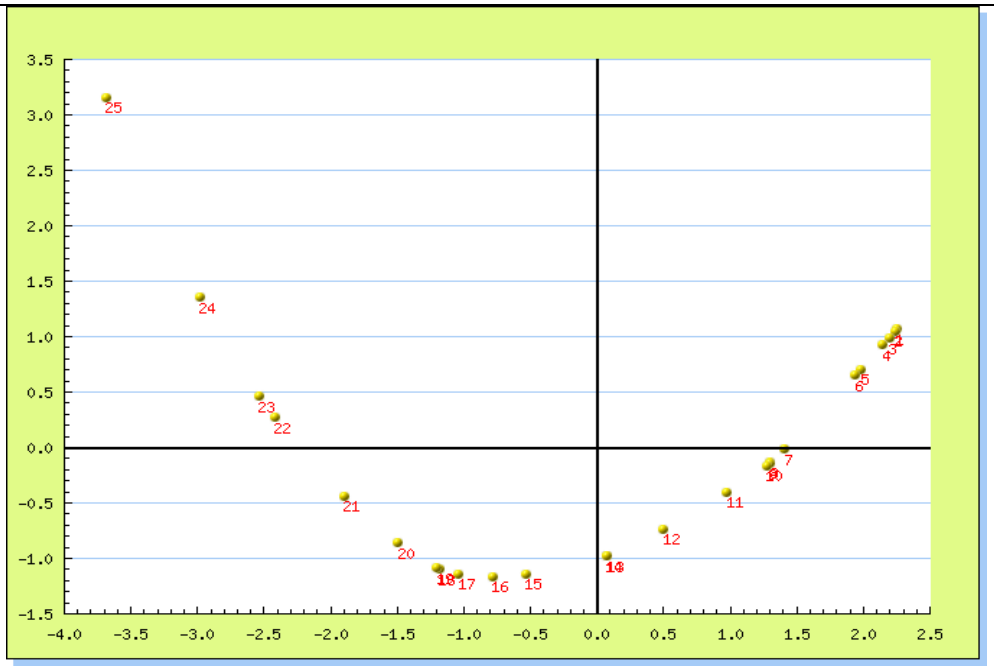


Figura 4.6 Escalado Multidimensional en nuestro sistema: materia Cálculo I

También fueron realizadas pruebas de rendimiento y tiempos de ejecución. Este algoritmo no presenta complicaciones matemáticas, y los datos que se deben analizar no son muy extensos (diferencia de promedio entre profesores, materias y unidades), por lo cual su ejecución debe ser rápida.

Al realizar las pruebas, se obtuvo que el promedio de ejecución fue de 3 minutos y 5 segundos, tomando en cuenta que la cantidad de datos que se encuentran en la base de datos del CENACAD es grande. En la Figura 4.7 pueden ser apreciados distintas medidas de tiempo para este algoritmo.

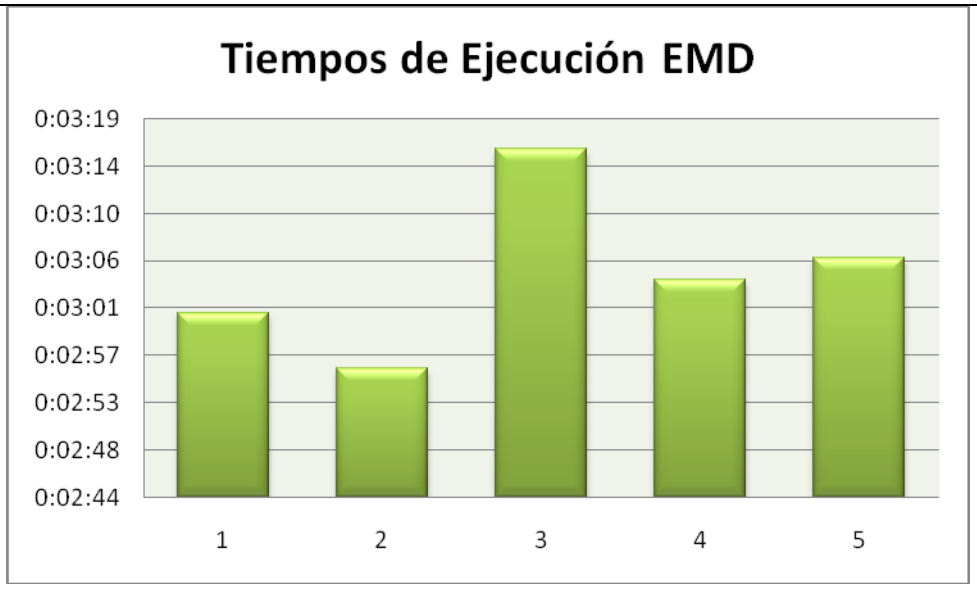


Figura 4.7 Tiempos de Ejecución del Escalado Multidimensional

En este capítulo se presentó la aplicación del Escalado Multidimensional de tres maneras distintas según los datos de las encuestas del CENACAD. Esto ha permitido observar las distancias relativas entre diferentes cursos ya sea de una misma materia o de un mismo profesor.

El siguiente capítulo mostrará la aplicación del Análisis Factorial en los datos del CENACAD.

De manera similar a este capítulo, se explicará paso a paso todos y cada uno de los elementos involucrados en la implementación del mismo.

CAPÍTULO 5

5 ANÁLISIS FACTORIAL

En el presente capítulo, se detalla el proceso de desarrollo e implementación del Análisis Factorial en los datos de las encuestas del CENACAD.

Como en los capítulos anteriores, se mostrará el correspondiente análisis, diseño, implementación y pruebas del algoritmo y los reportes obtenidos y presentados como resultado final en la página web.

El Análisis Factorial es frecuentemente usado para realizar reducciones de datos.

Reducir datos consiste en remover aquellos datos o factores que son redundantes y de esta manera hacer más pequeño el conjunto de factores que deben interpretarse[20]. Generalmente suelen presentarse datos en una matriz, éstos son generados o están contenidos en otros datos cuya importancia es mayor.

El Análisis Factorial según Zikmund [15] es un método de interdependencia, en el cual el objetivo principal es dar significado a un conjunto de variables o tratar de agrupar las cosas.

El Análisis Factorial puede dividirse en 2 tipos de análisis [16]:

- *Análisis Factorial Exploratorio*, en el cual no se conocen los factores que serán estudiados

- *Análisis Factorial Confirmatorio*, en el cual los factores son conocidos a priori, y se hace la asunción de que dichos factores describen a ciertas variables originales.

Se puede concluir en que el procedimiento para obtener un Análisis Factorial tiene mucha flexibilidad al usarse en el CENACAD.

5.1 Alcance de la Solución

Los cuestionarios de las encuestas generalmente tienen un gran número de preguntas, lo cual dificulta el análisis para el investigador, ya que con un gran número de variables (generalmente, 34 preguntas), se dificulta la interpretación de los resultados.

Además de esto, las preguntas están relacionadas entre sí, lo cual indica que tienen una alta covarianza, por lo cual es factible reducir el tamaño del problema y simplificar así el análisis.

El Análisis Factorial que se propone realizar, trata de reducir el número de variables a un pequeño número de factores que expliquen el mayor porcentaje de varianza de los datos, con un mínimo error.

Una vez extraídos los factores principales, se dará una interpretación a éstos factores, relacionándolos de cierta forma con una o más preguntas (variables iniciales).

El Análisis Factorial también servirá para validar el modelo de los cuestionarios, y la división de preguntas por áreas. En el caso ideal, los factores deberían agrupar a las preguntas según las áreas a las que pertenecen.

5.2 Análisis de la Solución

El Análisis Factorial propuesto en el presente trabajo de investigación se realizará por paralelo. Es decir, por cada paralelo se busca descubrir con cuántos factores se pueden describir los datos, extraer estos factores, y determinar cuáles son las preguntas más importantes y que explican la mayor cantidad de varianza de los datos.

Para determinar el número de factores necesarios para describir los datos, se usará la Regla de Kaiser [10], la cual nos indica que se deben "*conservar solamente aquellos factores cuyos valores propios (eigenvalues) son mayores a la unidad*". Esto se explicará con mayor detalle en la sección 5.4 Diseño del Reporte de este mismo capítulo.

Para extraer los factores se usará el método del Factor Principal, el cual permitirá estimar la matriz de carga de las variables sobre los

factores. Este método está basado en Componentes Principales, y es ampliamente usado por los paquetes estadísticos informáticos.

Finalmente, para obtener las relaciones entre factores y preguntas, será aplicada una rotación varimax a los factores, para determinar con mayor facilidad qué pregunta corresponde a qué factor, y mostrarlo en el reporte.

5.3 Diseño de la Aplicación

El primer paso para realizar el análisis es obtener todas las respuestas de los cuestionarios que pertenecen a un paralelo dado. Con estos datos se organiza una matriz de *nxp* (donde *n* es el número de estudiantes que evaluaron la materia y *p* es el número de preguntas que tiene el formulario). Esta matriz generada, será la analizada y a la que se le aplicará el procedimiento para reducir el número de variables *p* a un número de factores *r*, donde r<p[6].

A esta matriz de *nxp* (muestras vs variables, estudiantes vs preguntas), se le aplica el método del Factor Principal, iterando hasta converger en una solución para la matriz de carga.

De esta matriz de carga se extraen solamente los factores con valor propio (eigenvalue) mayores a 1. Este proceso es clave, pues determina el número de factores con el que se va a trabajar, y el porcentaje de varianza explicada por el conjunto de factores elegidos.

La matriz de carga también se denomina matriz de *“comunalidades”*, y tiene la particularidad de no ser única, es decir, no existe una solución única de ella para el método Factorial. Bajo este concepto, cuando se obtiene una solución a esta matriz, se trata de buscar una optimización mediante la Rotación de Factores, que como se explicó en la sección anterior, se realizará por medio del método Varimax.

Las nuevas comunalidades obtenidas a través de la rotación de factores se guardan en la base de datos, así como los puntajes de los formularios sobre los nuevos factores. Estos datos pueden servir para trabajar futuros algoritmos, como el de clusterización que se explica en el capítulo siguiente.

En la Figura 5.1 se puede observar el esquema completo del Análisis Factorial, desde el primer paso, que es obtener los datos a analizar, hasta el último, que es guardar los resultados del análisis en la base.

Figura 5.1 Diagrama de Flujo del Análisis Factorial

5.4 Diseño e Interpretación del Reporte

La estructura del reporte es similar en estructura a la del Análisis de Correspondencia, manteniendo así una consistencia y coherencia visual en todos los reportes del sistema.

Lo primero que se muestra en el reporte es una tabla con los factores obtenidos junto con su valor propio y el porcentaje de varianza que explican. Todos los factores con valor propio mayor a 1 son resaltados, pues son éstos los que van a ser extraídos. Se puede observar también cuál es el porcentaje de varianza que explican los factores, con lo cual se puede estar seguro que se está reduciendo el tamaño del problema con una mínima pérdida de información (Tabla 5.1).

Factor	Valor	Varianza	Varianza acumulada
1	13.219	0.568	0.568
2	4.816	0.207	0.775
3	2.075	0.089	0.864
4	1.000	0.043	0.907
5	0.676	0.029	0.936
6	0.451	0.019	0.956
7	0.297	0.013	0.969
8	0.249	0.011	0.979
9	0.167	0.007	0.986
10	0.147	0.006	0.993
11	0.097	0.004	0.997
12	0.051	0.002	0.999
13	0.019	0.001	1.000
14	0.001	0.000	1.000
15	-0.035	-0.002	1.000
16	-0.048	-0.002	1.000
17	-0.052	-0.002	1.000
18	-0.062	-0.003	1.000
19	-0.069	-0.003	1.000
20	-0.079	-0.003	1.000
21	-0.088	-0.004	1.000
22	-0.091	-0.004	1.000
23	-0.099	-0.004	1.000
24	-0.107	-0.005	1.000
25	-0.110	-0.005	1.000
26	-0.121	-0.005	1.000
27	-0.127	-0.005	1.000
28	-0.140	-0.006	1.000
29	-0.151	-0.006	1.000
30	-0.158	-0.007	1.000
31	-0.174	-0.007	1.000
32	-0.183	-0.008	1.000
33	-0.262	-0.011	1.000

Tabla 5.1 Tabla de factores con sus valores propios

De igual manera, se muestra el gráfico de sedimentación (Figura 5.2), que no es más que el gráfico de los factores y su correspondiente valor propio. Este gráfico, en inglés denominado

“Scree Plot”, fue propuesto por Catell (1966) [11], y también puede ser usado para determinar el número de factores, es la denominada “Prueba del Codo”, en la cual se traza una línea por encima de los factores menores, y todos los factores que queden encima de la línea son extraídos. En el reporte, la línea se dibuja en la abscisa y=1, que es la línea que determina en este caso cuáles factores se extraen y cuáles no.

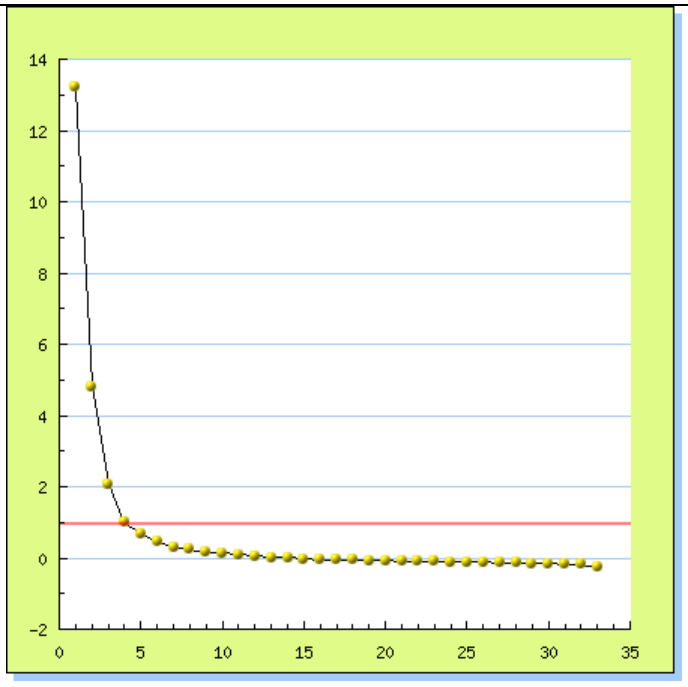


Figura 5.2 Gráfico de Sedimentación para elegir el número número de factores

Luego de esto se muestra una tabla que corresponde a la matriz de comunalidades, es decir, la relación que tiene cada factor con las preguntas. Aquí se puede ver cuáles preguntas están más

relacionadas con los factores, mediante sus pesos. Mientras mayor sea el peso o la comunalidad de esa pregunta sobre el factor, significa que mayor es la relación entre ambos. En la tabla siempre se resaltan los valores cuyo valor sea mayor de 0.5, lo cual nos ayuda a observar los como se agrupan las preguntas en base a los factores. (Tabla 5.2)

Pregunta	Factor 1	Factor 2	Factor 3	Factor 4
Asistí regularmente a clases	0.1519	1.1054	0.0230	0.1526
Asistí puntualmente a clases	0.1544	1.0919	0.0802	0.0560
Permanezco en clase durante toda la sesión programada	0.0345	1.0163	0.2571	0.0151
Participé de forma activa en clases, solicitando aclarar dudas, respondiendo preguntas, aportando con ejemplos	0.1309	1.1199	0.0999	0.0330
En general juzgo que mi esfuerzo y dedicación fueron apropiados	0.1383	1.0040	0.0273	0.1095
De acuerdo con sus respuestas anteriores, usted evalúa el desempeño del profesor(a) como:	0.4580	0.2165	0.3384	0.1033
Al inicio del término proporcionó y explicó a los estudiantes la programación y políticas del curso.	0.4316	0.1769	0.3597	0.1401
Refleja una adecuada preparación de sus clases.	0.4044	0.1182	0.3725	0.2539
Cumple con la programación propuesta al inicio del curso.	0.4545	0.1283	0.3459	0.1459
Relaciona los conocimientos previos de los estudiantes para el desarrollo de nuevos contenidos.	0.5175	0.1827	0.3695	0.1896
Presenta los contenidos de la clase de una manera comprensible.	0.6640	0.1004	0.4291	0.1611
Enfatiza durante la clase los puntos principales de los temas que expone.	0.4909	0.1269	0.3537	0.3500
Utiliza material de apoyo didáctico para reforzar los contenidos de las clases.	0.2906	0.1756	0.2813	0.4591
Presenta ejemplos apropiados para la comprensión de lo tratado en clase.	0.4417	0.1263	0.2315	0.5203
Promueve el razonamiento de los temas tratados.	0.7389	0.1077	0.2651	0.1677
Desarrolla los contenidos de la materia con un ritmo apropiado.	0.7511	0.2034	0.1223	0.2865
Fomenta el trabajo en equipo.	0.4581	0.2220	0.2012	0.3692
Tiene predisposición para aclarar dudas y ofrecer asesorías dentro y fuera de clases.	0.2286	0.0470	0.6407	0.1285
Facilita la participación activa de los estudiantes en clase	0.4561	0.1537	0.4298	0.2022
Es respetuoso y cordial en el trato con los estudiantes.	0.1164	0.0332	0.7784	0.1171
Estimula en la clase la formación de valores éticos y las buenas costumbres de los estudiantes.	0.0984	0.0540	0.7254	0.3199
Realiza evaluaciones periódicas (deberes, lecciones, proyectos, pruebas, etc.).	0.7161	0.2481	0.0630	0.3466
Formula claramente las preguntas en las evaluaciones.	0.3979	0.1312	0.1890	0.5159
Los temas en las evaluaciones son representativos del contenido del curso.	0.2373	0.1544	0.2464	0.6397
Califica procedimientos y resultados en las evaluaciones de los temas de examen.	0.3220	0.1062	0.1795	0.6826
Cumple con las políticas de evaluación señaladas para el curso.	0.2532	0.1884	0.2957	0.3947
Hace conocer los resultados de las evaluaciones periódicas en plazos oportunos a sus estudiantes.	0.2913	0.1227	0.3022	0.4119
Asiste puntualmente a clases (llega y se retira dentro del tiempo reglamentario).	0.4824	0.1533	0.2246	0.1630
Asiste regularmente a clases (frecuencia).	0.2226	0.0792	0.3435	0.3208
Contesta en forma satisfactoria las preguntas formuladas en clase.	0.2684	0.0696	0.5155	0.5038
Asigna actividades que requieren investigación por parte de los estudiantes.	1.0343	0.1934	0.0197	0.0550
Organiza durante la clase actividades de autoaprendizaje.	1.0429	0.2422	0.0800	0.1118
Promueve en el estudiante el pensamiento crítico.	0.7076	0.0686	0.2027	0.1977

Tabla 5.2 Comunalidades, o pesos que tienen los factores sobre las preguntas

Finalmente, se presenta un vínculo para ver el siguiente reporte, el cual será analizado en detalle en el siguiente capítulo, pero que utiliza los resultados de este Análisis Factorial como datos de entrada para su estudio.

5.5 Plan de Pruebas

Para probar la validez de este análisis, fue también implementado en Matlab 6.1 y los resultados ploteados en SPSS 13, obteniendo los mismos datos y probando así que el algoritmo estaba correctamente implementado, como lo muestran las Figuras Figura 5.3 Gráfico de Sedimentación obtenido mediante el SPSS 13.0 y Figura 5.4 Gráfico de Sedimentación mostrado en el reporte del CENACAD.

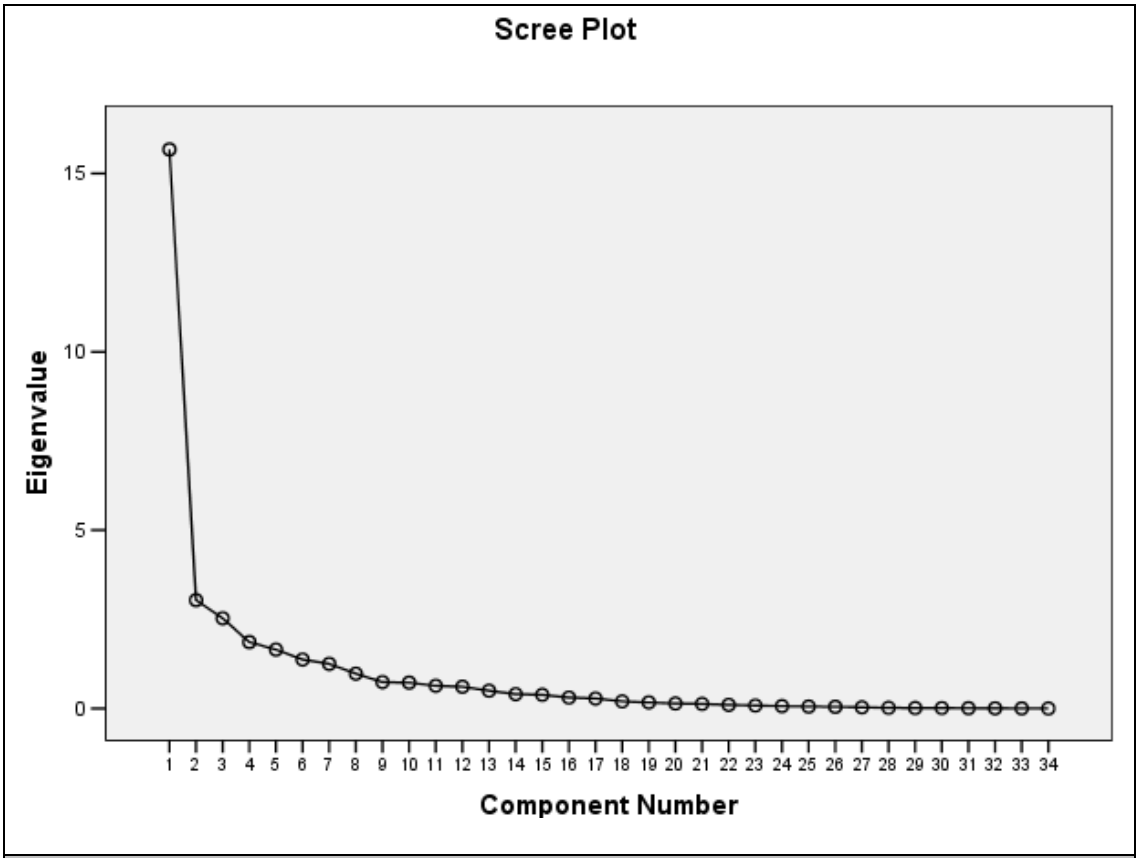


Figura 5.3 Gráfico de Sedimentación obtenido mediante el SPSS 13.0

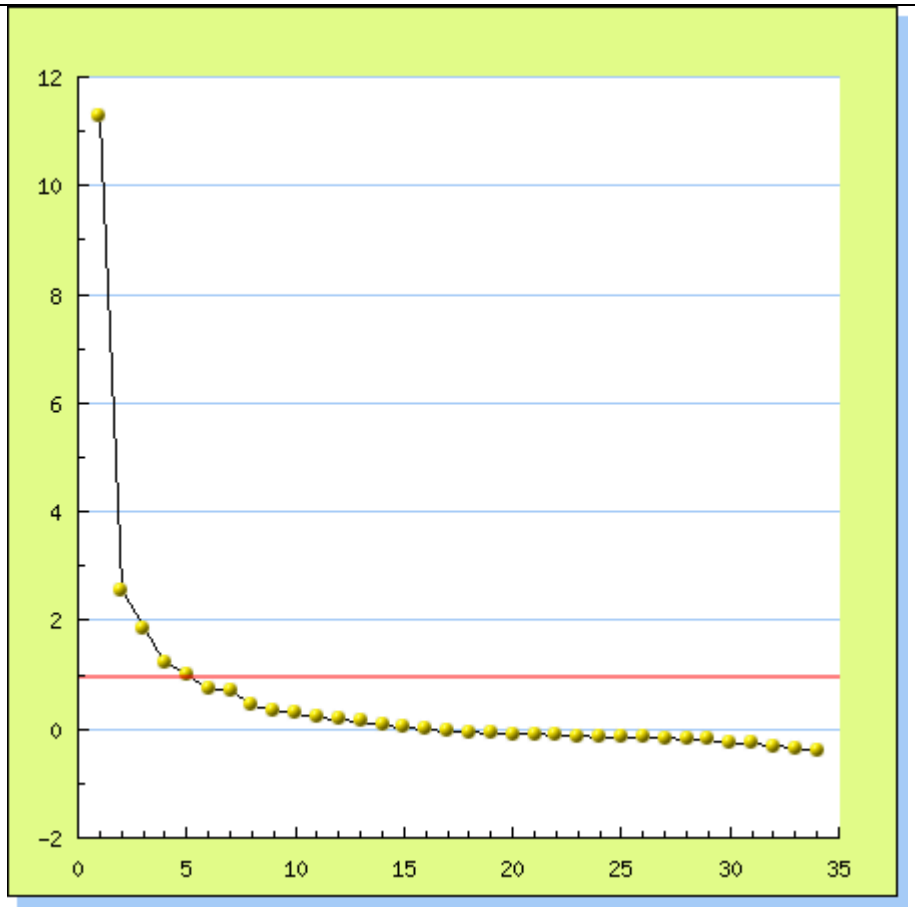


Figura 5.4 Gráfico de Sedimentación mostrado en el reporte del CENACAD

Fueron realizadas también pruebas de rendimiento, pues este análisis es el más complejo (matemáticamente hablando) y el que más cálculos necesita. Teniendo en cuenta la gran cantidad de datos que maneja el CENACAD, los tiempos de ejecución requerían atención, pues podían resultar considerablemente más altos que aquellos de análisis anteriores.

En la Figura 5.5 podemos ver gráficamente los tiempos de ejecución de este análisis, tomando en cuenta todos los datos del CENACAD.

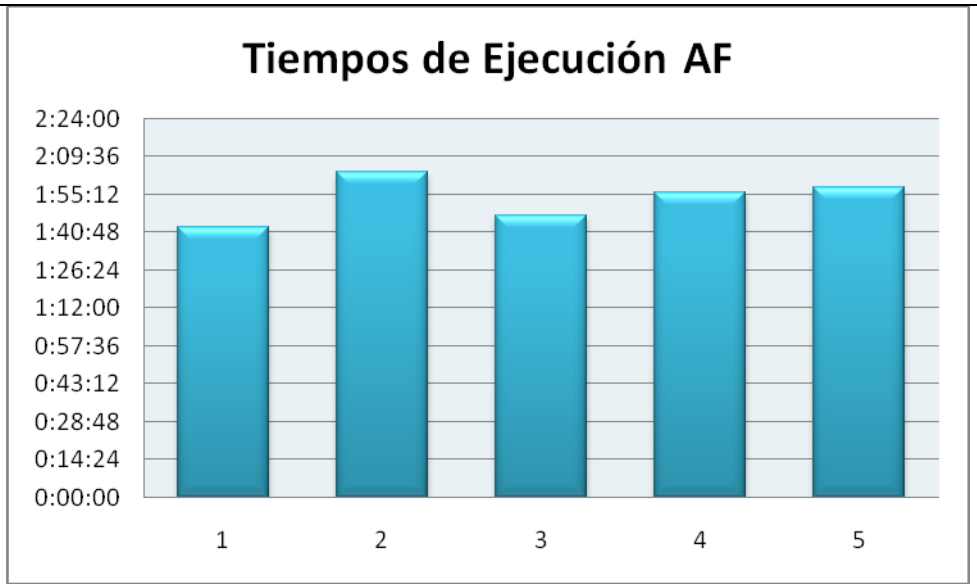


Figura 5.5 Tiempos de Ejecución del Análisis Factorial

Como se puede observar, el tiempo promedio de ejecución es de 1 hora y 53 minutos, lo cual es comprensible, al contar con un gran número de cálculos matemáticos. Sin embargo, y a pesar de parecer mucho tiempo comparado con los anteriores análisis, estos tiempos de ejecución son aceptables, teniendo en cuenta que el análisis con todos los datos del CENACAD sólo se realiza una vez, de ahí en adelante sólo se analizan las encuestas nuevas. Por lo tanto, podemos concluir que el algoritmo es eficiente y cumple su propósito.

En el presente capítulo, fueron presentados todos los detalles del proceso seguido para implementar el Análisis Factorial en las encuestas, y la manera correcta de interpretar sus reportes y resultados.

En el siguiente capítulo se hablará sobre el Análisis de Conglomerados o Clusterización, el cual aprovecha los resultados obtenidos por el presente análisis, tomándolos como datos de entrada para su algoritmo.

CAPÍTULO 6

6 ANÁLISIS DE CONGLOMERADOS

En el presenta capítulo, se explicará el Análisis de Conglomerados o Clusterización, y su aplicación en las encuestas del CENACAD.

A continuación se detallarán todos los elementos involucrados en la implementación de la Clusterización, el análisis previo, las decisiones de diseño, y la implementación y reporte de los resultados.

Con el Análisis de Conglomerados se puede lograr que una muestra grande de datos se reduzca objetivamente, y de este modo se puede obtener información más clara de grupos específicos. Con esto los datos son más concretos y se obtiene una descripción que puede ser comprendida de mejor manera para quienes deben tomar decisiones en base a estos grupos.

Con lo anteriormente expuesto puede concluirse que no existirá una única o definitiva solución al problema presentado en un análisis de clusterización, pues los conocimientos del investigador y el comportamiento de lo estudiado jugarán un papel muy importante a la hora de decidir soluciones.

El análisis de clusterización debe cumplir con las siguientes etapas[17]:

Selección de la muestra de datos

La selección de la muestra a ser utilizada debe de ser el máximo número de datos a ser estudiados para que exista mayor confiabilidad en los resultados.

Selección y transformación de las variables a utilizar

Deben encontrarse variables que sean representativas al estudio, no se deben elegir variables irrelevantes indiscriminadamente pues esto podría dar parte a la inclusión de datos atípicos en los resultados.

En cuanto a la transformación de las variables se debe de tomar en cuenta si:

- Al afectar una de las variables esto puede interferir en decisiones posteriores.
- La estandarización por variable resulta útil para futuras mediciones de distancia que puedan afectar los resultados de los análisis y no se recomienda estandarizarlas si esto refleja algo natural de los individuos estudiados.
- La estandarización por encuestado elimina patrones de los sujetos en estudio, ofreciendo poca o mucha relevancia a los mismos.
- La factorización de las variables resulta más conveniente que trabajar con la muestra total.
- El tipo de escala de medida puede afectar a etapas posteriores del análisis.

En esta tesis ha sido implementada la factorización de las variables ya que al inicio del análisis se factorizan las variables existentes (preguntas) para realizar el estudio de manera óptima.

Selección del concepto de distancia o similitud

Las medidas de similitud / distancia definen cercanía, no covariación (conexión entre las variables). Entre algunas de las medidas de distancia existentes tenemos:

- Distancia Euclídea: medida que representa la disimilaridad o distancia para datos continuos. Esta distancia se puede expresar como la raíz cuadrada de la suma de los cuadrados de las diferencia entre los valores de los elementos[25].

$$d_{ij} = \sqrt{\sum_{k=1}^{t} (X_{ik} - X_{jk})^2}$$

- Distancia Manhattan: esta distancia también conocida como City Block o distancia por manzanas, es una medida de disimilaridad o distancia para datos continuos. Está definida como suma de los valores absolutos de las diferencias entre los valores de los elementos[26]:

$$d_{ij} = \sum_{k=1}^{t} \left| X_{ik} - X_{jk} \right|$$

- Distancia Mahalanobis: medida estandarizada de la distancia euclídea. Los datos se estandarizan escalando las respuestas en términos de desviaciones típicas; es decir se ajustan las intercorrelaciones entre las variables[27].

$$d_{ij} = \left(X_i - X_j\right)' \sum^{-1} \left(X_i - X_j\right)$$

  En esta tesis se utilizará la medida de distancia euclídea que no es más que la raíz cuadrada de la resta de las longitudes entre 2 o más puntos.

Selección y aplicación del criterio de agrupación

En el análisis de clusterización se pueden seleccionar criterios de agrupación para los datos de 2 tipos:

- Los métodos jerárquicos,- que es aquel donde la agrupación se realiza mediante un proceso con fases de agrupación y desagrupación sucesivas.
- Los métodos no jerárquicos.- también llamados de K- Medias, que buscan un centro dentro de todos los datos presentados y van colocándolos en agrupaciones.

En este proyecto será utilizado un método no jerárquico llamado **Umbral Paralelo** donde se seleccionan simultáneamente centros de grupos y se agrupan aquellas semillas (valores) que están más cerca de un centro en el umbral.

Determinación de la estructura correcta

Para la elección del número de grupos correcto no existe un método estadístico como tal. Sin embargo, para encontrar el número de grupos que deben de considerarse se deben de tener en cuenta las variables iniciales, las observaciones realizadas a los datos y las conclusiones que se hayan dado en las etapas de agrupación

6.1 Alcance de la Solución

El presente análisis tratará de clasificar o agrupar todos los estudiantes dentro de un paralelo determinado. Esta clasificación se dará en base a las respuestas que hayan dado a los cuestionarios. La intención de este análisis es descubrir cuántos grupos de estudiantes existen en cada paralelo, y cómo está conformado cada grupo.

6.2 Análisis de la Solución

Lo primero que se debe analizar para realizar el Análisis de Conglomerados, es decidir en qué se va a basar el algoritmo para clusterizar a los estudiantes. La primera apreciación es basarse en las respuestas de las preguntas de cada encuesta, pero esto presenta algunos inconvenientes:

- Son 34 preguntas (generalmente), lo cual implica un número igual de variables para clusterizar, y por cada paralelo, representará un altísimo costo computacional para realizar el análisis.

- Las 34 preguntas no son independientes entre sí, muchas están relacionadas en mayor o menor grado, y si se realiza el análisis con estas variables, se introducirá ruido en el análisis al haber datos redundantes.

Otra alternativa que se presenta es clusterizar los datos respecto al promedio del formulario. El promedio es el puntaje que obtiene el formulario en base a unos pesos predefinidos. El problema de clusterizar en base al promedio es que se pierde mucha información, puesto que al fin y al cabo sólo se estaría tomando en cuenta 1

variable para el análisis, lo cual no puede representar de buena manera las 34 variables iniciales.

Finalmente, la clusterización se realizó en base a los factores obtenidos en el Análisis Factorial, el cual fue desarrollado por paralelo. Los puntajes de los formularios sobre los factores son los datos de entrada de la clusterización. Esto presenta varias ventajas respecto a las anteriores alternativas de agrupamiento:

- Reducido número de factores (en promedio salen 3 o 4 factores por paralelo)
- Reducción significativa de los datos, con pérdida mínima de información relevante.
- Los factores son independientes entre sí, por lo tanto no hay datos redundantes.

Un punto bastante importante en el Análisis de Conglomerados, es el número de grupos que se eligen para agrupar. Esta elección debe hacerse de tal manera que se minimice la suma de cuadrados de distancias (SCDG) entre cada elemento y el centroide de su grupo.

Este proceso es similar en concepto al proceso para escoger el número de factores en el capítulo anterior, es decir, es posible

realizar un gráfico de sedimentación para determinar el número óptimo de grupos de manera visual.

En el proceso de Clusterización que se implementó en el CENACAD, se usó un procedimiento sugerido por Hartigan (1975) [12], en el cual se realiza un test *F* de reducción de variabilidad, comparando la SCDG de *k* grupos, con la SCDG de *k+1* grupos, y se decide aumentar el número de grupos si y sólo si el valor de *F* es mayor a 10. El valor de *F* es obtenido mediante:

$$F = \frac{SCDG(G) - SCDG(G+1)}{SCDG(G+1) \Big/ (n-G-1)}$$

6.3 Diseño de la Aplicación

Como primer paso, y como es común en los análisis ya presentados, se extrae de la base los datos que van a ser clusterizados. Como se analizó en la sección anterior, los datos serán los puntajes sobre los factores principales obtenidos en el capítulo anterior.

A continuación se realiza la clusterización propiamente dicha, usando el algoritmo de las K-medias. Se inicia el proceso clusterizando los datos en 2 grupos, y se empieza a aumentar sucesivamente el número de grupos hasta alcanzar el criterio de optimalidad expuesto

en la sección anterior. Al final del algoritmo de las K-medias, se obtiene cuántos grupos existen en cada paralelo y a qué grupo pertenece cada formulario.

Finalmente estos datos se guardan en las tablas de hecho, para ser consultados por los reportes de las páginas.

A continuación se explica en detalle el proceso del algoritmo de las K-medias, el número de grupos se representa por k[7].

1. Se empieza el algoritmo asumiendo k=2
2. Se escogen los k primeros formularios distintos como centros de los grupos. Es decir, se asignan centros aleatorios a los grupos para empezar.
3. Uno por uno, se asignan todos los formularios al grupo cuyo centro está más cercano a él.
4. Finalmente, se recalculan los centros de los grupos.
5. Si los centros han cambiado respecto a la iteración anterior, se regresa a (2), sino, significa que cada formulario ya se encuentra en el grupo al que pertenece.
6. Se calcula el criterio de optimalidad para la elección del número de grupos. Si este criterio no se cumple, se regresa a

(2), pero aumentando el número de grupos. Si se cumple el criterio, se asume que el número de grupos es el correcto y el algoritmo se termina.

En la Figura 6.1, se muestra un detallado flujo sobre el proceso que sigue el algoritmo.

Figura 6.1 Diagrama de Flujo del Análisis de Conglomerados

6.4 Diseño e Interpretación del Reporte

Lo primero que se muestra en el reporte es una tabla que contiene los grupos en los que se dividió el paralelo, el número de estudiantes que contiene cada grupo, y el color con el que se lo va a representar en los siguientes gráficos. (Tabla 6.1)

Grupo	Símbolo	# estudiantes
1		8
2		4
3		6
4		3
5		4

Tabla 6.1 Muestra los grupos, su simbología en el gráfico, y el número de estudiantes que pertenecen al mismo

A continuación, se muestra en un gráfico bi-dimensional los centroides de los grupos (Figura 6.2). En el eje de las X se encuentran los factores extraidos, y en el eje de las Y se especifica el puntaje alcanzado por el centroide de ese grupo, en ese factor.

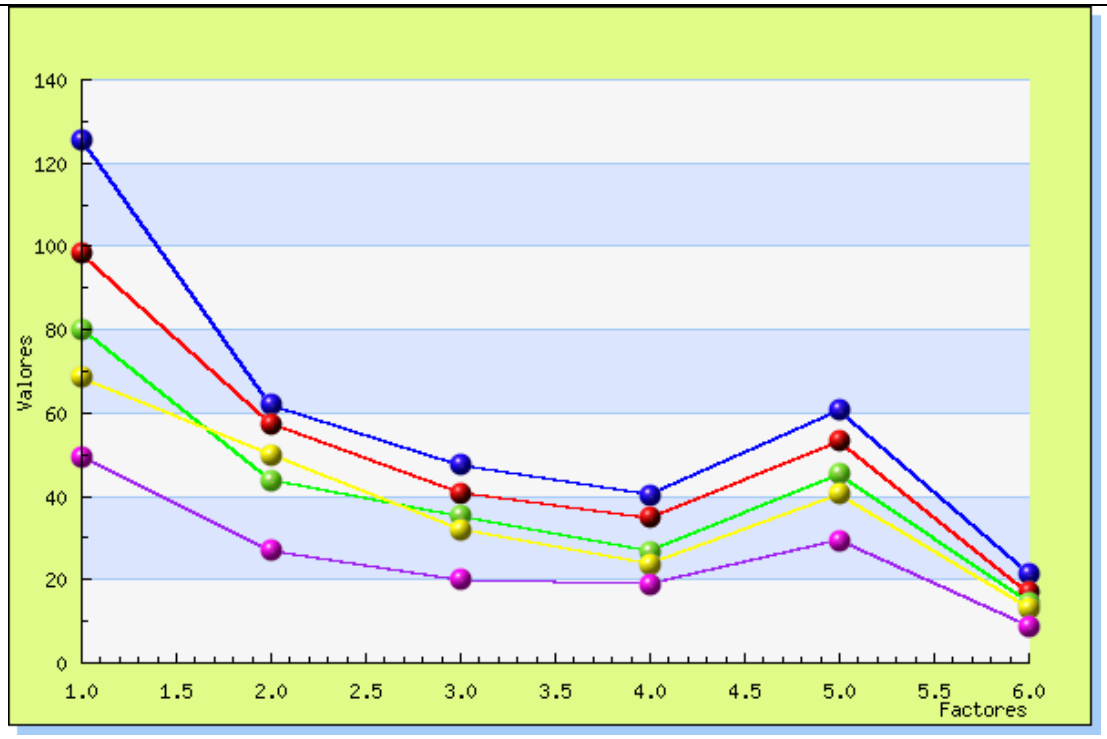


Figura 6.2 Gráfico de los centroides de los grupos, en base a los factores

Para poder mostrar de una manera clara los grupos de estudiantes y la estructura del curso respecto a los grupos, también se muestra un gráfico de los estudiantes, proyectados sobre los 2 primeros factores (los más importantes), y cada estudiante se pinta del color que le corresponde según el grupo al que pertenece. Esto permite visualizar de una manera clara cómo se ha llevado a cabo el proceso de categorización, e incluso permite evaluar la clusterización realizada. (Figura 6.3)

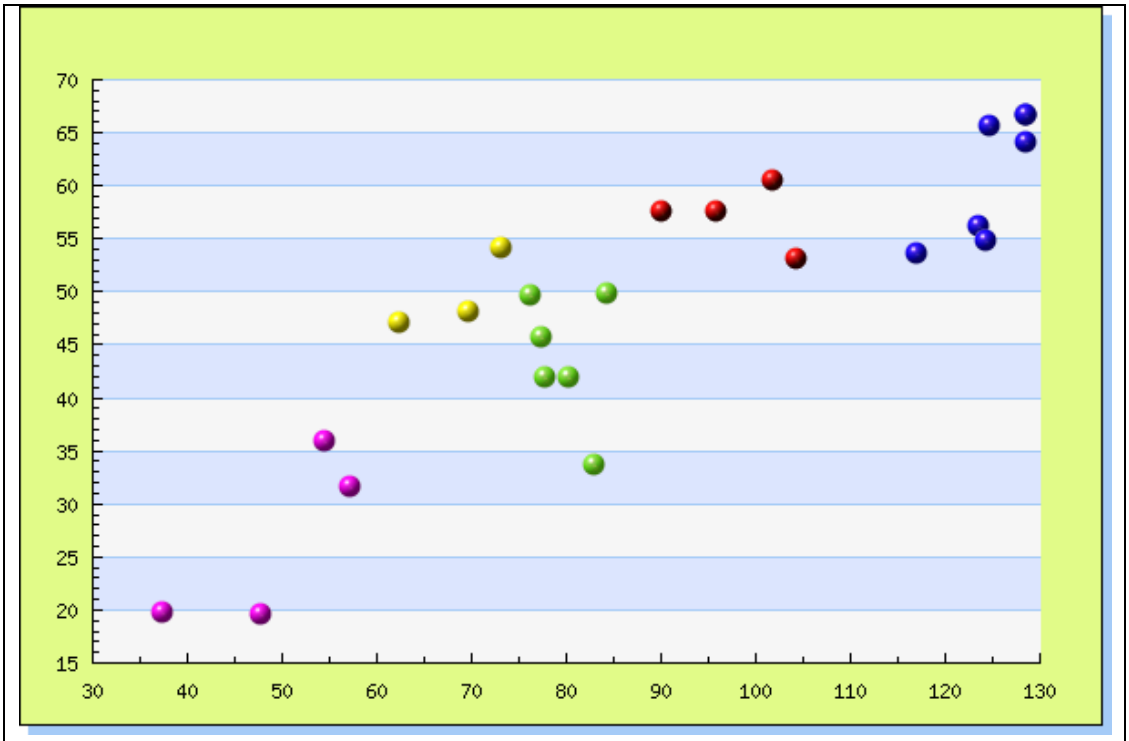


Figura 6.3 Gráfico de los formularios sobre los 2 factores principales

6.5 Plan de Pruebas

Para calcular la validez del algoritmo de clusterización, se realizaron pruebas con distinto número de grupos, para comprobar visualmente que los grupos estén siguiendo un orden lógico. En las siguientes imágenes, se puede observar el proceso de clusterización en cada iteración, esto quiere decir, que en cada gráfico se aumenta en uno el número de grupos, para observar el comportamiento del agrupamiento.

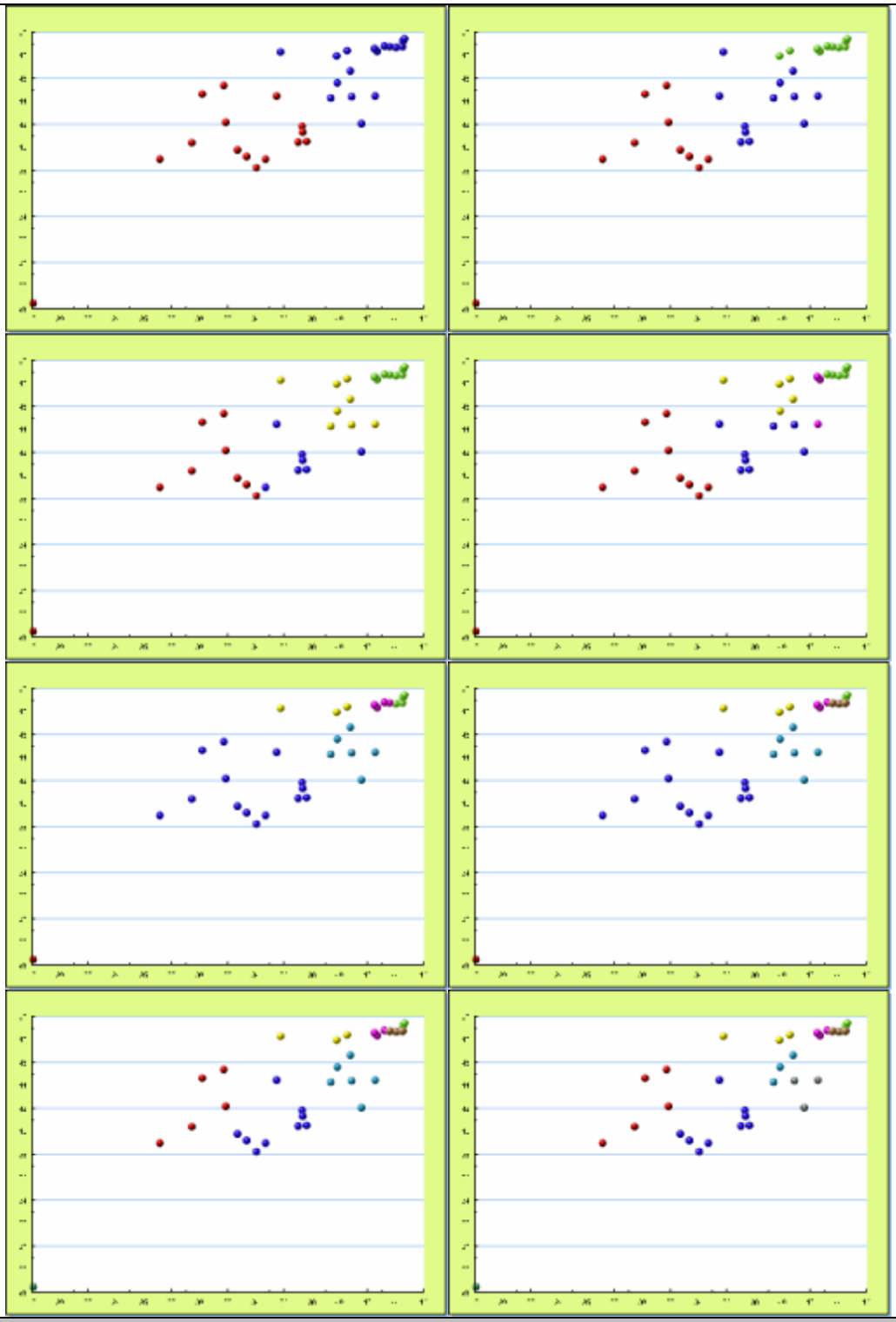


Figura 6.4 Pruebas de clusterización con distinto número de grupos: en cada gráfico se aumenta un grupo.

Como se observa en la Figura 6.4, el proceso de clusterización siguió un orden lógico, por lo cual se puede concluir que el algoritmo es válido y los resultados son confiables.

Al igual que en los algoritmos anteriores, se realizaron también pruebas de tiempos de ejecución, para determinar en promedio cuánto tiempo se tomaría en realizar el proceso de clusterización en todos los datos del CENACAD.

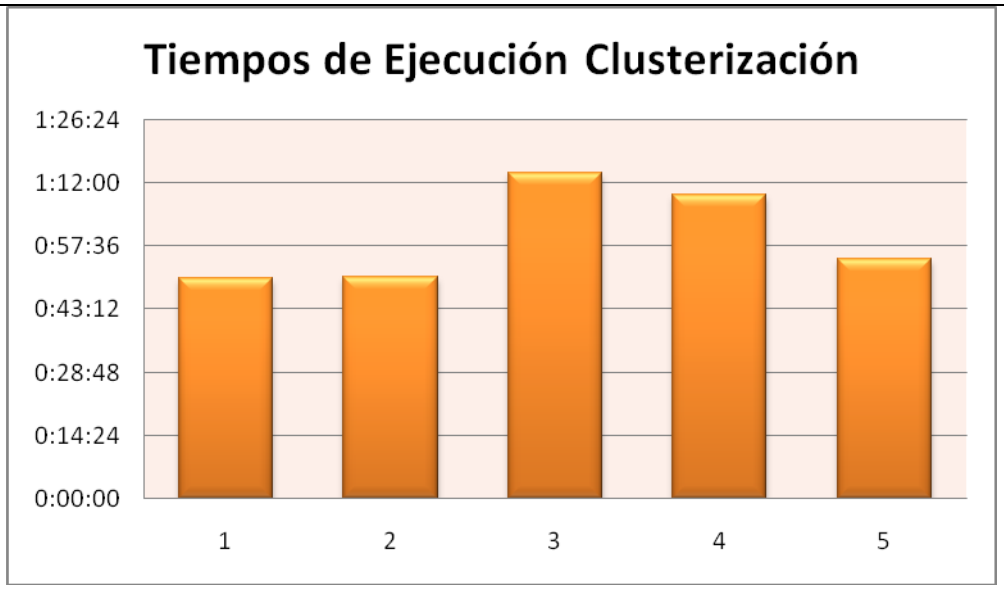


Figura 6.5 Tiempos de Ejecución del Análisis de Conglomerados

Como se puede observar en la Figura 6.5, el promedio de ejecución de todos los datos es de 1 hora, la mitad del método anterior, y consecuentemente aceptable para la cantidad de datos procesados.

En este capítulo, se analizó en detalle la aplicación de la Clusterización a las encuestas, sus resultados y conclusiones particulares.

Con este capítulo se acaba la serie de capítulos relacionados con técnicas estadísticas aplicadas a los datos. El próximo capítulo tratará acerca de las conclusiones que se han obtenido a través de todos los análisis efectuados.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. La combinación de las técnicas de estadística inferencial junto con la minería de datos implementadas en esta tesis permite generar información y elaborar reportes que ayuden en la toma de decisiones para mejorar el rendimiento educacional que por años ha mantenido la ESPOL.

2. La integración del Sistema desarrollado a la interfaz web actual del CENACAD permite a los profesores y directivos acceder y evaluar la información adicional sin tener que migrar los datos a otras herramientas de análisis estadístico ni aprender a usar dichas herramientas.

3. Los módulos desarrollados (análisis de correspondencia, escalado multidimensional, factorial y clusterización) reducen el número de variables que deben ser analizadas para emitir un juicio de valor por parte de cualquier directivo de alguna entidad educativa o empresarial, por lo podrían ser implementados en otras ramas tanto educativas como sociales.

4. Los reportes gráficos presentan la información de una manera más comprensible y amigable para el usuario que las tablas de resultados. Del desarrollo de este tema concluimos que es posible presentar gráficos/biplots que complementan los resultados mostrados numéricamente en las tablas de datos.

RECOMENDACIONES

1. Utilizar el sistema para tomar decisiones adecuadas que sirvan para la optimización de la evaluación de la información recogida por el CENADAD.

2. Realizar futuros módulos para incorporar otros estudios estadísticos como análisis discriminante, componentes principales, regresión múltiple.

3. En un futuro, dada la flexibilidad que brindan los métodos presentados y la estadística inferencial, incluir otras variables a ser estudiadas para aumentar la información disponible para ayudar en la toma de decisiones.

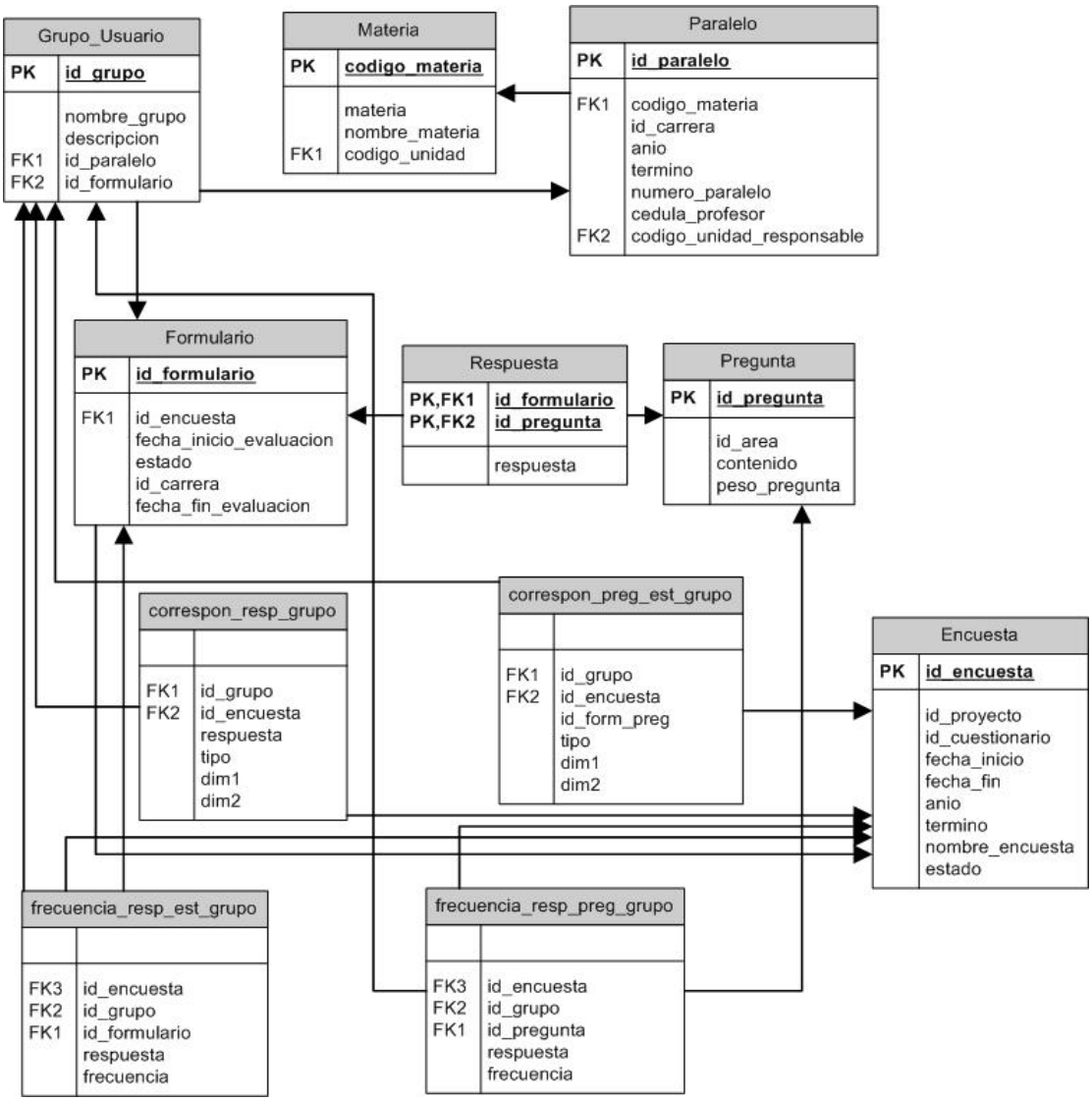
4. Presentar los módulos planteados al CSI, para aplicar estos análisis estadísticos a la base de datos de calificaciones de estudiantes, y poder determinar así por ejemplo factores que inciden en el éxito o fracaso de los estudiantes en una carrera específica.

5. Sugerimos medir la usabilidad y el tiempo tomado para el análisis de los reportes por parte de los usuarios.

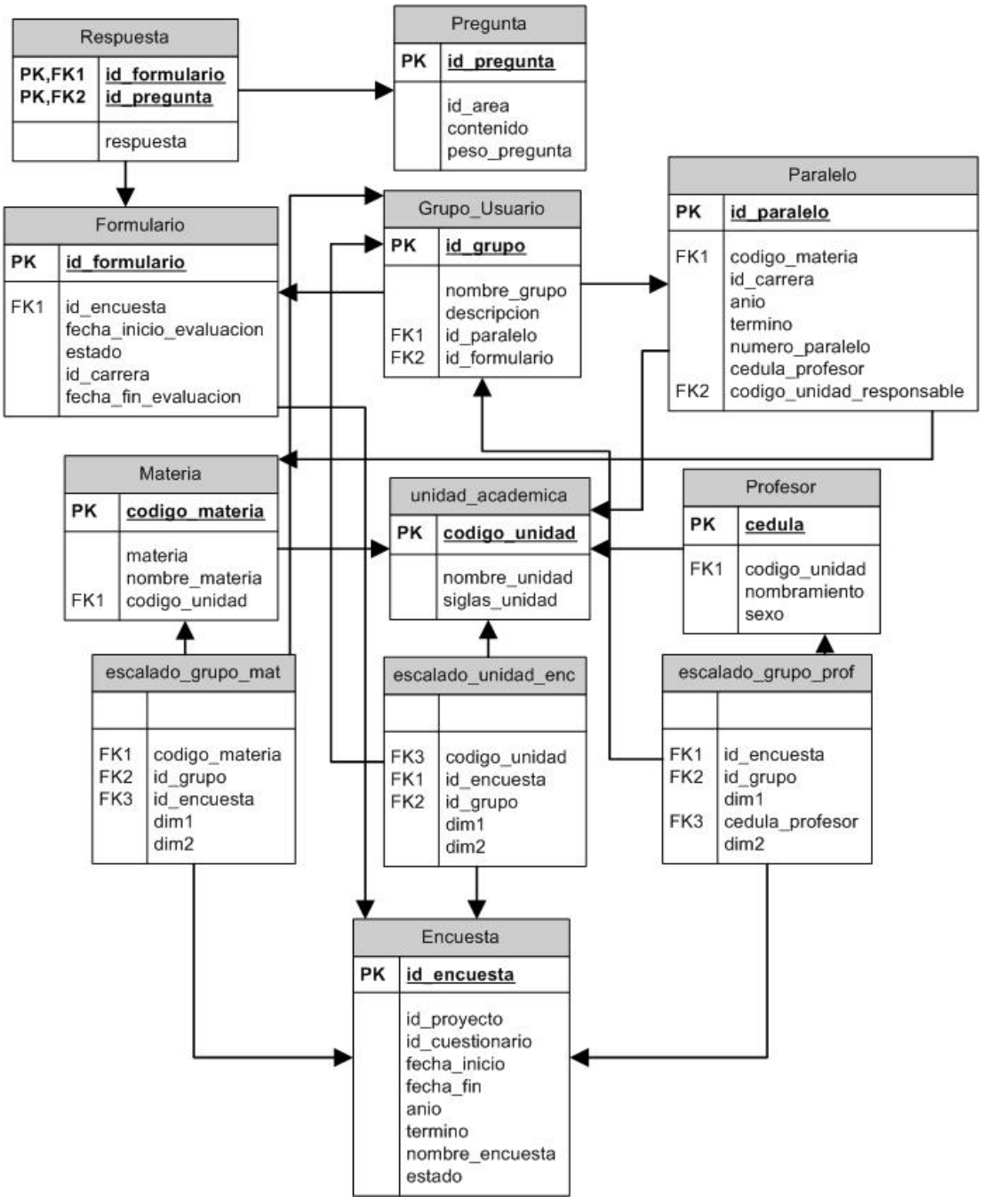
APÉNDICES

A APÉNDICE A: MODELOS LÓGICOS DE LOS ANÁLISIS IMPLEMENTADOS

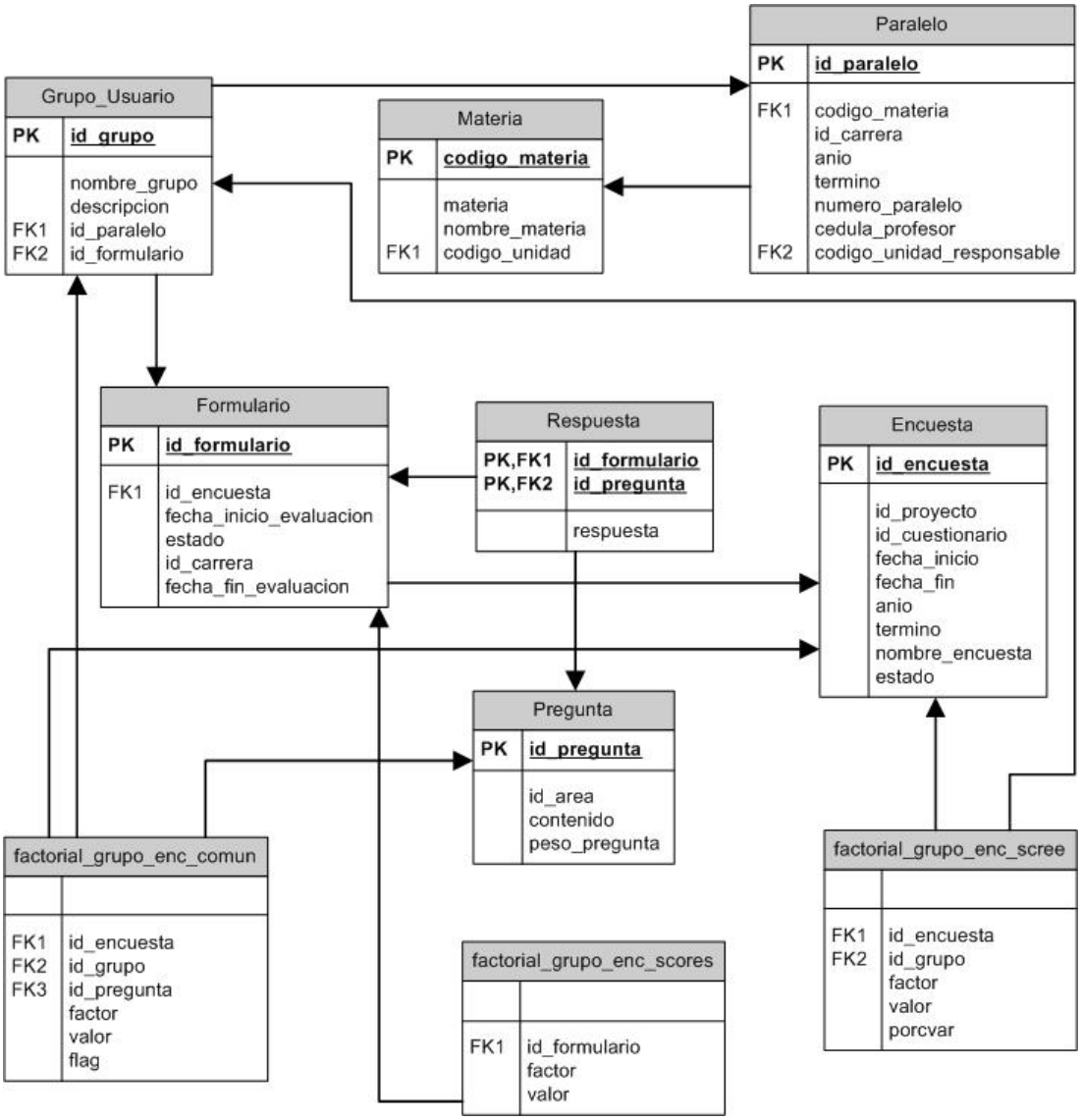
A.1 Análisis de Correspondencia



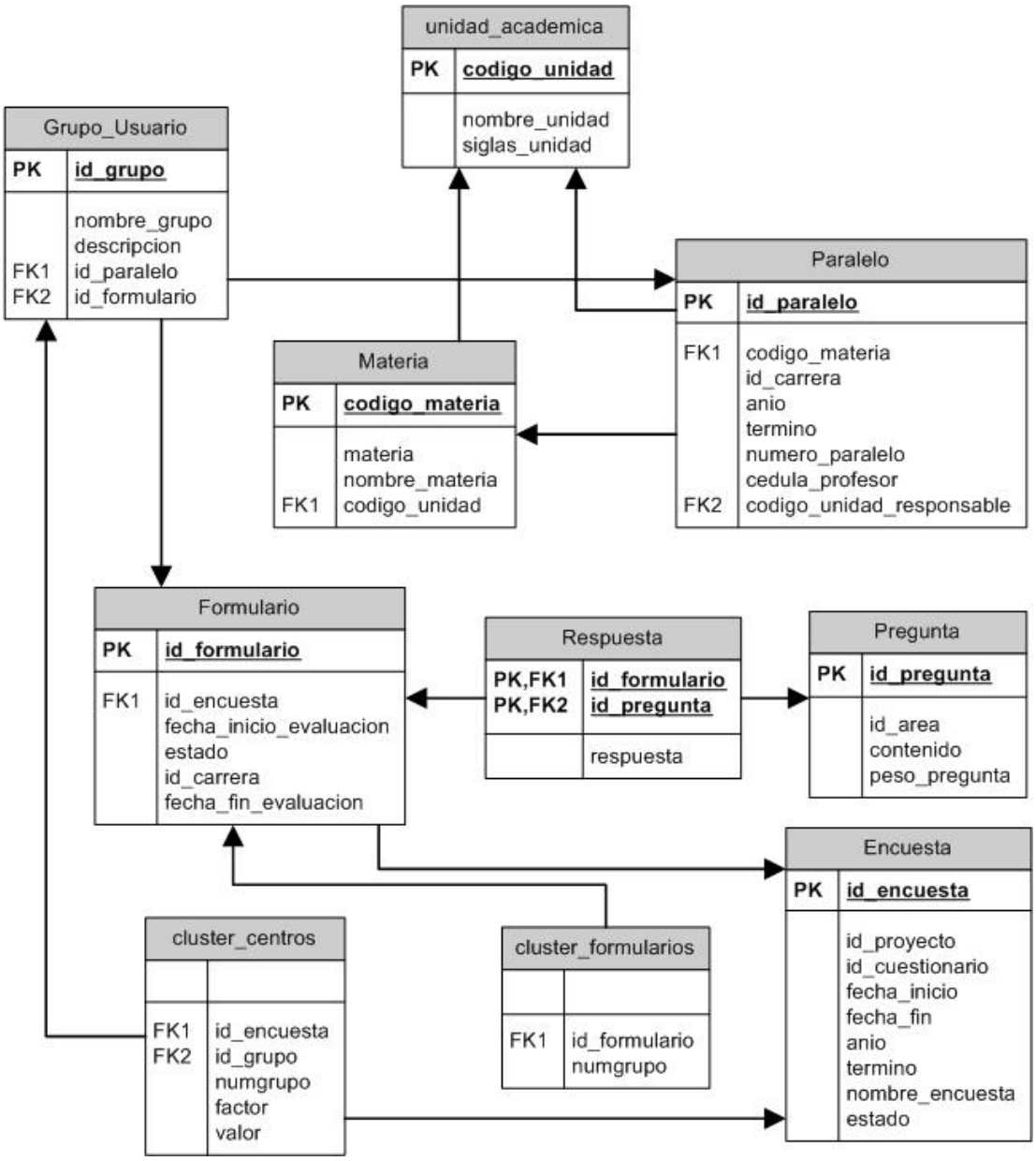
A.2 Escalado Multidimensional



A.3 Análisis Factorial



A.4 Análisis de Conglomerados



B APÉNDICE B: DICCIONARIO DE DATOS

B.1 Tablas de Reportes

Promedio de Preguntas por Grupo y Encuesta		
Nombre Físico: promedio_preg_grup_par_enc **Descripción:** Tabla que guarda los promedios de las preguntas de todos los estudiantes de un grupo (paralelo) en una encuesta específica.		
Campo	**Tipo de Dato**	**Descripción**
id_grupo	integer	Identificador del grupo
id_paralelo	integer	Identificador del paralelo
id_pregunta	integer	Identificador de la pregunta
id_encuesta	integer	Identificador de la encuesta
num_formularios	integer	Número de formularios evaluados
total_pregunta	double	Suma de todas las respuestas que ha obtenido
promedio_preg	double	Promedio de la pregunta

Promedio de Unidades por Encuesta		
Nombre Físico: promedio_unidad_enc **Descripción:** Tabla que guarda los promedios que cada unidad ha obtenido en una encuesta específica.		
Campo	**Tipo de Dato**	**Descripción**
id_unidad_responsable	integer	Identificador de la unidad académica
num_par_eval	integer	Número de paralelos que fueron evaluados
promedio_unid_enc	integer	Promedio que obtuvo la unidad en la encuesta
id_encuesta	integer	Identificador de la encuesta
suma_promedios	integer	Suma de todos los promedios
total_pregunta	double	Suma de todas las respuestas que ha obtenido
fecha_generada	double	Fecha en la que se generó esta información

Frecuenta de Respuestas por Pregunta		
Nombre Físico: frecuencia_resp_preg_grupo **Descripción:** Tabla que guarda la frecuencia dentro de un paralelo, de todas las preguntas.		
Campo	**Tipo de Dato**	**Descripción**
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo
respuesta	integer	Número de respuesta
frecuencia	integer	Frecuencia con la que apareció

Frecuenta de Respuestas por Estudiantes		
Nombre Físico: frecuencia_resp_est_grupo **Descripción:** Tabla que guarda la frecuencia dentro de un paralelo, de todas las preguntas.		
Campo	**Tipo de Dato**	**Descripción**
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo
Id_formulario	integer	Identificador del formulario
respuesta	integer	Número de respuesta
frecuencia	integer	Frecuencia con la que apareció

B.2 Tablas del Análisis de Correspondencia

Análisis de Correspondencia – Tabla de Respuestas		
Nombre Físico: correspon_resp_grupo **Descripción:** Tabla que guarda los puntajes de las 2 dimensiones más importantes de las respuestas.		
Campo	**Tipo de Dato**	**Descripción**
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
respuesta	integer	Alternativa de respuesta (1,2…5)
Tipo	integer	1: Estudiantes vs Respuestas 2: Preguntas vs Respuestas
dim1	Real	Puntaje en la primera dimensión (horizontal)
dim2	Real	Puntaje en la segunda dimensión

(vertical)

Análisis de Correspondencia – Tabla de Preguntas / Estudiantes		
Nombre Físico: correspon_preg_est_grupo **Descripción:** Tabla que guarda los puntajes de las 2 dimensiones más importantes de las preguntas o de los formularios.		
Campo	**Tipo de Dato**	**Descripción**
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
id_form_preg	integer	Identificador del formulario o de la pregunta
Tipo	integer	1: Estudiantes vs Respuestas 2: Preguntas vs Respuestas
dim1	real	Puntaje en la primera dimensión (horizontal)
dim2	real	Puntaje en la segunda dimensión (vertical)

B.3 Tablas del Escalado Multidimensional

Escalado Multidimensional – Tabla de Grupos por Materia		
Nombre Físico: escalado_grupo_mat **Descripción:** Tabla que guarda los puntajes del Escalado de todos los paralelos de una materia en particular		
Campo	**Tipo de Dato**	**Descripción**
cod_materia	char(12)	Código de la materia analizada
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
dim1	real	Puntaje en la primera dimensión (horizontal)
dim2	real	Puntaje en la segunda dimensión (vertical)

Escalado Multidimensional – Tabla de Grupos por Profesor
Nombre Físico: escalado_grupo_prof **Descripción:** Tabla que guarda los puntajes del Escalado de todos los paralelos que ha dictado un profesor en particular

Campo	Tipo de Dato	Descripción
cedula_profesor	char(15)	Identificador del profesor analizado
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
dim1	real	Puntaje en la primera dimensión (horizontal)
dim2	real	Puntaje en la segunda dimensión (vertical)

Escalado Multidimensional – Tabla de Unidades por Encuesta		
Nombre Físico: escalado_unidad_enc **Descripción:** Tabla que guarda los puntajes del Escalado de todas las unidades que se sometieron a una encuesta en particular		
Campo	**Tipo de Dato**	**Descripción**
codigo_unidad	char(8)	Código de la unidad académica
id_encuesta	integer	Identificador de la encuesta analizada
id_grupo	integer	Identificador del grupo evaluado
dim1	real	Puntaje en la primera dimensión (horizontal)
dim2	real	Puntaje en la segunda dimensión (vertical)

B.4 Tablas del Análisis Factorial

Análisis Factorial – Tabla de Comunalidades		
Nombre Físico: factorial_grupo_enc **Descripción:** Tabla que guarda las comunalidades de los factores obtenidos en cada pregunta.		
Campo	**Tipo de Dato**	**Descripción**
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
id_pregunta	integer	Identificador de la pregunta
factor	integer	Número de factor obtenido
valor	real	Puntaje del factor en la pregunta (comunalidad)
flag	smallint	1 si el valor es superior al criterio de selección de preguntas importantes, sino 0.

Análisis Factorial – Tabla de Puntajes		
Nombre Físico: factorial_grupo_enc_scores **Descripción:** Tabla que guarda los puntajes de cada formulario sobre los factores importantes.		
Campo	**Tipo de Dato**	**Descripción**
id_formulario	integer	Identificador del formulario
factor	integer	Número de factor obtenido
valor	real	Puntaje del formulario en el factor

Análisis Factorial – Tabla de Valores Propios		
Nombre Físico: factorial_grupo_enc_scree **Descripción:** Tabla que guarda los valores propios. Se usa para escoger el número óptimo de factores y para poder mostrar el gráfico de sedimentación.		
Campo	**Tipo de Dato**	**Descripción**
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
factor	integer	Número de factor obtenido
valor	real	Valor propio del factor
porcvar	real	Porcentaje de variación explicada por ese factor

B.5 Tablas del Análisis de Conglomerados

Análisis de Conglomerados – Tabla de Formularios		
Nombre Físico: cluster_formularios **Descripción:** Tabla que guarda en qué cluster se encuentra cada formulario.		
Campo	**Tipo de Dato**	**Descripción**
id_formulario	integer	Identificador del formulario
num_grupo	smallint	Número del grupo al que pertenece

Análisis de Conglomerados – Tabla de Centros del Cluster
Nombre Físico: cluster_centros **Descripción:** Tabla que guarda los valores centrales del cluster.

Campo	Tipo de Dato	Descripción
id_encuesta	integer	Identificador de la encuesta
id_grupo	integer	Identificador del grupo evaluado
numgrupo	integer	Número de grupo
factor	integer	Número de factor
valor	real	Valor central del factor en el grupo

REFERENCIAS BIBLIOGRÁFICAS

[1] **CENACAD**, Acerca de: <https://www.cenacad.espol.edu.ec/cenacad/index.php?module=Static&action=List&op=acerca>. 2007

[2] **ESPOL**, Mision: <http://www.espol.edu.ec/espol/main.jsp?urlpage=mision.jsp&id=7>. 2007

[3] **MOJAVI**, Mojavi, MVC Framework, < www.mojavi.org >, Febrero 2007

[4] **PEÑA, DANIEL,** "Análisis de Datos Multivariantes", McGraw-HILL. 2002, 171-180 p.

[5] **PEÑA, DANIEL,** "Análisis de Datos Multivariantes", McGraw-HILL. 2002, 195-206 p.

[6] **PEÑA, DANIEL,** "Análisis de Datos Multivariantes", McGraw-HILL. 2002, 347-348 p.

[7] **PEÑA, DANIEL,** "Análisis de Datos Multivariantes", McGraw-HILL. 2002, 217-230 p.

[8] **FERNÁNDEZ, FRANCISCO JAVIER, "**El uso del Análisis de Correspondencia Simple (ACS) como ayuda en la interpretación del dato en arqueología. Un caso de estudio.", Boletín Antropológico 20 (55). Agosto 2002, 687- 713 p.

[9] **CRIVISQUI, EDUARDO, "**Análisis Factorial de Correspondencia, un instrumento de investigación en ciencias sociales", Edición del

Laboratorio de Informática Social. Universidad Católica de Asunción. 1993.

[10] **KAISER, H.F.,** “The application of electronic computers to factor analysis”. *Educational and Psychological Measurement* (20), 1960, 141-151 p.

[11] **CATTELLL, R.B.**, “The Scree test for the number of factors”. *Multivariate Behavioral*, 1966

[12] **HARTIGAN, J.A,** “Clustering Algorithms”, 1975, NY: Wiley.

[13] **SALVADOR FIGUERAS, M,** "Análisis de Correspondencias", <http://www.5campus.com/leccion/correspondencias>. 2003, 1-2 p.

[14] **MODELO VISTA CONTROLADOR**, WIKIPEDIA LA ENCICLOPEDIA LIBRE, <http://es.wikipedia.org/wiki/Modelo_Vista_Controlador>, Julio 2007.

[15] **ZIKMUND, W.G**, “Investigación de Mercados”, Prentice Hall, 1998.

[16] **SALVADOR FIGUERAS, M y GARGALLO VALERO, P**, "Análisis Factorial", <http://www.5campus.com/leccion/factorial>. 2006, 5-7 p.

[17] **MAHÍA, R.**, “Introducción al Análisis Cluster”, <http://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF>. 2004, 3-6 p.

[18] **BELLIDO VASQUEZ, P**, “Estadísticas para marketing (1) El Análisis Factorial”, <http://www.ilustrados.com/publicaciones/EpyukylkEFAVidhbVk.php>. 2003.

[19] **SALVADOR FIGUERAS, M,** “Análisis de conglomerados o cluster” <http://www.5campus.org/leccion/cluster>. 2001.

[20] **SPSS 13.0 MANUAL**, Factor Analysis – “A Tutorial-Introduction to Data Reduction through Factor Analysis”. 2006

[21] **ABDI, H**, “Factor Rotations in Factor Analyses”, in “Encyclopedia of Social Sciences Research Methods”. 2003.

[22] **PARK, T,** “About the Varimax Criterion for Orthogonal Rotation” <http://www.stat.ufl.edu/~tpark/>. 2003.

[23] **GONDAR, J.E.,** “Análisis Factorial”. < http://www.estadistico.com/ arts.html?20011119>. 2001.

[24] **PERE, J.,** “Aplicaciones del Análisis Factorial en el desarrollo, evaluación y validación de instrumentos psicométricos”. <http://www.benitoarias.com/personal/tutoriales/af.htm>. 2002.

[25] **EUCLÍDEA,** DICCIONARIO ESTADÍSTICO, <http://www.estadistico .com/dic.html?p=379>. 2004

[26] **BLOQUE DE CIUDAD,** DICCIONARIO ESTADÍSTICO, <http://www. estadistico.com/dic.html?p=85>. 2004

[27] **DISTANCIA DE MAHALANOBIS,** DICCIONARIO ESTADÍSTICO, <http://www.estadistico.com/dic.html?p=1220>. 2004