Proyecto de Grado

Presentado ante la Ilustre Universidad de Los Andes como requisito parcial para obtener el Título de Ingeniero de Sistemas

Análisis Exploratorio de Datos Multivariantes de Vigilantes Universitarios

Por

Br. Mónica Liseth Abreu Valecillos
Tutor: Prof. Luís Alfonso Dávila

Noviembre 2006

Análisis Exploratorio de Datos Multivariantes de Vigilantes Universitarios

Br. Mónica Liseth Abreu Valecillos

Proyecto de Grado — Investigación de Operaciones

Resumen

El presente trabajo es el producto de una aplicación de la Metodología CRISP-DM de Minería de Datos para el Análisis Exploratorio de Datos Multivariantes de Vigilantes Universitarios, utilizando las Técnicas de Análisis de Conglomerados de K-Medias y Análisis Discriminante Paso a Paso para la obtención del modelo descriptivo y predictivo respectivamente, que pueda ser utilizado por la Dirección de Vigilancia de la Universidad de Los Andes como herramienta de soporte en la toma de decisiones para la selección de personal. El modelo descriptivo permite el agrupamiento del personal vigilante según patrones de comportamiento en tres grupos, identificados como de Alto, Mediano y Bajo Desempeño. El modelo predictivo permite identificar las variables que mas contribuyen en la clasificación de los vigilantes, asignación definitiva por ratificación o movimiento de un individuo ubicado en la frontera de dos grupos y la obtención de una función discriminante para la ubicación de casos nuevos anónimos. Complementariamente, se identifica para cada grupo el patrón de comportamiento presentando en forma gráfica las características más comunes de las variables analizadas.

Palabras Clave: Minería de Datos, Evaluación del Desempeño de Personal, Análisis Multivariante de Datos.

A mi familia, por su apoyo y confianza.

Mamá, este logro es tuyo. Te Amo.

Índice

Índice de Tablas

Índice de Figuras

Agradecimientos

A Dios Padre Todopoderoso, por cuidarme, ser mi guía y fortaleza día a día. Enseñándome a valorar la importancia de la vida.

A mi Madre, por tu amor, por tus enseñanzas, por el apoyo incondicional, por tu confianza, por ser una mamá ejemplar. Dios te bendiga. Te Amo.

A mi Hermano, por tu apoyo, por tu hija mi hermosa sobrina con quien compartimos momentos maravillosos. Que Dios los Bendiga. Los Amo.

A Dulmar Salcedo y Nakari Díaz, gracias amigas por su apoyo incondicional. Que Dios las colme de salud y bendiciones. Éxito en el camino que aún falta por recorrer.

A mi Abuela y a mis Tías, por sus bendiciones, por su apoyo, por el estímulo y la confianza depositada en mí. Las quiero...

A Gustavo, porque siempre me has apoyado, por tu confianza, por tu amistad, por tu amor, por formar parte de mi vida. Que Dios te bendiga. Te Amo.

Al Profesor Luís Alfonso Dávila, que además de brindarme sus conocimientos, apoyo y confianza es un amigo y un ser humano maravilloso. Que dios lo bendiga.

Al Profesor Felipe Pachano, no sólo por la enseñanza impartida sino también por su amistad. Lo aprecio mucho.

Al Dr. Omar Uzcategui, por su receptividad y colaboración en el desarrollo de este proyecto.

Al Lic. Manuel Rodríguez, Lic. Adriana Machado, T.S.U José Luís Almeida, y al Ing. German Moreno, por su apoyo para la recolección de los datos utilizados en el presente trabajo.

A Todas Aquellas Personas, que de una u otra manera me han apoyado y con quienes he compartido estos momentos de mi vida.

Introducción

Una de las principales preocupaciones del hombre siempre ha sido la clasificación o categorización; el desarrollo de técnicas estadísticas para la diferenciación entre grupos de individuos ha despertado siempre un fuerte interés por parte de disciplinas tan disímiles como la economía, medicina, finanzas, biología, entre otras. Un campo en el que se ha querido utilizar este tipo de herramientas es en el campo de los recursos humanos para la administración de personal, con el propósito de categorizar a sus empleados en distintos grupos y de este modo entender mejor su comportamiento y contar con un soporte a la toma de decisiones para la captación y selección de personal.

En esta materia los datos e información ha ido en aumento en cantidad y variedad, así como también, ha mejorado substancialmente su almacenamiento en bases de datos y otras fuentes, de donde puede ser extraída para analizar los mismos para obtener información útil para la organización.

En muchas ocasiones, el método tradicional de convertir los datos en conocimiento consiste en el análisis e interpretación realizada en forma manual. Esta forma de actuar puede resultar lenta y altamente subjetiva. De hecho el análisis manual es impracticable en dominios donde el volumen de datos es muy grande, sin la adecuada técnica y herramienta para el procesamiento.

En la última década ha surgido un conjunto de herramientas y técnicas para analizar datos y que tiene su origen en la estadística, algo lógico teniendo en cuenta que la materia prima de esta disciplina son precisamente los datos. Este conjunto de herramientas y técnicas mas conocida como minería de datos que busca obtener información intencional (conocimiento) mas que información extensional (datos), donde este conocimiento no es, generalmente, una parametrización de ningún modelo preestablecido por el usuario, sino que es un modelo novedoso y original , extraído completamente por la herramienta.

En el presente trabajo se pretende, precisamente, realizar una aplicación de Minería de Datos utilizando una metodología ampliamente utilizada en diferentes organizaciones del mundo [HJ04], para el Análisis exploratorio de Datos Multivariantes de Datos de Vigilantes Universitarios, con la finalidad de obtener conocimiento útil para la toma de decisiones en materia de selección de personal.

El mismo esta compuesto por cinco capítulos:

En el capítulo 1 se presenta la identificación del proyecto, motivación, definición del problema y los objetivos.

En el capítulo 2 se presenta el marco teórico sobre Minería de Datos, tipos de modelos, relación con otras disciplinas, aplicaciones.

En el capítulo 3 se presenta la metodología utilizada para el desarrollo del proyecto.

En el capítulo 4 se presentan las técnicas de Minería de Datos utilizadas para la clasificación de los individuos.

En el capítulo 5 se presentan los resultados obtenidos en cada una de las fases de la metodología CRISP-DM.

Por último, se presentan las conclusiones y recomendaciones.

Capítulo 1

Identificación del Proyecto

En el presente capítulo, se identifica el proyecto de investigación aplicada, desarrollada, presentando brevemente los antecedentes, la motivación para su realización, las restricciones en el área de aplicación, así como la definición del problema, el objetivo general y específico.

1.1 Identificación de la Organización en Estudio

La Dirección de Vigilancia de la Universidad de Los Andes, fue creada inicialmente como Sección de Vigilancia, en el año 1977, permaneciendo adscrita a la dirección de personal hasta diciembre del 2005, fecha en que se elevó a Dirección de Seguridad y Protección Interna de la Universidad de Los Andes por resolución del Consejo Universitario llevando bajo su responsabilidad la seguridad de los miembros de la comunidad universitaria, los visitantes y sus bienes e inmuebles, así como también sus espacios territoriales de los estados Mérida, Táchira, Trujillo y Barinas.

Para marzo del presente año contaba con un personal Ordinario activo de 280 vigilantes y un personal eventual de aproximadamente 700 vigilantes distribuidos en cinco turnos en las diferentes dependencias universitarias. Cuenta también con:

- 1 Director.
- 1 Jefe de Operaciones.
- Equipo Canino.
- Personal de Supervisión.
- Personal de Vigilancia.
- Cuerpo de Bomberos Universitarios.
- Grupos de Rescate.

La Dirección de Seguridad y Protección Interna de la U.L.A. trabaja para el resguardo de la comunidad universitaria y el apoyo a los organismos de protección y prevención de desastres naturales. En este sentido, plantea su misión, visión y objetivo general de la siguiente manera: [SU06]

1.1.1 Misión

"Proteger la integridad de la comunidad universitaria, en cuanto a sus trabajadores, ambiente instalaciones y equipos, así como a todos los miembros de la comunidad universitaria, en materia de seguridad interna; brindar apoyo a las comunidades vecinas mediante la ejecución de programas y proyectos de seguridad que involucren la protección del recinto universitario y su entorno, tomando en cuenta las características de la Universidad de Los Andes, que permitan educar y concienciar, para que realicen sus actividades en un ambiente de trabajo óptimo, con espíritu de pertenencia institucional y con proyección hacia la comunidad".

1.1.2 Visión

"Alcanzar la seguridad interna en cada una de las áreas de la Universidad, estableciendo niveles de responsabilidad y participación efectiva; a través de planes y programas adaptados a las nuevas exigencias de la comunidad universitaria, para optimizar la calidad y productividad en el trabajo, con un mínimo riesgo a la salud, al ambiente y al patrimonio universitario. Así mismo, proyecta lograr que la universidad tenga una imagen de alto nivel tanto interno como externo, en su condición de propiciadora de un ambiente seguro para realizar las actividades de docencia, investigación y extensión que le son propias".

1.1.3 Objetivo General de la Dirección de Vigilancia

"Salvaguardar el patrimonio universitario, diseñando y ejecutando programas de prevención de seguridad interna, protección ambiental, prevención y combate de incendios

a fin de resguardar la salud física, mental, social de los miembros de la comunidad universitaria y comunidad circunvecinas e igualmente la custodia y protección de los bienes universitarios".

1.2 Motivación

Los conocimientos adquiridos durante los últimos semestres de la Carrera de Ingeniería de Sistemas, sobre tratamiento estadístico de datos, motivó la realización de un proyecto, utilizando Técnicas y una Metodología de Minería de Datos (Data Mining) que permitiera descubrir y/o predecir patrones de comportamiento, descubrir tendencias y regularidades y, en general, el tratamiento estadístico-matemático de la información disponible para la obtención de conocimiento de soporte a la toma de decisiones. Para ello, se contactó a la Dirección de Vigilancia de la Universidad de los Andes con el fin de obtener datos e información del personal vigilante con los cuales realizar el trabajo de Minería de Datos.

1.3 Definición del Problema

La Dirección de Vigilancia de la Universidad de los Andes, en su interés por resguardar la seguridad e integridad de todas las personas e instalaciones del recinto universitario, ha considerado de gran utilidad la propuesta de realización de un análisis exploratorio del comportamiento del personal de vigilancia (fijos y eventuales) en las actividades llevadas a cabo, en pro de mejorar la calidad del servicio.

La idea general es la realización de un estudio orientado a descubrir patrones y modelos de comportamiento del personal de vigilancia, cuyos resultados sirvan de soporte al establecimiento de criterios, estrategias y toma de decisiones inherentes a la Dirección de Vigilancia con el propósito de fortalecer las actividades de captación, selección y ubicación de personal.

1.4 Objetivo General del Proyecto

- Realizar un análisis exploratorio de datos multivariantes del personal de vigilancia (datos personales, antecedentes, estudios realizados, ubicación, procedencia, etc.), utilizando Técnicas y una Metodología de Minería de Datos (Data Mining), para ofrecer a la Dirección de Vigilancia de la Universidad de Los Andes una herramienta de soporte en la toma de decisiones en materia de administración de este tipo de recursos humanos.

1.5 Objetivos Específicos

- Recopilar datos de interés para el estudio.
- Depurar los datos multivariantes.
- Realizar un análisis preliminar utilizando técnicas tradicionales de tratamiento de datos.
- Aplicar Técnicas Específicas de Minería de Datos.
- Analizar los resultados de la aplicación.
- Presentar los resultados utilizando representaciones gráficas convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de los datos.

1.6 Delimitaciones

El desarrollo de este Proyecto de Grado, estará delimitado a utilizar datos sólo del personal ordinario. Dentro de las diversas técnicas de análisis multivariantes serán aplicadas el Análisis de Conglomerados de K-medias y el Análisis Discriminante como modelo descriptivos y predictivos respectivamente, con el objeto de analizar los datos para extraer conocimiento útil, comprensible y novedoso. En las variables *permisos* y *amonestaciones* se considera el número de veces en que estos se presentan, no así las causas que las producen.

Capítulo 2

Introducción a la Minería de Datos

En este capítulo se hace una breve introducción de la Minería de datos, presentando algunas definiciones, describiendo brevemente los modelos descriptivos y predictivos que permiten el análisis de los datos, así como también las disciplinas que contribuyen a la minería de datos y las diversas áreas en la que se puede aplicar.

2.1 Minería de Datos

La Minería de Datos o Data Mining es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Se fundamenta en varias disciplinas tradicionales como la estadística, inteligencia artificial, bases de datos y otras áreas de la informática. Es capaz de extraer patrones, descubrir tendencias y regularidades, predecir comportamientos y, en general, de sacar provecho a la información computarizada, generalmente heterogénea y en grandes cantidades. Permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones. [HJ04]

La definición del concepto de Data Mining (DM) puede variar entre unos investigadores y otros. Por ejemplo, los estadísticos, analistas de datos y la comunidad de sistemas de gestión de la información adoptan mayoritariamente este término para referirse al *proceso genérico correspondiente a las técnicas y herramientas de investigación usadas para extraer información útil de una base de datos*. Dentro de estas técnicas podemos considerar todos aquellos métodos matemáticos, técnicas y software para el análisis inteligente de los datos y búsqueda de patrones o tendencias en los mismos aplicados de forma iterativa e interactiva.

Dentro de las definiciones que se pueden encontrar en la literatura relacionada se muestran algunas de las más significativas:

- ``Data Mining es la exploración y análisis, mediante métodos automáticos o semiautomáticos, de grandes cantidades de datos para descubrir reglas o patrones significativos" [BG97]

- ``Data Mining es el proceso analítico diseñado para explorar grandes cantidades de datos (típicamente relacionados con el mercado o los negocios) con el fin de investigar patrones consistentes y/o relaciones sistemáticas entre variables y, a continuación, validar los resultados aplicando modelos detectados para nuevos subgrupos de datos" [LE01]

- ``Data Mining es el conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos" [PF91]

- ``Data Mining es el descubrimiento eficiente de información valiosa, no obvia, de una gran colección de datos" [BJ96]

- ``La Minería de Datos es un proceso analítico diseñado para explorar grandes volúmenes de datos con el objeto de descubrir patrones y modelos de comportamiento o relaciones entre diferentes variables. Esto permite generar conocimiento para dar soporte a la toma de decisiones en los procesos fundamentales de una organización" [IN03]

- Se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos

> almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización. [CB00]

Dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos...), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. [HJ04]

Una vez analizadas estas definiciones se puede inferir que:

- Se utilizan grandes cantidades de datos.
- Se emplean técnicas de diversas disciplinas como la estadística, inteligencia artificial, entre otras.
- Se obtienen patrones de comportamiento, modelos y tendencias.
- Tiene como finalidad generar conocimiento útil y novedoso para la toma de decisiones.

2.2 Tipos de Modelos

Este conocimiento generado por la Minería de datos puede ser en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Sin embargo, existen muchas formas diferentes de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos.

En la práctica, los modelos pueden ser de dos tipos: *predictivos y descriptivos*. *Los* ***modelos predictivos*** pretenden estimar valores futuros o desconocidos de variables de

interés, que denominamos variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que nos referimos como variables independientes o predictivas. Por ejemplo, un modelo predictivo seria aquel que permite estimar la demanda de un nuevo producto en función del gasto en publicidad.

*Los **modelos descriptivos***, en cambio, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, una agencia de viaje desea identificar grupos de personas con unos mismos gustos, con el objeto de organizar diferentes ofertas para cada grupo y poder así remitirles esta información; para ello analiza los viajes que han realizado sus clientes e infiere un modelo descriptivo que caracteriza estos grupos.

Entre las técnicas predictivas encontramos la clasificación y la regresión, mientras que el agrupamiento, las reglas de asociación y las correlaciones son técnicas descriptivas.

2.3 Relación con otras Disciplina

La minería de datos es un campo interdisciplinario que se ha desarrollado en paralelo a partir de otras tecnologías. Por ello, la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas. Véase la Figura 2.1.

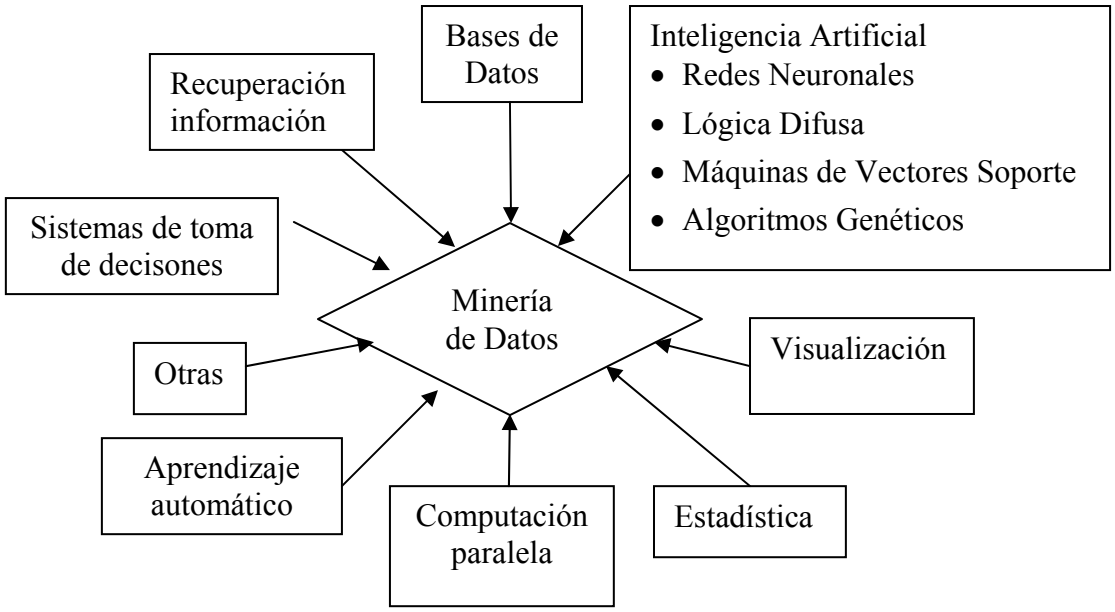

Figura 2.1 Disciplinas que contribuyen a la Minería de Datos. Fuente: [HJ04] [DL05]

Se pueden destacar como disciplinas más influyentes las siguientes: [HJ04] [DL05]

- **Las bases de datos:** los almacenes de datos y el procesamiento analítico en línea (OLAP) son conceptos que tienen una gran relación con la minería de datos, si bien en este último caso no se trata de obtener informes desarrollados a base de agregar los datos de cierta manera compleja, sino de extraer conocimiento novedoso y comprensible.

- **La recuperación de información:** consiste en obtener información desde datos textuales, por lo que su desarrollo histórico se ha basado en el uso efectivo de bibliotecas (recientemente digitales) y en la búsqueda por Internet. Una tarea usual es localizar documentos a partir de palabras claves, lo cual puede verse como un proceso de clasificación de los documentos en función de estas palabras clave. Usando para ello medidas de similitud entre los documentos y la consulta.

- **La estadística:** muchos de los conceptos, algoritmos y técnicas que se utilizan en minería de datos, han sido proporcionados por esta disciplina, como por ejemplo, la

media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, las técnicas bayesianas, entre otros. Incluso algunos paquetes de análisis estadístico se distribuyen y/o comercializan como herramientas de minería de datos.

- **El aprendizaje automático:** corresponde al área de la inteligencia artificial que se encarga de desarrollar algoritmos y programas capaces de aprender, y constituye, junto con la estadística, el centro del análisis inteligente de los datos. Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: la máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema.

- **Los sistemas para la toma de decisiones:** son herramientas y sistemas informatizados que ayuda a los individuos y a las organizaciones en la solución de problemas y en la toma de decisiones. El objetivo es suministrar la información necesaria para efectuar decisiones efectivas en el ámbito empresarial o en tareas de diagnóstico. Herramientas como los árboles de decisión provienen de esta área.
- **La visualización de datos:** el uso de técnicas de visualización permite al usuario descubrir, intuir o entender patrones que serían más difíciles de “ver" a partir de descripciones matemáticas o textuales de los resultados. Existen técnicas de visualización, como, por ejemplo, las gráficas (diagramas de barras, gráficas de dispersión, histogramas, etc.), las basadas en píxeles (cada dato se representa como un único píxel), las jerárquicas (dividiendo el área de representación en regiones dependiendo de los datos) y muchas otras.

- **Inteligencia Artificial:** rama de la ciencia de la computación que estudia la resolución de problemas no algorítmicos mediante el uso de cualquier técnica de computación disponible, sin tener en cuenta la forma de razonamiento subyacente a los métodos que se apliquen para lograr esa resolución.

- **Otras disciplinas:** dependiendo del tipo de datos utilizados o del tipo de aplicación, la minería de datos emplea de igual forma técnicas de otras disciplinas

como el lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, etc.

2.4 Aplicaciones

Las técnicas de minería de datos se han ido integrando en las actividades del día a día, convirtiéndose en algo habitual. Estas técnicas y métodos de minería de datos han sido tradicionalmente utilizados en áreas como la publicidad, ya que han permitido aumentar la receptividad de ofertas. Sin embargo, ésta no es la única área en la que se pueden aplicar. Se pueden encontrar ejemplos en todo tipo de aplicaciones: financieras, seguros, medicina, políticas, económicas, educación, procesos industriales, entre otros. A continuación se mencionan algunas de las áreas en las que puede ser usada la minería de datos. [HJ04]

2.4.1 Aplicaciones financieras y banca

En estas áreas la minería de datos permite descubrir patrones de fraude en la utilización de tarjetas de crédito, estimar el gasto en tarjeta de crédito por grupos, realizar una evaluación de riesgos en el otorgamiento de créditos, entre otras.

2.4.2 Análisis de mercado, distribución y, en general, comercio

El uso de técnicas de minería de datos en estas áreas permite la estimación de costos, inventarios y ventas, evaluación de campañas publicitarias para la captación de clientes, clasificación de clientes, etc.

2.4.3 Seguros y Salud Privada

La aplicación de métodos de minería de datos en este ámbito de paso a la identificación de patrones de comportamiento para clientes con posibilidad de riesgo, estimar el número de clientes que pretenden ampliar su póliza.

2.4.4 Educación

Las técnicas de minería de datos permiten la realización de estudios para la selección o captación de estudiantes en distintos niveles educativos, poder estimar y/o predecir cuanto será el tiempo de permanencia en la institución, además de la detección de abandonos y de fracaso.

2.4.5 Otras Áreas

En el campo de los recursos humanos la aplicación de técnicas de minería de datos permite la clasificación y selección de empleados. En política, las técnicas de minería de datos pueden ser usadas para realizar estudios que permitan conocer la preferencia hacia un candidato, así como el diseño de campañas políticas. En deportes, la organización y/o planificación de actividades. En agendas personales y correos electrónicos, análisis del empleo del tiempo, clasificación y distribución automática de correo.

Como se refleja en todas estas aplicaciones, la utilización de la minería de datos puede ayudar a entender mejor el ámbito donde se desenvuelve la organización a fin de mejorar la toma de decisiones.

Capítulo 3

Metodología de Aplicación de la Minería de Datos

En el capítulo anterior se ha visto que la minería de datos es un proceso que permite comprender mejor a los individuos y organizaciones, a través del conocimiento extraído de los datos, el contexto en el que deben actuar y tomar decisiones. Este proceso consta de una secuencia iterativa de etapas o fases. En este capítulo se presenta una metodología de aplicación como modelo de referencia y una guía para llevar a cabo un proceso de minería de datos, el cual puede ser adaptado a las necesidades de cada organización.

3.1 Fases de un Proyecto de Minería de Datos

Las fases o etapas para la realización de un proyecto de minería de datos, independientemente de la técnica específica de minería de datos usada, requieren de un conjunto de datos que generalmente están almacenados en una base de datos.

Para la realización de un proyecto de Minería de Datos se han propuesto diversas metodologías entre las que podemos citar: CRISP-DM [CC00], SEMMA [SAS01], CRITIKAL [IT99] [SA99], "5A's" [SPSS01], de las cuales fue seleccionada para la realización del presente proyecto la Metodología CRISP-DM, por ser una de las más aplicadas. Las razones fundamentales son debidas a su generalización y practicidad, además de su libre utilización, como se muestra en el resultado obtenido por [KD02] en la siguiente encuesta.

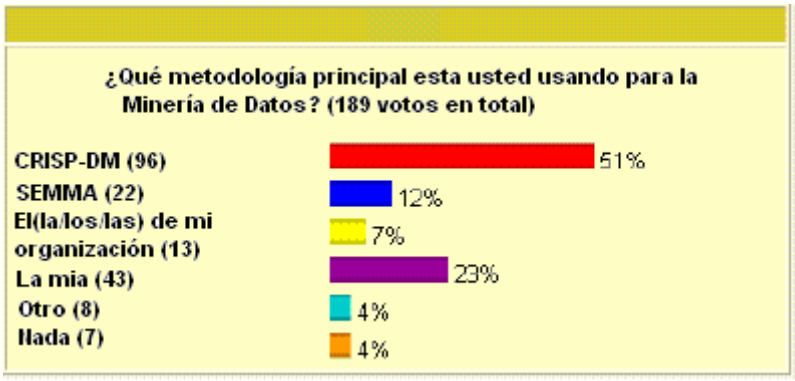

Figura 3.1. Encuesta realizada en [KD02] sobre la Metodología usada en Agosto del 2002

3.2 Metodología CRISP-DM

CRISP-DM (CRoss-Industry Standard Process for Data Mining) [CC00], es una metodología para el desarrollo de proyectos de Minería de Datos. El responsable de esta metodología es el consorcio CRISP-DM, formado por importantes empresas europeas y estadounidenses que poseen una amplia experiencia en proyectos de análisis de datos relacionados con muy diversos campos de la industria.

Esta metodología de minería de datos es desarrollada bajo un procedimiento jerárquico consistente de un conjunto de actividades o tareas explicadas en cuatro niveles: etapas, tareas generales, tareas específicas e instancias de proceso.

En el primer nivel (el nivel más alto) se encuentran un número de etapas, las cuales están organizadas en varias tareas generales presentadas en un segundo nivel. Este segundo nivel procura ser lo más general posible para poder cubrir, de esta manera todas las posibles situaciones, razón por la cual es denominado nivel genérico. Las tareas generales debe cubrir completamente el proceso de análisis y sus posibles aplicaciones.

El nivel de las tareas específicas (tercer nivel) permite describir como se deberían desarrollar las tareas generales en ciertas situaciones específicas.

Así por ejemplo, en el segundo nivel puede existir una tarea general llamada "preparación de los datos". El tercer nivel describe cómo se diferencia esta tarea de unas situaciones a otras, por ejemplo el reemplazo de valores faltantes en una variable numérica y el reemplazo de valores faltantes o ausentes en una variable categórica.

Un gran número de estas tareas pueden ser desarrolladas en la práctica sin tener un orden específico y en ciertas ocasiones será necesario volver a retomar tareas previas a fin de repetir ciertas acciones. Englobar todos los posibles caminos a lo largo del proyecto implica trabajar con un modelo considerablemente complejo, lo cual no es lo que pretende cubrir el modelo de procedimiento.

Por último se tiene, el nivel de instancias de proceso (cuarto nivel), que consiste en levar a cabo un conjunto de acciones y decisiones como resultado del desarrollo del proyecto de minería de datos. Una instancia de proceso describe de manera organizada las tareas definidas en los niveles más altos, representando así, más lo que sucede en un proceso particular, que lo que sucede en general.

El modelo de referencia de la metodología CRISP-DM describe las etapas, tareas y sus salidas para el desarrollo de un proyecto de minería de datos.

3.2.1 Contexto del Proyecto

En el CRISP-DM el paso entre el nivel general y el especializado es dirigido por el contexto del proyecto. Actualmente se distinguen cuatro contextos diferentes: [MB05]

- Conocimiento específico del área en la que se desarrollará el proyecto.
- Compresión del tipo de problema para describir de manera específica el (los) objetivo(s) que el proyecto llevará a cabo.
- Abarcar temas específicos que describan los diversos desafíos técnicos que puedan ocurrir durante el proceso.

- Especificar qué herramientas y/o qué técnicas serán aplicadas en el desarrollo del proyecto.

3.2.2 Proyección

El CRISP-DM distingue 2 tipos diferentes de proyecciones entre los niveles genérico y especializado: [MB05]

- La aplicación del modelo genérico del proceso para el desarrollo de un solo proyecto, es una *proyección para el presente* si se pretende proyectar las tareas generales y sus descripciones solo para este proyecto en particular, convirtiéndose esta en una proyección sencilla para posiblemente un solo uso.

- Cuando se especializa el modelo genérico del proceso de acuerdo a un contexto predefinido, este tiene una *proyección para el futuro*, puesto que podrá ser utilizado mas adelante en situaciones similares.

Por tanto, el tipo de proyección adecuada depende de las diversas necesidades y situaciones de cada organización.

3.2.3 Cómo Proyectar

La estrategia básica para proyectar el modelo genérico del proceso al nivel especializado es la misma para todos los tipos de proyecciones: [MB05]

- Estudiar contextos específicos.
- Descartar cualquier detalle que no sea aplicable en dicho contexto.
- Agregar detalles específicos al contexto.
- Adaptar contenidos genéricos de acuerdo a características específicas del contexto.
- Renombrar contenidos genéricos para posiblemente una mayor claridad.

3.3 Ciclo de vida de un Proyecto de Minería de Datos

El ciclo de vida de un Proyecto de Minería de Datos consta de seis fases, como se muestra en la figura 3.2.

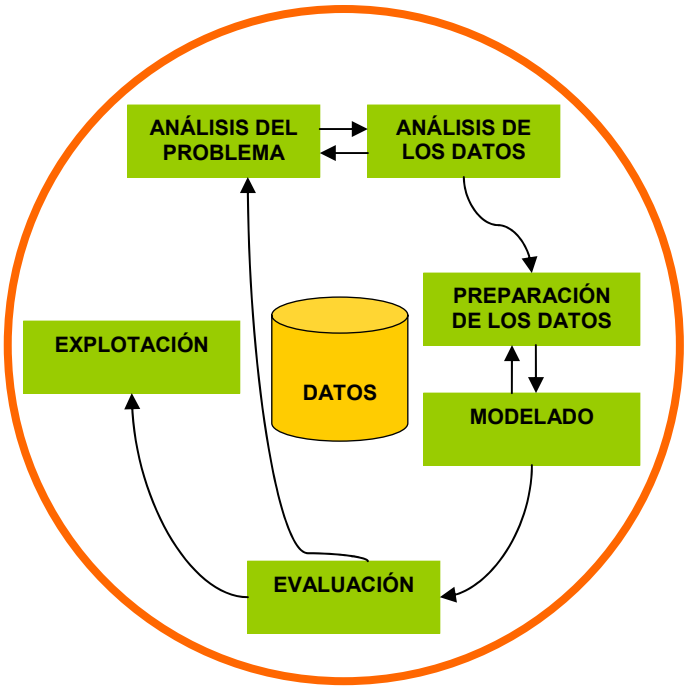

Figura 3.2. Fases del Modelo de referencia CRISP-DM.
Fuente: [MB05], Apuntes de la Asignatura "Minería de Datos", Pág. 33.

El orden de las mismas no es estricto, ya que frecuentemente a lo largo del desarrollo del proyecto, es necesario volver atrás en numerosas ocasiones, dependiendo de los resultados obtenidos en las fases previas. Las flechas indican las relaciones más habituales entre las fases. El círculo exterior simboliza la naturaleza cíclica del data mining, ya que la solución a la que finalmente se llega puede conducir al planteamiento de nuevas cuestiones que den origen a otros proyectos.

3.3.1 Análisis del Problema

Fase inicial que incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos y en una planificación.

3.3.2 Análisis de los Datos

Recolección inicial de datos para familiarizarse con ellos, identificar su calidad y descubrir las relaciones entre los más evidentes para las primeras hipótesis de relaciones ocultas entre ellos.

3.3.3 Preparación de los Datos

Esta etapa incluye la selección de tablas, registros y atributos de la base de datos que serán utilizados en el proyecto, así como su transformación y preparación para las herramientas de modelado.

3.3.4 Modelado

Normalmente existen varias técnicas para el mismo problema y cada una exige una entrada de datos particular, por ello es necesario interactuar con la fase anterior para adecuar la base de datos de trabajo a la técnica de modelado.

3.3.5 Evaluación

Una vez creado el modelo se debe evaluar el rendimiento del mismo teniendo en cuenta que se han introducido todos los criterios de negocio. Se debe dar el visto bueno final a la aplicación del modelo de Minería de Datos.

3.3.6 Explotación

Se trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización, difundir informes sobre el conocimiento extraído, etc.

Capítulo 4

Análisis y Selección de Técnicas de Minería de Datos

En el presente capítulo, se presentan las diversas técnicas que pueden ser aplicadas en la minería de datos. Se describen las seleccionadas para llevar el desarrollo de este proyecto.

4.1 Análisis de Técnicas de Minería de Datos

La clasificación de las técnicas y algoritmos de minería de datos puede ser efectuada de múltiples formas. En la práctica, quizá, una de las clasificaciones más interesantes de los algoritmos de minería de datos es la que corresponde con su función. Como es lógico, ésta dependerá de la definición que adoptemos con respecto al proceso de minería de datos.

Se pueden clasificar según la función que desempeñan fundamentalmente en:

- **Clasificadores**: Que clasifican datos en clases predefinidas.
- **Algoritmos de regresión:** A partir de los datos generan una función predictora.
- **Descubrimiento de Reglas de Asociación:** Búsqueda de relaciones entre variables.
- **Modelado de Dependencias:** Generación de modelos que "expliquen" las dependencias entre atributos.
- **Clusterizado o agrupamiento:** Búsqueda de conjuntos en los que agrupar los datos cuando las clases son desconocidas.
- **Aprendizajes basados en casos:** Se basa en indexar y recordar los casos más significativos, de forma que los nuevos casos sean clasificados según el descriptor más próximo.
- **Compactación:** Búsqueda de descripciones más compactas de los datos. Técnicas de reducción de dimensión.
- **Detección de desviaciones**: Búsqueda de desviaciones importantes de los datos respecto a valores anteriores.
- **Sumarización:** Describe las propiedades que comparten aquellas observaciones que pertenecen a una misma clase.

- **Técnicas de Minería Datos Aplicados a Datos Secuenciales.** Orientadas a la búsqueda de relaciones en datos que transcurren secuencialmente.

Se consideran también los algoritmos que pueden ayudar a las tareas previas de preprocesado y preparación de los datos, se puede añadir:

- Técnicas de visualización multivariante.
- Algoritmos de detección y eliminación de espurios.
- Algoritmos de detección de datos ausentes y rellenados de los mismos.
- Otros algoritmos para el tratamiento y preprocesado de la información.

Además de las técnicas ya mencionadas existen también:

- **Las Redes Neuronales,** que pueden servir no sólo para desarrollar modelos clasificadores o predictores, sino también como algoritmos de segmentación, proyectores, filtros, etc.

- **Los Algoritmos Genéticos,** que pueden ser usados para el agrupamiento, la clasificación y las reglas de asociación, así como para la selección de atributos. En cualquiera de estos casos, se comienza con un modelo o solución inicial y, a través de múltiples iteraciones, los modelos se combinan para crear nuevos modelos. Para ello, se usa una función de adaptación, que selecciona los mejores modelos que sobreviran o serán cruzados.

- **Máquinas de Vectores Soporte,** que pertenecen a la familia de clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidos por funciones núcleo o Kernel) con un sesgo inductivo muy particular (maximización del margen).

- **Análisis Multivariante.** El análisis multivariante es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariantes en el sentido de que hay varias variables medidas para cada individuo ú objeto estudiado.

  Su razón de ser radica en un mejor entendimiento del fenómeno objeto de estudio obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir.

Tipos de Técnicas Multivariantes

Se pueden clasificar en tres grandes grupos:

1. Métodos de Dependencia

Suponen que las variables analizadas están divididas en dos grupos: las variables dependientes y las variables independientes. El objetivo de los métodos de dependencia consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma.

Se pueden clasificar en dos grandes subgrupos según que la variable(s) dependiente(s) sea(n) cuantitativas o cualitativas.

Si la variable dependiente es cuantitativa algunas de las técnicas que se pueden aplicar son las siguientes:

- **Análisis de Regresión**
  Es la técnica adecuada si en el análisis hay una o varias variables dependientes métricas cuyo valor depende de una o varias variables independientes métricas.

- **Análisis de Supervivencia**

  Es similar al análisis de regresión pero con la diferencia de que la variable independiente es el tiempo de supervivencia de un individuo ú objeto.

- **Análisis de Varianza**

  Se utilizan en situaciones en las que la muestra total está dividida en varios grupos basados en una o varias variables independientes no métricas y las variables dependientes analizadas son métricas. Su objetivo es averiguar si hay diferencias significativas entre dichos grupos en cuanto a las variables dependientes se refiere.

- **Correlación Canónica**

  Su objetivo es relacionar simultáneamente varias variables métricas dependientes e independientes calculando combinaciones lineales de cada conjunto de variables que maximicen la correlación existente entre los dos conjuntos de variables.

Si la variable dependiente es cualitativa algunas de las técnicas que se pueden aplicar son las siguientes:

- **Análisis Discriminante**

  El objetivo del análisis discriminante es proporcionar reglas de asignación y clasificación óptimas de individuos a una de las clases de una clasificación preestablecida.

- **Modelos de regresión logística**

  Son modelos de regresión en los que la variable dependiente es no métrica. Se utilizan como una alternativa al análisis discriminante cuando no hay normalidad.

2. Métodos de Interdependencia

Estos métodos no distinguen entre variables dependientes e independientes y su objetivo consiste en identificar qué variables están relacionadas, cómo lo están y por qué.

Se pueden clasificar en dos grandes grupos según que el tipo de datos que analicen sean métricos o no métricos.

Si los datos son métricos se pueden utilizar, entre otras, las siguientes técnicas:

- **Análisis Factorial y Análisis de Componentes Principales**
  Se utiliza para analizar interrelaciones entre un número elevado de variables métricas explicando dichas interrelaciones en términos de un número menor de variables denominadas factores (si son inobservables) o componentes principales (si son observables).

- **Escalas Multidimensionales**
  Su objetivo es transformar juicios de semejanza o preferencia en distancias representadas en un espacio multidimensional. Como consecuencia se construye un mapa en el que se dibujan las posiciones de los objetos comparados de forma que aquéllos percibidos como similares están cercanos unos de otros y alejados de objetos percibidos como distintos.

- **Análisis Cluster**
  Su objetivo es clasificar una muestra de entidades (individuos o variables) en un número pequeño de grupos de forma que las observaciones pertenecientes a un grupo sean muy similares entre sí y muy disimilares del resto. A diferencia del Análisis Discriminante se desconoce el número y la composición de dichos grupos.

Si los datos son no métricos se pueden utilizar, además de las Escalas Multidimensionales y el Análisis Cluster, las siguientes técnicas:

- **Análisis de Correspondencias**
  Se aplica a tablas de contingencia multidimensionales y persigue un objetivo similar al de las escalas multidimensionales pero representando simultáneamente las filas y columnas de las tablas de contingencia.

- **Modelos log-lineal**
  Se aplican a tablas de contingencia multidimensional y modelan relaciones de dependencia multidimensional de las variables observadas que buscan explicar las frecuencias observadas.

3. Métodos Estructurales

Suponen que las variables están divididas en dos grupos: el de las variables dependientes y el de las independientes. El objetivo de estos métodos es analizar, no sólo como las variables independientes afectan a las variables dependientes, sino también cómo están relacionadas las variables de los dos grupos entre sí.
Analizan las relaciones existentes entre un grupo de variables representadas por sistemas de ecuaciones simultáneas en las que se suponen que algunas de ellas (denominadas constructos) se miden con error a partir de otras variables observables denominadas indicadores.

Los modelos utilizados constan, por lo tanto, de dos partes: un modelo estructural que especifica las relaciones de dependencia existente entre las constructos latentes y un modelo de medida que especifica como los indicadores se relacionan con sus correspondientes constructos.

4.2 Selección de Técnicas de Minería de Datos

Las características de los datos de Vigilantes utilizados en este proyecto permitirían que como técnicas de Minería de Datos se utilicen Redes Neuronales, Máquinas de Vectores Soporte, todas ellas permitirían obtener modelos descriptivos y predictivos. Sin embargo, para el desarrollo de este proyecto se empleará el Análisis de Conglomerados y el Análisis Discriminante, por ser estas herramientas más amigables y funcionales para llevar a cabo nuevos experimentos, sin ameritar ser empleadas por personal especializado y sin mayores conocimientos en la aplicación de las mismas, puesto que sólo sería necesario tener conocimientos básicos en el manejo del software para el análisis estadístico SPSS. SPSS para Windows proporciona un poderoso sistema de análisis estadístico y de gestión de datos en un entorno gráfico, utilizando menús descriptivos y cuadros de diálogo sencillos que realizan la mayor parte del trabajo. La mayoría de las tareas se pueden llevar a cabo simplemente situando el puntero del ratón en el lugar deseado y pulsando en el botón.

En cambio las redes neuronales, la lógica difusa y las máquinas vectores soporte, pueden resultar ser muchos más complejos y requerir por tanto un conocimiento más detallado y profundo. Incluso, una red neuronal, puede requerir de varios días para su entrenamiento.

Entre las herramientas de Minería de Datos más usadas en la aplicación de estas técnicas se encuentran:

- **El NeuroSolutions 5.0:** es una herramienta gráfica para el desarrollo de redes neuronales, que combina una interfaz de diseño modular y basada en iconos con la implementación de procedimientos de aprendizaje avanzados y optimización genética. El resultado es un entorno prácticamente ilimitado para el diseño de redes neuronales para investigación y para la resolución de problemas reales. [AE05]

- **Matlab 7.0:** Es un poderoso entorno de cálculo técnico integrado que combina el cálculo numérico, gráficos avanzados y visualización, y un lenguaje de programación de alto nivel. [JG05]

- **Weka 3.4-2:** extensa colección de algoritmos de Máquinas de conocimiento desarrollados por la Universidad de Waikato (Nueva Zelanda) implementados en Java; útiles para ser aplicados sobre datos mediante las interfaces que ofrece. Además Weka contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización. Weka está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla. Esta aplicación es de libre distribución (licencia GPL) y se destaca por la cantidad de algoritmos que presenta así como por la eficiencia de los mismos. [DG00]

4.3 Análisis de Conglomerados

El análisis de conglomerados (cluster analisys, en terminología inglesa) es una técnica multivariante que permite agrupar los casos o variables de un archivo de datos en función del parecido o similaridad existente entre ellos. De este modo el análisis cluster o de conglomerados divide las observaciones en grupos basándose en la proximidad o lejanía de unas con otras. Las observaciones muy cercanas deben caer dentro del mismo cluster y las muy lejanas deben caer en clusters diferentes, de manera que las observaciones dentro de un cluster sean homogéneas y lo más diferentes posibles de las contenidas en otros clusters. [PL04]

Existen tres métodos de Análisis de Conglomerados:

- **Análisis de Conglomerados de K-Medias:** Este procedimiento intenta identificar grupos de casos relativamente homogéneos basándose en las características seleccionadas. El uso del procedimiento Análisis de conglomerados de K-medias trabaja con datos continuos y ofrece una serie de funciones únicas que se detallan a continuación:

- ✓ Posibilidad de guardar las distancias desde los centros de los conglomerados hasta los distintos objetos.
- ✓ Posibilidad de leer los centros de los conglomerados iniciales y guardar los centros de los conglomerados finales desde un archivo SPSS externo.

- **Análisis de conglomerados jerárquico:** Este procedimiento intenta identificar grupos relativamente homogéneos de casos (o de variables) basándose en las características seleccionadas, mediante un algoritmo que comienza con cada caso (o cada variable) en un conglomerado diferente y combina los conglomerados hasta que sólo queda uno. Debido a que el análisis de conglomerados jerárquico es un método exploratorio, los resultados deben considerarse provisionales hasta que sean confirmados mediante otra muestra independiente.

De estos métodos se seleccionó el Análisis de Conglomerados de K-Medias por qué permite elegir uno de los dos métodos disponibles para clasificar los casos: la actualización de los centros de los conglomerados de forma iterativa o sólo la clasificación. También se puede solicitar los estadísticos F de los análisis de varianza. El tamaño relativo de estos estadísticos proporciona información acerca de la contribución de cada variable a la separación de los grupos.

4.3.1 Análisis de Conglomerados de K-Medias

El análisis de conglomerados de K medias es un método de agrupación de casos que se basa en las distancias existentes entre ellos en un conjunto de variables. El procedimiento comienza seleccionando los k casos más distantes entre sí, requiriendo que el usuario especifique previamente el número de conglomerados que desea obtener. Y a continuación se inicia la lectura secuencial del archivo de datos asignando cada caso al centro más próximo y actualizando el valor de los centros a medida que se van incorporando nuevos casos. Una vez que todos los casos han sido asignados a uno de los K conglomerados, se

inicia un proceso iterativo para calcular los centroides finales de esos K conglomerados. [UCA02]
Existe la posibilidad de utilizar la técnica de manera exploratoria, clasificando los casos e iterando para encontrar la ubicación de los centroides, o sólo como técnica de clasificación, clasificando los casos a partir de centroides conocidos suministrados por el usuario.

En este trabajo se empleará la clasificación cuando los centros iniciales de los diversos clusters son desconocidos. El método consiste en que el sistema procede a actualizar los centros de cada cluster inicial a través de un proceso iterativo hasta llegar a unos finales que son los que se utilizaran para clasificar los individuos. La estrategia consiste en seleccionar de entrada tantos individuos distintos como clusters se hayan demandado, de modo que estos individuos/clusters iniciales tengan una distancia máxima entre ellos y puedan por tanto servir como estimadores iniciales. El algoritmo procede entrando nuevos casos y comparando la distancia de estos casos respecto a los centros iniciales que son temporales. Si la distancia mínima de estos individuos respecto al centro de un cluster inicial es mayor que la distancia entre los centros de los dos clusters más próximos, el centro más próximo al individuo es reemplazado por éste. De un modo iterativo el sistema calcula los centros de los clusters finales.

4.4 Análisis Discriminante

El análisis discriminante ayuda a identificar las características que diferencian (discriminan) a dos o más grupos y a crear una función capaz de distinguir con la mayor precisión posible a los miembros de uno u otro grupo. [UCA02]

Obviamente, para llegar a conocer en qué se diferencian los grupos se necesita disponer de la información (cuantificada en una serie de variables) en la que se supone que se diferencian. El análisis discriminante es una técnica estadística qué permiten identificar que variables diferencian a los grupos y cuántas de estas variables son necesarias para alcanzar la mejor clasificación posible. La pertenencia a los grupos, conocida de antemano, se utiliza como variable dependiente (una variable categórica con tantos valores discretos

como grupos). Las variables en las que se supone que se diferencian los grupos se utilizan como variables independientes o variables de clasificación (también llamadas variables discriminantes). Éstas, deben ser variables cuantitativas continuas o, al menos, admitir un tratamiento numérico con significado.

Por último el objetivo del análisis discriminante es encontrar la combinación lineal de las variables independientes que mejor permite diferenciar (discriminar) a los grupos. Una vez encontrada esa combinación lineal (la función discriminante) podrá ser utilizada para clasificar nuevos casos. Se trata de una técnica de análisis multivariante que es capaz de aprovechar las relaciones existentes entre una gran cantidad de variables independientes para maximizar la capacidad de discriminación.

Existen dos métodos distintos para que las variables independientes puedan incorporarse a la función discriminantes, estos son:

- **Inclusión Forzosa (Introducir independientes juntas).** Este método de *Inclusión Forzosa* de variables permite construir la función discriminante incorporando todas las variables independientes incluidas en el análisis. Los únicos estadísticos que se obtienen con esta estrategia se refieren al ajuste global de la función discriminante; no se obtienen estadísticos referidos a la significación individual de cada coeficiente discriminante.

- **Método de inclusión por pasos.** Permite obtener información sobre la significación y contribución individual de cada variable en la Función Discriminante. En este método de *inclusión por pasos,* las variables independientes van siendo incorporadas paso a paso a la Función Discriminante y, de esta manera, es posible, por un lado, construir una función utilizando únicamente aquellas variables que realmente son útiles para la clasificación y, por otro, evaluar la contribución individual de cada variable al modelo discriminante. Por lo que, ha sido éste el método seleccionado para llevar a cabo el Análisis Discriminante.

El estadístico utilizado como método para la selección de las variables a ser incluidas en el modelo es la lambda de Wilks, la cual permite comprobar la hipótesis nula de que los centroides (medias multivariantes de los grupos) son iguales. La variabilidad entre los grupos irá aumentando a medida que estos se vayan separando más y más y la variabilidad en los grupos se irá haciendo cada vez menor respecto a la variabilidad total, disminuyendo así el valor del cociente. De modo que, valores cercanos a cero indicarán una gran diferencia entre los grupos, mientras que valores cercanos a 1 indicarán un gran perecido entre ellos.

$$\lambda = \left(\frac{SumadeCuadradosIntragrupos\ F_1}{SumadeCuadradosTotal\ F_1} \right) \left(\frac{SumadeCuadradosIntragrupos\ F_2}{SumadeCuadradosTotal\ F_2} \right)$$

Cada variable independiente candidata a ser incluida en el modelo se evalúa mediante el estadístico F que mide el cambio que se produce en el valor de la lambda de Wilks al incorporar cada una de las variables al modelo. Obtenido el valor del estadístico F para cada variable, se incorpora al modelo la variable a la que le corresponde el mayor valor F (o, lo que es lo mismo, la que produce el mayor cambio en la lambda de Wilks):

$$F = \left(\frac{n-g-p}{g-1} \right) \left(\frac{1 - \lambda_{p+1} / \lambda_p}{\lambda_{p+1} / \lambda_p} \right)$$

Donde n es el número de casos, g es el número de grupos, p es el número de variables independientes en el modelo antes del paso actual, λ_p es la lambda de Wilks que corresponde al modelo antes de incluir la variable que se está evaluando y λ_{p+1} es la lambda de Wilks que corresponde al modelo después de incluir esa variable.

Independientemente del método utilizado, en el procedimiento de *inclusión por pasos* siempre se comienza seleccionando la mejor variable independiente desde el punto de vista de la clasificación (es decir, la variable independiente en la que más se diferencian los

grupos). Sin embargo, esta variable sólo es seleccionada si cumple el criterio de entrada. Se selecciona la variable independiente que, cumpliendo el criterio de entrada, más contribuye a conseguir que la Función Discriminante diferencie a los grupos. Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son evaluadas nuevamente para determinar si cumplen o no el criterio de salida. Si alguna variable de las ya seleccionadas cumple el criterio de salida, es expulsada del modelo.

El software para el análisis estadístico SPSS establece los criterios de *entrada y salida* utilizados para incorporar o eliminar variables. De acuerdo con estos criterios, sólo son incluidas en el modelo aquellas variables que contribuyen a discriminar significativamente a los grupos. Una variable pasa a formar parte de la Función Discriminante si el valor del estadístico F es mayor que 3.84 (valor de *entrada*). Y es expulsada de la función si el valor del estadístico F es menor que 2.71 (valor de *salida*).

Capítulo 5

Desarrollo de la Aplicación

En este capítulo se desarrollan cada una de las fases y las salidas generadas al aplicar la metodología (Modelo y guía de referencia CRISP-DM) para llevar a cabo el proyecto de minería de datos, para el Análisis Exploratorio de Datos Multivariantes de Vigilantes Universitarios.

5.1 Análisis del Problema

En pro de mejorar la calidad del servicio de vigilancia en las diferentes dependencias universitarias, en el presente trabajo se pretende realizar un análisis exploratorio del personal vigilante ordinario adscritos a la Dirección de Vigilancia de la Universidad de los Andes, con miras a ofrecer una herramienta, basada en técnicas de Minería de Datos, que pueda ser utilizada por esta dirección como soporte a la toma de decisiones en la selección de personal.

5.2 Análisis de los Datos

Los datos para la realización de este proyecto fueron suministrados por la Dirección de Vigilancia, complementados por la Dirección de Personal de la Universidad de los Andes, los cuales corresponden a datos pertenecientes al personal de vigilancia ordinario, distribuido éste en cinco turnos y en las diferentes dependencias Universitarias. Se cuenta con 329 registros, compuestos por un conjunto de atributos que permiten identificar a cada uno de los individuos, de los cuales se consideraron las siguientes variables: Fecha de Ingreso, Fecha de Nacimiento, Profesión, Grado de Instrucción, Lugar de Nacimiento, Número de Hijos, Estado Civil, Permisos, Amonestaciones, Estatura.

La revisión y validación de los datos permitió detectar inconsistencias y datos faltantes. En la Tabla 5.1 se presentan los resultados obtenidos.

Tabla 5.1. Inconsistencias y Datos faltantes

Variables	Registros vacíos (% de casos)	Inconsistencias (% de casos)
Fecha de Ingreso	0	0
Fecha de Nacimiento	4.3	0.3
Profesión	17.6	0
Grado de Instrucción	1.8	0
Lugar de Nacimiento	3.3	0
Número de Hijos	0	0
Estado Civil	0.9	0
Permisos	0	0
Amonestaciones	0	0
Estatura	57.1	0

De estas variables, a su vez, se puede visualizar una distribución más detallada de la Profesión y Estado Civil, en las figuras 5.1 y 5.2.

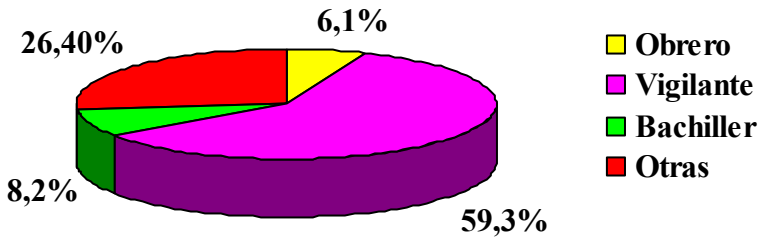

Figura 5.1. Profesión

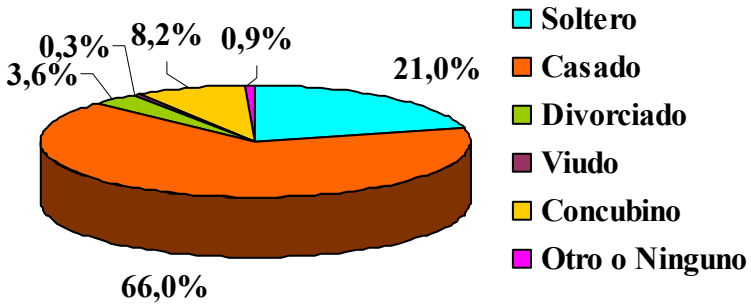

Figura 5.2. Estado Civil

5.3 Preparación de los Datos

En esta etapa se realizó el tratamiento de datos faltantes, la transformación de las variables de estudio, la creación de nuevas variables y se evalúa la independencia entre cada par de variables, para la aplicación eficiente de las técnicas de minería de datos seleccionadas.

5.3.1 Tratamiento de Datos Faltantes

Los valores faltantes, perdidos o ausentes fueron reemplazados de la siguiente manera:

En las variables cuantitativas como la edad y la estatura, el reemplazo se efectuó por la media y en el caso de variables categóricas se sustituyó por el valor moda.

5.3.2 Transformación de Atributos

La Transformación de datos es un proceso que modifica la forma de los mismos. Debido a que las técnicas de Minería de Datos seleccionadas, es decir, análisis de conglomerados de k-medias y análisis discriminante, sólo permite trabajar con valores numéricos o cuantitativos, se realizó una transformación automática para recodificar las variables cualitativas en categorías numéricas. Por otro lado, las variables Fecha de Ingreso y Fecha

de Nacimiento fueron transformadas para obtener Años de Servicio y Edad respectivamente, tomando como fecha actual el 02/08/2006. Una muestra de esta transformación se presenta en la figura 5.3.

VistaMinable - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window H

6 : ZNRO_HIJOS -0,013313653524201

	AÑOS_SERVICIO	EDAD	PROFESION
1	30,4684931507	57,0301369863	372
2	29,0849315068	49,0000000000	372
3	28,4136986301	54,3863013699	372
4	28,4136986301	61,0986301370	372
5	27,3315068493	49,1479452055	372
6	26,9205479452	47,1315068493	372
7	26,7561643836	54,5945205479	372
8	26,5780821918	64,3972602740	372
9	26,5780821918	52,9013698630	372
10	26,4438356164	49,8657534247	372
11	26,3315068493	57,9643835616	372
12	25,9561643836	53,6931506849	372
13	25,7808219178	51,2904109589	412
14	25,3315068493	50,3452054795	372
15	25,2493150685	48,0027397260	372
16	25,1671232877	47,8904109589	372

Figura 5.3. Transformación de Atributos

5.3.3 Creación de Nuevas Variables

Para las técnicas de análisis multivariantes estudiadas es necesario normalizar todos los atributos al mismo rango, debido a que estos métodos emplean algoritmos basados en distancias, por lo que es preciso que todas las variables estén en una misma escala. En este sentido todas las variables fueron transformadas en variables normales estandarizadas con media cero y varianza 1, creando así las nuevas variables llamadas en algunos casos variables tipificadas. Esta transformación fue realizada con el paquete estadístico SPSS

12.0, obteniéndose los resultados que se muestran en la figura 5.4, donde las nuevas variables o variables estandarizadas se identifican anteponiéndoles la letra z.

VistaMinable - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : AÑOS_SERVICIO 30,4684931507

	ZAÑOS_SERVICIO	ZEDAD	ZPROFESION	ZGRADO_INSTR
1	1,27307	,72415	-,27975	-,48298
2	1,11291	-,12475	-,27975	-,56180
3	1,03521	,44466	-,27975	-,56180
4	1,03521	1,15424	-,27975	-,27280
5	,90993	-,10911	-,27975	-,48298
6	,86236	-,32227	-,27975	-,56180
7	,84333	,46667	-,27975	-,56180
8	,82271	1,50295	-,27975	-,56180
9	,82271	,28768	-,27975	-,56180
10	,80717	-,03322	-,27975	-,53553
11	,79417	,82291	-,27975	1,93417
12	,75072	,37138	-,27975	-,56180
13	,73042	,11738	,07534	-,27280
14	,67841	,01746	-,27975	1,93417
15	,66889	-,23017	-,27975	-,56180
16	,65938	-,24204	-,27975	-,48298

Figura 5.4. Creación de Nuevas Variables

5.3.4 Correlación Bivariada de Pearson

A fin de conocer si existe o no una relación entre variables, se emplea el procedimiento de Correlación Bivariada de Pearson. La Tabla 5.2 muestra estos resultados. El primer valor de cada celda es el coeficiente de correlación de Pearson entre cada par de variables, el mismo puede oscilar entre -1 y 1, indicando un valor 0 una relación nula o independencia entre las variables; 1, una relación perfecta y positiva, y -1, una relación perfecta y negativa. En esta tabla puede apreciarse con excepción de las variables edad y años de servicio que no se observa dependencia significativa entre pares de las restantes variables. Esta correlación se realiza, puesto que uno de los supuestos del análisis discriminante es la independencia de las variables, las cuales son las que contribuyen a diferenciar más a los grupos a medida que entran al modelo. Sin embargo, debido a que es evidente que

mientras más años de servicio se tiene más edad y el coeficiente de correlación de Pearson entre estas dos variables no es tan alto, las mismas son incluidas en el análisis.

Tabla 5.2. Matriz de Correlaciones de Pearson

		AñosServicio	MEANEDAD,ALL)	PROFESION	GRADO_INSTR	LUGAR DE NACIMIENTO	NRO_HIJOS	EDO_CIVIL	PERMISOS	AMONESTACIONES	MEANESTATURA,ALL)
AñosServicio	Correlación de Pearson	1	,680**	-,422**	-,023	,049	,464**	,231**	-,040	,228**	-,107
	Sig. (bilateral)	.	,000	,000	,680	,379	,000	,000	,471	,000	,054
	N	324	324	324	324	324	324	324	324	324	324
MEANEDAD,ALL)	Correlación de Pearson	,680**	1	-,318**	-,002	-,021	,378**	,174**	-,077	,123*	-,092
	Sig. (bilateral)	,000	.	,000	,969	,710	,000	,002	,167	,027	,098
	N	324	324	324	324	324	324	324	324	324	324
PROFESION	Correlación de Pearson	-,422**	-,318**	1	-,147**	-,026	-,203**	-,068	-,073	-,133*	-,009
	Sig. (bilateral)	,000	,000	.	,008	,637	,000	,225	,188	,016	,866
	N	324	324	324	324	324	324	324	324	324	324
GRADO_INSTR	Correlación de Pearson	-,023	-,002	-,147**	1	,095	,033	-,013	,025	-,032	,110*
	Sig. (bilateral)	,680	,969	,008	.	,087	,556	,812	,650	,565	,047
	N	324	324	324	324	324	324	324	324	324	324
LUGAR DE NACIMIENTO	Correlación de Pearson	,049	-,021	-,026	,095	1	-,011	-,021	,014	,049	,000
	Sig. (bilateral)	,379	,710	,637	,087	.	,841	,712	,806	,379	,995
	N	324	324	324	324	324	324	324	324	324	324
NRO_HIJOS	Correlación de Pearson	,464**	,378**	-,203**	,033	-,011	1	,238**	-,036	,085	-,072
	Sig. (bilateral)	,000	,000	,000	,556	,841	.	,000	,518	,127	,196
	N	324	324	324	324	324	324	324	324	324	324
EDO_CIVIL	Correlación de Pearson	,231**	,174**	-,068	-,013	-,021	,238**	1	,103	,113*	-,088
	Sig. (bilateral)	,000	,002	,225	,812	,712	,000	.	,063	,042	,113
	N	324	324	324	324	324	324	324	324	324	324
PERMISOS	Correlación de Pearson	-,040	-,077	-,073	,025	,014	-,036	,103	1	,017	-,039
	Sig. (bilateral)	,471	,167	,188	,650	,806	,518	,063	.	,755	,482
	N	324	324	324	324	324	324	324	324	324	324
AMONESTACIONES	Correlación de Pearson	,228**	,123*	-,133*	-,032	,049	,085	,113*	,017	1	-,107
	Sig. (bilateral)	,000	,027	,016	,565	,379	,127	,042	,755	.	,054
	N	324	324	324	324	324	324	324	324	324	324
MEANESTATURA,ALL)	Correlación de Pearson	-,107	-,092	-,009	,110*	,000	-,072	-,088	-,039	-,107	1
	Sig. (bilateral)	,054	,098	,886	,047	,995	,196	,113	,482	,054	.
	N	324	324	324	324	324	324	324	324	324	324

5.4 Modelado

Una vez analizados y preparados los datos, se procedió a la aplicación de las técnicas multivariantes, Análisis de Conglomerados de K-medias y el Análisis Discriminante, obteniendo los siguientes resultados:

5.4.1 Análisis de Conglomerados de K-Medias

El objetivo principal de la utilización de este método es la agrupación de los vigilantes en tres grupos para obtener una clasificación preliminar según la proximidad o lejanía de los casos en estudio, de manera que las observaciones dentro de un conglomerado sean homogéneas y lo más diferentes posibles de los contenidos en otros conglomerados.

Se calculan los centros iniciales de los conglomerados, y se itera hasta que el desplazamiento de estos centros sea mínimo. Una vez seleccionados los centros de los conglomerados, cada caso es asignado al conglomerado de cuyo centro se encuentra más próximo.

La Tabla 5.3 contiene los centros iniciales de los conglomerados, es decir, los valores que corresponden, en las variables utilizadas, a los 10 casos que han sido elegidos como centros respectivos de los conglomerados solicitados.

Tabla 5.3. Centros Iniciales de los Conglomerados

	Conglomerado		
	1	2	3
Puntua: AñosServicio	1,13194	0,08977	0,89598
Puntua: MEAN(EDAD,ALL)	-0,27825	0,02354	2,35155
Puntua(PROFESION)	-0,27975	-0,27975	-0,27975
Puntua(GRADO_INSTR)	-0,56180	1,93417	-0,56180
Puntua: LUGAR DE NACIMIENTO	0,73624	1,71312	-2,47930
Puntua(NRO_HIJOS)	0,47337	0,47337	0,96006
Puntua(EDO_CIVIL)	2,93016	2,93016	-0,07928
Puntua(PERMISOS)	-0,52977	5,32647	-0,94807
Puntua(AMONESTACIONES)	9,52055	-0,33375	-0,33375
Puntua: MEAN(ESTATURA,ALL)	-0,60152	0,00000	0,00000

Se ha logrado la convergencia en la iteración 16. El cambio máximo de coordenadas absolutas para cualquier centro es de 0,000, como se puede apreciar en la tabla 5.4.

Tabla 5.4. Historial de Iteraciones

	Cambio en los centros de los conglomerados		
Iteración	1	2	3
1	3,034	4,171	3,766
2	2,021	,680	,100
3	1,457	,342	,083
4	,876	,109	,059
5	,408	,000	,028
6	,520	,000	,043
7	,425	,000	,043
8	,562	,124	,088
9	,570	,161	,111
10	,712	,095	,267
11	,421	,146	,236
12	,228	,145	,144
13	,113	,079	,075
14	,107	,000	,081
15	,099	,060	,077
16	,000	,000	,000

En la Tabla 5.5 se tienen las distancias entre pares de conglomerados finales que permite constatar cuán próximos o alejados están unos de otros.

Tabla 5.5. Distancias entre los Centros de los Conglomerados Finales

Conglomerado	1	2	3
1		3,144	2,693
2	3,144		2,618
3	2,693	2,618	

En la Tabla 5.6 se puede apreciar los resultados del análisis de varianza que permite constatar la variabilidad entregrupos e intragrupos. Conglomerado Media Cuadrática es la variabilidad entre los diferentes conglomerados, equivalente a la media cuadrática entregrupos del análisis de varianza. El Error Media Cuadrática es la variabilidad intragrupos. El cociente entre una y otra es el valor de F. Los grados de libertad de una y

otra corresponden al número de conglomerados menos uno y número de individuos menos número de conglomerados respectivamente. En nuestro caso estos valores corresponden a gl=2 y gl=326 respectivamente. Finalmente se tiene el grado de significación de la F. Cuanto mayor sea la F para una variable y menor el nivel de significación de la misma, mayor será su contribución en la diferenciación de los grupos. En este caso la variable que ofrece mayor dispersión entre los diferentes conglomerados es el número de permisos con una F = 256,213 y nivel de significación Sig. = 0,000. Es importante destacar que los resultados de este análisis se interpretan desde una perspectiva meramente descriptiva ya que los conglomerados han sido formados de modo que maximicen la diferencia de los individuos entre lo mismos y por tanto no puede utilizarse como un test de contraste de medias entre diferentes grupos.

Tabla 5.6. Análisis de Varianza

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntua: AñosServicio	94,998	2	0,427	321	222,433	0,000
Puntua: MEAN(EDAD,ALL)	64,861	2	0,616	321	105,216	0,000
Puntua(PROFESION)	15,109	2	0,899	321	16,803	0,000
Puntua(GRADO_INSTR)	5,874	2	0,960	321	6,121	0,002
Puntua: LUGAR DE NACIMIENTO	2,149	2	0,984	321	2,183	0,114
Puntua(NRO_HIJOS)	39,144	2	0,748	321	52,303	0,000
Puntua(EDO_CIVIL)	24,441	2	0,867	321	28,197	0,000
Puntua(PERMISOS)	99,910	2	0,390	321	256,213	0,000
Puntua(AMONESTACIONES)	10,246	2	0,950	321	10,789	0,000
Puntua: MEAN(ESTATURA,ALL)	7,455	2	0,974	321	7,656	0,001

Por último, la Tabla 5.7 informa sobre el número de casos asignado a cada conglomerado.

Tabla 5.7. Número de casos en cada Conglomerado

Conglomerado	1	125,000
	2	39,000
	3	160,000
Válidos		324,000
Perdidos		0,000

Una vez obtenidos los conglomerados se efectúa una categorización de los mismos de la siguiente manera:

Grupo 1: elevados años de servicio, instrucción primaria aprobada, menor número de permisos, mayor edad.

Grupo 2: años de servicio moderado, instrucción primaria incompleta, mayor número de permisos, edad moderada.

Grupo 3: pocos años de servicio, instrucción media, número de permisos moderado, menos edad.

En este sentido y de acuerdo a las características presentadas por cada uno de ellos, estos se clasifican en grupos de Alto, Bajo y Mediano desempeño a 125, 39 y 160 vigilantes respectivamente.

Como una muestra de este agrupamiento a continuación se presenta una muestra de los primeros 23 casos utilizados en el análisis con indicación del conglomerado al que ha sido asignado cada caso y la distancia euclídea existente entre cada caso y el centro de su conglomerado. Los casos 1, 2, y 3 pertenecen al *conglomerado* 1, mientras que el caso 4 pertenece al *conglomerado* 3 y así sucesivamente. Ver tabla 5.8.

Tabla 5.8. Pertenencia a los Conglomerados

Número de caso	Conglomerado	Distancia
1	1	2,206
2	1	3,211
3	1	1,045
4	3	3,637
5	1	2,309
6	1	1,742
7	1	1,136
8	1	1,438
9	1	2,094
10	1	2,674
11	2	3,098
12	2	2,308
13	1	2,725
14	1	2,340
15	1	1,555
16	1	1,681
17	2	1,927
18	3	2,493
19	1	2,744
20	3	2,431
21	2	4,980
22	3	1,524
23	1	2,278

5.4.2 Análisis Discriminante

Como se explicó en el capítulo 4 para la obtención del modelo predictivo (Función Discriminante) se utilizará el análisis discriminante paso a paso. Este método toma como referencia un criterio o variable independiente, en este caso los grupos establecidos previamente a través del Análisis de Conglomerados de K-Medias.

En cada paso se informa de la variable que ha sido incorporada al modelo y, en su caso, de la variable o variables que han sido expulsadas. En este caso, todos los pasos llevados a cabo han sido de incorporación de variables: en el primer paso, la variable **permisos**; en el segundo, la variable **años de servicio**; etc. Así hasta un total de ocho variables, como se muestra en la tabla 5.9.

Tabla 5.9. Variables Introducidas/eliminadas

Paso	Introducidas	Lambda de Wilks							
						F exacta			
		Estadístico	gl1	gl2	gl3	Estadístico	gl1	gl2	Sig.
1	Puntua(PERMISOS)	0,385	1	2	321,000	256,213	2	321,000	0,000
2	Puntua: AñosServicio	0,161	2	2	321,000	238,898	4	640,000	0,000
3	Puntua(EDO_CIVIL)	0,144	3	2	321,000	174,335	6	638,000	0,000
4	Puntua(GRADO_INSTR)	0,132	4	2	321,000	139,084	8	636,000	0,000
5	Puntua(NRO_HIJOS)	0,124	5	2	321,000	116,593	10	634,000	0,000
6	Puntua: MEAN(EDAD,ALL)	0,118	6	2	321,000	100,476	12	632,000	0,000
7	Puntua: LUGAR DE NACIMIENTO	0,115	7	2	321,000	87,925	14	630,000	0,000
8	Puntua: MEAN(ESTATURA,ALL)	0,111	8	2	321,000	78,310	16	628,000	0,000

El estadístico *lambda de Wilks*, corresponde al método elegido para conocer una a una, las variables que se van incorporando a la función discriminante tras evaluar, su grado de contribución individual a la diferenciación de los grupos. En este método de selección de variables, cada variable independiente candidata a ser incluida en el modelo se evalúa mediante el *estadístico F*, que mide el cambio que se produce en el valor de la lambda de Wilks al incorporar cada una de las variables al modelo. Obtenido el valor del estadístico F para cada variable, se incorpora al modelo la variable a la que le corresponde el mayor valor F (o, lo que es lo mismo, la que produce el mayor cambio en la lambda de Wilks). Se observa que el valor del estadístico lambda de Wilks va disminuyendo en cada paso, lo cual es síntoma de que, conforme se van incorporando variables al modelo, los grupos van estando cada vez menos solapados. En la columna *F exacta* se encuentra el valor transformado de la lambda de Wilks y su significación.

La Tabla 5.10 se encuentra dividida por cada uno de los pasos. En cada paso se mencionan las variables incorporadas al modelo hasta ese momento y, para cada variable, el nivel de *tolerancia*, el valor del *estadístico F* que permite valorar si la variable debe o no ser expulsada (*F para eliminar*) y la *lambda de Wilks* global que se obtendría si se eliminara la variable del modelo.

Tabla 5.10. Variables Incluidas en el Análisis

Paso		Tolerancia	F para eliminar	Lambda de Wilks
1	Puntua(PERMISOS)	1,000	256,213	
2	Puntua(PERMISOS)	,995	256,826	,419
	Puntua: AñosServicio	,995	223,037	,385
3	Puntua(PERMISOS)	,983	255,600	,374
	Puntua: AñosServicio	,985	204,471	,328
	Puntua(EDO_CIVIL)	,980	19,283	,161
4	Puntua(PERMISOS)	,982	254,628	,344
	Puntua: AñosServicio	,946	219,014	,314
	Puntua(EDO_CIVIL)	,973	20,484	,149
	Puntua(GRADO_INSTR)	,956	13,524	,144
5	Puntua(PERMISOS)	,982	253,472	,322
	Puntua: AñosServicio	,933	155,675	,246
	Puntua(EDO_CIVIL)	,968	16,796	,137
	Puntua(GRADO_INSTR)	,941	15,598	,136
	Puntua(NRO_HIJOS)	,960	10,490	,132
6	Puntua(PERMISOS)	,981	250,889	,306
	Puntua: AñosServicio	,791	67,900	,169
	Puntua(EDO_CIVIL)	,965	17,251	,131
	Puntua(GRADO_INSTR)	,935	16,671	,131
	Puntua(NRO_HIJOS)	,958	9,198	,125
	Puntua: MEAN(EDAD,ALL)	,814	7,746	,124
7	Puntua(PERMISOS)	,964	257,233	,302
	Puntua: AñosServicio	,781	68,144	,164
	Puntua(EDO_CIVIL)	,965	17,100	,127
	Puntua(GRADO_INSTR)	,929	16,123	,126
	Puntua(NRO_HIJOS)	,958	9,057	,121
	Puntua: MEAN(EDAD,ALL)	,811	7,510	,120
	Puntua: LUGAR DE NACIMIENTO	,963	5,036	,118
8	Puntua(PERMISOS)	,963	256,209	,293
	Puntua: AñosServicio	,779	68,359	,160
	Puntua(EDO_CIVIL)	,965	16,509	,123
	Puntua(GRADO_INSTR)	,926	14,485	,122
	Puntua(NRO_HIJOS)	,958	9,002	,118
	Puntua: MEAN(EDAD,ALL)	,811	7,376	,117
	Puntua: LUGAR DE NACIMIENTO	,963	5,037	,115
	Puntua: MEAN(ESTATURA,ALL)	,988	4,417	,115

Puesto que las variables utilizadas en este trabajo presentan un coeficiente de correlación cercano a cero, lo cual indica que existe independencia entre las mismas, la tolerancia no

disminuye sensiblemente en el momento en que se incorpora una nueva variable al modelo. En el paso 1 puede observarse que el nivel de tolerancia para la variable **permisos** es 1 pues, al estar sola, no existen variables que puedan explicar nada de ella. En el segundo paso, al incorporarse la variable **años de servicio** al modelo, la tolerancia baja a 0.995, lo cual refleja que no existe una alta correlación entre ésta y la variable permisos (la correlación entre las dos variables es de 0.005). De la misma manera, la variable **estado civil** no correlaciona tanto con la variable **permisos** y **años de servicio**: al incorporarse al modelo en el tercer paso, su tolerancia sólo baja hasta 0.980.

Asimismo, se puede apreciar en la tabla 5.11, que las variables **profesión** y **amonestaciones** quedan excluidas del modelo con un valor de lambda de Wilks = 0.111 y 0.110 respectivamente y con un valor F = 0.151 y F = 2.795, no cumpliéndose el criterio de entrada que exige que la F sea mayor que 3.84.

Tabla 5.11. Variables no Incluidas en el Análisis

Paso		Tolerancia	Tolerancia mín.	F para introducir	Lambda de Wilks
0	Puntua: Años Servicio	1,000	1,000	222,433	,419
	Puntua: MEAN(EDAD,ALL)	1,000	1,000	105,216	,604
	Puntua(PROFESION)	1,000	1,000	16,803	,905
	Puntua(GRADO_INSTR)	1,000	1,000	6,121	,963
	Puntua: LUGAR DE NACIMIENTO	1,000	1,000	2,183	,987
	Puntua(NRO_HIJOS)	1,000	1,000	52,303	,754
	Puntua(EDO_CIVIL)	1,000	1,000	28,197	,851
	Puntua(PERMISOS)	1,000	1,000	256,213	,385
	Puntua(AMONESTACIONES)	1,000	1,000	10,789	,937
	Puntua: MEAN(ESTATURA,ALL)	1,000	1,000	7,656	,954
1	Puntua: Años Servicio	,995	,995	223,037	,161
	Puntua: MEAN(EDAD,ALL)	1,000	1,000	103,347	,234
	Puntua(PROFESION)	,994	,994	16,771	,349
	Puntua(GRADO_INSTR)	1,000	1,000	6,006	,371
	Puntua: LUGAR DE NACIMIENTO	,984	,984	4,720	,374
	Puntua(NRO_HIJOS)	,998	,998	52,256	,290
	Puntua(EDO_CIVIL)	,989	,989	28,148	,328
	Puntua(AMONESTACIONES)	,997	,997	11,182	,360
	Puntua: MEAN(ESTATURA,ALL)	,999	,999	7,512	,368
2	Puntua: MEAN(EDAD,ALL)	,823	,819	7,662	,154
	Puntua(PROFESION)	,898	,898	,147	,161
	Puntua(GRADO_INSTR)	,962	,958	12,351	,149
	Puntua: LUGAR DE NACIMIENTO	,973	,973	6,230	,155
	Puntua(NRO_HIJOS)	,980	,977	11,668	,150
	Puntua(EDO_CIVIL)	,980	,980	19,283	,144
	Puntua(AMONESTACIONES)	,990	,988	3,339	,158
	Puntua: MEAN(ESTATURA,ALL)	,992	,988	6,624	,154

Tabla 5.11. Variables no Incluidas en el Análisis. (Continuación)

3	Puntua: MEAN(EDAD,ALL)	,822	,815	7,848	,137
	Puntua(PROFESION)	,897	,892	,043	,143
	Puntua(GRADO_INSTR)	,956	,946	13,524	,132
	Puntua: LUGAR DE NACIMIENTO	,973	,966	6,012	,138
	Puntua(NRO_HIJOS)	,974	,966	8,462	,136
	Puntua(AMONESTACIONES)	,989	,978	2,894	,141
	Puntua: MEAN(ESTATURA,ALL)	,992	,979	5,991	,138
4	Puntua: MEAN(EDAD,ALL)	,815	,797	9,023	,125
	Puntua(PROFESION)	,870	,870	,191	,132
	Puntua: LUGAR DE NACIMIENTO	,967	,938	5,462	,128
	Puntua(NRO_HIJOS)	,960	,933	10,490	,124
	Puntua(AMONESTACIONES)	,989	,939	2,767	,130
	Puntua: MEAN(ESTATURA,ALL)	,989	,941	4,584	,129
5	Puntua: MEAN(EDAD,ALL)	,814	,791	7,746	,118
	Puntua(PROFESION)	,870	,864	,161	,124
	Puntua: LUGAR DE NACIMIENTO	,967	,925	5,261	,120
	Puntua(AMONESTACIONES)	,987	,926	2,972	,122
	Puntua: MEAN(ESTATURA,ALL)	,989	,929	4,530	,121
6	Puntua(PROFESION)	,868	,746	,083	,118
	Puntua: LUGAR DE NACIMIENTO	,963	,781	5,036	,115
	Puntua(AMONESTACIONES)	,983	,781	3,328	,116
	Puntua: MEAN(ESTATURA,ALL)	,989	,788	4,414	,115
7	Puntua(PROFESION)	,868	,737	,084	,115
	Puntua(AMONESTACIONES)	,982	,773	3,174	,112
	Puntua: MEAN(ESTATURA,ALL)	,988	,779	4,417	,111
8	Puntua(PROFESION)	,866	,736	,151	,111
	Puntua(AMONESTACIONES)	,978	,769	2,795	,110

Una vez conocidas las variables incluidas en el análisis, las cuales son las que más contribuyen a diferenciar a los grupos en la Función Discriminante la tabla 5.12 refleja la clasificación obtenida de los 30 primeros individuos de la muestra y que en resumen aporta la siguiente información:

- Número de casos: el número correspondiente a cada individuo en la base de datos.

- Grupo real: grupo al que pertenecen los individuos. Los individuos 1, 2 y 3 pertenecen al grupo 1, mientras que el individuo número 4 pertenece al grupo 3 y así sucesivamente.

- Grupo pronosticado: grupo al que son asignados los individuos de acuerdo con la Función Discriminante. Si de acuerdo con la misma el individuo está bien clasificado, no tendrá nada después del número del grupo destino. Si en cambio está mal clasificado, aparecerán dos asteriscos a su derecha, signo inequívoco de su incorrecta clasificación. La función discriminante lo ha clasificado en el grupo en el que su pertenencia tiene una mayor probabilidad a posteriori.

- Grupo mayor: grupo al que tienen mayor probabilidad de pertenecer un individuo. Puede o no coincidir con el grupo real.

    - ✓ Probabilidad condicional
    - ✓ Grados de libertad
    - ✓ Probabilidad a posteriori
    - ✓ Distancia de mahalanobis al cuadrado respecto al centroide del grupo.

En la columna del segundo grupo mayor se observan los grupos al que pertenece cada individuo en segundo lugar en sentido probabilístico.

Las dos últimas columnas muestran las puntuaciones discriminantes de los individuos para las dos funciones discriminantes.

Tabla 5.12. Estadísticos por Casos

			Grupo mayor						Segundo grupo mayor			Puntuaciones discriminantes	
				P(D>d | G=g)									
	Número de casos	Grupo real	Grupo pronosticado	p	gl	P(G=g | D=d)	Distancia de Mahalanobis al cuadrado hasta el centroide	Grupo	P(G=g | D=d)	Distancia de Mahalanobis al cuadrado hasta el centroide	Función 1	Función 2
Original	1	1	1	,015	2	,876	8,421	2	,120	10,062	,886	3,095
	2	1	1	,745	2	,998	,589	3	,002	13,835	2,614	,015
	3	1	1	,785	2	,995	,486	3	,005	11,671	1,760	1,065
	4	3	1**	,134	2	,810	4,014	3	,369	5,509	,075	1,110
	5	1	1	,574	2	,971	1,110	3	,028	8,885	1,138	1,070
	6	1	1	,423	2	,918	1,719	3	,062	7,044	1,686	-,891
	7	1	1	,890	2	,972	,233	3	,028	7,801	1,634	,026
	8	1	1	,669	2	,988	,805	3	,012	10,057	2,162	-,477
	9	1	1	,587	2	,903	1,066	3	,097	6,023	,928	,566
	10	1	1	,441	2	,839	1,636	3	,161	5,428	1,349	-,737
	11	2	2	,143	2	,999	3,899	1	,001	19,539	-,667	3,959
	12	2	1**	,044	2	,593	6,244	2	,233	5,782	-,050	1,897
	13	1	1	,262	2	,859	2,677	3	,341	4,484	1,151	-1,036
	14	1	1	,509	2	,814	1,351	3	,186	4,796	,786	,320
	15	1	1	,589	2	,854	1,069	3	,146	5,079	,932	,221
	16	1	1	,309	2	,755	2,349	3	,243	5,111	,488	,762
	17	2	2	,127	2	,840	4,132	1	,137	10,092	-,244	2,685
	18	3	1**	,414	2	,715	1,763	3	,285	4,092	1,006	-,543
	19	1	1	,944	2	,884	,115	3	,016	8,836	1,617	,476
	20	3	3	,238	2	,817	2,870	2	,044	6,129	-,883	,700
	21	2	2	,000	2	1,000	29,884	3	,000	90,046	-3,806	8,061
	22	3	3	,403	2	,789	1,820	1	,211	3,967	,372	-,827
	23	1	1	,475	2	,884	1,490	3	,116	6,053	1,490	-,734
	24	1	1	,604	2	,983	1,008	3	,017	9,621	2,110	-,597
	25	3	3	,274	2	,895	2,589	1	,105	6,377	,493	-1,671
	26	3	3	,319	2	,641	2,283	1	,359	2,948	,421	-,395
	27	3	3	,569	2	,911	1,129	1	,089	5,296	,096	-,975
	28	3	3	,311	2	,628	2,333	1	,372	2,885	,416	-,344
	29	3	3	,500	2	,999	1,385	2	,001	12,939	-2,124	-,780
	30	1	1	,316	2	,527	2,307	3	,473	3,015	,682	-,449

5.4.3 Modelos Predictivos Obtenidos

Como resultado de la aplicación del análisis discriminante paso a paso, se obtuvo además de la clasificación mostrada en la tabla 5.12, dos modelos predictivos (Funciones Discriminantes) las cuales se analizarán a continuación:

$$F_1 = -0.003 + 1.058 * AñosServicio + 0.317 * Edad - 0.326 * GradoInstnucción - 0.128 * LugarNac. + 0.309 * NroHijos + 0.365 * EdoCivil - 0.802 * NroPermisos - 0.176 * Estatura$$

$$F_2 = 0.004 + 0.419 * AñosServicio + 0.170 * Edad - 0.176 * GradoInstnucción + 0.187 * LugarNac. + 0.106 * NroHijos + 0.172 * EdoCivil - 1.405 * NroPermisos - 0.100 * Estatura$$

Los autovalores de las dos funciones que componen el modelo son desiguales, como se aprecia en la tabla 5.13. La primera función explica el 62.9% de la variabilidad disponible en los datos (*% de varianza*), mientras que la segunda función explica el 37.1%.

Tabla 5.13. Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	2,568[a]	62,9	62,9	,848
2	1,514[a]	37,1	100,0	,776

Los coeficientes no estandarizados de cada variable para cada una de las funciones discriminantes se muestran en la tabla 5.14. A partir de estos coeficientes y de los valores de cada individuo en las variables independientes es posible obtener las dos funciones discriminantes y a partir de ellas las puntuaciones discriminantes para cada uno de los individuos en ambas funciones.

Tabla 5.14. Coeficientes no estandarizados de las Funciones Discriminantes

	Función	
	1	2
Puntua: AñosServicio	1,058	,419
Puntua: MEAN(EDAD,ALL)	,317	,170
Puntua(GRADO_INSTR)	-,326	-,176
Puntua: LUGAR DE NACIMIENTO	-,128	,187
Puntua(NRO_HIJOS)	,309	,106
Puntua(EDO_CIVIL)	,365	,172
Puntua(PERMISOS)	-,802	1,405
Puntua: MEAN(ESTATURA,ALL)	-,176	-,100
(Constante)	-,003	,004

Puesto que se obtienen tantas funciones discriminantes como grupos menos uno, en este estudio por tener tres categorías de individuos, son dos el número de funciones obtenidas. Las funciones discriminantes se extraen de manera jerárquica, de tal forma que la primera función explica el máximo posible de las diferencias entre los grupos, la segunda función explica el máximo de las diferencias todavía no explicadas, y así sucesivamente hasta alcanzar el 100% de las diferencias existentes. Por lo que serán estas, las Funciones Discriminantes utilizadas para clasificar según el desempeño en los servicios de vigilancia a nuevos individuos, a fin de predecir en qué categoría de desempeño laboral se encuentra una persona. En este sentido, estas funciones sirven de soporte al establecimiento de criterios y toma de decisiones para fortalecer las actividades de selección de personal.

5.5 Evaluación

Hasta aquí se ha llevado a cabo el proceso de construcción o estimación del modelo. Para valorar la capacidad predictiva del modelo estimado se debe prestar atención a los resultados de la clasificación, según sus probabilidades previas o a priori.

5.5.1 Clasificación con Probabilidades Previas Iguales

Al clasificar utilizando el software para el análisis estadístico SPSS fijando probabilidades previas iguales para cada grupo se obtienen los resultados de la tabla 5.15 y 5.16.
En la tabla 5.15, puede observarse una clasificación de 125, 39 y 160 casos en los grupos 1, 2 y 3 respectivamente.

Tabla 5.15. Probabilidades Previas Iguales para los Grupos

Número inicial de casos	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
1	,333	125	125,000
2	,333	39	39,000
3	,333	160	160,000
Total	1,000	324	324,000

En la tabla 5.16 se observa el resultado de la aplicación de la función discriminante. Puede apreciarse como en el grupo 1 que inicialmente lo conformaban 125 casos, se "movieron" 2 casos al grupo 2 y 4 casos al grupo 3. Así mismo, del grupo 2, se "movió" 1 caso al grupo 3 y del grupo 3 se "movieron" 7 casos al grupo 1. De donde puede concluirse que esta función discriminante clasifica correctamente el 95.67% de los casos agrupados originalmente con el análisis de conglomerados de k-medias.

Tabla 5.16. Resultados de la Clasificación

		Número inicial de casos	Grupo de pertenencia pronosticado			Total
			1	2	3	
Original	Recuento	1	119	2	4	125
		2	0	38	1	39
		3	7	0	153	160
	%	1	95,2	1,6	3,2	100,0
		2	,0	97,4	2,6	100,0
		3	4,4	,0	95,6	100,0

Esta tabla indica que se ha clasificado correctamente el 95.67% de los casos, lo cual, comparado con el 33% (probabilidades previas iguales para todos los grupos) esperable en

una clasificación completamente al azar, puede interpretarse como una mejora considerable.

En el grupo de individuos de bajo desempeño (grupo 2) se consigue el porcentaje más alto de clasificación correcta, 97.4%, frente a un porcentaje del 95.2% en el grupo de alto desempeño (grupo 1) y del 95.6% en el grupo de mediano desempeño (grupo 3).

Como complemento se presenta en la figura 5.5 el mapa territorial el cual representa el territorio (espacio) que corresponde a cada uno de los grupos en el plano definido por las dos funciones discriminantes: la primera función en el eje de las abscisas y la segunda función en el eje de ordenadas.

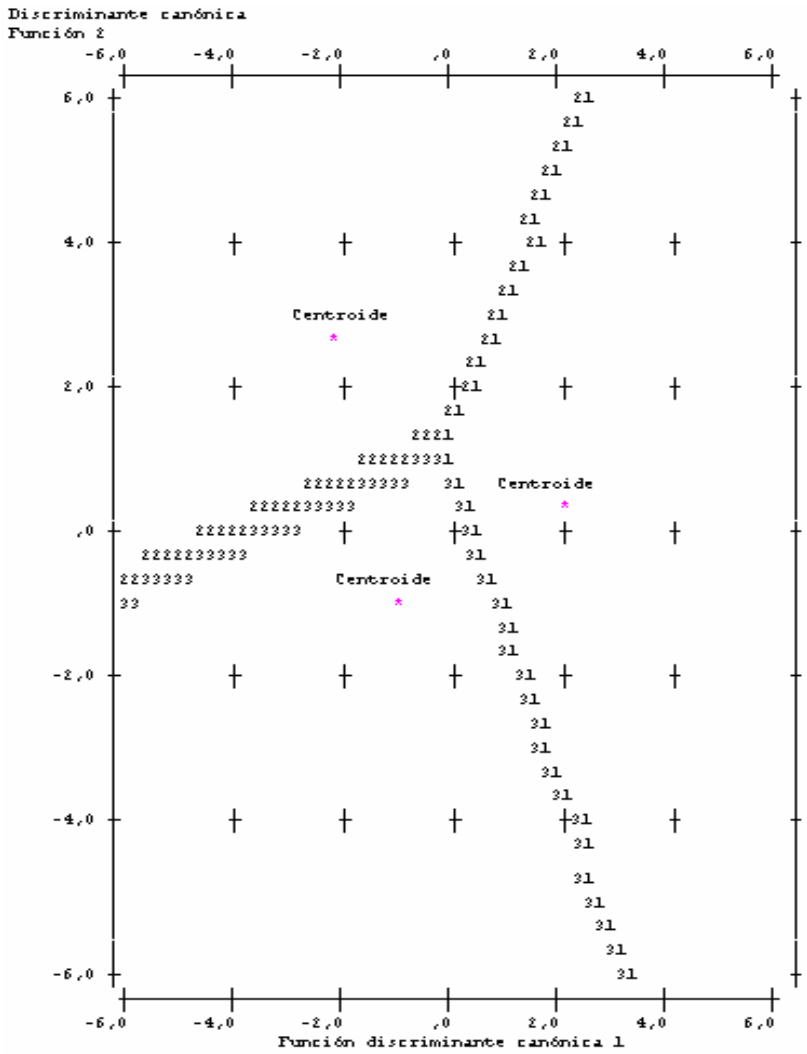

Figura5.5. Mapa Territorial

Los centroides de cada grupo están representados por asteriscos. Las secuencias de números que aparecen dividiendo el plano en territorios son los límites o fronteras impuestos por la regla de clasificación. Los números 1, 2,3 identifican el grupo al que corresponde cada territorio. Hay que tener en cuenta que, puesto que la regla de clasificación cambia al cambiar las probabilidades previas, si se cambian esas probabilidades también cambiarán las fronteras de los territorios (el efecto concreto es que las fronteras se alejan del centroide del grupo al que se le asigna mayor probabilidad).

El mapa territorial también se utiliza para clasificar individuos futuros. Para conocer el grupo pronosticado de un individuo cualquiera (es decir, el grupo en el que será clasificado), basta con representar en el mapa territorial el punto definido por sus puntuaciones discriminantes en ambas funciones. El grupo pronosticado es aquel al que corresponde el territorio en el que queda ubicado el punto.

Al observar la disposición de los tres territorios sobre el mapa, resulta fácil anticipar que los individuos con puntuaciones altas en la primera función discriminante serán clasificados en el grupo de alto desempeño (grupo 1), mientras que los individuos con puntuaciones próximas a cero o negativas en esa función serán clasificados en el grupo de bajo desempeño (grupo 2) o de mediano desempeño (grupo 3). En la segunda función discriminante, si la puntuación del individuo es positiva será clasificado en el grupo de bajo desempeño, mientras que si la puntuación en esa función es negativa será clasificado en el grupo de mediano desempeño.

La figura 5.6 muestra el diagrama de dispersión de todos los casos utilizados en el análisis sobre el plano definido por las dos funciones discriminantes. En la gráfica se representan también las posiciones de los centroides de grupo.

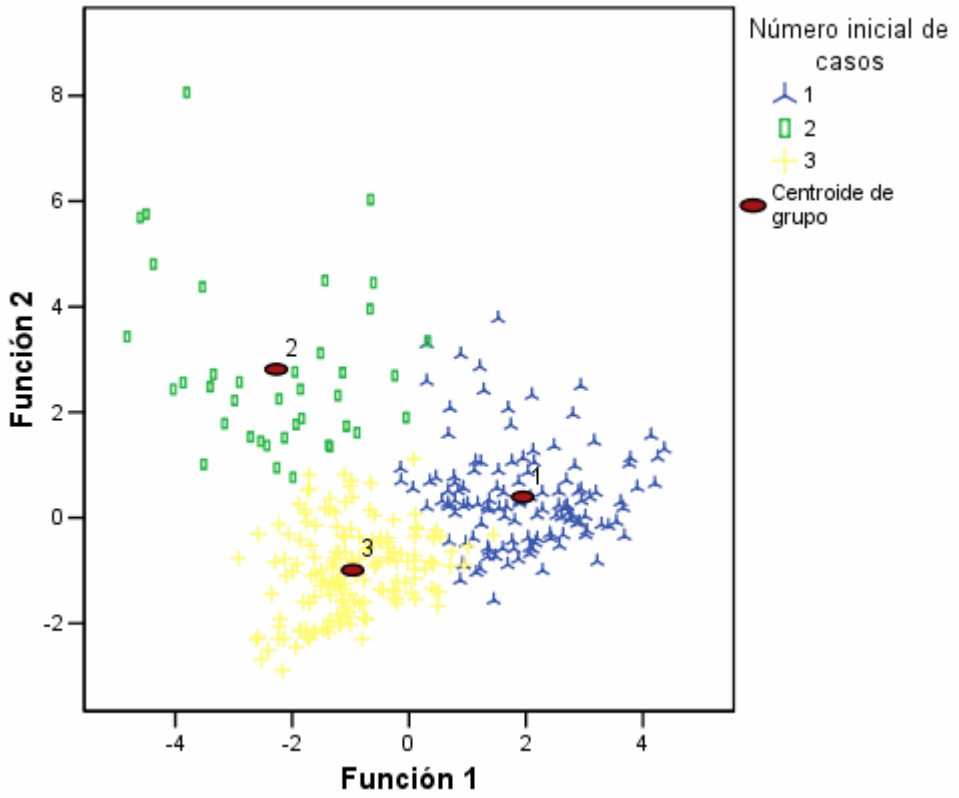

Figura 5.6. Diagrama de Dispersión de los Tres Grupos en las Dos Funciones Discriminantes

5.5.2 Clasificación con Probabilidades Previas según el Tamaño de los Grupos

Repitiendo el análisis con las probabilidades previas o a priori calculadas a partir del tamaño de los grupos (Ver tabla 5.17) la matriz de confusión ofrece los resultados que muestra la tabla 5.18.

Tabla 5.17. Probabilidades Previas según el Tamaño de los Grupos

Número inicial de casos	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
1	,386	125	125,000
2	,120	39	39,000
3	,494	160	160,000
Total	1,000	324	324,000

Tabla 5.18. Resultados de la Clasificación

		Número inicial de casos	Grupo de pertenencia pronosticado			Total
			1	2	3	
Original	Recuento	1	119	1	5	125
		2	1	35	3	39
		3	4	0	156	160
	%	1	95,2	,8	4,0	100,0
		2	2,6	89,7	7,7	100,0
		3	2,5	,0	97,5	100,0

Puede apreciarse que el porcentaje de clasificación correcta es el mismo 95.67% a pesar de que las probabilidades previas están calculadas según el tamaño de los grupos. Sin embargo, al variar los territorios con la nueva regla de clasificación, ha disminuido el porcentaje de clasificación correcta del grupo más pequeño (los de bajo desempeño), mientras que en el grupo más numeroso (los de mediano desempeño) ha aumentado el porcentaje de clasificación.

5.5.3 Capacidad Predictiva de la Función Discriminante

Para evaluar la capacidad predictiva de la función discriminante se llevó a cabo una validación cruzada, que consiste en:

- Seleccionar, de la muestra original, un subconjunto aleatorio de casos (muestra de validación).
- Estimar la función discriminante con los casos restantes (muestra de entrenamiento).

- Utilizar esa función para clasificar los casos de la muestra de validación.

La validación cruzada consiste, por tanto, en clasificar casos con una función que no incluye información sobre ellos.

Esta es realizada tomando en cuenta las siguientes consideraciones:

- Probabilidades previas calculadas a partir del tamaño de los grupos
- Se crea una variable de selección (es decir, una variable en la que aproximadamente el 90% los casos tiene el valor 1 y el resto el valor cero).
- El valor de la variable de selección identifica a los casos que serán incluidos en el análisis (muestra de entrenamiento).

La tabla 5.19 contiene las matrices de confusión correspondientes a los casos seleccionados (muestra de entrenamiento) y a los no seleccionados (la muestra de validación).

Tabla 5.19. Resultados de la Clasificación

			inicial de casos	Grupo de pertenencia			Total
				1	2	3	
Casos seleccionados	Original	Recuento	1	113	1	5	119
			2	1	33	3	37
			3	3	0	139	142
		%	1	95,0	0,8	4,2	100,0
			2	2,7	89,2	8,1	100,0
			3	2,1	0,0	97,9	100,0
Casos no seleccionados	Original	Recuento	1	6	0	0	6
			2	0	1	1	2
			3	1	0	17	18
		%	1	100,0	0,0	0,0	100,0
			2	0,0	50,0	50,0	100,0
			3	5,6	0,0	94,4	100,0

En la muestra de entrenamiento se obtiene una tasa de acierto del 95.63% y, en la de validación, una tasa de acierto del 92.30%. Se espera por tanto, que la función

discriminante obtenida clasifique correctamente al 92.30% de los futuros casos nuevos que se intenten clasificar.

Se repite el proceso de validación cruzada invirtiendo las muestras de entrenamiento y validación del primer análisis. Se puede comprobar que el porcentaje de clasificación correcta en la nueva muestra de entrenamiento es del 84.61% y del 85.90% en la nueva muestra de validación. Ver tabla 5.20.

Tabla 5.20. Resultados de la Clasificación

			inicial de casos	Grupo de pertenencia			Total
				1	2	3	
Casos seleccionados	Original	Recuento	1	3	0	3	6
			2	0	2	0	2
			3	1	0	17	18
		%	1	50,0	0,0	50,0	100,0
			2	0,0	100,0	0,0	100,0
			3	5,6	0,0	94,4	100,0
Casos no seleccionados	Original	Recuento	1	85	10	24	119
			2	0	37	0	37
			3	2	6	134	142
		%	1	71,4	8,4	20,2	100,0
			2	0,0	100,0	0,0	100,0
			3	1,4	4,2	94,4	100,0

Basándose en estos resultados, se puede concluir que, si utilizamos cualquiera de las dos funciones obtenidas para clasificar nuevos casos, se puede esperar que el porcentaje de clasificación correcta se encuentre por encima del 85.90%.

5.6 Explotación

En esta etapa de la metodología CRISP-MD, se presentan los resultados finales a partir del conocimiento obtenido de los datos.

5.6.1 Identificación de los Grupos Obtenidos

Inicialmente utilizando el Análisis de Conglomerados de K-Medias se obtuvieron tres grupos identificados como: (Ver Figura 5.7)

- Alto desempeño: 125 casos
- Mediano desempeño: 160 casos
- Bajo desempeño: 39 casos

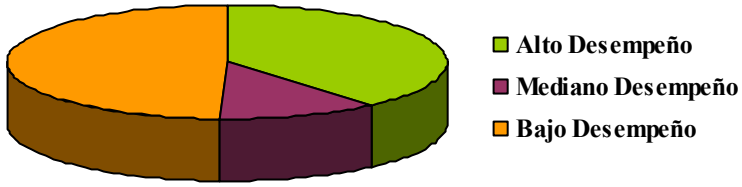

Figura 5.7. Número de casos en cada conglomerado

Los cuales como producto de la aplicación del Análisis Discriminante con probabilidades previas según el tamaño de los grupos se transformaron en: (Ver Figura 5.8)

- Alto desempeño: 124 casos
- Mediano desempeño: 164 casos
- Bajo desempeño: 36 casos

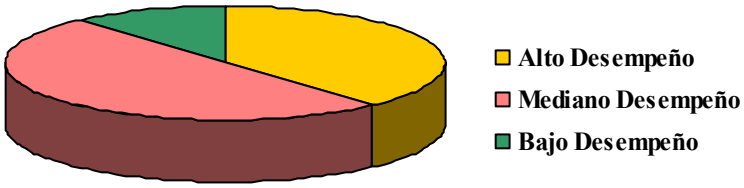

Figura 5.8. Resultados de la clasificación

Resultando de mayor importancia las variables años de servicio, edad, grado de instrucción, lugar de nacimiento, número de hijos, estado civil, número de permisos y estatura, siendo éstas las variables independientes que más contribuyen a diferenciar a los grupos en la Función Discriminante y pudiéndose excluir las variables profesión y amonestaciones.

- **Casos de Alto Desempeño**

  En estos casos se puede apreciar en la tabla 5.21 como características comunes entre los grupos las siguientes:

Tabla 5.21. Casos de Alto Desempeño. Características Comunes

Variables	Características Comunes	% de casos
Años de Servicio	18	8
Edad	50	2.4
Profesión	Vigilante	78.4
Lugar de Nacimiento	Estado Mérida	29.6
Número de Hijos	3	23.2
Estado Civil	Casado	73.6
Amonestaciones	0	73.6
Estatura	1.71	45.6

Entre las variables independientes que permiten diferenciar a los grupos, se encuentran el *Grado de Instrucción* y el *Número de Permisos*. (Ver tabla 5.22)

Tabla 5.22. Casos de Alto Desempeño. Características Distintivas

Variables	Características Distintivas	% de casos
Grado de Instrucción	Primaria Aprobada	54.4
Permisos	3	14.4

- **Casos de Mediano Desempeño**

Como puede apreciarse, la tabla 5.23 presenta como características similares en este grupo las siguientes:

Tabla 5.23. Casos de Mediano Desempeño. Características Comunes

Variables	Características Comunes	% de casos
Años de Servicio	18	11.9
Edad	50	5.6
Profesión	Vigilante	73.8
Lugar de Nacimiento	Estado Mérida	47.5
Número de Hijos	2	30.6
Estado Civil	Casado	61.3
Amonestaciones	0	95.6
Estatura	1.71	68.1

La tabla 5.24 muestra las variables independientes que contribuyen a la diferenciación de los grupos.

Tabla 5.24. Casos de Mediano Desempeño. Características Distintivas

Variables	Características Distintivas	% de casos
Grado de Instrucción	Diversificada	26.9
Permisos	5	16.3

- **Casos de Bajo Desempeño**

Las características más comunes en estos casos se reflejan en la tabla 5.25.

Tabla 5.25. Casos de Bajo Desempeño. Características Comunes

Variables	Característica Comunes	% de casos
Años de Servicio	18	15.4
Edad	50	5.1
Profesión	Vigilante	84.6
Lugar de Nacimiento	Estado Mérida	51.3
Número de Hijos	3	35.9
Estado Civil	Casado	66.7
Amonestaciones	0	82.1
Estatura	1.71	51.3

En la tabla 5.26 se encuentran las variables independientes que más contribuyen a diferenciar los grupos.

Tabla 5.26. Casos de Bajo Desempeño. Características Distintivas

Variables	Característica Distintivas	% de casos
Grado de Instrucción	Primaria Incompleta	25.6
Permisos	17	10.3

5.6.2 Evaluación de Casos Futuros o Anónimos

Con el propósito de conocer a qué grupo pertenece un caso o individuo futuro, éste se evalúa en la Función Discriminante creada a partir de los coeficientes no estandarizados (ver tabla 5.14). Una vez conocidas las puntuaciones (valores de las funciones discriminantes para cada individuo), el individuo futuro se clasificará en el grupo cuyo centroide esté más cerca de la puntuación discriminante considerada. En este sentido la

tabla 5.27 muestra la ubicación de los centroides de cada grupo. La primera función distingue fundamentalmente el grupo 1 (cuyo centroide se encuentra en la parte positiva) de los grupos 2 y 3 (cuyos centroides se encuentran en la parte negativa), mientras que la segunda función permite distinguir entre los dos grupos que han quedado más próximos en la primera.

Tabla 5.27. Funciones en los Centroides de los Grupos

Número inicial de casos	Función	
	1	2
1	1,946	,394
2	-2,273	2,814
3	-,966	-,993

Estos valores (las posiciones de los centroides de grupo) se pueden apreciar en el siguiente grafico, donde se muestra el diagrama de dispersión de los tres grupos, que permite situar la posición de los casos y los centroides sobre las dos funciones discriminantes simultáneamente. (Ver Figura 5.9)

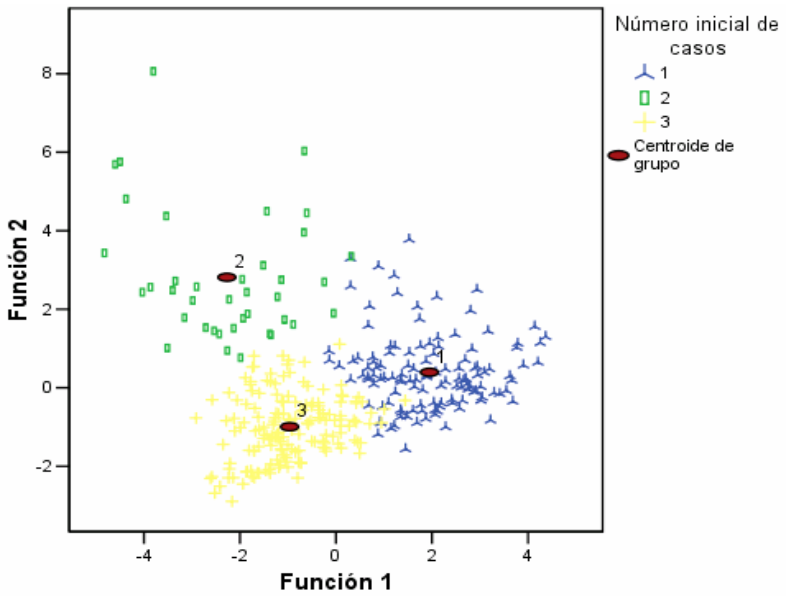

Figura 5.9. Diagrama de Dispersión de los Tres Grupos con sus respectivos centroides

De los 329 registros 5 de ellos fueron excluidos de los diversos análisis para ser considerados como casos anónimos y de esta manera determinar en qué grupo será clasificado. La tabla 5.28 muestra los registros que presentan los valores de las variables independientes que más contribuyen a diferenciar a los grupos.

Tabla 5.28. Casos Anónimos

	1	2	3	4	5
Años de Servicio	26,5780821918	18,2493150685	17,8383561644	17,8383561644	19,7561643836
Edad	52,9013698630	37,9890410959	54,1342465753	50,1800366460	55,1150684932
Grado de Instrucción	4	4	10	99	4
Lugar de Nacimiento	10	63	98	63	79
Número de Hijos	8	0	2	0	5
Estado Civil	2	1	2	1	2
Número de Permisos	2	4	20	4	5
Estatura	1,71	1,71	1,75	1,71	1,70

Para evaluar un individuo en la Función Discriminante, previamente se deben estandarizar los datos. Esto es debido a que el Análisis de Conglomerados de k-medias permite agrupar los casos en función del parecido o similitud existente entre ellos comparando la distancia de estos casos respecto a los centros iniciales, por lo que si las variables utilizan diferentes escalas, los resultados podrían ser equívocos.

En resumen para clasificar a un individuo futuro o anónimo, se pueden seguir los siguientes pasos:

- Calcular las variables estandarizadas utilizando la media y desviación estándar obtenida de la muestra de entrenamiento.

$$Z_{\text{var}\,iable} = \frac{Valor\,\text{Re}\,al - Media}{Desviación Es\,\text{tan}\,dar}$$

- Calcular el valor de las Funciones Discriminantes para el individuo.

$$F_1 = -0.003 + 1.058 * AñosServicio + 0.317 * Edad - 0.326 * GradoInstnucción - 0.128 * LugarNac. + 0.309 * NroHijos + 0.365 * EdoCivil - 0.802 * NroPermisos - 0.176 * Estatura$$

$$F_2 = 0.004 + 0.419 * AñosServicio + 0.170 * Edad - 0.176 * GradoInstnucción + 0.187 * LugarNac. + 0.106 * NroHijos + 0.172 * EdoCivil - 1.405 * NroPermisos - 0.100 * Estatura$$

- Si el valor la primera función (F_1) es positivo, el individuo es clasificado en el grupo de *Alto desempeño* (grupo 1), de lo contrario, si el valor de la segunda función (F_2) es positivo, el individuo es clasificado en el grupo de *Bajo desempeño* (grupo 2), de lo contrario es clasificado en el grupo de *Mediano Desempeño* (grupo 3).

Así por ejemplo, considerando el individuo número 1 de la tabla 5.28 se conocerá a que grupo pertenece: Alto, Mediano o Bajo Desempeño.

Tabla 5.29. Datos del Individuo a Evaluar

Variables	Valores Reales	Valores Estandarizados
AñosServicio	26,57808219	0,788446318
Edad	52,90136986	0,283933719
GradoInstruccion	4	-0,55464457
LugarNac	10	-2,096680451
NroHljos	8	2,378153338
EdoCivil	2	-0,088891524
Permisos	2	-0,681424656
Estatura	1,71	0,060719948

Tabla 5.30. Resultados de la clasificación

Valores de las Funciones Discriminantes	
F1=	2,60859226
F2=	-0,638512785
Resultado de la Clasificación	
Grupo	**Desempeño**
1	Alto
2	Bajo
3	Mediano

Como puede apreciarse en la tabla 5.30 el valor de la primera función (F_1), es positivo, por lo que el nuevo caso o individuo es clasificado en el grupo de Alto Desempeño.

Conclusiones

Una vez culminado el presente trabajo, se considera haber alcanzado el objetivo propuesto, obteniendo un modelo descriptivo y predictivo que permiten describir e identificar patrones de desempeño y comportamiento y, a partir de esta clasificación previa poder predecir y/o estimar a qué categoría pertenece un individuo, y de esta manera, proporcionar una herramienta que sirva de soporte a la toma de decisiones en la elección de personal.

A través de las herramientas de análisis de conglomerados de K-Medias y el análisis discriminante para el análisis exploratorio de datos multivariantes de vigilantes universitarios se ha obtenido, una combinación lineal de las variables independientes (Funciones Discriminantes) que permiten diferenciar (discriminar) a los grupos clasificados por sus características en los datos de alto, mediano y bajo desempeño. Encontradas estas funciones discriminantes, podrán ser utilizadas para clasificar según el desempeño en los servicios de vigilancia a nuevos individuos, a fin de predecir en qué categoría de desempeño laboral se encuentra una persona. Los individuos con puntuaciones discriminantes positivas en la primera función son clasificados en el grupo de alto desempeño (grupo 1), mientras que en la segunda función los individuos con puntuaciones positivas o negativas son clasificados en el grupo de bajo desempeño (grupo 2) o de mediano desempeño (grupo 3) respectivamente.

Como patrones de comportamiento se encontró que los casos clasificados de Alto desempeño presentan como características predominantes, instrucción primaria y pocos permisos. Así mismo los casos clasificados de Mediano desempeño presentan una instrucción media y un número de permisos moderado, mientras los clasificados en bajo desempeño reflejan como características una instrucción primaria incompleta y muchos permisos.

Este conocimiento generado puede utilizarse como soporte al establecimiento de criterios, estrategias y toma de decisiones para fortalecer las actividades de selección de personal.

Recomendaciones

Aumentar el volumen y la variedad de la información que se encuentra disponible en la base de datos, y repetir la aplicación de estas técnicas de Minería de Datos a fin de mejorar los resultados.

Experimentar con otras técnicas de clasificación, a fin de tener varios resultados que se puedan comparar, obteniendo así una mayor certeza en la clasificación.

Referencias Bibliográficas

[AE05] Aertia Software. (2005), NeuroSolutions de NeuroDimension. Disponible en: http://www.aertia.com/productos.asp?pid=218&pg=pr

[BG97] Berry, M.; Gordon, L. (1997), Data Mining Techniques For Marketing, Sales and Customer Support, Editorial Wiley.

[BJ96] Bigus, J. P. (1996), Data Mining with Neural Networks, McGraw Hill.

[CC00] Chapman, P.; Clinto, J. (2000), 'CRISP-DM 1.0. Step-by-step Data Mining Guide', Disponible en: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[CB00] Clark, P.; Boswell, R. (2000), Data Mining. Practical Machine Learning Tools and Techniques with Java Implementatios, Morgan Kaufmann.

[DL05] Dávila, L. A. (2005), Seminario de Investigación de Operaciones, EISULA.

[DG00] Diego García Morate. (2000), Manual de Weka, http://www.dia.fi.upm.es/~concha/SPAM/morate.pdf

[EI04] España. Inc., S. (2004), 'SPSS Base 13.0 Manual del Usuario', Disponible en: http://www.ualberta.ca/AICT/RESEARCH/NumStatsServers/SPSSusersguide13.pdf.

[FG00] Fayyad, U.; Grinstein, G. (2000), Information Visualization in Data Mining and Knowledger Discovery, Academic Press, UK.

[HA00] Hair, A. (2000), Análisis Multivariante, Prentice Hall.

[HA97] Hair, J. (1997), Análisis Multivariante, Mc.Graw Hill.

[HJ04] Hernández, J, et al. (2004), Introducción a la Minería de Datos, Pearson Educación S.A.

[IN03] Inflexa. (2003), 'Qué es minería de datos'. Disponible en: http://www.inflexa.com/jsp/template.jsp?pag=mineria-datos.htm&mnu=mnumineria.htm

[IT99] IT Innovation Centre. (1999), 'Critikal. European Project for large scale Data Mining', Disponible en: http://www.attar.com/pages/critikal.htm

[JG05] Javier García de Jalón, José Ignacio Rodríguez, Jesús Vidal. (2005), Aprenda Matlab 7.0 como si estuviera en primero. Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid. Disponible en: http://mat21.etsii.upm.es/ayudainf/aprendainf/Matlab70/matlab70primero.pdf

[KD02] KDNuggets. (2002), 'Portal de Minería de Datos', Disponible en http://www.kdnuggets.com.

[LE01] Libro electrónico sobre Algoritmos de Data Mining. (2001), Disponible en: http://www.statsoft.com/textbook/stathome.html

[MB05] Martínez, F.; Bienvenido, J. (2005), Apuntes de la Asignatura Minería De Datos, Universidad de la Rioja. Área de Proyectos de Ingeniería. Departamento de Ingeniería Mecánica.

[PD02] Peña, D. (2002), Análisis Multivariante de datos, Mc-Graw Hill.

[PF91] Piatetski-Shapiro, G.; Frawley, W. (1991), Knowledge Discovery in Databases, AAAI/MIT Press.

[PL04] Pérez López, C. (2004), Técnicas de Análisis Multivariante de Datos, Pearson Educación S.A. España.

[SAS01] SAS. (2001), 'Semma. a proven data mining process', Disponible en: http://www.sas.com/products/miner/semma.html

[SA99] Scout ; Al-Attar ; Schneider ; Nisbet ; Barth ; Schwarz, H. (1999), CRITIKAL Final Report, Departament of ECS. University of Southampton.

[SU06] SIVULA. (2006), 'SIVULA (Sistema de Información Vigilancia ULA)', Disponible en: http://www.vigilancia.ula.ve.

[SPSS01] SPSS (2001), Clementine 6.0: Users Guide, SPSS.

[UCA02] UCA. (2002), 'SPSS 10. Guía para el Análisis de Datos', Disponible en: http://www2.uca.es/serv/ai/formación/spss/Pantalla/verguía.pdf

[V03a] Visauta, B. (2003a), Análisis Estadístico con SPSS para Windows, Vol Estadística Básica. Segunda Edición, Mc-Graw Hill.

[V03b] Visauta, B. (2003b), Análisis Estadístico con SPSS para Windows, Vol. II. Estadística Multivariante. Segunda Edición, Mc-Graw Hill.