UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

Técnicas de Inteligencia Computacional Aplicadas a la Determinación Óptima de Grupos en Minería de Datos: Un Caso de Estudio

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN INGENIERÍA (COMPUTACIÓN)

P R E S E N T A:

FÁTIMA LIZETH RODRÍGUEZ ERAZO

DIRECTOR DE TESIS:

DR. ÁNGEL FERNANDO KURI MORALES

México, D.F. **2007.**

Los triunfos no son nada si no hay con quien compartirlos.

Es por eso que dedico este trabajo a Dios y a mi familia.

Alcanzar esta meta ha sido el producto de mucho esfuerzo, pero llegar hasta aquí no hubiera sido posible sin el apoyo de tantas personas buenas que Dios puso en mi camino. Quiero enviar un "gracias" muy especial a mis padres por su amor, a Roberto por su paciencia y apoyo y al Dr. Ángel Kuri por sus consejos y dedicación a este trabajo y a ésta que fue su alumna y que ha aprendido mucho con esta experiencia.

No puedo nombrarlos a todos, pero vaya para ustedes mis maestros, amigos, compañeros y colegas mi más sincero agradecimiento.

Fátima Rodríguez Erazo

Índice

Índice de Tablas

Índice de Figuras

Capítulo 1 Introducción

El presente documento tiene por objetivo presentar el desarrollo y los resultados del trabajo de investigación realizado para la formulación, aplicación (a un caso de estudio) y validación de una metodología para el análisis de grupos en conjuntos grandes de datos[1].

El desarrollo y resultados del proyecto se acompañan con un marco teórico base para la metodología formulada y con información general de la definición inicial del proyecto.

Este documento incluye 6 capítulos. El capítulo 1 (Introducción) contiene información general del planteamiento inicial del proyecto, así como un marco de conocimiento general de la situación actual al momento de iniciarse este trabajo, es decir, el estado del arte. Se hace mención de trabajos recientes realizados en el área de minería de datos y tratamiento de grandes bases de datos.

En el capítulo 2 (Marco Teórico) se incluye un marco teórico que describe brevemente las herramientas base para la definición de la metodología propuesta y que fueron utilizadas en el caso de estudio.

El detalle de la metodología formulada se presenta en el capítulo 3 (Metodología). Se definen sus pasos y la forma como puede ser aplicada a un caso de estudio.

El caso de estudio que sirvió como marco de aplicación y prueba[2] para la metodología formulada se detalla en el capítulo 4 (Caso de Estudio). Comprende información general de la empresa propietaria de los datos, una descripción del problema tratado, información referente a la forma como la metodología fue aplicada y los resultados obtenidos para el caso de estudio.

En el capítulo 5 (Análisis de Resultados) se presenta el análisis de los resultados obtenidos. El análisis parte de dos enfoques: validación de los resultados obtenidos y evaluación de la eficiencia del proceso. Se comparan los datos y la eficiencia resultante de la aplicación de la metodología propuesta contra los resultados de un agrupamiento sin la utilización de dicha metodología.

El capítulo 6 (Conclusiones) contiene las conclusiones generales de este trabajo, además de sugerencias para futuras investigaciones.

[1] Un resumen de la metodología propuesta y su aplicación al caso de estudio puede ser consultado en [KURI07].

[2] Datos de una empresa Latinoamericana multinacional.

1.1 Descripción del trabajo

Al inicio de este proyecto existían ciertas condiciones que motivaron el surgimiento del mismo y que sustentaron su desarrollo. Para poder comprender la motivación inicial se presenta en esta sección una descripción del proyecto partiendo desde un marco general de antecedentes, con base en el cual, se formula el planteamiento del problema, los objetivos planteados y la justificación del trabajo. Además se enumeran los principales resultados esperados para la conclusión del proyecto.

1.1.1 Antecedentes

En la era digital actual los volúmenes de información manejados por bases de datos, almacenes de datos (también llamados Data Warehouses) y otras fuentes de datos crecen desmesuradamente, haciendo cada vez más difícil su análisis y la obtención de información valiosa.

La minería de datos se ha vuelto una importante herramienta para el análisis de grandes volúmenes de datos, tomando cada vez más popularidad en su utilización y mejorando día con día las técnicas, algoritmos y procesos utilizados para trabajar en el análisis de fuentes de datos.

La minería de datos es el resultado de combinar las disciplinas de estadística, tecnología de bases de datos, reconocimiento de patrones, inteligencia artificial y visualización, para poder descubrir conocimiento a partir de un conjunto grande de datos. Comprende varias tareas de análisis, entre las cuales se encuentra la tarea de formar agrupamientos de elementos a partir de los datos que los representan, con el fin de conocer mejor los elementos en estudio y formar una base para su clasificación. Esta tarea de agrupamiento (conocida también como clustering), es una de las tareas más importantes y difundidas de la minería de datos.

A pesar de que la minería de datos proporciona herramientas valiosas para la extracción de conocimientos de grandes volúmenes de datos, a medida que las dimensiones de las fuentes de datos van aumentando, los resultados se obtienen con mayor esfuerzo, tiempo y en ocasiones pérdida de la calidad del resultado. Es por esto que el análisis de grandes conjuntos de datos se ha convertido en uno de los 10 problemas principales con que se enfrenta la minería de datos [YANG06] y se considera un área de vital importancia donde deberán existir más esfuerzos e investigaciones que permitan alcanzar resultados óptimos [FAYYAD03].

Muchos esfuerzos se han realizado para mejorar el desempeño de la minería de datos: optimizaciones de algoritmos, formulación de nuevos algoritmos, optimización del uso de recursos hardware trabajando con procesamiento paralelo, distribuido y con cómputo en malla (del inglés grid computing). También se han aplicado técnicas de reducción de las dimensiones de conjuntos de datos haciendo uso de técnicas estadísticas como muestreo, análisis multivariante, entre otros.

La tarea de agrupamiento en minería de datos también presenta múltiples algoritmos, optimizaciones y nuevas técnicas que tratan de mejorar el desempeño de la agrupación de elementos en un conjunto enorme de datos.

Existen múltiples documentos que muestran datos de casos de estudio de agrupamientos sobre diferentes elementos como preferencias de clientes para productos alimenticios [COZ06], archivos [FORMAN05], datos financieros [BENSMA04], código de software [KANELL06], etc. Cada uno de ellos presenta de manera general el trabajo que se realizó y los resultados que se obtuvieron. Existen también múltiples documentos que presentan las técnicas disponibles para realizar minería de datos. Técnicas ampliamente conocidas se presentan en los libros de minería de datos y en libros de agrupamiento, en artículos se presentan también mejoras a las técnicas ya conocidas, alternativas de trabajo con dichas técnicas e incluso combinación de dos o más técnicas para conformar una nueva técnica de trabajo.

Con tantas herramientas disponibles para el profesional que desee realizar agrupamientos sobre un caso real, existe el problema de cómo comenzar a trabajar y cómo aplicar estas herramientas más adecuadamente. Más concretamente, las limitaciones de la minería de datos están más relacionadas a datos y personas que a tecnología disponible [SEIFER04].

1.1.2 Planteamiento del Problema

Se requiere de una metodología sistemática, basada en técnicas de inteligencia computacional[3], que permita analizar conjuntos grandes de datos (específicamente realizando análisis de agrupamiento de elementos) reduciendo el esfuerzo necesario, el sesgo en el tratamiento de los datos y optimizando los resultados de acuerdo a un mínimo de datos fuente.

1.1.3 Objetivos del trabajo

Objetivo general

Proponer, aplicar y validar una metodología sistemática y eficiente para el agrupamiento de elementos en el análisis de grandes volúmenes de datos que, utilizando técnicas de inteligencia computacional, reduzca el sesgo en el tratamiento de los datos y obtenga resultados aceptables con el mínimo de esfuerzo y datos disponibles.

Objetivos específicos

- Proponer la metodología general para la determinación de grupos en grandes bases de datos.
- Aplicar la metodología propuesta en un caso de estudio real de agrupamiento de clientes.
- Validar los resultados obtenidos con la metodología propuesta.

[3] *Inteligencia Computacional* es una rama de la Ciencia de la Computación que estudia problemas para los cuales no hay algoritmos computacionales efectivos [1]. Incluye técnicas como: redes neuronales, lógica difusa, programación evolutiva.

1.1.4 Justificación y relevancia

Este trabajo se enfoca, fundamentalmente, en conocer, aplicar y sistematizar el uso de varias de las técnicas existentes para minería de datos a través de una metodología que, valiéndose de un análisis no sesgado (es decir, sin presunciones a priori), permita analizar de manera cercana al óptimo los agrupamientos de un conjunto grande de datos.

Con la formulación de una metodología para el análisis de grupos de una manera cercana al óptimo, en lugar de presentarse una herramienta más de minería de datos se plantea una forma de utilizar esas herramientas tan ampliamente difundidas con una visión objetiva e independiente de percepciones subjetivas.

Existen múltiples casos de estudio que pueden servir como referencia pero que no formalizan el trabajo realizado en cada uno de ellos. Existen, igualmente, múltiples técnicas de minería de datos para agrupamiento de datos y es mucha la bibliografía que presenta estas técnicas, pero en ella no se hace un especial énfasis en la metodología que podría seguirse para analizar los datos.

1.1.5 Resultados esperados

- Metodología de análisis definida y probada en un caso de estudio.
- Obtener una confiabilidad de la metodología con un intervalo de confianza del orden del 5%.
- Obtener resultados aceptables para el caso de estudio, que satisfagan las necesidades de la empresa interesada.

1.2 Estado del Arte

Se han realizado diferentes esfuerzos por mejorar la forma en que la información es analizada a partir de grandes conjuntos de datos, ésta es precisamente la tarea de la minería de datos, que con sus variadas técnicas de estadística, computación, bases de datos, inteligencia artificial y visualización de datos, busca facilitar al investigador la obtención de información útil y relevante.

La utilización de minería de datos en el área de bases de datos ha sido de gran importancia para mejorar la forma en que las empresas trabajan y aprovechan la información generada por sus operaciones. De esto mismo se desprende el interés de muchos investigadores por optimizar los procesos actuales de minería de datos.

1.2.1 Análisis de grandes bases de datos

Cuando las bases de datos eran pequeñas (con cantidades de registros en el orden de decenas o centenas), analizar su contenido era fácil y podía realizarse incluso de manera manual. En estas circunstancias, casi cualquier algoritmo de minería de

datos puede obtener un desempeño aceptable con uso mínimo de recursos computacionales. Sin embargo, en el caso de las grandes bases de datos (cantidades de registros en miles, millones, billones, etc. y con dimensiones de cientos o miles de atributos) el desempeño y el uso de recursos computaciones se vuelve un punto crucial al momento de procesar los datos.

Con el aumento en el tamaño de las fuentes de datos crece la necesidad de tiempo de procesamiento y de accesos a datos. El tiempo de procesamiento depende de las características de hardware de la computadora utilizada para el proceso y del tipo de algoritmo empleado para analizar los datos; sin embargo, todo algoritmo dependerá, finalmente, del acceso a datos.

El tiempo de acceso a datos es el tiempo que transcurre desde que se solicita cierta información de memoria secundaria hasta que estos datos están disponibles. El tiempo de acceso a disco duro es mayor al de memoria y se mide en milisegundos (ms). Cubre el intervalo desde el momento en que se emitió la solicitud de acceso hasta el momento en que se recibe la información que indica el éxito (o fracaso) de la operación [2]. Durante este intervalo la unidad de disco mueve el cabezal de lectura/escritura sobre la superficie del disco hasta hacerlo llegar a la pista adecuada (tiempo de posicionamiento), sitúa el cabezal en su posición y espera a que los sectores correspondientes giren bajo el cabezal (tiempo de rotación) para ejecutar la lectura o escritura efectiva.

El tiempo de acceso de disco duro varía según las características de hardware. Un tiempo menor a 3 ms se considera rápido y un tiempo mayor se considera lento.

Si el análisis de una base de datos de 1 millón de registros y 100 variables requiriera de un único acceso a disco por cada dato a analizar (valor de una variable en un registro), se necesitarían 100 millones de accesos y tomaría 300 mil segundos (aproximadamente equivalente a 83.3 horas). Estos tiempos de procesamiento tan largos han hecho necesario buscar mejoras al proceso de minería de datos.

Al problema del tiempo de análisis de una gran base de datos se suma el problema de encontrar datos faltantes que alteren la exactitud de los resultados. Es por esto que las investigaciones de minería de datos trabajan a partir de dos enfoques principales: mejorar la eficiencia y mejorar la efectividad [KEIM99].

Mejorar la eficiencia implica analizar la base de datos haciendo el mejor uso de los recursos computacionales disponibles. Consecuentemente, significa mejorar los algoritmos de minería de datos para que obtengan una respuesta en el menor tiempo posible.

El segundo enfoque, mejorar la efectividad, busca analizar la base de datos y encontrar la respuesta más cercana a la exacta. Esto implica que los algoritmos de minería de datos sean capaces de encontrar una respuesta correcta a pesar de problemas en las fuentes de datos, tal es el caso de “ruido” en los datos y/o datos faltantes [KEIM99].

1.2.1.1 Preprocesamiento de datos

El preprocesamiento de datos implica preparar las fuentes de datos realizando labores de limpieza, transformación y reducción de los datos. Este paso de la minería de datos es el que puede llegar a tomar más tiempo de todo el proceso [FAYYAD03]. Se busca con este enfoque, mejorar la calidad de la "vista minable", de manera que los algoritmos de minería de datos que se ejecuten sobre los datos obtengan resultados con mayor eficiencia y/o efectividad. Esto se logra a través de la reducción de la vista minable con la que se procederá a realizar el procesamiento de la minería de datos, a fin de que los algoritmos trabajen con una fuente de datos más pequeña que represente el mismo comportamiento y características que la fuente de datos original. También se logra al transformar los datos continuos de la vista minable en un formato discreto que optimice el trabajo de los algoritmos de aprendizaje. Existen cuatro líneas de trabajo para el preprocesamiento de los datos [3]:

- Selección de características
- Selección de instancias
- Compactación de datos
- Discretización.

Selección de características. Este proceso consiste en reducir las características o atributos del conjunto de datos mediante la remoción de datos irrelevantes, redundantes o con ruido [LIU05]. Se han realizado muchas investigaciones en este tema [SAS06], [VU06], [ZHANG07], existiendo a la fecha una vasta cantidad de algoritmos disponibles. Una clasificación de estos algoritmos se presenta en [LIU05] y un resumen de las técnicas más usadas para reducción de dimensionalidad se presenta en [FODOR02].

La selección de características ha demostrado ser un proceso importante para áreas como estadística, reconocimiento de patrones, aprendizaje automático y minería de datos [SAS06].

En [LIU05] se resume el proceso de selección de características en 4 pasos que se esquematizan en la figura 1.

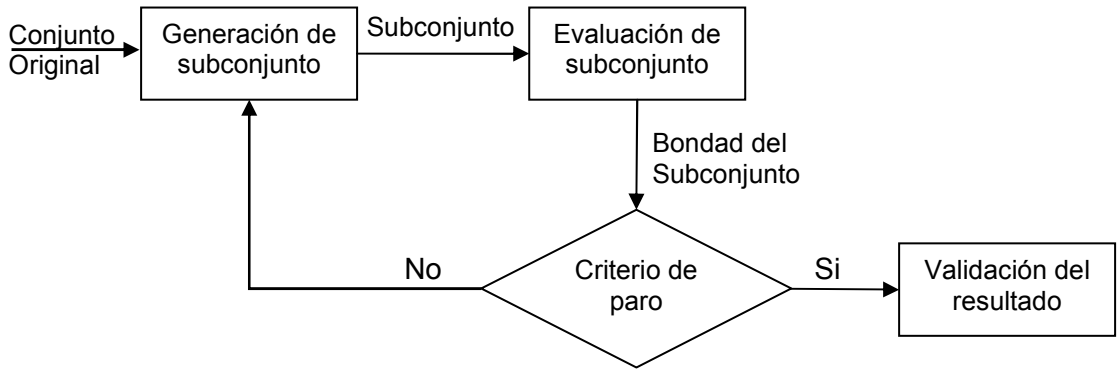


Figura 1: Proceso de Selección de Características

El primer paso del proceso, la generación de subconjunto, realiza una búsqueda de un subconjunto de características que será evaluado de acuerdo a algún criterio predefinido. Si el nuevo subconjunto resulta ser mejor que el anterior, reemplaza al mejor subconjunto encontrado y se realiza otra búsqueda de un mejor subconjunto. El proceso se detiene cuando se alcanza un criterio de paro establecido. En algunas ocasiones el resultado es validado según conocimiento a priori o con diferentes pruebas al conjunto de datos.

Selección de instancias. También llamada reducción de casos. Aunque está muy relacionada con un proceso de muestreo, la selección de instancias realmente importante para la minería de datos es aquella que provee un conjunto menor de instancias que presenta similares características que el conjunto original y que por lo tanto, puede ser usado para tareas de minería de datos.

Existen actualmente muchas formas de seleccionar instancias a partir de un conjunto de datos. En [LIU02] se presentan las técnicas más conocidas y se incluye una referencia a las mayores líneas de investigación y desarrollo del tema. Las técnicas de selección de instancias se clasifican en: muestreo, métodos asociados con clasificación, métodos asociados con agrupamiento y etiquetamiento de instancias.

La investigación para encontrar mejores formas de selección de instancias continúa abierta [ZHU06], [BRIGHT02], [PALMER02].

Compactación de datos. El término “compactación de datos” (del inglés *data squashing*) fue introducido por DuMouchel y otros autores en el artículo “Squashing Flat Files Flatter” [DUMOUC99] de 1999. En este artículo se define la compactación de datos como una forma de resumir, de manera efectiva, un conjunto grande de datos o instancias en una versión más pequeña, teniendo las mismas variables que la versión original de datos.

La metodología es motivada por la estadística y consiste, en términos generales, en obtener un conjunto n' más pequeño que el conjunto n original, en donde los puntos

de datos son seleccionados por un algoritmo que simula la estructura estadística del conjunto original de datos [HAND01-12]. Cada elemento del conjunto compactado de datos posee un peso y la suma de estos pesos es igual al número de elementos en el conjunto original de datos. Además, un elemento del conjunto compactado puede o no ser miembro del conjunto original de datos.

Según [LIU02] existen dos maneras de obtener la compactación de datos: 1) modelo libre y 2) modelo dependiente. El primero se basa en la adopción de momentos[4] para asegurar que los datos originales y los datos compactados son suficientemente parecidos. El segundo, asume un modelo estadístico para la compactación de datos. Divide los datos usando agrupamientos basados en probabilidad y luego selecciona elementos datos que permitan imitar la distribución supuesta de los datos.

Se han realizado diferentes investigaciones para mejorar los algoritmos de compactación de datos [DUMOUC03], [OWEN03] y para aplicar esta técnica en diferentes áreas.

Discretización. Es el proceso de cuantificar atributos continuos de un conjunto de datos. La aplicación de un proceso de discretización a un conjunto de datos puede extender significantemente los límites de un algoritmo de aprendizaje [HUSSAI02].

El proceso de discretización se puede resumir en cuatro pasos:

1. Ordenar el conjunto de valores de la variable a ser transformada.
2. Evaluar puntos de corte para la separación de intervalos adyacentes.
3. De acuerdo a algún criterio, dividir o unir intervalos de valores continuos.
4. Detener el proceso al llegar a un criterio de paro.

Se han realizado diferentes investigaciones sobre este tema, desarrollándose métodos con enfoque de aprendizaje supervisado y no supervisado (aprendizaje en donde no se conoce el resultado deseado). Algunos de los algoritmos conocidos a la fecha son clasificados en [HUSSAI02].

1.2.1.2 Procesamiento de datos

La minería de datos comprende diferentes tareas y cada una de ellas ha permitido el desarrollo de investigaciones en cuanto a algoritmos de minería de datos que ofrezcan cada vez un mejor desempeño y mejores resultados. Las investigaciones se han orientado, principalmente, a dos líneas de trabajo: a la innovación y optimización de algoritmos y a la instrumentación de técnicas de procesamiento que mejoren el desempeño de los algoritmos.

La línea de investigación que dedica tiempo a la optimización de algoritmos de minería de datos busca nuevos algoritmos o mejorar los ya existentes para disminuir el orden de complejidad de los mismos, de manera que trabajen mejor al enfrentarse

[4] Son medidas estadísticas que describen una población, tal como la media, mediana, curtosis, etc.

con grandes grupos de datos. La otra línea de investigación busca aumentar el desempeño de los algoritmos existentes mediante el uso de herramientas adicionales o métodos novedosos de trabajo computacional, como es el caso de cómputo paralelo, distribuido y en malla, a lo que ha dado por llamarse súper cómputo.

Aquí se presentan únicamente algunos de los avances en cuanto a agrupamiento de datos.

1.2.2 Técnicas de agrupamiento para análisis de grandes bases de datos

"Agrupamiento" se define en [HAN03] como el proceso de agrupar datos en clases o *grupos* de tal forma que los objetos dentro de un grupo tengan un alto grado de similitud entre ellos, pero sean muy distintos de objetos en otros grupos.

Motivados por la importancia de esta tarea de minería de datos se han dedicado muchos esfuerzos para alcanzar un mejor desempeño y tiempo de respuesta de los algoritmos de agrupamiento. Se han realizado muchos trabajos que intentan resolver el problema de agrupamiento sobre grandes bases de datos relacionales y también sobre almacenes de datos [PADMAN03], [YANG99].

En [JAIN99] y en [BERKHI02] se hace un resumen de las diferentes técnicas para agrupamiento y sus algoritmos más conocidos, agrupándolos en:

- Algoritmos de agrupamiento jerárquico
- Algoritmos de segmentación
- Agrupamiento de vecinos más cercanos
- Agrupamiento difuso
- Representación de grupos
- Agrupamiento por redes neuronales artificiales
- Métodos de agrupamiento evolutivo
- Métodos basados en búsquedas.

Cada una de estas categorías presenta una gran variedad de algoritmos. Dos de los algoritmos más populares y aplicados a grandes bases de datos, son el de K-medias y el BIRCH [ZHANG96]. Es por esto que muchos de los esfuerzos han sido encaminados a mejorar la eficiencia de estos algoritmos.

Algunas de las modificaciones o mejoras más importantes realizadas a los algoritmos de agrupamiento incluyen la utilización de conceptos como densidad (como en [PETER03]) y geometría, con el fin de reducir la complejidad de los algoritmos originales. También se han sumado esfuerzos por mejorar las técnicas de divide y vencerás para lograr aplicar los algoritmos sobre grandes conjuntos de datos. Algunos esfuerzos se presentan en [CHENG06] y [JAGADI99].

Otros métodos intentan trabajar únicamente con parte de los datos utilizando técnicas de muestreo [GUHA98].

También se ha incorporado la teoría de algoritmos genéticos para la mejora de algoritmos existentes como: "Genetic K-means algorithm" [KRISHN99] y MSGKA [TSAI02] que utiliza un algoritmo conocido como "multiple-searching genetic algorithm" (MSGA).

Se han diseñado nuevos algoritmos que prometen eficiencia para el tratamiento de grandes volúmenes de datos, como son:

- CLARANS [RAYMON94]
- ACE [PETER03]
- COPLE [YANG99]

En la actualidad hay algunos algoritmos de agrupamiento matemáticamente eficientes que alcanzan una complejidad de O(n); sin embargo, cuando se trata de grandes cantidades de información, aún con esta complejidad se requiere de millones de operaciones para llegar al resultado buscado. Es por esto que otros esfuerzos se han dedicado a realizar adaptaciones de los algoritmos existentes a entornos paralelos y distribuidos, como es el caso de:

- Parallel Unsupervised K-Windows [TASOUL03]: Es una adaptación del algoritmo K-Windows[5].
- Adaptación del algoritmo ISODATA para ser ejecutado en un vector de supercomputadoras [RICCAR98].
- PARCLE [ZHO03]: es una mejora, y adaptación para entornos paralelos de BIRCH.
- K-medias para entornos distribuidos como el [VAIDYA03] que es una versión de K-medias para entornos donde los datos se encuentran segmentados verticalmente sobre un ambiente distribuido.

Estas propuestas plantean el uso de entornos homogéneos o supercomputadoras, lo que representa en sí una dificultad.

Una alternativa más general en esta área es la que se hace en [FORMAN00], en donde se presenta una propuesta de cómo paralelizar algoritmos de a) Agrupamiento iterativos y b) Basados en centros.

Puede, fácilmente, detectarse que existen dos corrientes de investigación: innovación y optimización de algoritmos y adaptación de los algoritmos existentes a entornos de súper cómputo.

Este trabajo se enfoca en buscar una alternativa que no dependa del súper cómputo ni de la generación de nuevos algoritmos para agrupamiento, si no que trabaje con herramientas "simples" y proporcione buenos resultados.

[5] Es una mejora al algoritmo de K-medias usando *ventanas* de datos. Para más información véase [ALEVIZ02]

Capítulo 2
Marco Teórico

En este capítulo se presentan las bases teóricas más importantes en las que se apoyó este trabajo. Se incluye un breve resumen de la teoría de minería de datos, especialmente en lo referente al agrupamiento de elementos. Adicionalmente se definen brevemente técnicas estadísticas que sirven en el análisis del caso de estudio tratado.

2.1 Minería de datos

La minería de datos engloba un conjunto de técnicas provenientes de diferentes áreas (estadística, inteligencia artificial, bases de datos, procesamiento de imágenes, entre otras), que son utilizadas con el objetivo de analizar grandes volúmenes de datos en búsqueda de conocimiento nuevo, insospechado y relevante para el propietario de los datos, presentando dicho conocimiento en formas novedosas, entendibles y de utilidad.

Muchas veces el concepto de minería de datos se usa como sinónimo del término "descubrimiento de conocimiento en bases de datos" (definido en [FAYYAD96] como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos"), abreviado como KDD, por sus siglas en inglés. En otros casos el concepto de minería de datos se refiere únicamente a una fase dentro del proceso de KDD, como se muestra en la figura 2 (tomada de [FAYYAD96]).

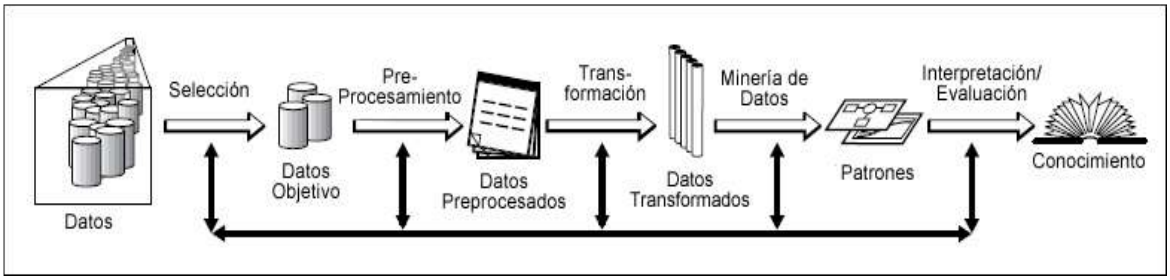


Figura 2: Proceso KDD

Como se observa en la figura 2, en este enfoque la minería de datos solamente se refiere a la extracción de patrones a partir de los datos transformados. Para este estudio tomaremos el concepto de minería de datos como un sinónimo de

"descubrimiento de conocimiento en bases de datos", pues ha sido la forma más generalizada de utilización de este concepto en la actualidad.

2.1.1 Proceso general de minería de datos

El proceso que se sigue para un proyecto de minería de datos puede variar de acuerdo a las características del problema, pero en general el proceso puede dividirse en fases de trabajo. En la actualidad, uno de los procesos de minería de datos ampliamente aceptado fue definido por el proyecto CRISP-DM (por las siglas en inglés para "Cross Industry Standard Process for Data Mining") [4]. Este proceso se esquematiza en la figura 3.

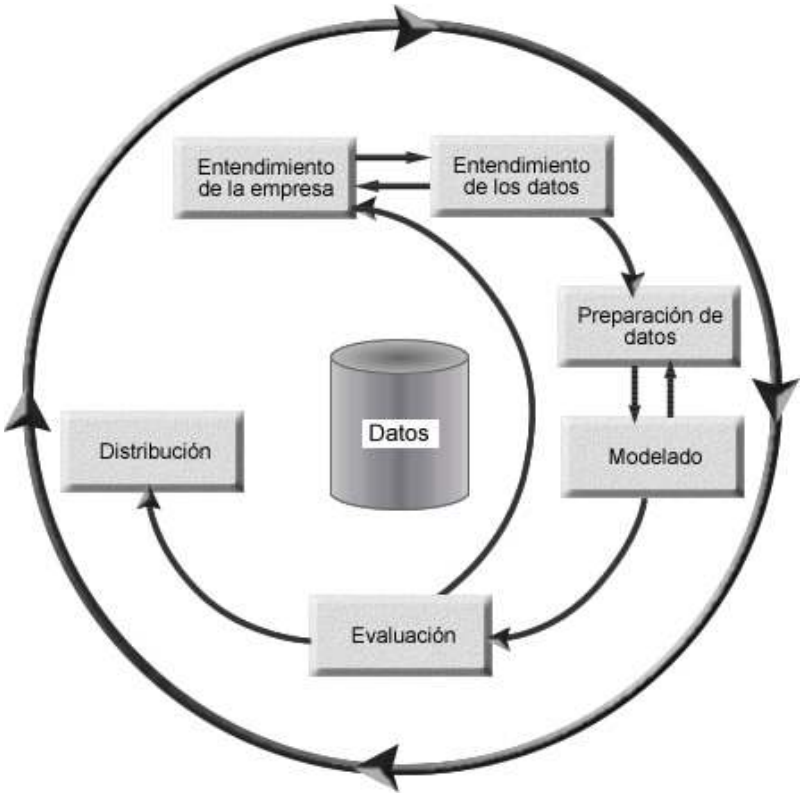


Figura 3: Proceso CRISP-DM

El modelo CRISP-DM define 6 fases cuya secuencia no es rígida. depende de los resultados obtenidos en cada fase. Las flechas de la figura indican las dependencias más importantes y frecuentes entre fases. El círculo externo simboliza la naturaleza cíclica de la minería de datos.

A partir de la documentación del modelo CRISP-DM se presenta una breve descripción de cada fase.

1 *Entendimiento de la empresa*: esta fase se centra en entender los objetivos del proyecto y de la empresa.

2 *Entendimiento de los datos*: esta fase recolecta los datos de estudio e intenta obtener una familiarización con los mismos para identificar problemas de calidad de los datos y, en algunos casos, formular una primera hipótesis de la información oculta.

3 *Preparación de datos*: incluye todas las actividades necesarias para construir el conjunto final de datos que servirá para realizar las tareas de minería de datos (este conjunto procesado de datos también es llamado "vista minable").

4 *Modelado*: varias técnicas de modelado son seleccionadas y aplicadas al conjunto de datos de trabajo para obtener uno o varios modelos que cumplan con los objetivos del proyecto.

5 *Evaluación*: se evalúan los modelos obtenidos y se revisan los pasos seguidos para construir dichos modelos, de manera que se garantice el apropiado cumplimiento de los objetivos del proyecto.

6 *Distribución*: en esta fase el conocimiento obtenido es organizado y presentado en una forma que el cliente puede utilizar. Esta fase puede ser tan simple como generar un reporte o tan compleja como implementar un proceso repetible de minería de datos a través de la empresa.

De las múltiples tareas de modelado que abarca la minería de datos, nos enfocaremos en la tarea de agrupamiento.

2.1.2 Agrupamiento

El análisis de grupos tiene su propia importancia y ha sido de gran valía para muchas áreas como la minería de datos, estadística, procesamiento de imágenes, biología, aprendizaje automatizado, entre otros.

El análisis de grupos, visto como una tarea de minería de datos, puede ser usado como una herramienta para obtener una visión descriptiva de los datos, que permita observar las características de los grupos resultantes y dé paso a futuros análisis.

Existen muchas técnicas para realizar la tarea de agrupamiento de minería de datos, pero todos los algoritmos de formación de grupos realizan su labor valorando las diferencias entre los objetos de estudio, basándose en los valores de los atributos de cada objeto y midiendo la similitud entre éstos a través de una medida de distancia. La medida de distancia más conocida y utilizada es la distancia euclidiana, definida por la siguiente fórmula [JAIN99].

$$d_2(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2}$$

$$d_2(X_i, X_j) = \left\| X_i - X_j \right\|$$

Donde *d* es el número de dimensiones para los objetos X_i y X_j.

La elección de la medida de similitud entre elementos es sólo uno de los pasos del proceso de agrupamiento. Este proceso puede ser resumido en los siguientes pasos [JAIN99]:

1 *Representación de elementos*. Se refiere al número de grupos, número de elementos disponibles y al número, tipo y escala de las características convenientes para el algoritmo de agrupamiento.

2 *Definición de una medida de proximidad entre elementos.* Se define una métrica que sea apropiada para el dominio de los datos.

3 *Agrupamiento*. La salida del agrupamiento puede ser concreta (segmentación de los elementos en grupos) o difusa (cada elemento posee una variable del grado de membresía o pertenencia a cada uno de los grupos resultantes).

4 *Abstracción de datos*. Es el proceso de extraer una representación simple y compacta de los datos. Se realiza sólo si es necesario.

5 *Evaluación de los grupos*. Este paso, al igual que el anterior, se realiza sólo si es necesario. Es un análisis que utiliza un criterio específico de optimalidad; normalmente se trata de un criterio subjetivo. Una estructura de grupos se considera válida si no pudo presentarse como resultado del azar o como una consecuencia de la utilización de un algoritmo de agrupamiento.

La figura 4 presenta un esquema de la secuencia de los primeros tres pasos, descritos anteriormente, para la tarea de agrupamiento.

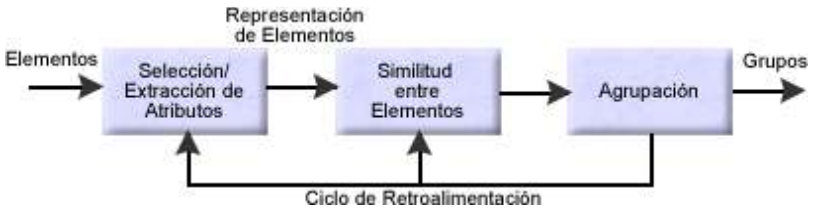


Figura 4: Proceso general de agrupamiento

Existen muchos algoritmos para realizar la tarea de agrupamiento. En este documento se describen únicamente el algoritmo “Fuzzy C-Means” y el de “Mapas Auto Organizados (SOM)” que fueron utilizados en el proyecto. Para mayor información de otros algoritmos de agrupamiento refiérase a [BERKHI02], [JAIN99] y [KEIM99].

2.1.2.1 Fuzzy C-Means

Este algoritmo fue propuesto por Dunn en 1973 en [DUNN73] y mejorado por J. C. Bezdek en 1981 en su libro “Pattern Recognition with Fuzzy Objective Function Algorithms“. Posteriormente, se han hecho algunas adaptaciones y mejoras al algoritmo original.

Se trata de un algoritmo de agrupamiento difuso que, dada una cantidad de centros o grupos para los datos de estudio, aplica lógica difusa para distribuir los elementos en cada grupo.

Lógica difusa y agrupamiento.

La lógica difusa (del inglés fuzzy logic[6]) "permite modelar conocimiento impreciso y cualitativo, así como transmitir, manejar incertidumbre y soportar, en una extensión razonable, el razonamiento humano de una forma natural" [HERNAN04]. La idea central de la lógica difusa es que no provee resultados en un conjunto discreto de valores, si no en un conjunto continuo.

Un conjunto difuso es aquel donde una función de pertenencia o membresía asigna a cada elemento un grado de pertenencia al grupo, dentro del intervalo [0,1] (donde 0 implica no pertenencia y 1 es pertenencia). Esta forma de ver los conjuntos permite solapamiento en las fronteras de los grupos.

Un algoritmo de agrupamiento difuso asocia cada patrón o elemento del conjunto de elementos con todos los grupos usando una función de membresía. La salida de estos algoritmos es un agrupamiento, pero no una partición[7] [JAIN99].

Para conocer más sobre algoritmos de agrupamiento difuso refiérase a [BARALD99-I] y [BARALD99-II].

Algoritmo.

El procedimiento general se enfoca en encontrar los centros de los grupos V_i (i=1,2,...c) y la función de membresía ($\mu_{i,k}$) que define el grado en que cada uno de los n elementos o patrones pertenece a cada grupo. El número de grupos *c* es definido a priori.

El agrupamiento difuso debe cumplir las siguientes ***condiciones generales***:

- La sumatoria total de las membresías de un elemento k es igual a 1.

$$\sum_{i=1}^{c} \mu_{i,k} = 1$$, donde c es el número de grupos evaluados.

- No existen grupos vacíos.

$$\sum_{i=1}^{n} \mu_{i,k} > 0$$, donde n es el número de elementos evaluados.

6 También traducido como "lógica borrosa".
7 Dividir un conjunto de elementos en particiones implica grupos excluyentes.

Una versión simplificada del algoritmo incluye los siguientes pasos [KASABO98]:

1. Inicializar arbitrariamente los grados de membresía $\mu_{i,k}$ para i=1,2,...c y k=1,2,...n de tal manera que las condiciones generales se cumplan.
2. Calcular los valores para los centros de los grupos.

$$V_i = (\sum_{k=1}^{n} (\mu_{i,k})^2 x_k) / (\sum_{k=1}^{n} (\mu_{i,k})^2), \text{ para } i = 1, 2, \ldots, c$$

3. Actualizar el grado de membresía.

$$\mu_{i,k} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{i,k}}{d_{j,k}} \right)^{\frac{2}{m-1}}}, \quad \text{para} \quad d_{j,k} > 0, \ \forall\, i,k$$

Donde,

d_{ik} es la distancia euclidiana del elemento x_k al centro V_i

d_{jk} es la distancia euclidiana del elemento x_k al centro V_j.

m es un valor que determina qué tan difuso será el resultado del agrupamiento. Entre más grande es el valor de m, más difuso es el resultado de agrupamiento.

4. Si los valores calculados para V (los centros de los grupos) no son diferentes de los valores calculados en la iteración anterior (de acuerdo a un parámetro de error), entonces parar.

2.1.2.2 Mapas Auto organizados

Uno de los principales algoritmos del agrupamiento no supervisado es el correspondiente a los llamados Mapas auto organizados (SOM por sus siglas en inglés *Self Organizing Maps*) y fue propuesto por Kohonen en 1981 [KOHONE81].

El principal objetivo de un SOM es convertir la señal de entrada de un espacio de dimensiones arbitrarias a un espacio discreto de *n* dimensiones (típicamente una o dos) y desempeñar esta conversión de manera adaptativa en una forma topológicamente ordenada [HAYKIN99-9]. En ciertos casos a cada elemento del espacio discreto se le denomina una “neurona” y es por ello que los SOMs, frecuentemente, se asocian con las redes neuronales.

Arquitectura.

Los SOMs tienen dos capas: una capa de entrada y una de salida. La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta [WANG03]. Con este método

de aprendizaje la red puede, libremente, organizarse de acuerdo a similitudes en los datos de entrada, lo que conduce a un mapa de salida que representa dichos datos.

Además de la capa de entrada y la capa de salida, el SOM está compuesto por un arreglo con los valores de los pesos asignados a cada uno de los enlaces de una neurona en la capa de entrada a una neurona en la capa de salida. Estos enlaces son modificados a medida que la red es entrenada con los datos de entrada.

Un ejemplo de arquitectura para estas redes es esquematizada en la figura 5.

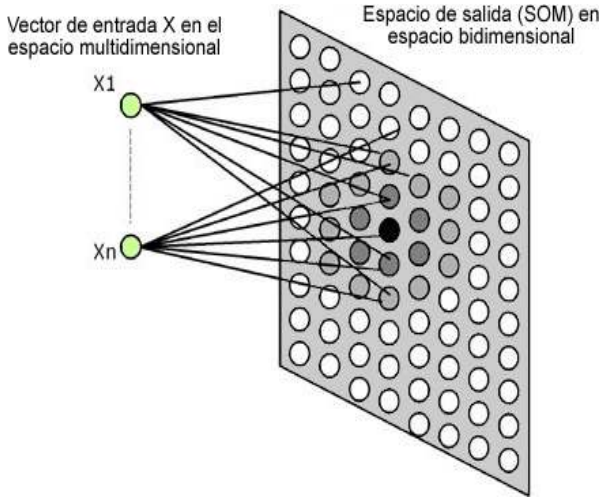


Figura 5: Arquitectura de un SOM

Parámetros.

Según [HAYKIN99-9] los parámetros esenciales del algoritmo son:

- Un espacio (continuo) de entrada de patrones de activación que son generados de acuerdo a una cierta distribución de probabilidad.
- Una topología de red en la forma de una malla de neuronas, la cual define un espacio de salida discreto.
- Una función de vecindad variante en el tiempo que es definida alrededor[8] de una neurona ganadora.
- Un parámetro de aprendizaje que toma un valor inicial y se acerca asintóticamente a cero.

Algoritmo.

El algoritmo puede resumirse en los siguientes pasos:

1. Todas las coordenadas de las neuronas se inicializan con valores aleatorios.
2. Se inicializa un contador de épocas n ← 1. Una época es la presentación de todas las muestras o elementos al SOM.

[8] En el sentido de la métrica elegida.

3. Se define la tasa de aprendizaje $\alpha(n)$ para la época *n*.
4. Se define la función de retroalimentación $r_{ik}(n)$ de la neurona i a la neurona *ganadora k* en la época *n*, dada por:

$$r_{ik}(n) = \exp(-\frac{{d_{ik}}^2}{\sigma(n)^2})$$

   En donde $\sigma(n)$ es la taza de aprendizaje para la época *n* y *d* es la distancia (euclidiana) entre la *i-ésima* neurona y la *k-ésima* neurona (o *neurona ganadora*).

5. Se define el factor de aprendizaje f_α. Este tiene un valor cercano a 1 (por ejemplo *0.99*)
6. Se define el factor radial f_σ. Este tiene, asimismo, un valor cercano a 1 (p.e. *0.995*).
7. Se presenta la muestra *x(t)* (del conjunto de entrenamiento) a la red.
8. Se determina cuál de las neuronas está más cerca de la muestra. Normalmente se calcula la distancia euclidiana entre el vector de pesos de las neuronas y el vector de entrenamiento. A esta neurona (la *k-ésima*) se le denomina la neurona *ganadora*.
9. Los vectores de peso de cada una de las neuronas *perdedoras* se modifican de acuerdo con:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(n) \bullet r_{ik}(n) \bullet (x(t) - w_{ij}(t))$$

10. Si se ha cumplido una época se actualizan los parámetros de aprendizaje, de acuerdo con:

$$\alpha(n+1) = \alpha(n) \bullet f_\alpha$$

$$\sigma(n+1) = \sigma(n) \bullet f_\sigma$$

11. Si se cumple algún criterio de convergencia terminar, si no ir al paso 7.

2.1.3 Determinación del Número de Grupos.

El número de grupos es un parámetro que provee el usuario. Ejecutar un algoritmo varias veces conduce a una secuencia de sistemas de agrupamiento. Cada sistema consiste de mayor granularidad y grupos menos separados. Por ejemplo, para el caso de K-medias, la función objetivo es monótonamente decreciente. Es por esto que determinar el número óptimo de grupos es un problema no trivial [BERKHI02].

Se han realizado muchas investigaciones en torno a la definición de un criterio para encontrar un número óptimo de grupos. Uno de los criterios más populares es el criterio del "codo" (*elbow criterion*) que se basa en el coeficiente de partición y en el coeficiente de entropía de la partición. Estos coeficientes sólo pueden ser calculados cuando se ha utilizado algún tipo de agrupamiento difuso.

El coeficiente de partición (PC) es una medida de qué tan compacto es un grupo. Se calcula utilizando la siguiente fórmula.

$$PC = \sum_{k=1}^{K} \sum_{i=1}^{c} \frac{(\mu_{ik})^2}{K}$$

Donde,

c = Número de centros

K = Número de elementos

$\mu_{i,k}$ = valor de membresía del elemento k al grupo i

El coeficiente de entropía de la partición (PE) mide la "información" contenida en cada agrupamiento. Estrictamente se mide qué tan difusos son los conjuntos. En la teoría estadística de la información cada evento recibe una probabilidad de aparición que está normalizada en el intervalo [0,1]. La información contenida en el evento i se define como $I(i) = -\log_2 P(i)$ y el valor esperado de la información de todos los eventos es $\sum I(i) \bullet P(i)$. A este valor se le denomina la "entropía" del sistema. De manera semejante, cuando se trabaja con conjuntos difusos, el grado de pertenencia al conjunto (también normalizado en [0,1]) toma el lugar de la probabilidad y se define, análogamente, la "entropía" del sistema difuso. Es este valor el que mide PE.

PE aumenta al aumentar el número de grupos, puesto que cada grupo se va disgregando en nuevos grupos. Este valor se calcula utilizando la siguiente fórmula.

$$PE = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{c} \mu_{ik} \ln(\mu_{ik})$$

El criterio del codo estipula que el mejor número de grupos (c) corresponde al punto en donde, simultáneamente, la tendencia de PE a incrementar y la tendencia de PC a decrecer cambian; en otras palabras, cuando la curvatura de las tendencias cambia, tal como lo muestra el gráfico de la figura 6. En él se indica, con un ovalo punteado, el punto donde se da el cambio en las tendencias de PC y PE y que por lo tanto es un número óptimo de grupos.

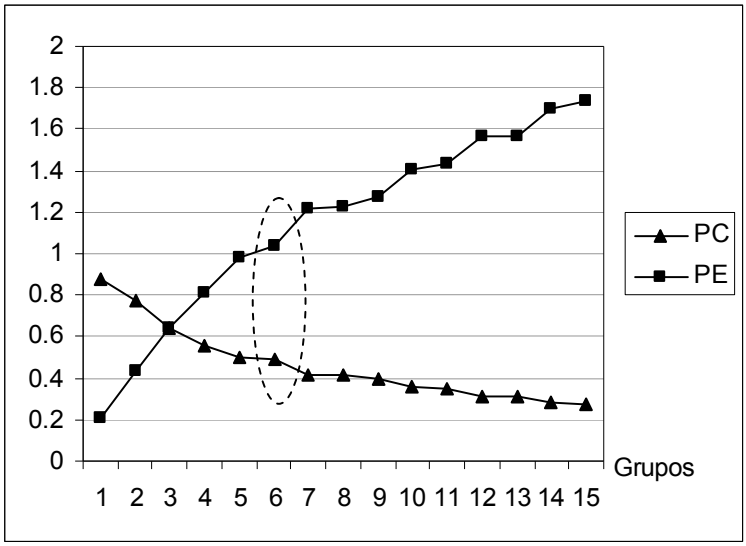


Figura 6: Criterio del codo

2.2 Análisis de datos

El análisis de datos consiste en transformar datos con el objetivo de extraer información útil [6]. Dependiendo del tipo de datos y el problema a resolver, puede incluir la aplicación de métodos estadísticos, ajuste de curvas, selección de subconjuntos basados en criterios específicos, entre otras técnicas. A diferencia de la minería de datos, el análisis de datos presta menor interés al descubrimiento de patrones ocultos e inesperados en los datos, que a la verificación o desaprobación de un modelo existente, o a la extracción de parámetros necesarios para la adaptación de un modelo (experimental) a la realidad.

La base del análisis de datos son la estadística y la metodología estadística, las cuales tratan dos tipos de problemas [7]:

- Resumir, describir y explorar los datos.
- Usar una muestra de datos para inferir la naturaleza del sistema que produce los datos.

Las secciones siguientes tratan sobre técnicas para análisis de datos que fueron utilizadas en el desarrollo del proyecto. Tal es el caso de, muestreo, selección de características y ajuste de curvas.

2.2.1 Muestreo

El muestreo es un método estadístico para seleccionar un cierto número de elementos a partir de una población, para ser incluidos en un conjunto de datos denominado “muestra”. Existen dos tipos principales de muestreo:

1. Muestreo probabilístico: los métodos de muestreo que se agrupan en esta clasificación se basan en el principio de equiprobabilidad, que significa que todos los miembros de la población poseen la misma probabilidad de formar parte de una muestra.
2. Muestreo no probabilístico o intencional: los métodos de muestreo que se agrupan bajo esta clasificación no consideran iguales a todos los elementos de la población y realizan la extracción de las muestras basándose en criterios predeterminados por el investigador.

Cuando no se sabe nada acerca de los datos, podemos suponer que cada elemento posee la misma importancia relativa dentro del conjunto y por lo tanto resultan más apropiados los métodos de muestreo probabilístico.

2.2.1.1 Métodos de muestreo probabilístico.

Existen cuatro principales métodos probabilísticos de muestreo, que se definen brevemente a continuación [HAND01-4].

Muestreo aleatorio simple. Es la forma más básica de muestreo. En este método, los n registros que componen la muestra son seleccionados de los N registros de la base de datos en tal forma que cada conjunto de n registros tiene una probabilidad igual de ser seleccionado.

Muestreo aleatorio sistemático. Esta forma de muestreo selecciona los elementos de la muestra en una forma ordenada. La forma de la selección depende del número de registros de la base de datos y el número de registros que se requiere para la muestra.

Muestreo aleatorio estratificado. En este método la población es dividida en subpoblaciones o estratos no sobrepuestos y una submuestra es seleccionada dentro de cada estrato para conformar la muestra total.

Muestreo aleatorio por conglomerados. En este método primero se divide la población en grupos convenientes para el muestreo. En seguida, se selecciona una porción de los grupos por un método sistemático o al azar. Finalmente, dentro de cada uno de los grupos seleccionados se toman todos o parte de sus elementos para conformar la muestra. Bajo este método, aunque no todos los grupos son muestreados, cada grupo tiene una igual probabilidad de ser seleccionado. Por lo tanto la muestra es aleatoria.

2.2.1.2 Determinación del tamaño de la muestra

Debe elegirse un tamaño de muestra tal que éste garantice que los datos obtenidos sean representativos de la población.

Lo primero que debe tomarse en cuenta para determinar el tamaño de la muestra son las características de la población. Si es posible asumir que la población a tratar posee un comportamiento *normal*, es decir, que el atributo a medir se comporta como una distribución normal, entonces puede utilizarse alguna fórmula estadística basa en las propiedades de la distribución normal para el cálculo del tamaño de la muestra.

Para ello deben tenerse en cuenta los siguientes criterios:

- El nivel de precisión, también llamado error de muestreo, es el rango en el cual se estima que se encuentra el valor verdadero de la población. Mientras menor sea el error aceptable, mayor deberá ser la muestra.
- El nivel de confianza, denotado como (1-α), es el grado de confianza o seguridad que se tendrá de que el verdadero valor del parámetro estimado en la población se sitúe en el intervalo de confianza[9] obtenido. Refleja la probabilidad de confianza de no rechazar una hipótesis nula verdadera.

La fórmula para calcular el tamaño de la muestra depende del tipo de análisis que se desee realizar, el tamaño de la población, tipo de muestreo a realizar, características del parámetro a estimar, entre otros. Así por ejemplo, para estimar la media de una población se puede usar la siguiente fórmula [CHOW03].

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{E^2}$$

Donde,

n = tamaño de la muestra.

$Z_{\alpha/2}$ = es el (α/2) cuantil de la distribución normal estándar. Valor que es tomado a partir de las tablas definidas para la distribución normal y que representa el área bajo la curva.

E = máximo error aceptable (precisión).

σ^2 = varianza de la población.

Para poder utilizar esta fórmula es necesario conocer la varianza de la población.

Otro ejemplo es el tratamiento de proporciones, en donde el atributo a estimar en una población grande se asume que posee una probabilidad de aparición. Dicha probabilidad de aparición se incluye en la fórmula para el cálculo de la muestra, tal como se presenta en al siguiente fórmula [5].

[9] Intervalo de confianza: intervalo de valores alrededor de un parámetro muestral en donde se espera que se sitúe el parámetro poblacional a estimar.

$$n = \frac{Z^2_{\alpha/2} pq}{E^2}$$

Donde,

n = tamaño de la muestra

$Z_{\alpha/2}$ = es el (α/2) cuantil de la distribución normal estándar.

p = proporción estimada del atributo que está presente en la población.

q = 1 - p

E = máximo error aceptable (precisión).

Una alternativa para tratar con una población que no puede ser asumida como una distribución normal, o de la cual no se conoce mucho, se encuentra en el uso de *validación cruzada.*

Validación cruzada

Es una herramienta estadística cuya motivación principal es validar un modelo sobre un conjunto de datos diferente al usado para la estimación de parámetros [HAYKIN99-4]. Consiste en dividir un conjunto de datos en dos subconjuntos (muestras), de manera que el análisis se realiza inicialmente sobre una muestra (conjunto de entrenamiento), de donde se obtiene un modelo que posteriormente es validado contra los datos restantes (conjunto de prueba o validación).

El tamaño de la muestra, en la validación cruzada, depende del método que se esté utilizando. Existen varias formas de validación cruzada, las más comunes son:

1. *Validación por retención* (holdout validation). Los elementos son seleccionados aleatoriamente para la muestra inicial para la validación de datos; los elementos restantes son conservados para datos de entrenamiento. Normalmente el 80% de los datos se destina para entrenamiento y el restante 20% para validación.
2. *Validación cruzada k-partes*. (k-fold cross validation). El conjunto original es dividido en k muestras. De estas muestras una es retenida como datos de validación para probar el modelo, y las restantes k-1 muestras son usadas como datos de entrenamiento. El proceso de validación cruzada se repite k veces, con cada una de las k muestras de manera que sean usadas exactamente una vez como datos de validación. En este método, cada elemento pasa sólo una vez a formar parte del conjunto de prueba y al conjunto de entrenamiento k-1 veces.
3. *Validación cruzada dejar-uno-fuera* (leave-one-out cross validation). Implica utilizar un único elemento del conjunto como dato de validación y los restantes elementos como datos de entrenamiento. Este proceso es repetido por cada elemento del conjunto de datos. Lógicamente esta es la validación más costosa.

2.2.2 Selección de características relevantes.

Este proceso no es igual que el de "limpieza de datos", más bien es un proceso posterior donde se hace una evaluación de los atributos de los datos de estudio para determinar aquellos que pasaran a formar parte de la vista minable. Se deben realizar, entonces, tareas como: eliminación de claves candidatas, eliminación de atributos dependientes y reducción de la dimensionalidad.

2.2.2.1 Análisis correlacional.

El análisis correlacional consiste en crear una matriz de correlaciones que presenta la relación entre pares de variables. Esta matriz se calcula utilizando el coeficiente de correlación.

Un coeficiente de correlación (generalmente identificado con "*r*") cuantifica la relación lineal entre las variables [MYATT07]. Su valor está en el rango de -1.0 a +1.0. Valores positivos indican una correlación positiva (el aumento en una variable significa aumento en la otra variable) y valores negativos indican una correlación negativa (el aumento en una variable indica decremento en la otra variable), un valor de cero indica que no hay ningún tipo de correlación entre las variables. Las figura 7, 8 y 9 muestran la correlación positiva, negativa y pobre, respectivamente.

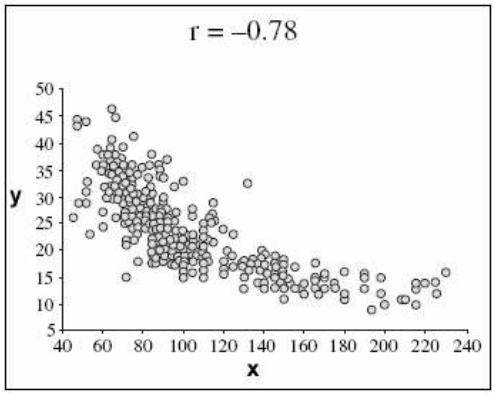


Figura 7: Correlación Negativa

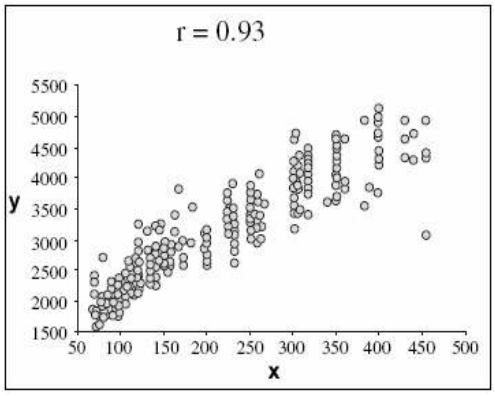


Figura 8: Correlación Positiva

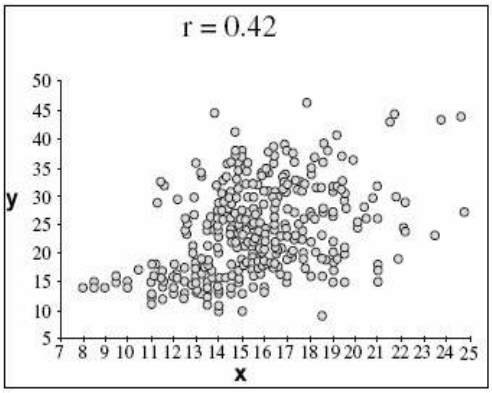


Figura 9: Correlación pobre

La fórmula para calcular el coeficiente de correlación es la siguiente [DAGOST86-5]:

$$r = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{(n-1)S_x S_y}$$

Dónde,

x, y: variables de estudio.

n: número de elementos.

$\bar{x}, \bar{y}$: promedio para las variables x e y.

S_x, S_y: error estándar para las variables x e y.

2.2.3 Prueba de bondad de ajuste.

En el análisis de datos, buscar un modelo matemático que pueda representar los datos en estudio es una tarea común. Para verificar si el conjunto de datos se puede ajustar, o afirmar que proviene del modelo dado, se utilizan pruebas estadísticas llamadas pruebas de bondad de ajuste.

En algunos casos es posible asumir que los datos corresponden a una función o distribución estadística conocida, si es así, la prueba de bondad de ajuste tiene por objetivo la aceptación o el rechazo del modelo. Existen varias pruebas estadísticas que se clasifican dentro de esta categoría. La **prueba Chi-cuadrada** es una de las pruebas más conocidas y aplicadas.

Encontrar el modelo que mejor represente un conjunto de datos no es un problema fácil, ya que existen innumerables funciones y distribuciones matemáticas conocidas que deberían ser probadas para encontrar la indicada. A esto se suma que el origen de los datos pueda estar asociado a un fenómeno para el cual no exista una función matemática conocida.

En este caso se recurre a una búsqueda del modelo matemático que mejor se ajuste procurando encontrar una buena aproximación y no el modelo exacto. Esta búsqueda se auxilia de pruebas de bondad de ajuste que son una medida de que tan acertado es un modelo dado para los datos en estudio. Un ejemplo de estas medidas son el **error estándar y el coeficiente de correlación**.

2.2.3.1 Chi-cuadrada

Esta prueba se basa en la hipótesis nula (H_o) de que no hay diferencias significativa entre la distribución muestral (distribución obtenida a partir de la muestra) y la teórica. La hipótesis alternativa siempre se enuncia como que los datos no siguen la distribución supuesta [8].

La prueba Chi-cuadrada (x^2) compara las frecuencias observadas con las frecuencias esperadas, calculadas de acuerdo a la hipótesis nula formulada. El valor de esta prueba se obtiene aplicando la siguiente fórmula [DAGOST86-3].

$$x^2 = \frac{\sum_{i=1}^{g} (O_i - E_i^2)}{E_i}$$

Dónde,

g : número de intervalos de clases en que se dividen las observaciones.

O_i : frecuencia observada en el intervalo de clase i.

E_i: frecuencia esperada en el intervalo de clase i.

Se acepta H_o cuando $x^2 < x_t^2 (r-1)(g-1)$. En caso contrario se rechaza.

Dónde,

La segunda parte de la inecuación se interpreta como el valor de chi-cuadrado en la tabla de valores de chi-cuadrado para la fila (r-1) y la columna (g-1).

t: representa el valor proporcionado por las tablas de chi-cuadrada, según el nivel de significancia elegido.

r: número de clasificaciones de los datos.

k: número de parámetros estimados a partir de los datos muestrales para obtener los valores esperados.

Cuanto más se aproxima a cero el valor de chi-cuadrada, más ajustados están los datos a la distribución dada.

2.2.3.2 Error estándar y coeficiente de correlación

Existen varios métodos para encontrar el modelo matemático que mejor se ajuste a un conjunto de datos. La regresión estadística es un ejemplo de ellos.

La regresión es utilizada para simular la relación existente entre dos o más variables. Cuando el problema implica una única variable independiente, la técnica estadística se denomina regresión simple. Cuando implica dos o más variables independientes, se denomina regresión múltiple [HAIR05]. Para este proyecto se utiliza únicamente regresión simple.

Ajustar los datos a un modelo dado utilizando regresión significa asegurar que una "función calidad", que es una función arbitraria que mide la divergencia entre los datos y el modelo, es minimizada. En este sentido, los parámetros del modelo son ajustados hasta que la función calidad llegue al valor más pequeño posible.

Buscar el mejor modelo de ajuste significa realizar el proceso de regresión por cada uno de los modelos en estudio y comparar luego los resultados. Para comparar el ajuste de los datos al modelo de regresión se puede utilizar el **error estándar** y el **coeficiente de correlación**.

El error estándar cuantifica la dispersión de los datos alrededor de la curva evaluada. A medida que la calidad del modelo incrementa, el error estándar se aproxima a cero. Este valor es calculado con la siguiente fórmula:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n - p}}$$

Dónde,

y_i : valor del dato evaluado.

n : número de puntos evaluados.

p : número de parámetros del modelo o función.

El coeficiente de correlación también es fundamental para un análisis de bondad de ajuste [DAGOST86-5].

A diferencia del coeficiente de correlación presentado en la sección 2.2.2.1, la prueba de correlación para verificar la bondad de ajuste se realiza entre el valor real de la variable dependiente (y_i) y el valor proporcionado por el modelo evaluado ($f(x_i)$) determinando que tan aproximados son los valores.

Mientras mejor sea el ajuste de los datos al modelo evaluado, el coeficiente de correlación tendrá un valor más cercano a 1.

Capítulo 3 Metodología

En este capítulo se define la metodología para agrupamiento diseñada y aplicada en el caso de estudio tratado en este proyecto. Además de definirse las fases que constituyen la metodología, también se incluye un apartado que muestra la aplicabilidad de dicha metodología.

3.1 Definición general de la metodología.

Según la metodología definida por CRISP-DM (presentada en la sección 2.1.1), existen seis fases principales para llevar a cabo un proyecto de minería de datos. De estas fases, la fase de modelado constituye el proceso principal para el descubrimiento de nuevo conocimiento a partir de los datos disponibles. Dicha fase la subdividimos en tres subfases:

1. Reducción del espacio de búsqueda
2. Determinación del número de grupos
3. Aplicación de algoritmo de agrupamiento

En las secciones siguientes se describe cada una de estas subfases.

3.2 Reducción del espacio de búsqueda

Este paso consiste en obtener una muestra que represente apropiadamente el conjunto original de datos [KURI07].

Para poder extraer un subconjunto representativo del conjunto original se siguen los siguientes pasos:

1. Realizar una reducción vertical en los datos.
2. Realizar una reducción horizontal sobre los datos.
3. Validar la representatividad

3.2.1 Reducción vertical.

Este proceso consiste en reducir la cantidad de variables a tratar.

Es llamado también selección de características o reducción de la dimensionalidad. Existen muchos métodos para llevar a cabo esta tarea (véase referencias en la sección 1.2.1.1 y en la sección 2.2.2), uno de ellos es la eliminación de variables altamente correlacionadas.

El hecho que una variable esté altamente correlacionada es función de criterios casuísticos. Para el caso de estudio, basándose en la experiencia, se consideró una correlación superior o igual a 0.75 (valor absoluto) como una correlación alta.

Una vez que la matriz de correlaciones (ver sección 2.2.2.1) es calculada, las correlaciones que presenta son revisadas para identificar conjuntos de variables altamente correlacionadas. De cada conjunto se elige una variable[10] y las demás son descartadas.

3.2.2 Reducción horizontal.

La reducción horizontal se refiere a la selección de filas de una base de datos, las cuales serán utilizadas como fuente de datos para el algoritmo de minería de datos. Es decir, significa extraer una muestra del conjunto original de datos.

La determinación del tamaño de la muestra depende de las características de los datos (sección 2.2.1.2). Si es posible asumir que la muestra se comporta de manera *normal*, entonces el cálculo del tamaño de la muestra se determina aplicando una fórmula basada en la distribución normal como las presentadas en la sección 2.2.1.2.

Si no es posible asumir un comportamiento normal (como sucede para la mayoría de bases de datos), entonces puede calcularse el tamaño de la muestra como si se tratara de la realización de una validación cruzada (sección 2.2.1.2).

La idea central de la validación cruzada es permitir la evaluación de un modelo obtenido a partir de un conjunto de datos (de entrenamiento) contra datos desconocidos (de prueba). Esto se logra al dividir el conjunto original en m conjuntos de entrenamiento que serán validados contra m conjuntos de prueba.

Como en este caso no se trata de realizar un entrenamiento, si no una validación de las mismas muestras, tomaremos la validación cruzada en k-partes (k-fold cross validation) de tal manera que cada parte o muestra sirva de conjunto de prueba.

Las muestras a tomar serán excluyentes y el tamaño de cada una será igual al 20% del conjunto original de datos (de acuerdo a lo reportado en [KEARNS96] tomar un r=0.2 es una opción adecuada), es decir k=5. Trabajar con muestras excluyentes permite que se analice el conjunto completo de datos sin que existan solapamientos en las muestras.

[10] En principio todas las variables tienen la misma significación estadística y cualquiera de ellas puede ser elegida como representativa del conjunto.

Cada muestra debe ser extraída haciendo uso de muestreo aleatorio, para asegurar que todos los elementos son considerados de manera equitativa y que por lo tanto las muestras extraídas también poseen pesos iguales.

3.2.3 Validación de la representatividad

Para asegurar que la muestra que será analizada es realmente representativa del conjunto original de datos, es necesario realizar pruebas que indiquen si el comportamiento de las variables que conforman cada elemento es estadísticamente equivalente en las diferentes muestras seleccionadas. Estas pruebas son pruebas de bondad de ajuste (véase sección 2.2.3).

En este paso, todas las variables altamente correlacionadas han sido descartadas, por lo que ya no existen relaciones lineales entre las variables. Se ha descartado entonces todas las relaciones poco complejas entre variables.

Averiguar el modelo matemático que mejor describe la relación entre las variables de estudio no es un problema trivial. El método usado aquí, es el análisis correlacional de alto grado y resulta más adecuado cuando no se conoce el tipo de curva que mejor ajusta al conjunto de datos. Para facilitar el análisis usamos regresión simple (véase sección 2.2.3.2).

3.2.3.1 Modelos a Evaluar

Encontrar el mejor modelo, por medio de regresiones, presupone evaluar una serie de modelos matemáticos para determinar cual es el que mejor se ajusta a los datos. En este estudio se evaluaron 34 modelos matemáticos considerados de los más comúnmente utilizados en aplicaciones del mundo real. Dichos modelos están divididos en familias, de acuerdo a sus características de comportamiento.

En esta sección se incluye una breve descripción de cada familia de modelos. En el anexo 1 se pueden consultar los modelos de cada familia y sus fórmulas matemáticas.

1. ***Familia exponencial***: los modelos exponenciales involucran funciones logarítmicas o exponenciales. Generalmente son curvas cóncavas o convexas, pero algunos de estos modelos pueden incluir un punto de inflexión y un máximo o mínimo.
2. ***Familia de potencias***: involucra la elevación de uno o más parámetros a la potencia de la variable independiente, o elevar la variable dependiente a la potencia de un parámetro dado. Esta familia es generalmente un conjunto de curvas convexas o cóncavas sin puntos de inflexión o máximos o mínimos.
3. ***Modelos basados en densidad***: esencialmente presenta dos tipos de respuesta: una relación “asintótica” y una “parabólica”. Si el comportamiento es tal que x incrementa, pero y se aproxima a un valor fijo, la relación mostrada es asintótica. Si el comportamiento es tal que hay un óptimo distinto al incrementar x, la relación es parabólica.

4. ***Modelos crecientes***: los modelos de esta familia se caracterizan por un crecimiento monótono desde algún valor fijo hacia una asíntota.
5. ***Familia Sigmoidal***: poseen un crecimiento con forma de S. Estas curvas inician en un punto fijo e incrementa su tasa de crecimiento monótonamente hasta alcanzar un punto de inflexión. Después de esto la taza de crecimiento se aproxima a un valor final asintóticamente. Esta familia es un subconjunto de la familia de modelos crecientes.
6. ***Otros***: además de las familias antes mencionadas existen otro modelos a ser evaluados que no se acomodan a las familias presentadas.

Además de estas familias de modelos se incluyen modelos lineales para su evaluación.

3.2.3.2 Algoritmo de validación

El proceso para validación de la representatividad de las muestras se resume en el siguiente algoritmo [KURI07]:

1. Seleccionar conjuntos de **"y"** variables para realizar las pruebas de bondad de ajuste (para regresión simple y=2). Como cada variable posee la misma importancia dentro del conjunto de variables a analizar, el orden en que estas sean tomadas no importa.
2. Realizar una búsqueda de la mejor función regresiva para el conjunto de variables seleccionadas, por cada una de las muestras. Se selecciona como mejor ajuste la función regresiva que presenta el coeficiente de correlación de regresión con el valor más cercano a 1.
3. Ejecutar el paso 1 y 2 hasta que no existan más variables que evaluar.
4. Verificar las similitudes estadísticas entre las funciones resultantes para cada conjunto de variable y muestra. Si cada conjunto de variables presenta comportamiento similar en todas las muestras entonces se toma como representativa cualquiera de las muestras evaluadas.

La función resultante a la que se ajusten mejor los conjuntos de variables no debe ser necesariamente la misma ni debe poseer los mismos valores de parámetros. Es el coeficiente de correlación el que determina qué tan bien se ha realizado el ajuste a un modelo y por lo tanto qué tan aceptable es dicho modelo.

Las combinaciones posibles de parejas de variables a analizar esta dada por:

$$L = \frac{Y(Y-1)}{2}$$

Donde,

L: número de parejas posibles.

Y: número total de variables.

Como el orden de las variables en las parejas no es importante, se descarta la mitad de las combinaciones.

Para asegurar que todas las variables pasen por un análisis de regresión sin repetir variables, sería necesario tomar aproximadamente Y/2 parejas.

3.2.3.3 Criterio de Paro

Evaluar todas las posibles parejas de variables significa dedicar muchos esfuerzos para este propósito. Para evitar evaluar todas las posibles parejas definimos un criterio de paro: el proceso puede detenerse cuando se alcance un nivel de confianza aceptable.

El nivel de confianza representa la probabilidad de que una validación exitosa sea el resultado de la casualidad. Un nivel de confianza bajo indica una validación exitosa de la muestra y un nivel de confianza alto indica que las pruebas realizadas no permiten asegurar una muestra válida.

Para poder realizar el cálculo de este valor de confianza debemos tener presentes las siguientes definiciones:

- Código de modelo: cada uno de los modelos evaluados es codificado con un número entero (en este caso del 1 al 34 por ser 34 los modelos a evaluar).
- Ajuste: llamaremos ajuste al conjunto de **m** códigos de modelos obtenidos como resultado del proceso de regresión efectuado para una pareja de variables en las **m** muestras sujetas al análisis.
- Maximalidad: número mayor de modelos iguales en un ajuste. Por ejemplo, un ajuste de {2, 5, 6, 5, 5} posee una maximalidad de 3.
- Ajustes similares: ajustes con la misma maximalidad.

Determinamos el nivel de confianza con la siguiente fórmula.

$$C = \prod_{i=1}^{k} P_i$$

Dónde,

C: valor de confianza

k: cantidad de parejas de variables a analizar

P_i: Probabilidad de aparición del ajuste i. Consideraremos esta probabilidad independiente de la probabilidad de aparición de los otros ajustes evaluados.

Definimos la probabilidad de aparición de un ajuste como la probabilidad de aparición de ajustes similares. Este valor se calcula dividiendo el número de ajustes similares entre el número total de posibles ajustes. Así,

$$P_i = \frac{\#\text{ ajustes similares}}{\#\text{ posibles ajustes}}$$

Un valor de P_i muy pequeño se interpreta como que el ajuste obtenido posee una baja probabilidad de aparición y por lo tanto, no es un resultado obtenido por casualidad.

El número de ajustes similares y de posibles ajustes se determina usando la teoría de combinaciones y combinaciones con repetición[11] (CR).

Ajustes similares.

Para determinar el número de ajustes similares, en donde la maximalidad es ***t,*** usamos la siguiente fórmula:

$$\#\,Ajustes\ Similares = \begin{cases} q & \text{para } t = m \\ \binom{q}{m} & \text{para } t = 1 \\ q\left[\sum_{i=\lceil h/t \rceil}^{h} v\binom{q-1}{i}\right] & \text{, en otro caso} \end{cases}$$

Dónde,

q = número de modelos a evaluar.

h = m-t

v = cantidad de combinaciones de i números menores o iguales a t que suman h.

Los tres casos que son tratados en esta fórmula se explican así,

- El primer caso de la fórmula (*t=m*), implica que el mismo modelo resultó adecuado para todas las muestras evaluadas. Entonces, el número de ajustes similares está dado por la cantidad de modelos a evaluar (en nuestro caso *q=34*).

[11] Una combinación es un arreglo de elementos en donde no interesa el orden de los elementos. En una combinación con repetición pueden existir elementos repetidos [9].

- El segundo caso de la fórmula (*t=1*), implica que todos los modelos son diferentes entre si. Este cálculo se realiza por medio de la fórmula para combinaciones para *q* modelos en m muestras.
- El tercer caso se resuelve partiendo del hecho que el modelo que más se repite (el que da el valor de la maximalidad al ajuste) está acompañado de **x** modelos diferentes. Se efectúa un cálculo de combinaciones en donde cada combinación representa un número de modelos diferentes que acompañan al modelo de la maximalidad.

Posibles ajustes.

La cantidad total de posibles ajustes está dada por la cantidad de combinaciones con repetición posibles para ***q*** modelos en ***m*** muestras. Entonces,

$$\#\,Posibles\ \ Ajustes = CR_q^m = \binom{q+m-1}{m} = \frac{(q+m-1)!}{(q-1)!m!}$$

Donde,

q: cantidad de distribuciones o modelos a probar.

m: cantidad de muestras a analizar.

Ejemplos.

Con 34 modelos a probar en 5 muestras, la probabilidad de obtener un ajuste con maximalidad *t=5* es

$$P_i = \frac{34}{501942} = 6.77369E\text{ - }5$$

Para *t = 2*

$$\#\,Ajustes\ Similares = 34\left[\sum_{i=\lceil 3/2\rceil}^{3} v\binom{33}{i}\right] = 34\left[1\binom{33}{2} + 1\binom{33}{3}\right] = 203456$$

$$P_i = \frac{203456}{501942} = 0.405337$$

Como es lógico, la probabilidad de que existan 2 modelos iguales es mayor a la probabilidad de que todos los modelos sean iguales.

En la tabla 1 se muestran todos los valores de ajustes similares y de probabilidad con *m=5 y q=34*.

t	Ajustes Similares	Probabilidad
1	278256	0.55435887
2	203456	0.40533767
3	19074	0.03800041
4	1122	0.00223532
5	34	6.7737E-05
Total	501942	1

Tabla 1: Probabilidad de aparición de ajustes.

En el anexo 2 se incluye el código (en java 1.5) de un programa que calcula los valores de confianza a partir de un *m* y q dados.

3.3 Determinación del número de grupos

Luego de haber reducido el espacio de búsqueda, se procede a definir el número de grupos en los que se espera se divida el conjunto de datos. Este número de grupos puede ser definido de acuerdo a criterios como experiencias anteriores, requerimientos del proyecto, documentación existente, etc. Si no existen estos parámetros, se puede utilizar el proceso de determinación de grupos presentado en la sección 2.1.3, calculando el número de grupos por medio del criterio del codo.

3.4 Aplicación de algoritmo de agrupamiento.

Luego de determinar el número de grupos para el conjunto de datos se aplica un algoritmo de agrupamiento sobre la muestra de datos obtenida. Para el caso de estudio se utilizó un mapa auto organizado.

3.5 Aplicabilidad de la metodología

Antes de aplicar esta metodología es necesario estimar si el costo de su aplicación no será mayor al costo de resolver el problema con otra metodología. Para determinar esto se sugieren los siguientes pasos:

- Determinar el costo del proceso de agrupamiento en términos de acceso a datos.
- Determinar el tamaño de la muestra

- Determinar el número máximo de ajustes a probar en el proceso de validación de la muestra.
- Determinar el número de modelos a probar en el proceso de regresión.
- Determinar el número de variables a probar para la regresión.

Para mostrar la forma en que se comporta el costo de la metodología supondremos a) un tamaño de muestra del 20% de los datos, b) 34 modelos a evaluar, c) 5 ajustes de prueba y d) un proceso de agrupamiento que requiere al menos 100 accesos a cada dato del espacio de búsqueda.

Además supondremos que la reducción vertical siempre es efectuada. De tal manera que las estimaciones se basan en el costo producido por la reducción horizontal y la validación de la muestra.

El costo producido por la metodología es determinado a partir de la siguiente fórmula:

$$Costo\,\mathrm{Re}\,ducido = y * n * q * a + z * CostoOriginal$$

Donde,

y = número de variables a evaluar en cada proceso de regresión.

n = número total de tuplas del problema.

q = número de modelos a evaluar.

a = número de ajustes a probar.

z = tamaño de la muestra (porcentaje)

CostoOriginal = costo del proceso de agrupamiento cuando no se hace la reducción horizontal. Es igual al número de accesos que necesita el método por el número total de datos (total de tuplas * total de variables del problema).

La primera parte de la fórmula representa el costo de la validación y la segunda el costo de agrupamiento sobre la muestra.

En la figura 10 se muestra cómo varía la tasa de reducción de costo del agrupamiento de la metodología según el número de ajustes. A medida que el número de ajustes es mayor, la tasa de reducción de costos que proporciona la metodología disminuye hasta que su utilización no proporcione una reducción si no un aumento de costos

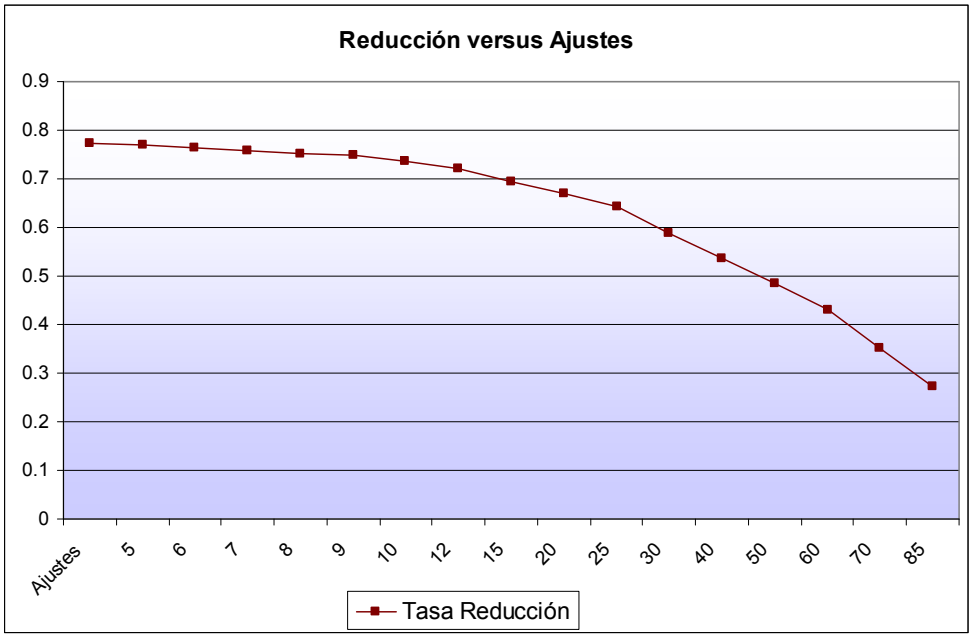


Figura 10: Gráfico tasa de reducción versus número de ajustes

La figura 11 muestra cómo, para el mismo ambiente, la cantidad de modelos a evaluar afecta la tasa de reducción. En este caso la tasa de reducción no es muy afectada por el aumento en el número de modelos a evaluar.

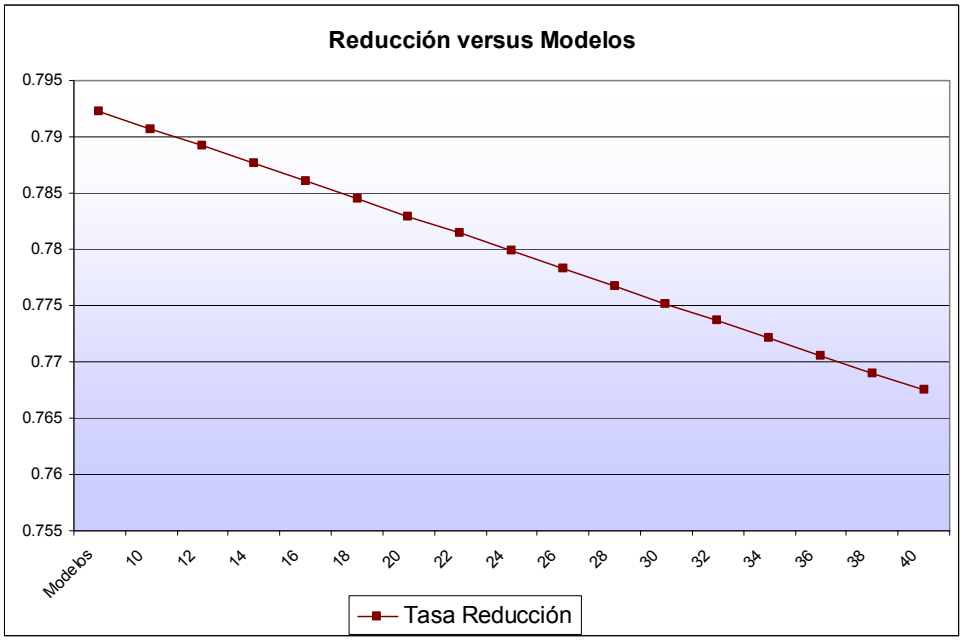


Figura 11: Gráfico tasa de reducción versus número de modelos

La aplicabilidad de la metodología depende de que tan costoso sea el proceso de agrupamiento a efectuar. La figura 12 muestra la relación entre el costo de validación y el costo original de agrupamiento y la relación entre la tasa de reducción y el costo original. Esta gráfica fue determinada considerando 5 muestras (20%) evaluadas con 5 ajustes y 34 modelos para los casos en que el método de agrupamiento variara de 1 a 150 accesos por cada dato.

La tasa de reducción muestra que para que exista una verdadera reducción de costo el método de agrupamiento debería requerir más de 3 accesos por cada dato. A medida que la cantidad de accesos aumenta la tasa de reducción se acerca más a 0.8 (valor de la reducción horizontal al descartar el 80% de los datos).

En el caso del costo de validación relativo, éste fue extraído dividiendo el costo de validación de la muestra entre el costo original del agrupamiento. Puede observarse en la gráfica que este costo se vuelve cada vez más pequeño a medida que la complejidad del método de agrupamiento aumenta.

Para nuestro supuesto de un método de agrupamiento superior a 100 accesos por dato, la aplicación de la metodología produce una reducción alta.

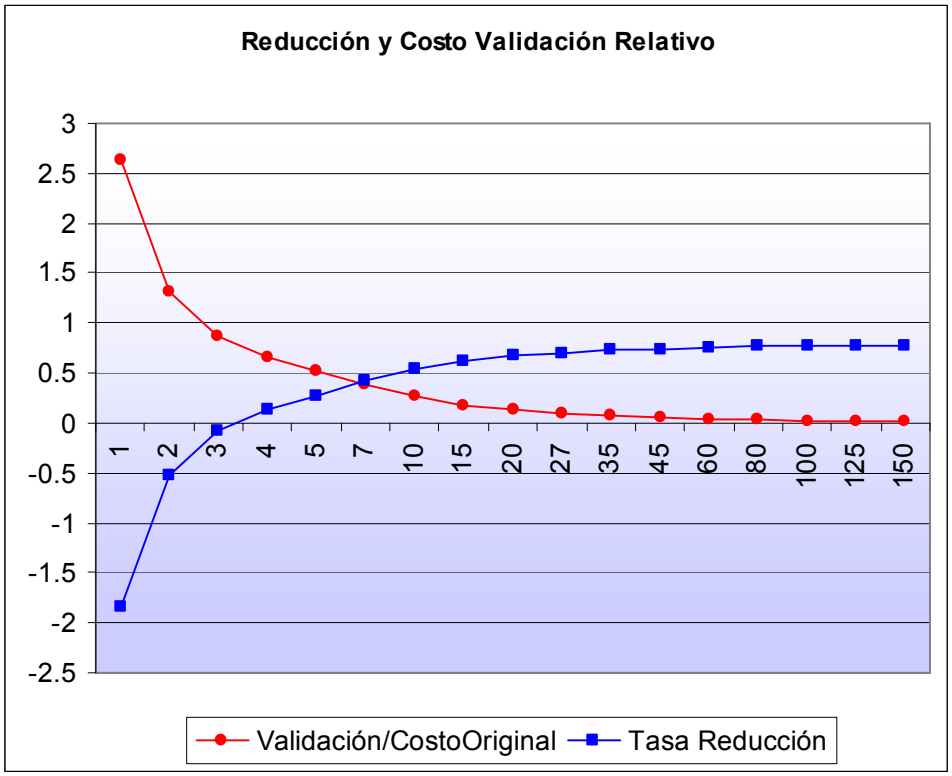


Figura 12: Gráfico comparativo de reducción y costo de validación

Este comportamiento de la tasa de reducción proporcionada por la metodología puede explicarse con el gráfico de la figura 13, que muestra como a medida que la complejidad del método de agrupamiento a utilizar aumenta, la diferencia entre el costo original y el costo de la metodología aumenta.

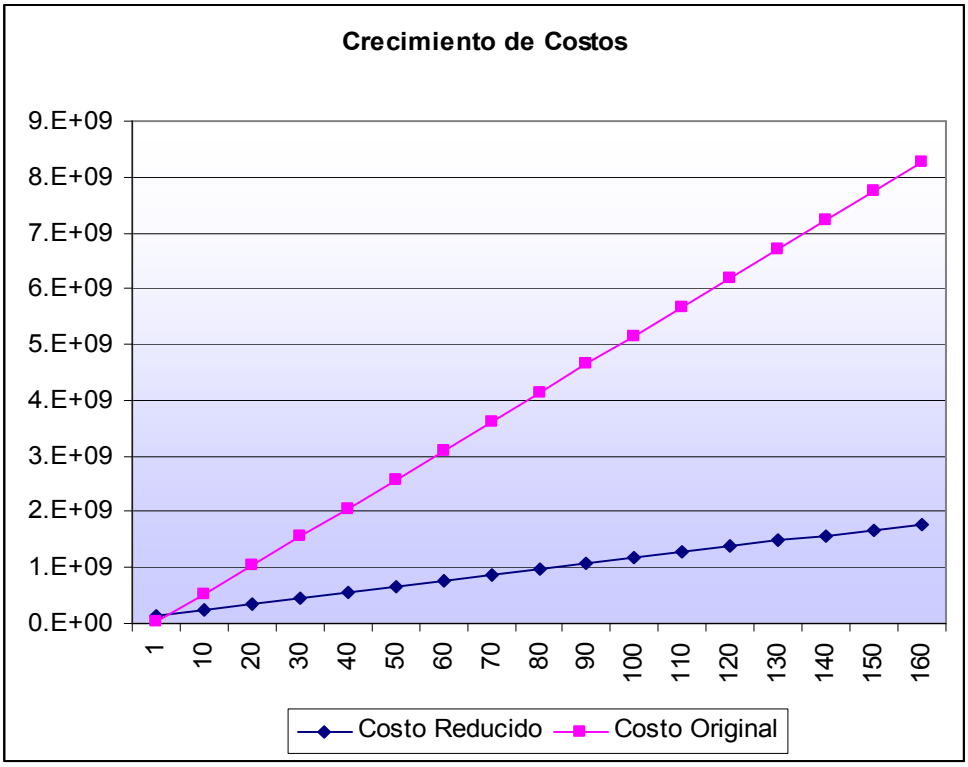


Figura 13: Gráfico comparativo de costos

Capítulo 4
Caso de Estudio

La metodología fue creada a partir de la necesidad presentada por un proyecto real de minería de datos orientado a determinar un agrupamiento de clientes para una empresa de gran tamaño.

En este capítulo se presenta el caso de estudio, detallando la orientación del proyecto, el desarrollo del mismo y los resultados obtenidos al aplicar la metodología definida en el capítulo anterior.

4.1 Entendimiento del negocio

El proyecto de minería de datos fue realizado para una gran empresa multinacional (una de las más grandes de Latinoamérica) a la que de aquí en adelante llamaremos "la empresa".

La empresa atiende a millones de clientes en la República Mexicana y a lo largo de varios países de Latinoamérica prestando servicios en el área de telecomunicaciones.

4.1.1 Objetivos del negocio

La empresa tiene por principal objetivo "consolidar su liderazgo en el mercado nacional, expandiendo su penetración de servicios de telecomunicaciones en todos los mercado posibles, para situarse como una de las empresas de más rápido y mayor crecimiento a nivel mundial."[12]

Para poder lograr este objetivo la empresa dedica muchos recursos para ofrecer un servicio de calidad al cliente. En este sentido, las relaciones de la empresa con sus clientes son importantes y necesitan de atención y un proceso continuo de mejoras, no sólo para ofrecer un buen servicio a los clientes actuales, si no también para lograr aumentar la cantidad de clientes de la empresa y el consumo de servicio que éstos hacen.

4.1.2 Recursos disponibles

Cada una de las áreas operativas de la empresa posee información referente a los clientes de la empresa y a los servicios prestados a éstos. Esta información es

12 Segmento tomado de la misión de la empresa

almacenada en varias bases de datos que son actualizadas por diferentes sistemas gestores de bases de datos, pero la mayoría de ellas son almacenadas y administradas con IBM “Universal Database” (UDB) versión 7.0.

La información se encuentra distribuida en diferentes servidores de la empresa situados en localidades separadas.

Para poder llevar a cabo el proyecto fue necesario una fase previa en la que se llevó a cabo una serie de reuniones de trabajo con personal administrativo e informático de la empresa, que tenían por objetivo revisar, valorar y seleccionar las diferentes fuentes de datos que posee la empresa a fin de determinar los datos de mayor valor para la caracterización de clientes de acuerdo a los objetivos del negocio.

Esta fase previa duró cerca de tres meses y requirió de muchas horas-hombre. El resultado final fue un listado de los datos de cliente que serían considerados. Además se dispuso que los datos recolectados fueran colocados en una base de datos de IBM: “Universal Database” (UDB) versión 7.0.

Para el proyecto se disponía de acceso a la base de datos resumida de clientes, equipo de cómputo compuesto por una computadora de escritorio y una herramienta de minería de datos de IBM: Intelligent Miner versión 6.1. Esta herramienta permite realizar varias tareas de minería de datos trabajando directamente sobre bases de datos UDB o sobre archivos de texto plano.

4.1.3 Objetivos del Proyecto de Minería de Datos

Para poder enfocar los esfuerzos de mejora de atención al cliente y, consecuentemente, obtener un incremento en las ventas, la empresa necesitaba contar con una caracterización de clientes basada en la información existente, por ejemplo, datos generales de cliente, servicios contratados, facturación, rangos de crédito, áreas de servicio, etc.

De acuerdo a lo anterior, el objetivo principal del proyecto de minería de datos es caracterizar los clientes de la empresa de manera que se identificaran tipos (grupos) de clientes atendidos. Esto permitiría, en un futuro cercano y de acuerdo a las características de los tipos identificados, mejorar los servicios prestados, ofrecer nuevos servicios y/o incrementar las ventas a clientes registrados y a nuevos clientes.

A partir de los diferentes datos registrados para cada cliente, se desea entonces, construir un modelo de agrupamiento que permita caracterizar a los tipos de clientes que atiende la empresa.

4.2 Entendimiento de los datos

Una vez establecidos los objetivos de la empresa y del proyecto de minería de datos se procedió con una fase de recolección de los datos necesarios para la realización de la tarea de agrupamiento.

4.2.1 Recolección de datos

La fuente de datos del proyecto está constituida por 400,000 registros de clientes.

Para obtener los datos del proyecto el *personal informático de la empresa* realizó el siguiente proceso:

1. Creación de la base de datos fuente para el proyecto. Según la lista de datos identificados como importantes para la empresa se diseñó y creó la estructura de la nueva base de datos.
2. Selección del grupo de clientes para el proyecto.
3. Extracción de los datos identificados de cada base de datos de la empresa.
4. Transformación de datos, en los casos que fuera necesario. La transformación era necesaria cuando los datos se encontraban en un formato diferente al de la estructura de la base de datos nueva, cuando eran el resultado de una operación matemática, etc.
5. Vaciado de datos en la base de datos fuente. Los datos extraídos de diferentes bases de datos se acomodaron a la estructura de la nueva base de datos.
6. Revisión de la base de datos nueva. Se revisó que la base de datos estuviera completa de acuerdo a la información extraída de las otras bases de datos.

Este proceso dio como resultado una base de datos en UDB v 7.0 con las tablas, registros y campos importantes para el proyecto de minería de datos, los cuales son descritos en la siguiente sección.

4.2.2 Descripción de datos

La base de datos fuente del proyecto se resume en 9 tablas de datos. Algunas de estas tablas contienen únicamente información de códigos que son utilizados en otras tablas de datos. La información contenida en la base de datos está representada en columnas de valores numéricos y alfanuméricos.

En la tabla 2 se presenta una breve descripción de las tablas que componen la fuente de datos para el proyecto.

Tabla	Columnas	Filas	Descripción
FACTURACION	25	400,000	Facturación y renta de los clientes
SERVINTERNET	121	400,000	Datos de contratación del servicio de Internet
CONTRATACION	49	400,000	Datos referentes a la contratación de servicio en forma de paquetes
CLIENTE	11	400,000	Datos generales del cliente
LOCAL	121	400,000	Información de consumo de servicios telefónicos locales
DIGITAL	85	400,000	Información de consumo de servicios telefónicos digitales
AREAS	2	73	Codificación de áreas de servicio
CREDITO	2	4	Codificación de calificación crediticia del cliente.
ANTIGUEDAD	3	183	Codificación de antigüedad del cliente

Tabla 2: Fuentes de datos para el proyecto.

Las principales tablas de esta fuente de datos poseen un campo CVE que identifica al cliente al que se refiere la información. El valor del campo CVE no corresponde al número telefónico asignado, si no a un código numérico asignado a cada cliente de la empresa. Las últimas 3 tablas no contienen el campo CVE mencionado, por lo que su relación con cada cliente debe ser inferida a partir de las relaciones con otras tablas de la base de datos.

4.2.3 Calidad de los datos.

Durante el proceso de recolección de los datos del proyecto, *el equipo técnico de la empresa* revisó las fuentes de datos originales; verificando que los datos a ser extraídos fueran fidedignos y consistentes con las reglas de integridad establecidas para cada base de datos de la empresa y para la base de datos del proyecto. Tal es el caso de rangos de valores, unicidad de llaves primarias, integridad referencial de llaves, etc.

4.3 Preparación de datos

Con la base de datos fuente se procedió a realizar una preparación de los datos, descartando datos redundantes, registros incompletos, datos inconsistentes o vacíos, entre otros.

4.3.1 Limpieza de datos

En esta fase se realizó una revisión minuciosa de los datos en busca de valores inconsistentes, incompletos, redundantes y nulos. Cada tabla fue revisada en busca de problemas con la información representada.

Esta fase determinó que,

1. Cada una de las tablas de análisis posee la misma cantidad de registros de CVE.
2. No existen datos faltantes o nulos
3. No hay datos inválidos según el dominio de cada campo.
4. No existen violaciones de integridad referencial.

4.3.2 Integración de datos

Para poder trabajar los datos de la base de datos como un todo, es necesario integrar la información de las diferentes tablas en una sola vista que incluya todos los campos a trabajar.

Para analizar estas tablas se procedió, primero, por analizar el tipo de campos de cada tabla y revisar las relaciones entre tablas.

La base de datos cuenta con 6 tablas principales que contienen datos referentes a clientes: FACTURACION, SERVINTERNET, CONTRATACION, LOCAL, DIGITAL y CLIENTE. La integración de estas tablas se realiza de manera directa al tomar un campo llave CVE como campo de enlace.

En el caso de las 3 tablas restantes: AREAS, CREDITO y ANTIGÜEDAD, la integración no puede realizarse de manera directa con respecto a un campo CVE, ya que éstas no poseen dicho campo. El proceso seguido para este caso es el siguiente:

1. La tabla AREAS puede ligarse por medio de un campo AREA a la tabla CLIENTE.
2. La tabla CREDITO posee únicamente dos campos y representa una codificación de las calificaciones crediticias. Como esta tabla en sí representa la codificación de valores alfanuméricos y el código correspondiente a cada cliente ya se encuentra en la tabla CLIENTE, no es necesario que esta tabla sea integrada a la vista de datos.
3. ANTIGUEDAD es una tabla que presenta una relación de muchos a muchos, por lo que no puede asociarse de manera única a cada cliente. Por esta razón, esta tabla es ignorada.

Este proceso dio como resultado una vista general de datos con 415 columnas.

4.3.3 Transformación de datos

Para poder trabajar con los datos era necesario convertir aquellos campos que fueran alfanuméricos en valores numéricos que permitieran el tratamiento matemático necesario (*ver SOMs*).

Para lograr esta transformación de datos se procedió con una codificación de los valores alfanuméricos que se encontraban en la fuente de datos. Dichos campos provenían principalmente de la tabla CLIENTE.

Los valores alfanuméricos fueron revisados y codificados en valores numéricos discretos, luego de asignar un valor numérico a cada aparición diferente del campo.

Para integrar las diferentes tablas en una sola fuente de datos se construyó una consulta SQL que realizara la unión de las tablas y la representación codificada de los campos alfanuméricos.

4.4 Modelado

Con los datos preparados para su análisis con técnicas de minería de datos se procedió a aplicar la fase de modelado; que en este caso, y de acuerdo a la metodología planteada, significa realizar primero una reducción del espacio de búsqueda sobre el cual trabajaran los algoritmos de agrupamiento.

4.4.1 Reducción del espacio de búsqueda

La reducción del espacio de búsqueda, como lo indica la sección 3.2 implica realizar una reducción vertical y horizontal sobre los datos, además de un proceso de validación de la representatividad de la muestra obtenida.

4.4.1.1 Reducción Vertical

Al aplicar un análisis correlacional se obtuvo una reducción significativa en la cantidad de variables que conformaban la vista de datos. Del total de variables originales (415) solamente se mantuvieron 129 variables consideradas significativas. En otras palabras, la reducción vertical obtenida equivale al **68.91%** (286/415) de las variables a considerar y por ende, del tamaño de la vista de datos.

En el anexo 3 se presenta una porción de la matriz de correlaciones obtenida.

La cantidad de variables significativas por cada una de las tablas originales se presenta en la tabla 3.

Tabla	Variables
FACTURACION	9
CLIENTE	9
SERVINTERNET	51
LOCAL	55
DIGITAL	5
Total	129

Tabla 3: Variables significativas por tabla de datos

Como puede observarse la mayor parte de las variables significativas proviene de 5 tablas de datos. Para el caso, la tabla CONTRATACION no aportó variables significativas, lo que puede explicarse a partir del hecho que la gran mayoría de los valores de sus variables (el 98.12%) presentaba un valor de 0.

4.4.1.2 Reducción Horizontal

Luego de efectuar la reducción vertical se procedió a ejecutar la reducción horizontal. En esta fase era necesario determinar el tipo de muestreo a realizar y el tamaño de la muestra a extraer.

Como no es posible asegurar que los datos del estudio constituyen proporciones, el criterio a tomar para el tamaño de la muestra es el uso de validación cruzada (como fue presentado en la sección 3.2.2) con un tamaño de muestra del 20% y una cantidad de 5 muestras excluyentes. Además, considerando que todos los registros de clientes se consideran de igual importancia, la forma de muestreo empleada fue muestreo aleatorio (véase sección 2.2.1.1).

Teniendo en consideración lo anteriormente expuesto, el tamaño de cada una de las 5 muestras equivale a 80,000 registros de datos (20% de 400,000). La selección de los elementos de cada muestra se realizó utilizando un muestreo aleatorio simple asignando a cada fila de la vista de datos un valor aleatorio entre 0 y 1 inclusive y tomando aquellos registros con un valor menor o igual a 0.2 para la primera muestra; registros con un valor entre 0.2 y 04 para la segunda muestra y así sucesivamente.

Este proceso representa una reducción del 80% sobre los datos originales y en conjunto con la reducción vertical conlleva a una reducción de aproximadamente el **93.8%** de la base de datos original (68.91% por reducción vertical y 80% por reducción horizontal en cada muestra).

Una vez obtenidas las muestras, se procede a realizar la validación de la representatividad de las mismas.

4.4.1.3 Validación de la representatividad

Para demostrar que las muestras extraídas son representativas del conjunto original de datos empleamos el mismo algoritmo presentado en [KURI07] y citado en la sección 3.2.3.2 de este documento.

Para efectuar el ajuste de los modelos a las diferentes variables que componen cada muestra (129 para este caso) se utilizó la herramienta Curve Expert v. 1.3 (http://curveexpert.webhop.biz/). Con esta herramienta se generaron las gráficas que se muestran en esta sección y en el anexo 4.

Para no realizar el proceso de verificación sobre todas las posibles combinaciones de parejas de variables se estableció como criterio de paro obtener un valor de confianza menor a 5E-9 con no más de 5 pruebas de ajuste. Este valor puede ser alcanzado si se encuentran dos parejas de variables que presenten el mismo modelo de comportamiento en las 5 muestras evaluadas.

Para obtener el valor de confianza se utiliza el procedimiento basado en combinaciones con repetición detallado en la sección 3.2.3.3.

En la tabla 4 se presenta por cada prueba (i), las variables analizadas (columna Var.), el resultado de ajuste para cada muestra (M1..M5) y el valor de confianza obtenido para la prueba.

i	Var.	Resultado del Ajuste					Confianza
		M1	M2	M3	M4	M5	
1	V1 - V10	Polinomio de 4° grado	Polinomio de 4° grado	Polinomio de 4° grado	Polinomio de 4° grado	Polinomio de 4° grado	6.77E-5
2	V2 - V20	MMF	Función racional	MMF	Polinomio de 4° grado	MMF	0.038
3	V3 - V30	Función racional	Polinomio de 4° grado	Polinomio de 4° grado	MMF	MMF	0.405
4	V4 - V40	Función Racional	Función Racional	Función Racional	Función Racional	Función Racional	6.77E-5
5	V5 - V50	Polinomio de 5° grado	Polinomio de 5° grado	Polinomio de 5° grado	Polinomio de 5° grado	MMF	2.23E-3
						Producto	**1.57E-13**

Tabla 4: Ajuste de muestras para 10 variables

Con las 5 pruebas realizadas se alcanzó un valor de confianza de 1.72E-13 que indica que los resultados obtenidos no son producto del azar, si no la afirmación de que las muestras son representativas.

Puede observarse en la tabla 4 que la prueba 3 produjo los resultados más deficientes obteniendo una coincidencia de 2 muestras. Para completar el análisis numérico de la confianza, se realiza también un análisis de las gráficas de los modelos obtenidos.

Las siguientes figuras muestran algunas de las gráficas resultantes para dos muestras al analizar 3 pares de variables. Las gráficas completas de las pruebas realizadas se presentan en el anexo 4.

Las figuras 14 y 15 muestran como el comportamiento para las variables V1 y V10 es similar en la muestra 2 y 3.

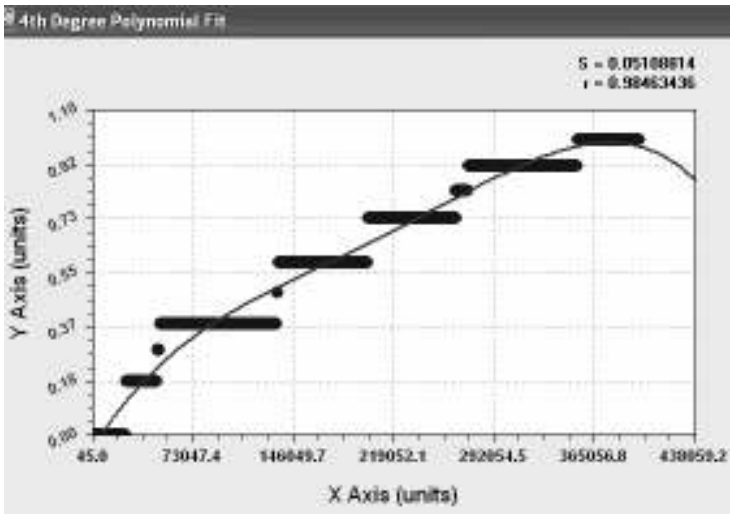


Figura 14: Ajuste regresivo. Modelo polinomial de 4° grado para la muestra 2

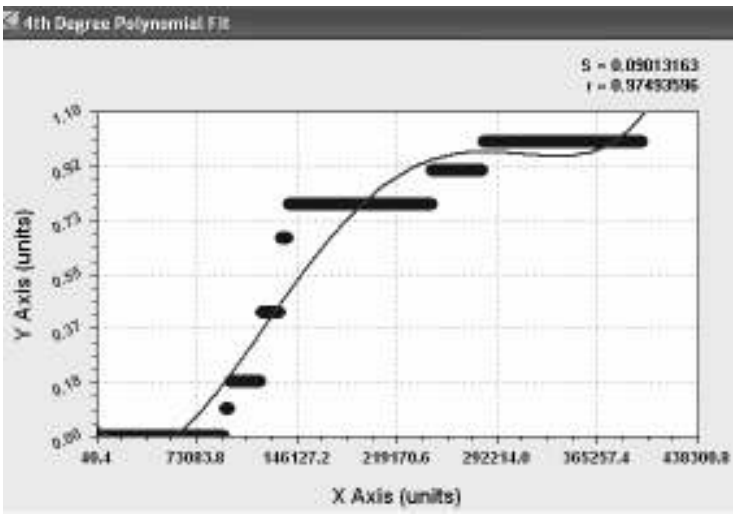


Figura 15: Ajuste regresivo. Modelo polinomial de 4° grado para la muestra 3

Las figuras 16 y 17 muestran el par de variables V2 y V20, para la muestra 1 y 5.

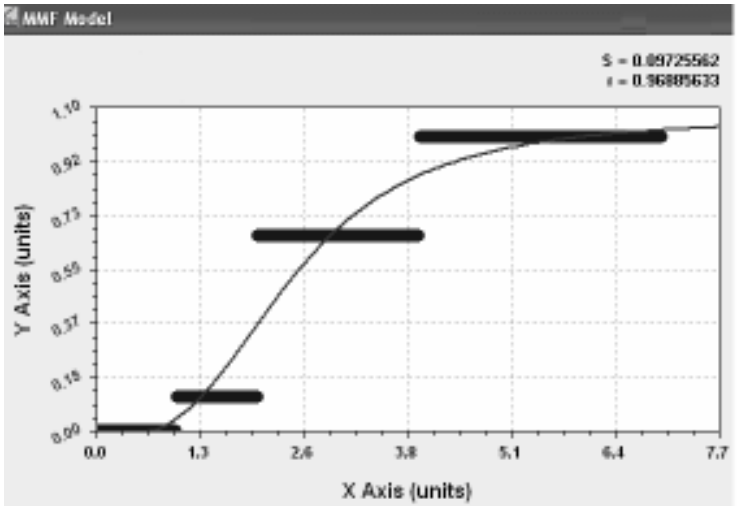


Figura 16: Ajuste regresivo. Modelo MMF para la muestra 1

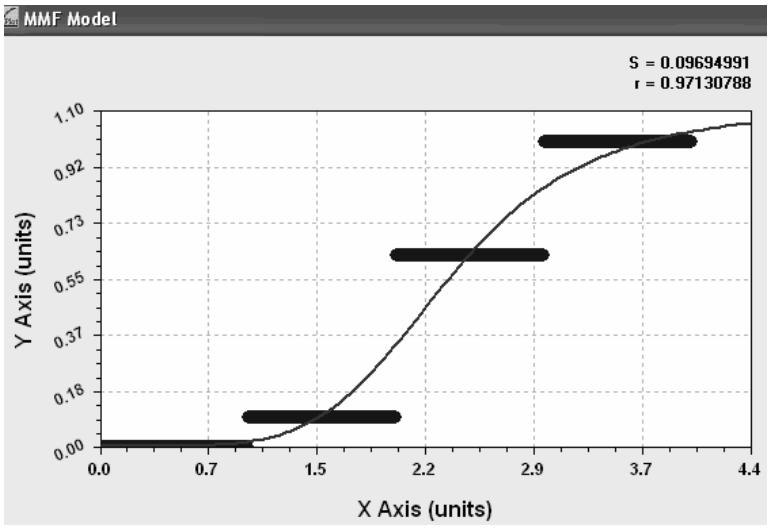


Figura 17: Ajuste regresivo. Modelo MMF para la muestra 5

Las figuras 18 y 19 muestran el par de variables V3 y V30 para la muestra 1 y 3. En este caso puede observarse que el comportamiento de las variables es muy similar en ambas muestras analizadas a pesar de que no se ajustaron al mismo modelo matemático.

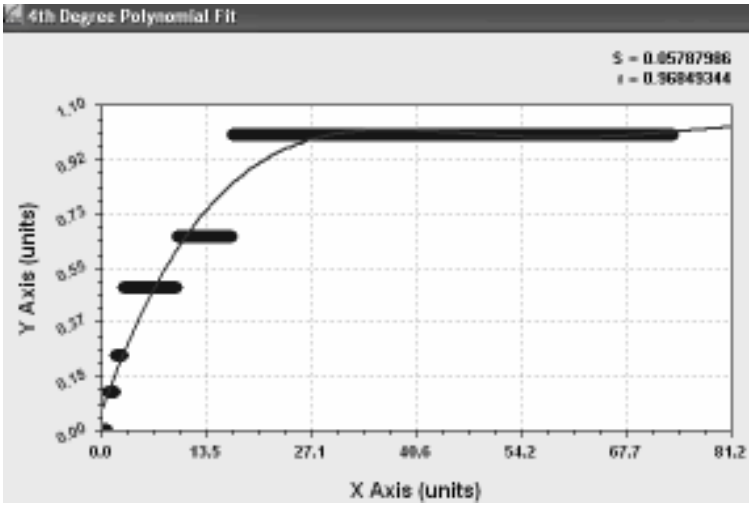


Figura 18: Ajuste regresivo. Función racional para la muestra 1

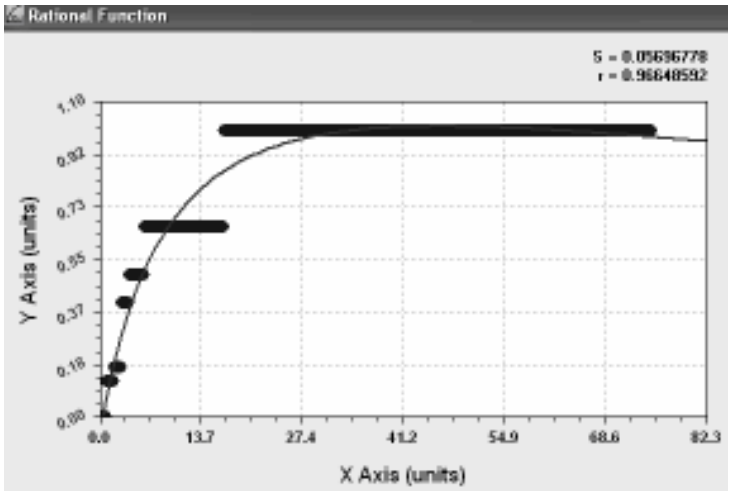


Figura 19: Ajuste regresivo. Polinomio de 4° grado para la muestra 3

Como puede observarse de las gráficas, incluso las variables que se ajustaron a diferentes modelos presentaron comportamiento similar. Es así que el análisis de las gráficas también comprueba la representatividad de las muestras.

Habiendo comprobado la representatividad de las muestras, se selecciona una de estas como la vista minable sobre la que se trabajará para obtener el modelo de minería de datos buscado.

4.4.2 Determinación del número óptimo de grupos

El método usado fue detallado en la sección 2.1.3. Se siguieron los siguientes pasos:

1. Aplicar el algoritmo de agrupamiento Fuzzy C Means para obtener modelos de grupos para *c = 2, 3,..., C*. Dónde *C* representa el valor más grande aceptable para el número de grupos.
2. Graficar los valores del coeficiente de partición (PC) y el coeficiente de entropía (PE) para cada modelo obtenido.
3. Determinar el número óptimo de grupos de acuerdo al criterio del codo.

Luego de aplicar este proceso se obtuvieron los valores para PC y PE que se presentan en la tabla 5.

Grupos	PC	PE
2	0.87925	0.20433
3	0.77000	0.43575
4	0.64171	0.63928
5	0.56007	0.81172
6	0.49803	0.98207
7	0.48874	1.03566
8	0.41314	1.21995
9	0.41385	1.22441
10	0.40040	1.27230
11	0.35936	1.40278
12	0.34936	1.43338
13	0.31431	1.56234
14	0.31522	1.56798
15	0.28717	1.69967
16	0.27687	1.73856

Tabla 5: Valores para Criterio del Codo

La figura 20 muestra el gráfico de estos valores y la ubicación del número óptimo de grupos se resalta con el óvalo. Como puede observarse, el número óptimo de grupos, de acuerdo a este criterio, es 6.

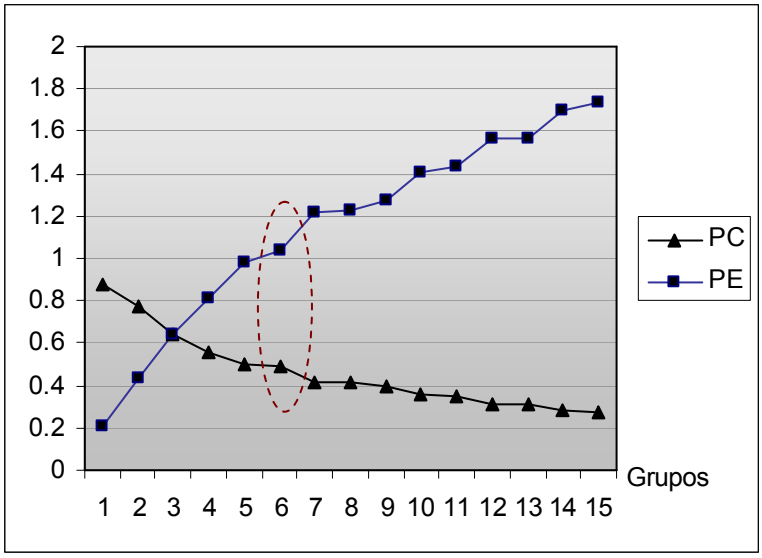


Figura 20: Criterio del Codo

Una vez encontrado el número de grupos se procede a aplicar el algoritmo de agrupamiento.

4.4.3 Agrupamiento

Para realizar el agrupamiento se usaron Mapas Auto Organizados (véase la sección 2.1.2.2).

A este algoritmo se le proporciona el número de grupos e indica la cantidad de neuronas a ser utilizadas. Como se desea que cada neurona represente un grupo de clientes, se definen 6 neuronas para la capa de salida.

El resultado proporcionado por el minero se presenta en la figura 21. Al lado izquierdo del gráfico se muestran los porcentajes de población asignados a cada grupo, ordenados de mayor a menor. Al lado derecho de cada banda que representa un grupo se puede ver un número que indica la neurona que fue asignada a ese grupo. Los gráficos dentro de la banda muestran las seis variables que aportaron más información a la definición del grupo.

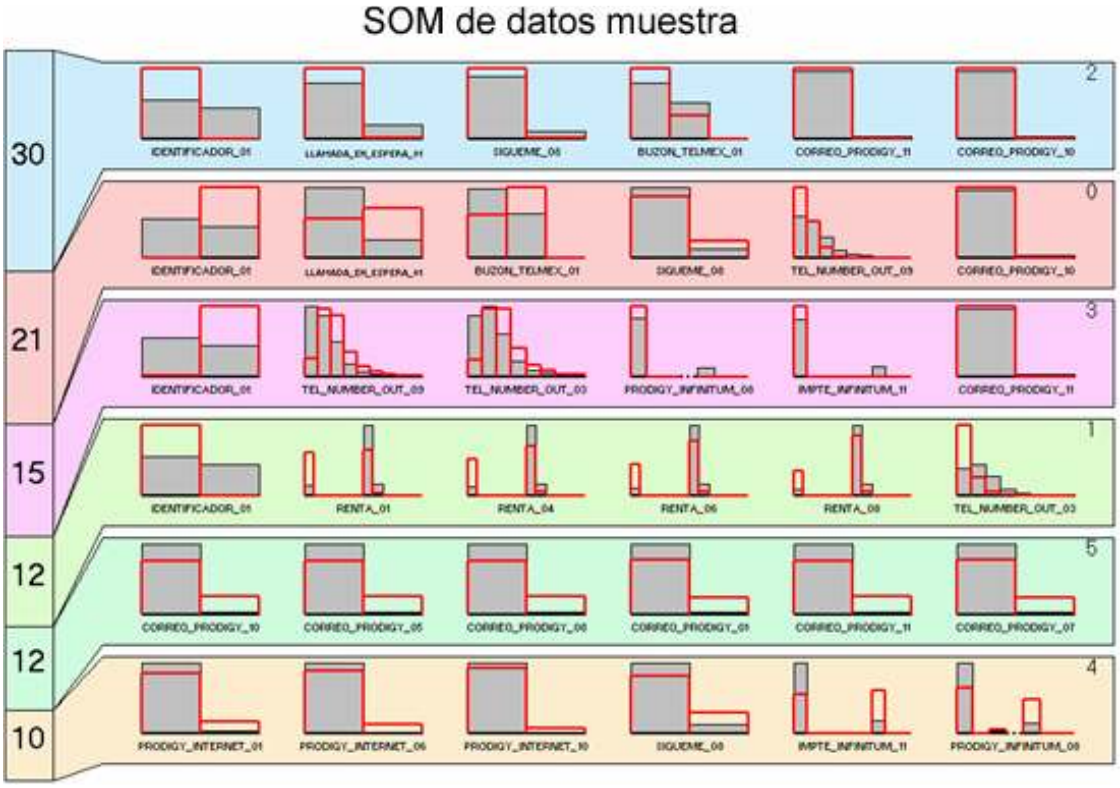


Figura 21: Resultado del agrupamiento sobre la muestra

4.5 Validación de resultados

Para asegurar la calidad de los resultados obtenidos se revisaron los datos utilizados para la generación del modelo. Se revisó también el proceso de agrupamiento aplicado a la muestra y las propiedades generales de los resultados proporcionados por el minero.

El modelo de minería de datos generado cumple con los objetivos del proyecto de minería de datos y está acorde a los objetivos de la empresa.

Capítulo 5
Análisis de Resultados

En el capítulo anterior se mostró la aplicación de la metodología propuesta al caso de estudio y los resultados que se obtuvieron. En este capítulo se presentan un análisis de dichos resultados.

El análisis se enfoca en dos aspectos principales: validez de los resultados obtenidos y eficiencia del proceso seguido.

Nuestra hipótesis para el análisis es la siguiente:

a) *La metodología proporciona resultados con más del 90% de confiabilidad*

b) *El proceso seguido muestra una eficiencia superior a la obtenida sin la aplicación de la metodología propuesta*

Para facilitar la referencia a los modelos analizados asignaremos nombres a cada uno. El primero es el modelo generado por la metodología, el cual fue construido a partir de una muestra de los datos originales; llamaremos a este modelo “Modelo 1”. El segundo modelo se obtiene al ejecutar el proceso de agrupamiento sobre el conjunto original de datos; a este modelo lo llamaremos “Modelo 2”.

5.1 Validación de resultados.

La validez de la metodología resulta de la comparación de resultados obtenidos con los dos modelos.

Los valores numéricos que nos permitirán validar los resultados surgen de los siguientes pasos:

- Reducir la vista de datos únicamente de manera vertical. (Reduciendo la cantidad de variables y no la de registros).
- Ejecutar un proceso de agrupamiento para obtener el Modelo 2.
- Comparar las distribuciones de población para el Modelo 1 y Modelo 2.
- Etiquetar[13] el conjunto original de datos y la muestra con el Modelo 1 y con el Modelo 2.
- Comparar la distribución de elementos obtenida a partir del etiquetamiento con ambos modelos.

[13] Asignar a cada elemento del conjunto un número que corresponde a cada uno de los grupos determinados.

5.1.1 Comparación del Modelo 1 y el Modelo 2

El mismo algoritmo de agrupamiento empleado para el Modelo 1 fue utilizado para obtener el Modelo 2. A los datos fuente se les aplicó la reducción vertical (eliminación de variables altamente correlacionadas). El resultado del agrupamiento proporcionado por el minero se presenta en la figura 22.

Puede observarse en la figura 22 que la distribución de población en los 6 diferentes grupos no es igual a la mostrada en la figura 21, que presenta los resultados obtenidos para el Modelo 1.

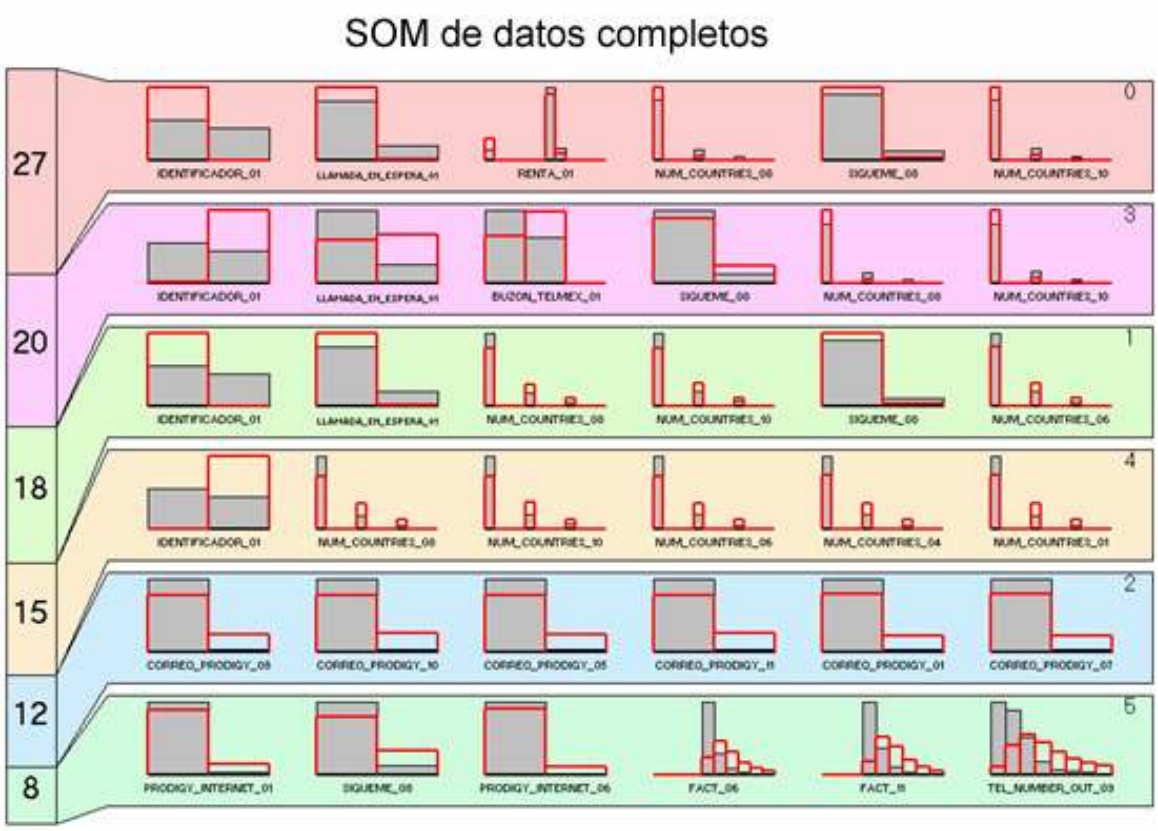


Figura 22: Resultado del agrupamiento sobre datos completos

La tabla 6 muestra los valores absolutos de las diferencias.

Los grupos han sido nombrados con letras para evitar confusión con los números de neuronas presentados en las gráficas del minero (figura 21 y 22).

Grupo	Modelo 1 (%)	Modelo 2 (%)	Diferencia (%)
A	30	27	3
B	21	20	1
C	15	18	3
D	12	15	3
E	12	12	0
F	10	8	2

Tabla 6: Comparación de Modelo 1 y Modelo 2

Los datos de la tabla 6 muestran que, aunque existen diferencias en las distribuciones de la población para los grupos, ningún grupo posee una diferencia superior al 3% (más aún, existe un grupo con 0% de diferencia en su distribución).

La suma de los elementos de la población que se encuentran distribuidos en grupos diferentes (en el modelo 2 con respecto al modelo 1) es cercana al 6%.

Si se observa el gráfico de agrupamiento para la muestra y el de agrupamiento para el conjunto completo de datos se puede notar que existen coincidencias entre las seis variables que aportan más información a la definición de cada grupo (el minero presenta en las franjas de cada grupo las seis variables más importantes). La tabla 7 muestra la cantidad de variables que coinciden por cada grupo entre las seis más importantes que definen al grupo.

Grupo	Variables coincidentes
A	3
B	4
C	1
D	1
E	6
F	3

Tabla 7: Coincidencias de variables

Puede observarse de la tabla 7 y la tabla 6 que el grupo “E” que posee la mayor cantidad de variables coincidentes, también es el que posee la menor diferencia en la distribución de población. Por otro lado, los grupos que poseen menos variables coincidentes son los que presentan una mayor diferencia en la distribución de la población.

De este análisis se destaca que, aunque cada grupo no posea el mismo orden de importancia para las variables que lo definen, los resultados de las distribuciones de población siguen siendo aceptables.

5.1.2 Representatividad de la muestra

Si la muestra obtenida es verdaderamente representativa del conjunto total de datos, el modelo obtenido a partir de ésta debería inducir una distribución de población similar al ser aplicado para etiquetar: a) conjunto muestral y b) el conjunto completo.

La tabla 8 muestra los resultados de la distribución de elementos etiquetados, luego de aplicar el Modelo 1 a la muestra y al conjunto completo.

Como es lógico, la distribución de la población para la muestra presenta los mismos valores de distribución del Modelo 1. E, interesantemente, estos valores resultan ser aproximadamente iguales al aplicar el Modelo 1 al conjunto de datos completo. Para poder observar las diferencias se han presentado los porcentajes obtenidos con una precisión de 2 decimales. De la tabla 8 puede observarse que el grado de error de la representatividad de la muestra es menor al 0.2% en cada grupo analizado. Podemos asegurar, pues, que la muestra obtenida es representativa del conjunto de datos completo.

Grupo	Muestra (%)	Datos completos (%)	Diferencia (%)
A	30.06	30.24	0.18
B	21.01	20.91	0.10
C	15.45	15.37	0.08
D	12.27	12.25	0.02
E	11.54	11.55	0.01
F	9.67	9.68	0.01

Tabla 8: Distribución de población con el Modelo 1

5.1.3 Validación cruzada

Como tercer paso de validación aplicaremos el Modelo 2 a los dos conjuntos de datos y procederemos a etiquetarlos.

La tabla 9 presenta una sección para los resultados del etiquetamiento de los datos de muestra y otra para los resultados del etiquetamiento del conjunto total de datos. Cada sección presenta la cantidad de individuos etiquetados en cada grupo usando el Modelo 1 y el Modelo 2, la cantidad de individuos diferentes entre estas dos agrupaciones y el porcentaje de la población total estudiada.

	Etiquetamiento para la muestra				Etiquetamiento para datos completos			
Gr	Modelo1	Modelo2	Difer.	% Poblac.	Modelo1	Modelo2	Difer.	%Poblac.
A	23906	21503	2403	3	120961	108084	12877	3
B	16712	16155	557	1	83655	81420	2235	1
C	12285	14327	2042	3	61471	72367	10896	3
D	9760	11828	2068	3	49013	59313	10300	3
E	9179	9580	401	1	46195	48356	2161	1
F	7687	6136	1551	2	38705	30460	8245	2

Tabla 9: Validación cruzada de los modelos

Como puede apreciarse en la tabla 9, los porcentajes poblacionales de la diferencia entre las agrupaciones son iguales para el caso de la muestra y el conjunto total de datos; lo que significa que tanto el Modelo 1 como el Modelo 2 arrojan distribuciones de población iguales sobre la muestra y sobre el conjunto completo de datos. Este hecho confirma la representatividad de la muestra y las diferencias entre los dos modelos, ya que la diferencia en la distribución de la población es similar a las diferencias presentadas en la tabla 6.

5.1.4 Conclusiones de la validación

1. La diferencia en las distribuciones de población que presentan los modelos 1 y 2 es menor o igual al 3% de la población en cada grupo.
   - *Conclusión*: el error observado es aceptable.
2. El porcentaje total de población distribuida de manera diferente en el Modelo 1 y 2 equivale a un 6%.
   - *Conclusión*: el error es aceptable.
3. Las variables más importantes que definen cada grupo en ambos modelos son similares.
   - *Conclusión*: El efecto producido por las variables de estudio no fue afectado por la extracción de la muestra.
4. El espacio de búsqueda reducido (muestra) es representativo del conjunto completo de datos.
   - *Conclusión*: El proceso seguido para la extracción fue correcto.

5.2 Eficiencia de la metodología

La eficiencia se refiere a la capacidad de lograr un resultado esperado a partir de la utilización del mínimo de recursos posible. Ahora comparamos la eficiencia del Modelo 1 y el Modelo 2 para determinar la pertinencia de la metodología propuesta.

Definimos la eficiencia como sigue:

$$Eficiencia = \frac{Exactitud}{Recursos}$$

Para efectos de este análisis consideraremos una ponderación del modelo completo de 100 análogamente, asignaremos al modelo muestral una de 94.

5.2.1 Eficiencia por registros

En este caso, consideramos a cada registro de cliente como un recurso. La eficiencia viene dada por la cantidad de registros que se empleó para obtener la exactitud resultante para cada modelo.

$$Eficiencia1 = \frac{94}{80{,}000} = 0.001175$$

$$Eficiencia2 = \frac{100}{400{,}000} = 0.00025$$

Como puede deducirse del resultado, la eficiencia lograda para el Modelo 1 es superior a la lograda con el Modelo 2.

*Cada elemento procesado en el Modelo 1 aportó **4.7 veces más valor** a la solución obtenida que los elementos procesados para el Modelo 2.*

5.2.2 Eficiencia por datos.

Si consideramos cada valor de campo de la base de datos como un recurso, entonces consideraremos tres ambientes: 1) El conjunto original de datos, 2) El conjunto reducido verticalmente y 3) El conjunto reducido vertical y horizontalmente. De esta forma, la eficiencia para cada ambiente es de,

$$Eficiencia1 = \frac{100}{166{,}000{,}000} = 6.024\text{E-}7$$

$$Eficiencia2 = \frac{100}{51{,}600{,}000} = 19.38\text{E-}7$$

$$Eficiencia3 = \frac{94}{10{,}320{,}000} = 91.08\text{E-}7$$

A los ambientes les corresponde: 1) 400,000 registros x 415 variables; 2) 400,000 registros x 129 variables y 3) 80,000 registros x 129 variables. Como se ve, el modelo correspondiente al ambiente 3 (Modelo 1) es **4.7** veces más eficiente que el del ambiente 2 (Modelo 2) y **15.13** veces más eficiente que el del ambiente 1.

*Este nivel de eficiencia se debe a que el modelo generado a partir de la muestra produce una exactitud del **94%** con sólo el **6.2%** de los datos originales (10,320,000/166,000,000).*

5.2.3 Eficiencia por trabajo realizado.

Además de medir la eficiencia de la metodología a partir del uso de memoria también puede medirse a partir del tiempo involucrado. Nos referiremos al uso de recursos asociados a la capacidad de cómputo como "esfuerzo".

Para determinar el esfuerzo realizado para encontrar un modelo, medimos la cantidad de veces que un dato fue considerado en los cálculos realizados.

Cada vez que un dato es requerido para efectuar cálculos, es necesario que el sistema operativo de la computadora, el sistema gestor de la base de datos, y el *hardware* realicen un proceso de extracción del dato solicitado de los dispositivos de almacenamiento de la computadora.

La combinación de todos estos factores da como resultado un ambiente de procesamiento que es el que determina el tiempo de recuperación de un elemento dato.

Consideramos que el Modelo 1 y el Modelo 2 fueron procesados en el mismo ambiente de procesamiento. Entonces, podemos obviar el cálculo del tiempo de recuperación de un elemento y tomar como parámetro del esfuerzo requerido la cantidad de veces que cada modelo accedió a datos.

Este valor se mide de manera aproximada tomando como base la complejidad de cada paso que compone el proceso de generación de los modelos. Dicha complejidad se mide de acuerdo a la cantidad de acceso a datos y registros.

Antes de calcular la eficiencia del proceso se necesita definir la complejidad para los algoritmos de agrupamiento utilizado: Fuzzy C-Means y SOM.

5.2.3.1 Complejidad de Fuzzy C-Means.

El cálculo de esta complejidad se basa en el algoritmo de Fuzzy C-Means presentado en la sección 2.1.2.1, centrándose en los accesos necesarios a datos.

Debe tenerse en cuenta que el algoritmo accede a tres matrices de datos:

1. Matriz de datos con **n** filas (número de elementos de la muestra) por **y** variables.
2. Matriz de centros con **c** filas por **y** variables.
3. Matriz de pertenencia con **n** filas por **c** columnas (una por cada centro).

El cálculo de los centros de lo grupos utiliza la siguiente fórmula:

$$V_i = (\sum_{k=1}^{n} (\mu_{i,k})^2 x_k) / (\sum_{k=1}^{n} (\mu_{i,k})^2), \text{ para } i = 1, 2, \ldots, c$$

Calcular el valor de todos los V equivale a:

$H_1 = c(2ny+n+y)$

De donde,

2ny resulta de los accesos necesarios para la primera sumatoria.

n corresponde a los accesos de la sumatoria del denominador.

y los accesos necesarios para la actualización de V_i

c la cantidad de V que serán calculadas.

El siguiente paso del algoritmos es el cálculo de los $\mu_{i,k}$

$$\mu_{i,k} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{i,k}}{d_{j,k}} \right)^{\frac{2}{m-1}}}, \quad \text{para} \quad d_{j,k} > 0, \forall i,k$$

Calcular todos los elementos de la matriz equivale a:

$H_2 = 4nyc^2+nc$

De donde,

$4nyc^2$ resulta de la sumatoria (que equivale a 4yc) en n*c ocasiones.

nc es la cantidad de elementos de $\mu_{i,k}$ que serán actualizados.

El total de esfuerzo para el algoritmo es entonces:

$H = e(H1 + H2) = e[c(2ny+n+y) + 4nyc^2+nc]$

Donde **e** es la cantidad de épocas[14] establecidas para el algoritmo.

[14] Época: presentación de todos los datos de entrenamiento al algoritmo.

Puede observarse que el esfuerzo total depende de la cantidad de elementos tratados, la cantidad de variables que componen a los elementos, el número de centros buscado y la cantidad de épocas establecidas para el algoritmo.

5.2.3.2 Complejidad de SOM

El cálculo de esta complejidad se basa en el algoritmo SOM presentado en la sección 2.1.2.2 centrándose en los accesos necesarios a datos.

Debe tenerse en cuenta que el algoritmo accede a los siguientes datos:

1. Matriz de datos con **n** filas (número de elementos de la muestra) por **y** variables.
2. Matriz de pesos de las conexiones de la neuronas con **y** filas por **c** columnas (número de neuronas de salida = número de centros).
3. Una matriz de valores para la función de retroalimentación $r_{ik}(n)$ de la neurona i a la neurona *ganadora k* en la época *n*.
4. Un arreglo de valores para α

La búsqueda de la neurona ganadora para un elemento presentado a la red toma

$H_1 = 2cy$

Ya que este proceso significa buscar la neurona más cercana (por distancia euclidiana) al elemento presentado. Cada cálculo de distancia toma 2y accesos (accesos para dos elementos del mismo número de variables) y se busca en las c neuronas que representan los centros de los grupos.

Para actualizar los pesos de las conexiones se usa la siguiente fórmula:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(n) \bullet r_{ik}(n) \bullet (x(t) - w_{ij}(t))$$

Calcular el peso de todos los w_{ij} implica realizar un esfuerzo de:

$H_2 = 5cy$

Ya que cada uno de los elementos de la fórmula anterior significa un acceso a datos y se deben calcular c*y valores de pesos.

Para calcular los valores de retroalimentación se usa la fórmula:

$$r_{ik}(n) = \exp(-\frac{{d_{ik}}^2}{\sigma(n)^2})$$

Calcular todos los valores de la matriz toma

$H_3 = cy(2y+1)$

Ya que el cálculo de la distancia toma 2y, el acceso a $\sigma(n)$ es 1 y este cálculo se realiza para cada uno de los c*y valores.

Para cada época, que es la presentación de todos los elementos a la red, se realiza un esfuerzo de:

$He = n(H_1+H_2) + H_3 = n\,(2cy+5cy) + 2cy^2 + cy = 7ncy + 2cy^2 + cy$

El esfuerzo total del algoritmo viene dado por:

$H = e(7ncy + 2cy^2 + cy)$

5.2.3.3 Esfuerzo realizado

El cálculo del esfuerzo necesario para cada paso del proceso se ha realizado de acuerdo a ciertos supuestos:

1. Los pasos de preprocesamiento de los datos no son tomados en cuenta ya que son comunes a la generación de ambos modelos.
2. La reducción vertical tampoco es tomada en cuenta porque también es común a ambos modelos.
3. La reducción horizontal implica extraer una muestra de filas (en este caso no se toman los valores de las variables si no las filas como entidades). Para cada fila se calcula un valor aleatorio, lo que implica un número de cálculos igual al número de filas. Asumimos que el costo de calcular un valor aleatorio no es significativo.
4. La validación de la muestra implica cálculos en torno a los valores de las variables consideradas, evaluando parejas de variables. Para el caso de estudio se evaluaron 5 parejas de variables en 5 diferentes muestras con 34 diferentes modelos. Como las muestras son excluyentes, probar con las 5 muestras significa probar todas las filas del conjunto original. El total de accesos entonces es igual a "n x v x q", con n igual al número de elementos del conjunto original, v igual al número de variables probadas y q igual al número de modelos probados.
5. La determinación del número de grupos se realiza ejecutando múltiples veces el algoritmo de Fuzzy C-Means. Tomamos la complejidad del algoritmo de acuerdo a lo establecido en la sección 5.2.3.1 y tomando en cuenta que cada ejecución del algoritmo tomó 50 épocas y el algoritmo se ejecutó varias veces con valores diferentes para los centros.
6. El último paso, el agrupamiento, posee una complejidad establecida por la fórmula deducida en la sección 5.2.3.2. Este algoritmo se ejecutó una sola vez.

En la tabla 10 se muestran las estimaciones de esfuerzo para cada paso por cada modelo generado. En este caso, "n" se refiere a los elementos datos (valores de

celdas en la base de datos). Para el caso del Modelo 1 n=80,000 filas por 129 variables, mientras que para el Modelo 2 n=400,000 filas por 129 variables.

Proceso	Modelo 1	Modelo 2
Validación	27,200,000	0
Número de grupos	3.2261E+12	1.61304E+13
Agrupamiento	2168202330	3612334110
Total	3.2283E+12	1.6134E+13

Tabla 10: Cálculos de esfuerzos para Modelo1 y Modelo2

La proporción de trabajo que fue necesario para realizar el segundo modelo, con respecto al primero, corresponde al **20%** (calculado por 1.613E+13/3.2283E+12). Esto significa que con el 20% de esfuerzo se alcanzó un 94% de exactitud en la solución encontrada.

Además, la eficiencia de estos modelos viene dada por:

$$Eficiencia1 = \frac{94}{3.2283\text{E+12}} = 2.9117\text{E-11}$$

$$Eficiencia2 = \frac{100}{1.6134\text{E+13}} = 6.19809\text{E-12}$$

De acuerdo a estos resultados puede determinarse que la eficiencia del primer modelo es mayor a la del segundo (aproximadamente **4.7** veces más eficiente).

5.2.4 Conclusiones de la eficiencia.

1. En cada análisis de eficiencia realizado, el Modelo 1 resulta ser más eficiente que los otros modelos.
2. El modelo es eficiente desde el punto de vista de valores datos (4.7 veces más que el Modelo 2), registros (4.7 veces más que el Modelo 2) y esfuerzo requerido (4.7 veces más que el Modelo 2).
3. El Modelo 1 requiere un esfuerzo significativamente menor al esfuerzo requerido por el Modelo 2 para llegar a una solución aceptable.

A partir de los resultados obtenidos con el análisis de la eficiencia se puede asegurar que la aplicación de la metodología significa la ejecución de un proceso más eficiente en cuanto a uso de recursos datos y a esfuerzo necesario.

Capítulo 6 Conclusiones

Para finalizar este documento y el trabajo de investigación realizado se incluye en este capítulo las conclusiones generales del proyecto y una mención de futuros trabajos que pueden realizarse siguiendo la misma línea de investigación planteada para este proyecto.

6.1 Conclusiones Generales

Luego del trabajo realizado, los resultados obtenidos y la valoración de los mismos, se llega a las siguientes conclusiones:

1. La reducción del espacio de búsqueda es susceptible de arrojar una muestra representativa del conjunto original de datos, lo cual permite obtener modelos válidos de agrupamiento.
2. La aplicación de la metodología propuesta produjo resultados excelentes con un alto grado de exactitud (94%).
3. El proceso seguido demostró ser mucho más eficiente en el uso de recursos datos (4.7 veces) y procesamiento requerido (4.7 veces) que un proceso que trabaja sobre los datos originales.
4. Al sólo haberse necesitado el 6.2% de los datos y un 10% del esfuerzo requerido para obtener una exactitud del 94%, la aplicación de la metodología demostró ser una opción que ofrece una excelente relación costo-beneficio.
5. Se probó que la metodología propuesta es aplicable a un caso de estudio real y por su generalidad y objetividad, sin duda podrá ser aplicada en otros proyectos de minería de datos.

6.2 Trabajos Futuros

Probar el método en otras tareas de minería de datos.

Verificar la aplicabilidad de la metodología en otras fases del proceso de minería de datos (por ejemplo limpieza de datos), y enriquecerla a partir de los resultados que se obtengan.

Realizar pruebas con otros métodos de agrupamiento. Ya que la metodología propuesta se definió de manera tan general que no se encuentra ligada a un algoritmo de agrupamiento específico, lo anterior es factible. Por ejemplo, aplicar

técnicas de agrupamiento de cuantización vectorial (en donde se mezclan los conceptos de los SOMs con la lógica difusa).

Sistematizar el análisis del caso tratado en este trabajo usando técnicas estadísticas que nos permitan determinar el grado de eficiencia del método en general. El análisis de los resultados mostró tan alto grado de eficiencia del proceso seguido que es factible esperar que la dedicación de más esfuerzos al estudio y aplicación de esta metodología sea muy rentable.

En particular, sería deseable determinar el conjunto óptimo de funciones de regresión en las que se basa el proceso de validación.

Se sugiere investigar métricas distintas a la euclidiana aplicada en este trabajo. Por ejemplo, es factible proponer otras métricas de la familia L_i (aquí solamente exploramos L_2). Asimismo, podrían explorarse métricas como la de Mahalanobis, en la que las correlaciones se consideran explícitamente.

Explorar otras métricas o métodos para ser usadas como criterio de paro en la fase de validación de la representatividad de las muestras.

Es posible explorar, también, otros métodos de reducción de la dimensionalidad de los conjuntos de datos (tales como el Análisis de Componentes principales, por ejemplo).

Explorar los resultados con otros métodos de ajuste de curva. Explorar también los resultados al aplicar métodos de regresión múltiple.

Probar la tolerancia de la metodología a la presencia de ruido en los datos de estudio.

Evaluar la posibilidad de incorporar la metodología propuesta a herramientas de software para minería de datos. Desarrollar, si es factible, una herramienta que sirva de guía para el usuario en la aplicación de la metodología.

Bibliografía

[ALEVIZ02] P. Alevizos, B. Boutsinas, D. Tasoulis, M.N. Vrahatis, "Improving the Orthogonal Range Search k-windows Algorithm". Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence. IEEE (2002), 239-245.

[BARALD99-I] Andrea Baraldi; Palma Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognition-Part I". IEEE Transactions on Systems, Man and Cybernetics. IEEE, Vol. 29 Issue 6 (1999), 778-785.

[BARALD99-II] Andrea Baraldi; Palma Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognition-Part II". IEEE Transactions on Systems, Man and Cybernetics. IEEE Vol. 29 Issue 6 (1999), 786-801.

[BENSMA04] H. Bensmail; R. P. De Gennaro, "Cluster Analysis of Imputed Financial Data Using an Augmentation-Based Algorithm". Statistical Data Mining and Knowledge Discovery. Chapman & Hall/CRC Press (2004), 513-528.

[BERKHI02] Pavel Berkhin, "Survey of Clustering Data Mining Techniques". Accrue Software, Inc (2002), 1-56.

[BRIGHT02] Henry Brighton; Chris Mellish, "Advances in Instance Selection for Instance-BasedLearning Algorithms". Data Mining and Knowledge Discovery, Springer, Vol 6. Issue 2 (2002), 153-172.

[CHENG06] David Cheng; Ravi Kannan; Santosh Vempala; Grant Wang, "A Divide-and-Merge Methodology for Clustering". ACM Transactions on Database Systems, ACM, Vol 31. Issue 4 (2006), 196-205.

[CHOW03] Shein-Chung Chow; Jun Shao; Hansheng Wang, "Sample Size Calculations in Clinical Research" 1a edición. Capítulo 1, Taylor & Francis Group (2003).

[COZ06] Juan José del Coz; Jorge Díaz; Antonio Bahamonde; Eric Dransfield; Costas Stamataris; Demetrios Zygoyianis; Tyri Valdimarsdottir; Edi Piasentier; Geoffrey Nute; Alan Fisher, "Learning the Reasons Why Groups of Consumers Prefer Some Food Products". Proceedings of the 6th Industrial Conference on Data Mining. Springer LNAI (2006), 297-309.

[DAGOST86-3] Ralph B. D'Agostino; Michael A. Stephens, "Goodness-of-Fit Techniques" 1a edición. Capítulo 3, Marcel Dekker Inc (1986)

[DAGOST86-5] Ralph B. D'Agostino; Michael A. Stephens, "Goodness-of-Fit Techniques" 1a edición. Capítulo 5, Marcel Dekker Inc (1986)

[DUMOUC03] William DuMouchel; Deepak K. Agarwal, "Application of Sampling and Fractional Factorial Designs to Model-Free Data Squashing.". Proceedings of the international conference on Knowledge discovery and data mining, ACM (2003), 511-516.

[DUMOUC99] William DuMouchel; Chris Volinsky; Theodore Jhonson; Corinna Cortes; Daryl Pregibon, "Squashing Flat Files Flatter". Proceedings of the international conference on Knowledge discovery and data mining, ACM (1999), 6-15

[DUNN73] J.C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters”. Journal of Cybernetics, Vol 3 (1973), 32-57.

[FAYYAD03] Usama M. Fayyad; Gregory Piatetsky-Shapiro; Ramasamy Uthurusamy, "Summary from the KDD-03 panel -- Data Mining: The Next 10 Years". SIGKDD Explorations. ACM ,Vol 5. Issue 2 (2003), 191-196.

[FAYYAD96] Usama Fayyad; Gregory Piatetsky-Shapiro; Padhraic Smyth, "The KDD Process for extracting Useful Knowledge from Volumes of Data". Communications of the ACM ,Vol 39. Issue 11 (1996), 27-34.

[FODOR02] Imola K. Fodor, "A survey of dimension reduction techniques". CiteSeer (2002). citeseer.ist.psu.edu/fodor02survey.html

[FORMAN00] George Forman, Bin Zhang, "Distributed data Clustering can be efficient an exact". ACM SIGKDD Explorations Newsletter. ACM, Vol 2. Issue 2 (2000), 34-38

[FORMAN05] George Forman; Kave Eshghi; Stephane Chiocchetti, “Finding Similar Files in Large Documents Repositories”. Proceedings of the KDD 2005. ACM (2005), 394-400.

[GUHA98] Sudipto Guha; Rajeev Rastogi; Kyuseok Shim, "CURE: An Efficient Clustering Algorithm for Large Databases". Proceedings of the 1998 International Conference on Management of Data, ACM (1998), 73-84.

[HAIR05] Joseph F. Hair, Jr.; Rolph E. Anderson; Ronald L. Tatham; William C. Black, "Análisis Multivariante" 5a edición. Capítulo 4, Pearson. Prentice Hall (2005)

[HAN03] Jiawei Han; Micheline Kamber, "Data Mining: Concepts and Techniques" 1a edición. Capítulo 8, Morgan Kaufmann Publishers(2001)

[HAND01-4] David Hand; Heikki Mannila; Padhraic Smyth, "Principles of Data Mining" 1a edición. Capítulo 4, MIT Press (2001)

[HAND01-12] David Hand; Heikki Mannila; Padhraic Smyth, "Principles of Data Mining" 1a edición. Capítulo 12, MIT Press (2001)

[HAYKIN99-4] Simon Haykin, "Neural Networks. A Comprehensive Foundation." 2a edición. Capítulo 4, Prentice Hall (1999)

[HAYKIN99-9] Simon Haykin, "Neural Networks. A Comprehensive Foundation" 2a edición. Capítulo 9, Prentice Hall (1999)

[HERNAN04] José Hernández Orallo; Ma. José Ramírez Quintana; César Ferri Ramírez, "Introducción a la Minería de Datos" 1a edición. Capítulo 15, Pearson. Prentice Hall (2004).

[HUSSAI02] Huan Liu; Farhad Hussain; Chew Lim Tan; Manoranjan Dash, "Discretization: An Enabling Technique". Data Mining and Knowledge Discovery, Springer, Vol. 6 Issue 4 (2002), 393-423.

[JAGADI99] H. V. Jagadish; Laks V. S. Lakshmanan; Divesh Srivastava, "Snakes and Sandwiches: optimal Clustering Strategies for a Data Warehouse". Proceedings of the 1999 International Conference on Managament of Data, ACM (1999), 37-48.

[JAIN99] A.K. Jain; M.N. Murty; P.J. Flynn, "Data Clustering: A Review". ACM Computings Surveys, ACM, Vol 31. Issue 3 (1999), 264-323.

[KANELL06] Yiannis Kanellopoulos; Thimios Dimopulos; Christos Tjortjis; Christos Makris, “Mining Source Code Elements for Comprehending Object-Oriented Systems and Evaluating their Maintainability”. SIGKDD Explorations. ACM, Vol. 8 Issue 1 (2006), 33-40

[KASABO98] Nikola K. Kasabov, "Foundations of Neural Networks, Fuzzy Systems, and" 1a edición. Capítulo 3, MIT Press (1998).

[KEARNS96] Michael Kearns, "A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split". Advances in Neural Information Processing Systems. MIT Press, Vol. 8 (1996), 183-189.

[KEIM99] Daniel A. Keim; Alexander Hinneburg, “Tutorial 3. Clustering Techniques for Large Data Sets - From the Past to the Future”. Tutorial notes of the 5th ACM SIGKDD, ACM (1999), 141-181.

[KOHONE81] Teuvo Kohonen, “Automatic Formation of Topological Maps of Patterns in a Self-Organizing System”. Proceedings of the 2nd International Conference on Image Analysis (1981), 214-220.

[KRISHN99] K. Krishna and M. Narasimha Murty, "Genetic K-means algorithm". IEEE Transactions on Systems, Man, and Cybernetics. IEEE, Vol 29. Issue 3 (1999), 433-439.

[KURI07] Ángel Kuri-Morales; Fátima Rodríguez, "A Search Space Reduction Methodology for Large Databases: A Case Study". Proceedings of the 7th Industrial Conference on Data Mining. Springer LNAI (2007), 199-214.

[LIU02] Huan Liu; Hiroshi Motoda, "On Issues of Instance Selection". Data Mining and Knowledge Discovery. ACM, Vol 6. Issue 2 (2002), 115-130.

[LIU05] Huan Liu; Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering". IEEE Transactions on knowledge and Data Engineering, Vol 17. Issue 4 (2005), 491-502.

[MYATT07] Glenn J. Myatt, "Making Sense of Data. A Practical Guide to Exploratory Data Analysis and Data Mining" 1a edición. Capítulo 5, John Willey & Sons (2007)

[OWEN03] Art Owen, "Data Squashing by Empirical Likelihood". Data Mining and Knowledge Discovery, Springer, Vol 7. Issue 1 (2003), 101-113.

[PADMAN03] Sriram Padmanabhan; Leslie Cranston, "Multi-Dimensional Clustering: A New Data Layout Scheme in DB2". Proceedings of the 2003 International Conference on Management of Data, ACM (2003), 637-641.

[PALMER02] Christopher R. Palmer; Christos Faloutsos, "Density Biased Sampling: An Improved Method for Data Mining and Clustering". Proceedings of the 2006 International Conference on Management of Data, ACM (2000), 82-92.

[PETER03] William Peter; John Chiochetti; Clare Giardina, "New Unsupervised Clustering Algorithm for Large Datasets". Proceedings of the ninth International Conference on Knowledge Dicovery and Data Mining, ACM (2003), 643-648.

[RAYMON94] Raymond T. Ng; Jiawei Han, "Efficient and Effective Clustering Methods for Spatial Data Mining". Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers (1994), 144-155.

[RICCAR98] G.A. Riccardi; P.H. Schow, "Adaptation of the ISODATA clustering algorithm for vector supercomputer execution". Proceedings of the Supercomputing 88. IEEE Vol. 2 (1988), 141-150.

[SAS06] SAS, Proceedings of the Second Workshop on Feature Selection for Data Mining: Interfacing Data Mining and Statistics (2006).

[SEIFER04] Jeffrey W. Seifert, "Data Mining: An Overview". Congressional Research Service-The Library of Cong (2004)

[TASOUL03] Dimitris K. Tasoulis Panagiotis D. Alevizos, Basilis Boutsinas, and Michael N. Vrahatis, "Parallel Unsupervised k-Windows: An Efficient Parallel Clustering Algorithm". Springer LNCS, Vol 2763/2003 (2003), 336-344.

[TSAI02] Cheng-Fa Tsai; Zhi-Cheng Chen; Chun-Wei Tsa, "MSGKA: an efficient clustering algorithm for large databases". IEEE International Conference on Systems, Man and Cybernetics. IEEE, Vol 5 (2002).

[VAIDYA03] Jaideep Vaidya, Chris Clifton, "Privacy Preserving K-Means Clustering over Vertically Partitioned Data". Proceedings of the ninth International Conference on Knowledge Discovery and Data Mining. ACM (2003), 206-215.

[VU06] Khanh Vu; Kien A. Hua; Hao Cheng; Sheau-Dong Lang, "A Non-Linear Dimensionality-Reduction Technique for Fast Similarity Search in Large Databases". Proceedings of the 2006 International Conference on Management of Data, ACM (2006), 527-538.

[WANG03] John Wang, "Data Mining: Opportunities and Challenges" 1a edición. Capítulo 14, Idea Group, Inc (2003)

[YANG06] Quiang Yang; Xindong Wu, "10 Challenging problems in Data Mining Research". International Journal of Information Technology, Vol. 5, No. 4 (2006), 597-604.

[YANG99] Yiling Yang; Xudong Wan; Jinyuan You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data". Proceedings of the eighth ACM SIGKDD, ACM (2002), 682-687.

[ZHANG07] Daoqiang Zhang; Zhi-Hua Zhou; Songcan Chen, "Semi-Supervised Dimensionality Reduction". Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM (2007).

[ZHANG96] Tian Zhang, Raghu Rama krishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases". Proceedings of the 1996 International Conference of Management of Data, ACM (1996), 103-114.

[ZHO03] Bing Zho, Jun-Yi Shen, Qin-Ke Peng, "PARCLE: A Parallel Clustering Algorithm for Cluster System". Proceedings of the International Conference on Machine Learning and Cybernetics. IEEE, Vol 1 (2003), 4-8.

[ZHU06] Xingquan Zhu; Xindong Wu, "Scalable Representative Instance Selection and Ranking". Proceeding of the 18th international conference on pattern recognition, IEEE (2006), 352-355.

Referencias Web

[1] http://www.fizyka.umk.pl/publications/kmk/06-CIdef.pdf

[2] http://www.terra.es/tecnologia/glosario/ficha.cfm?id_termino=1417

[3] http://www.lsi.us.es/redmidas/IIreunion/trans/prepro.pdf

[4] http://www.crisp-dm.org

[5] http://edis.ifas.ufl.edu/PD006

[6] http://en.wikipedia.org/wiki/Data_analysis

[7] http://www.vias.org/tmdatanaleng/cc_what_is_it.html

[8] http://www.itch.edu.mx/academic/industrial/estadistica1/cap04.html#cuatro_pruebas_chi

[9] http://mundomatematico.webcindario.com/historia/textos/Combinatoria.pdf

Anexo 1: Modelos matemáticos

La siguiente tabla presenta 34 diferentes modelos matemáticos utilizados para pruebas de bondad de ajuste.

Familia	Modelo	Ecuación
	Lineal	$y = a + bx$
	Cuadrático	$y = a + bx + cx^2$
	Polinomio de Orden n	$y = a + bx + cx^2 + dx^3 + \ldots$
Familia Exponencial	Exponencial	$y = ae^{bx}$
	Exponencial modificado	$y = ae^{b/x}$
	Logarítmico	$y = a + b\ln x$
	Log Recíproco	$y = \frac{1}{a + b\ln x}$
	Modelo de Presión de Vapor	$y = e^{a+b/x+c\ln x}$
Familia de Potencias	Potencia	$y = ax^b$
	Potencia Modificada	$y = ab^x$
	Potencia Desplazada	$y = a(x-b)^c$
	Geométrico	$y = ax^{bx}$
	Geométrico Modificado	$y = ax^{b/x}$
	Raíz	$y = ab^{1/x}$
	Modelo de Hoerl	$y = ab^x x^c$
	Modelo Modificado de Hoerl	$y = ab^{1/x} x^c$
Modelos basados en densidad	Recíproco	$y = \frac{1}{ax + b}$
	Recíproco Cuadrático	$y = \frac{1}{a + bx + cx^2}$
	Modelo de Bleasdale	$y = (a + bx)^{-1/c}$
	Modelo de Harris	$y = \frac{1}{(a + bx^c)}$
Modelos Crecientes	Tasa de Saturación	$y = \frac{ax}{b + x}$

Familia	Modelo	Ecuación
	Asociación Exponencial 2	$y = a(1 - e^{-bx})$
	Asociación Exponencial 3	$y = a(b - e^{-cx})$
Modelos Sigmoidales	Relación de Gompertz	$y = ae^{-e^{b-cx}}$
	Logístico	$y = \frac{a}{1 + be^{-cx}}$
	Modelo de Richards	$y = \frac{a}{(1 + e^{b-cx})^{1/d}}$
	Modelo MMF	$y = \frac{ab + cx^d}{b + x^d}$
	Modelo de Weibull	$y = a - be^{-cx^d}$
Otros	Hiperbólico	$y = a + b/x$
	Sinusoidal	$y = a + b\cos(cx + d)$
	Capacidad de Calor	$y = a + bx + c/x^2$
	Gaussiano	$y = ae^{\frac{-(x-b)^2}{2c^2}}$
	Función Racional	$y = \frac{a + bx}{1 + cx + dx^2}$

Anexo 2: Programa para cálculo de valores de confianza

```
package comb;

public class Combinar {
// Clase principal del programa

    public static void main(String[] args) {
       // Programa principal
       // Recibe dos argumentos para su ejecución
       int m; //Primer parámetro: cantidad de muestras a evaluar
       int q; //Segundo parámetro: cantidad de modelos a evaluar
       int total;
       int comSimilar;
       Calculador cal;

       if (args.length>=2){
          m=Integer.parseInt(args[0]);
          q=Integer.parseInt(args[1]);
          if ((m>q)||(q<=0)||(m<=0))
             System.out.print("***Valores de parámetros incorrectos.");
          else {
             cal=new Calculador();
             total=cal.combinat(q+m-1, m);
             System.out.println("****************************************");
             System.out.println("| t | Ajustes -->  Probabilidad ");
             //El siguiente ciclo se encarga de extraer todos los valores
             //para las probabilidades de ajustes
             //t representa el número de ajustes iguales en una prueba de m
             //muestras
             for (int t=1;t<=m;t++){
                comSimilar=cal.similares(t, m, q);
                System.out.print("| "+Integer.toString(t)+" | ");
                System.out.print(Integer.toString(comSimilar)+"  -->  ");
                System.out.println(Double.toString((double)comSimilar/
                (double)total));
             }
             System.out.println("****************************************");
             System.out.println(" Combinaciones ="+Integer.toString(total));
             System.out.println("****************************************");
          }
       }
       else
          System.out.print("***No ha indicado suficientes parámetros***");
    }
}
```

```
import java.math.BigInteger;

public class Calculador {
//Clase que se encarga de realizar todos los cálculos matemáticos.

   public int contar(int suma, int num, int pos){
   //Función que se encarga de contar el número de combinaciones de pos
   //números menores o iguales a num que suman suma
      int m=0;
      suma=suma-num;
      m=pos-1;
      if ((suma<m)||(m<0))
         return 0;
      else{
         if ((suma==0)&& (m==0)){
            return 1;
         }
         for (int i=num;i>0;i--){
            if (contar(suma, i, m)>0)
               return 1;
         }
         return 0;
      }
   }

   public int calculoV(int t, int h, int pos){
      int com=0;
      for (int j=t;j>0;j--){
         com=com+this.contar(h,j,pos);
      }
      return com;
   }

   public int combinat(int num, int pos){
      //Función para extraer un combinatorio
      int valor=0;
      if (num>=pos)
         valor=this.factorial(num).divide(this.factorial(num-
         pos).multiply(this.factorial(pos))).intValue();
      return valor;
   }

   private BigInteger factorial(int num){
      if ((num>1))
         return factorial(num-1).multiply(new
         BigInteger(Integer.toString(num)));
      return new BigInteger("1");
   }

   public int similares(int t, int m, int q)
{     //Función que se encarga de implementar la fórmula para
      //determinación de ajustes similares
      int similar=0;
      if ((t>0) && (t<=m) && (q>0)){
```

```
        if (t==1){
          similar=this.combinat(q, m);
        }
        else{
          if (t==m){
            similar=q;
          }
          else{
            int h;
            int limh;
            h=m-t;
            if (h % t==0)
                  limh=h /t;
            else
                  limh=h / t +1;

            for (int i=limh;i<=h;i++){
                  int com=this.calculoV(t, h, i);
                  similar=similar+com*this.combinat(q-1, i);
            }
            similar=q*similar;
          }
        }
      }

      return similar;
      }
}
```

Resultado para m=5 y q=34

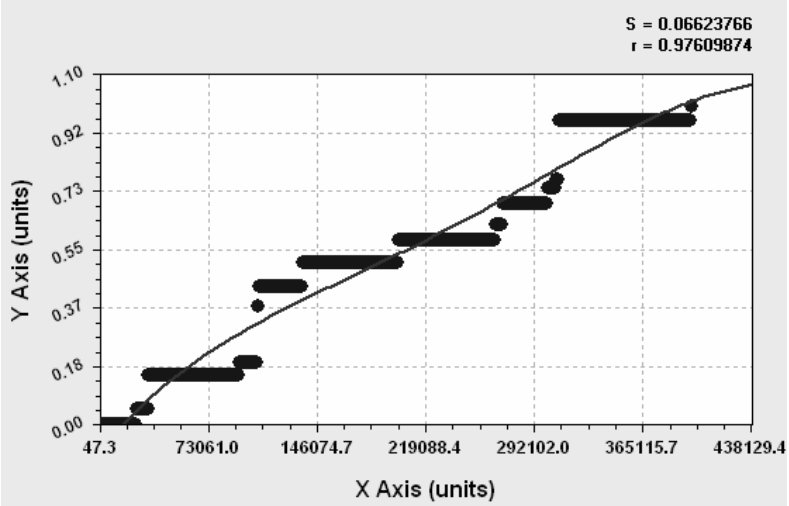
```
************************************************
| t | Ajustes -->  Probabilidad
| 1 | 278256  -->  0.5543588701483438
| 2 | 203456  -->  0.40533766849556324
| 3 | 19074  -->  0.038000406421459056
| 4 | 1122  -->  0.002235318024791709
| 5 | 34  -->  6.7736909842173E-5
************************************************
 Combinaciones = 501942
************************************************
```

Anexo 3: Matriz de correlaciones

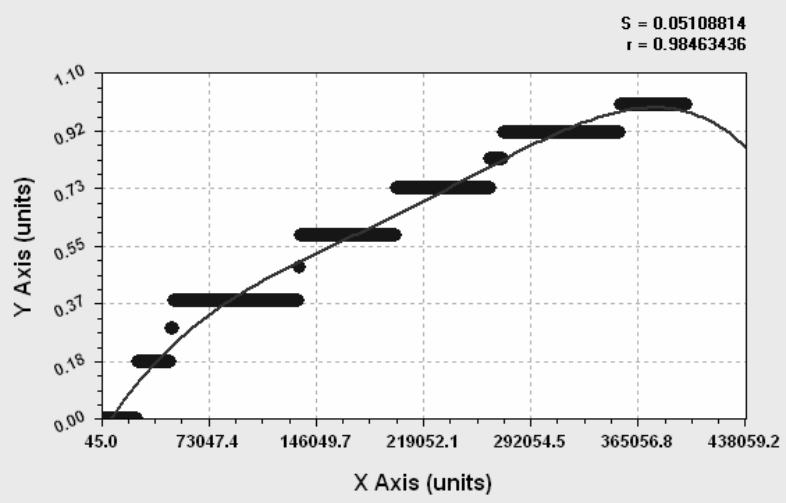
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
V1	1.000	0.008	0.006	0.012	0.007	-0.019	-0.147	0.019	0.005	-0.010	0.003	0.004
V2	0.008	1.000	-0.100	0.015	0.001	0.026	-0.026	0.034	0.006	0.059	0.023	0.008
V3	0.006	-0.100	1.000	0.092	0.113	-0.198	-0.195	0.038	0.017	0.005	0.035	0.013
V4	0.012	0.015	0.092	1.000	0.027	-0.057	-0.012	-0.012	0.004	-0.021	0.015	0.003
V5	0.007	0.001	0.113	0.027	1.000	-0.144	-0.055	0.010	0.006	-0.029	0.013	0.006
V6	-0.019	0.026	-0.198	-0.057	-0.144	1.000	0.118	0.092	-0.001	0.407	0.040	0.000
V7	-0.147	-0.026	-0.195	-0.012	-0.055	0.118	1.000	-0.006	-0.005	0.036	-0.004	-0.007
V8	0.019	0.034	0.038	-0.012	0.010	0.092	-0.006	1.000	-0.006	-0.022	0.004	0.005
V9	0.005	0.006	0.017	0.004	0.006	-0.001	-0.005	-0.006	1.000	-0.012	-0.003	0.008
V10	-0.010	0.059	0.005	-0.021	-0.029	0.407	0.036	-0.022	-0.012	1.000	0.069	-0.005
V11	0.003	0.023	0.035	0.015	0.013	0.040	-0.004	0.004	-0.003	0.069	1.000	0.001
V12	0.004	0.008	0.013	0.003	0.006	0.000	-0.007	0.005	0.008	-0.005	0.001	1.000
V13	0.001	0.036	-0.012	-0.040	-0.037	0.332	0.045	0.346	0.063	0.599	0.179	0.056
V14	0.003	0.040	-0.023	-0.028	-0.028	0.217	0.006	0.124	0.020	0.347	0.052	0.016
V15	0.009	0.017	0.020	-0.007	0.005	0.051	-0.001	0.507	0.064	-0.002	0.002	0.001
V16	0.937	0.046	-0.022	-0.035	-0.039	0.364	0.028	-0.004	0.881	0.690	0.044	-0.006
V17	0.007	0.889	0.022	-0.002	0.009	0.046	0.008	0.265	0.065	0.801	0.087	0.014
V18	0.010	0.016	0.952	-0.006	0.007	0.051	-0.004	0.542	0.073	-0.011	0.906	0.005
V19	0.004	0.010	0.013	0.971	0.006	0.000	-0.007	0.010	0.006	-0.010	-0.002	0.938
V20	0.001	0.037	-0.022	-0.027	0.820	0.200	0.010	0.120	0.018	0.329	0.052	0.017
V21	0.012	0.018	0.022	-0.005	0.005	0.976	-0.004	0.514	0.074	-0.009	0.009	0.004
V22	0.007	0.014	0.022	-0.002	0.007	0.036	0.962	0.239	0.062	-0.038	0.081	0.014
V23	0.006	0.007	0.018	0.005	0.006	-0.003	-0.007	0.857	0.777	-0.013	-0.003	0.028
V24	0.007	0.010	0.012	0.003	0.006	-0.001	-0.007	0.018	0.003	-0.011	-0.001	0.630
V25	0.003	0.037	-0.026	-0.027	-0.027	0.209	0.006	0.120	0.014	0.307	0.049	0.013
V26	0.010	0.019	0.018	-0.004	0.007	0.045	-0.009	0.464	0.067	-0.012	0.010	0.004
V27	0.007	0.010	0.012	0.003	0.007	-0.003	-0.007	0.027	0.018	-0.011	0.001	0.545
V28	0.003	0.040	-0.021	-0.027	-0.028	0.224	0.006	0.130	0.015	0.312	0.054	0.019
V29	0.008	0.017	0.015	-0.004	0.007	0.045	-0.005	0.463	0.068	-0.009	0.017	0.006
V30	0.008	0.012	0.015	-0.003	0.008	0.022	0.000	0.196	0.056	-0.035	0.070	0.014
V31	0.016	0.028	0.027	-0.007	0.011	0.054	-0.012	0.719	-0.001	-0.015	0.050	0.008
V32	0.007	0.020	0.031	0.014	0.013	0.022	-0.008	0.010	-0.001	0.021	0.719	0.005
V33	0.012	0.020	0.017	-0.004	0.007	0.037	-0.006	0.441	0.068	-0.010	0.021	0.006
V34	0.009	0.006	0.017	0.007	0.007	-0.009	-0.008	0.014	0.585	-0.012	-0.002	0.026
V35	0.008	0.007	0.010	0.004	0.007	-0.005	-0.008	0.035	0.022	-0.011	0.001	0.472
V36	0.002	0.036	-0.022	-0.026	-0.028	0.218	0.006	0.123	0.011	0.284	0.045	0.017
V37	0.021	0.019	0.013	-0.005	0.008	0.012	-0.016	0.521	0.000	-0.004	0.080	0.006
V38	0.009	0.004	0.8323	0.008	0.008	-0.019	-0.012	0.010	0.391	-0.011	0.075	0.018
V39	0.011	0.008	0.008	0.002	0.005	-0.012	-0.008	0.031	0.018	-0.010	0.001	0.361
V40	0.003	0.037	-0.019	-0.031	-0.024	0.781	0.008	0.135	0.013	0.801	0.047	0.023
V41	0.013	0.010	0.010	-0.003	0.004	0.013	-0.007	0.298	0.043	-0.006	0.038	0.006
V42	0.009	0.018	0.024	0.012	0.012	0.015	-0.009	0.013	0.000	0.027	0.486	0.007
V43	0.004	0.040	-0.019	-0.029	-0.022	0.231	0.007	0.134	0.020	0.280	0.049	0.883
V44	0.006	0.010	0.012	-0.001	0.009	-0.012	0.004	0.142	0.042	-0.035	0.061	0.011

Anexo 4. Gráficas de comparación de variables.

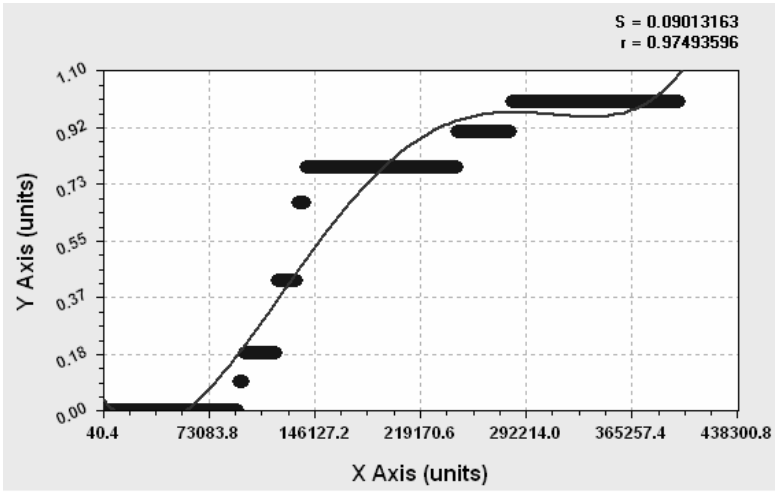
1. Análisis de variables V1 y V10



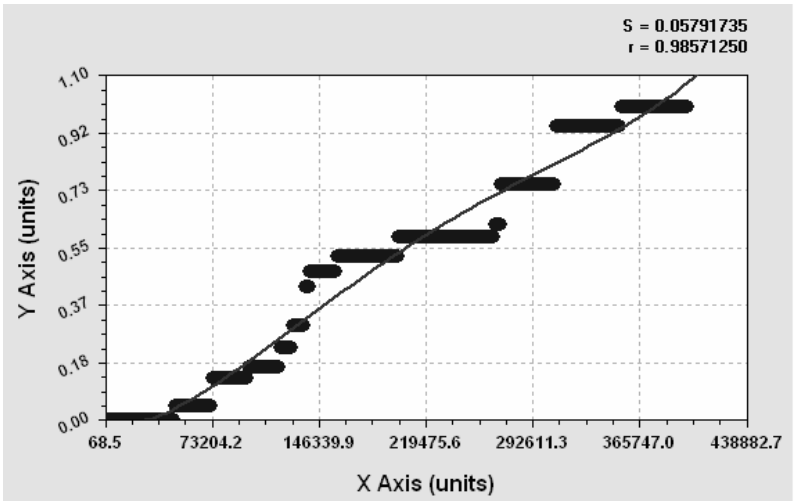
Muestra: 1
Modelo:
Polinomio de 4° grado
S = 0.066
r = 0.976



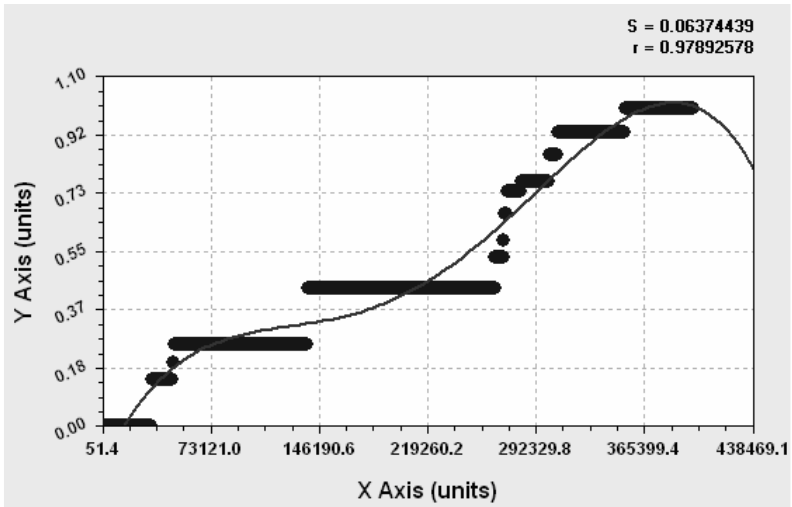
Muestra: 2
Modelo:
Polinomio de 4° grado
S = 0.051
r = 0.985



Muestra: 3
Modelo:
Polinomio de 4° grado
S = 0.090
r = 0.975

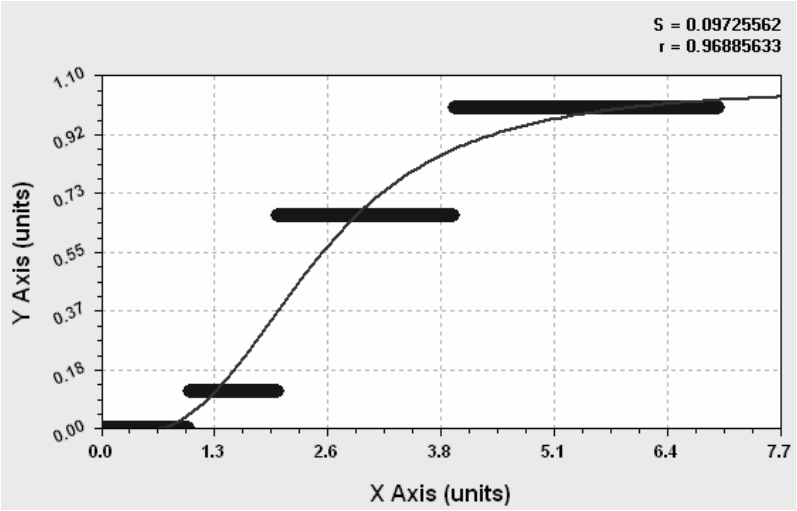


Muestra: 4
Modelo:
Polinomio de 4° grado
S = 0.058
r = 0.985

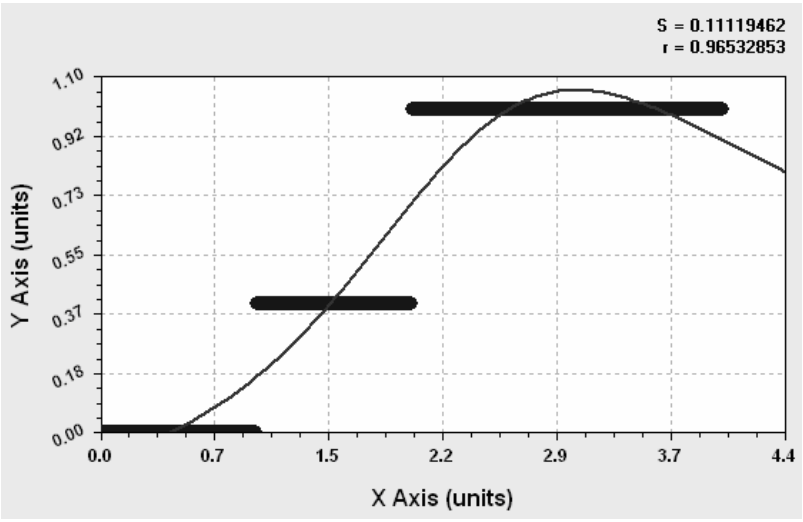


Muestra: 5
Modelo:
Polinomio de 4° grado
S = 0.063
r = 0.979

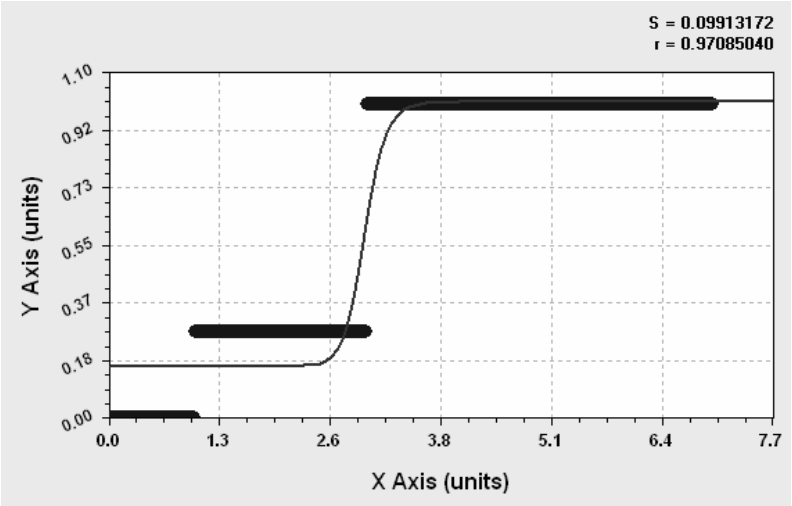
2. Análisis de variables V2 y V20.



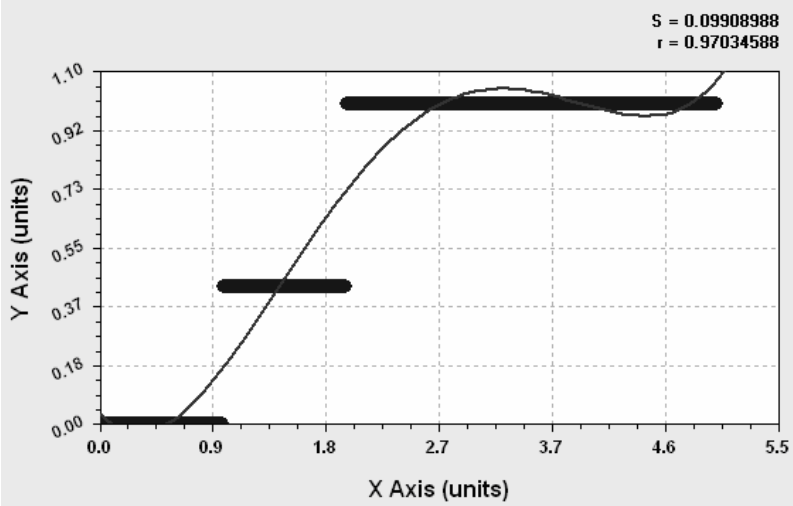
Muestra: 1
Modelo: MMF
S = 0.097
r = 0.968



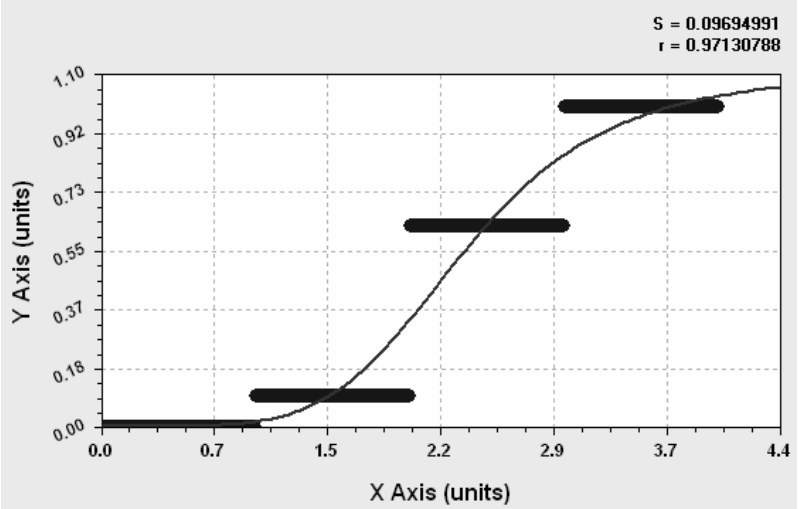
Muestra: 2
Modelo: Función Racional
S = 0.111
r = 0.965



Muestra: 3
Modelo: MMF
S = 0.099
r = 0.971

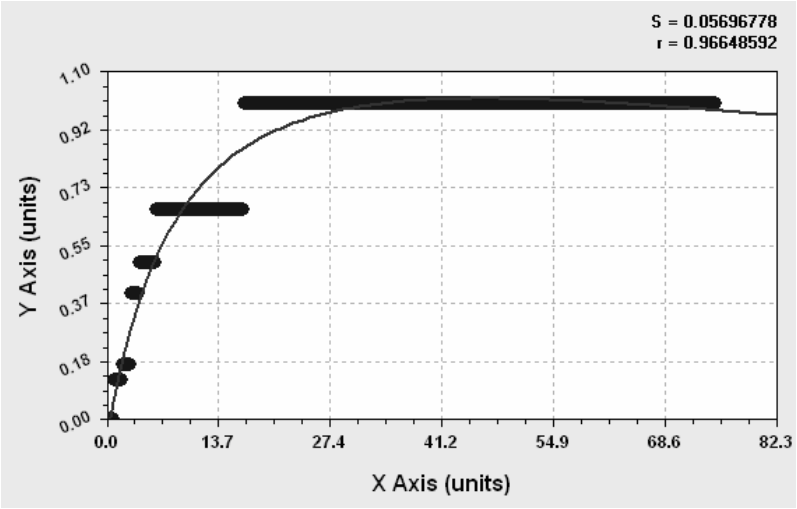


Muestra: 4
Modelo:
Polinomio de 4° grado
S = 0.099
r = 0.970

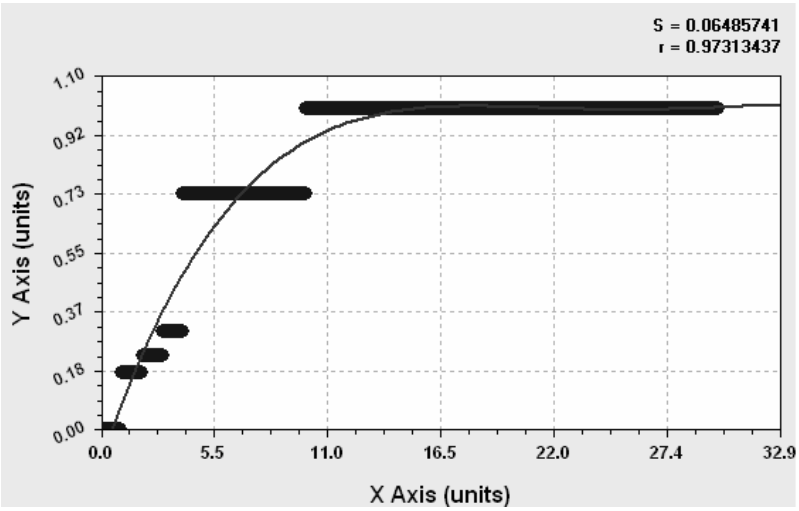


Muestra: 5
Modelo: MMF
S = 0.097
r = 0.971

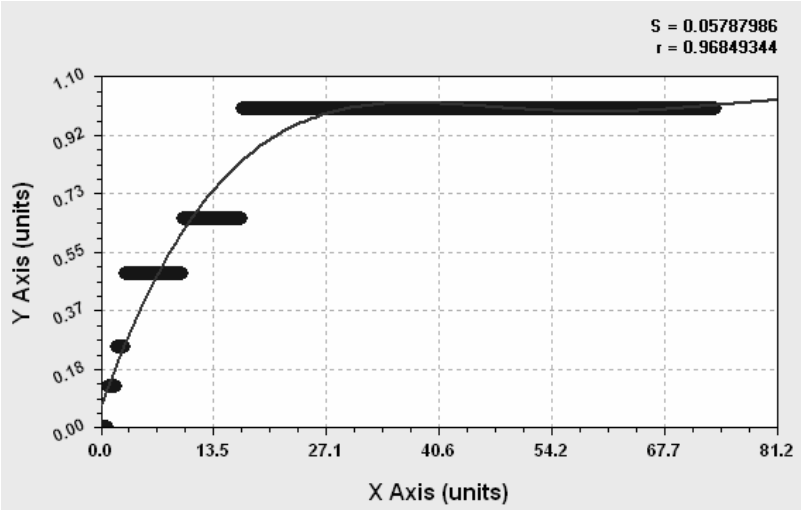
3. Análisis de variables V3 y V30.



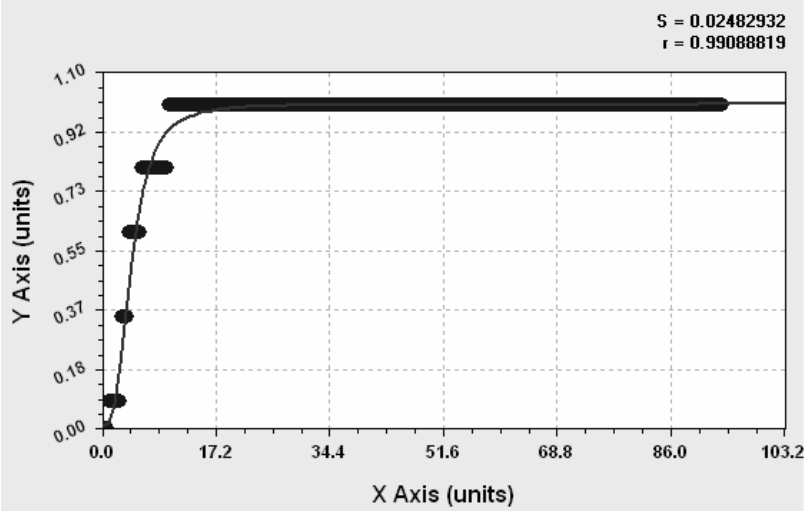
Muestra: 1
Modelo: Función Racional
S = 0.057
r = 0.966



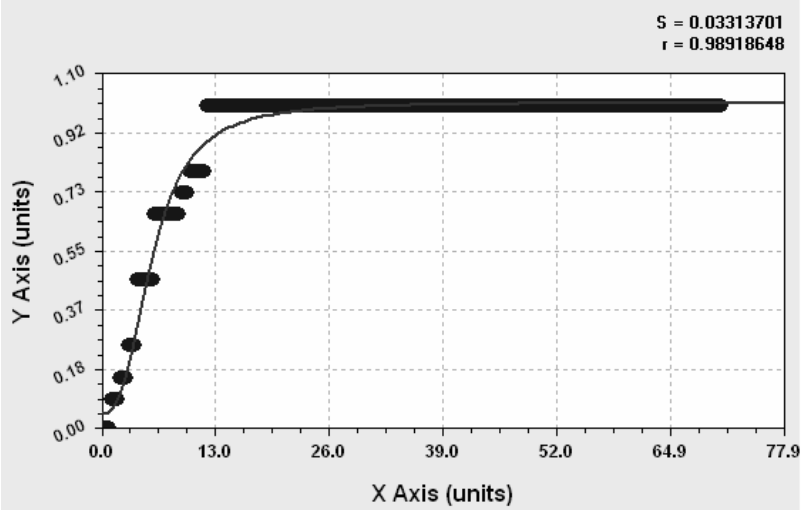
Muestra: 2
Modelo:
Polinomio de 4° grado
S = 0.065
r = 0.973



Muestra: 3
Modelo:
Polinomio de 4° grado
S = 0.058
r = 0.968

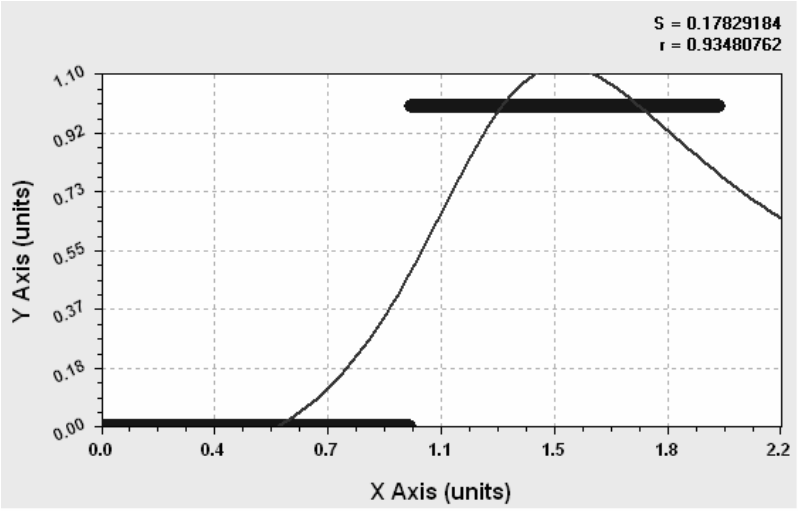


Muestra: 4
Modelo: MMF
S = 0.025
r = 0.991

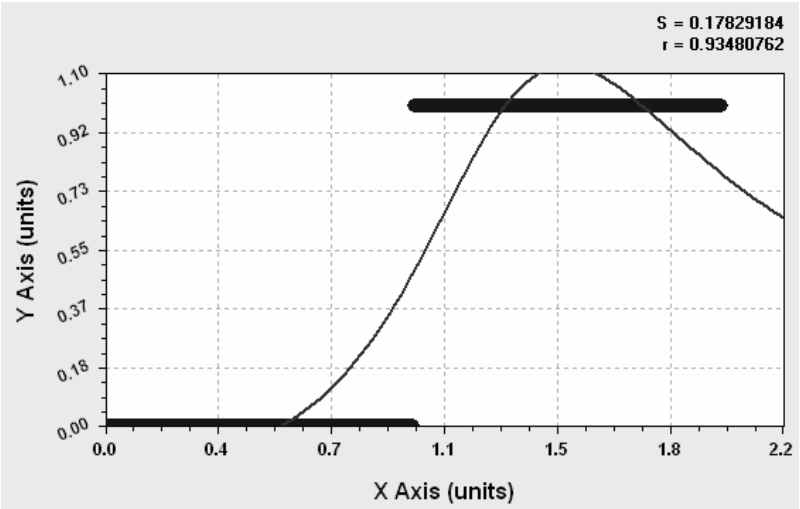


Muestra: 5
Modelo: MMF
S = 0.033
r = 0.989

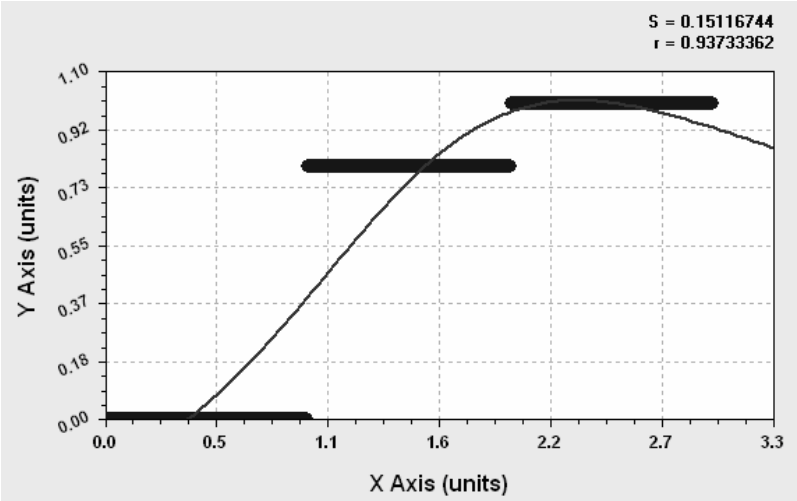
4. Análisis de variables V4 y V40.



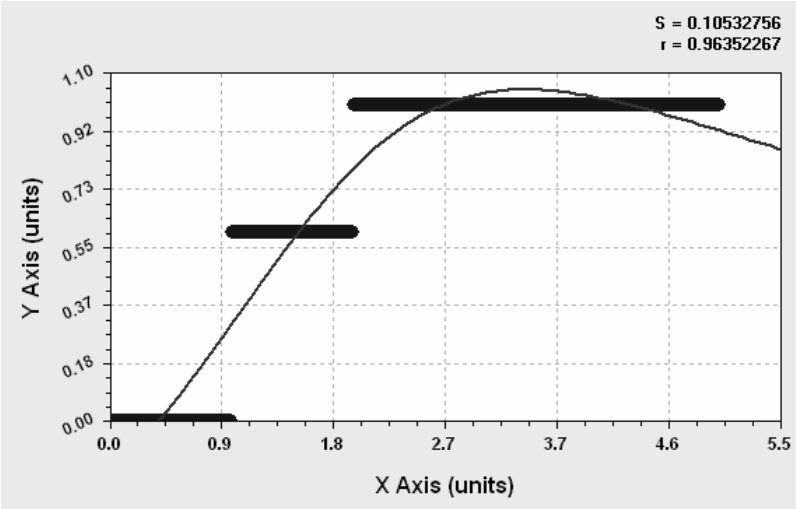
Muestra: 1
Modelo: Función Racional
S = 0.178
r = 0.935



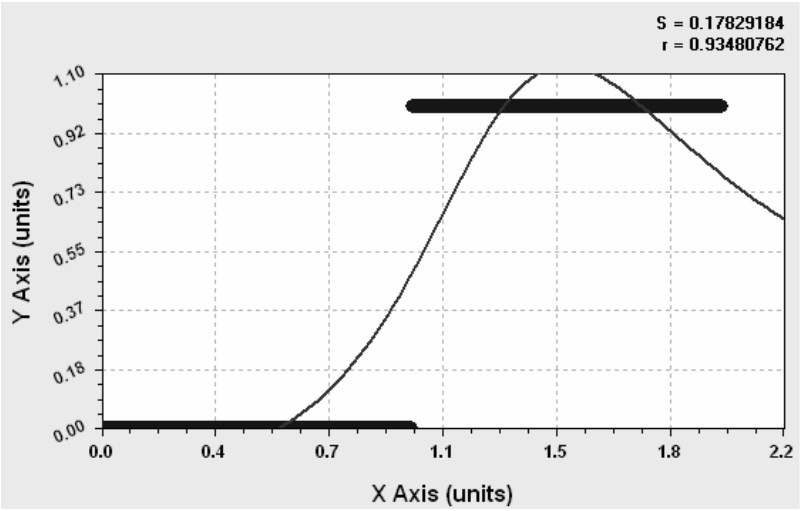
Muestra: 2
Modelo: Función Racional
S = 0.178
r = 0.935



Muestra: 3
Modelo: Función Racional
S = 0.151
r = 0.937

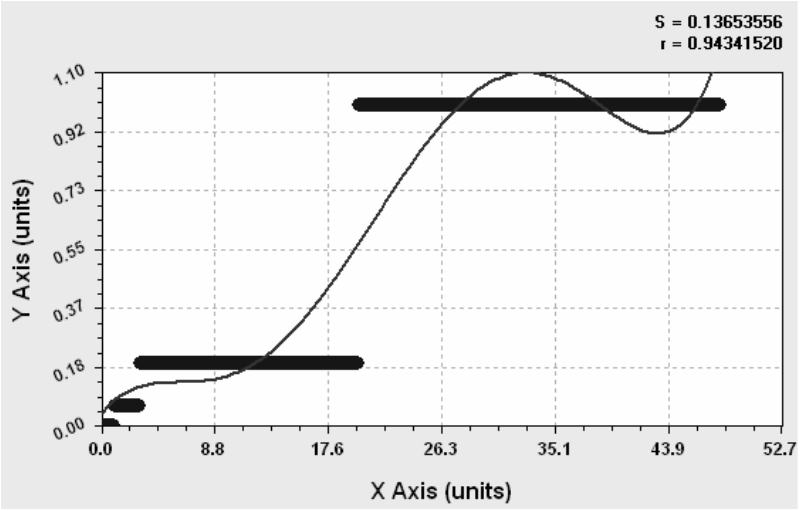


Muestra: 4
Modelo: Función Racional
S = 0.105
r = 0.963

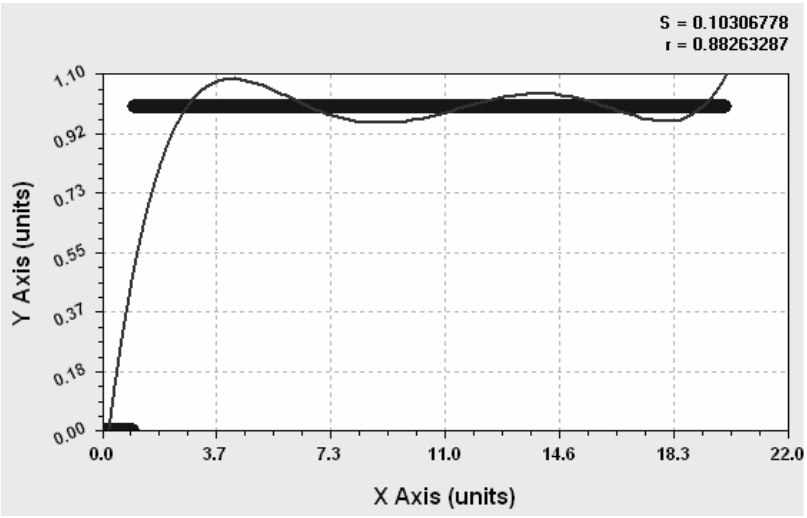


Muestra: 5
Modelo: Función Racional
S = 0.178
r = 0.935

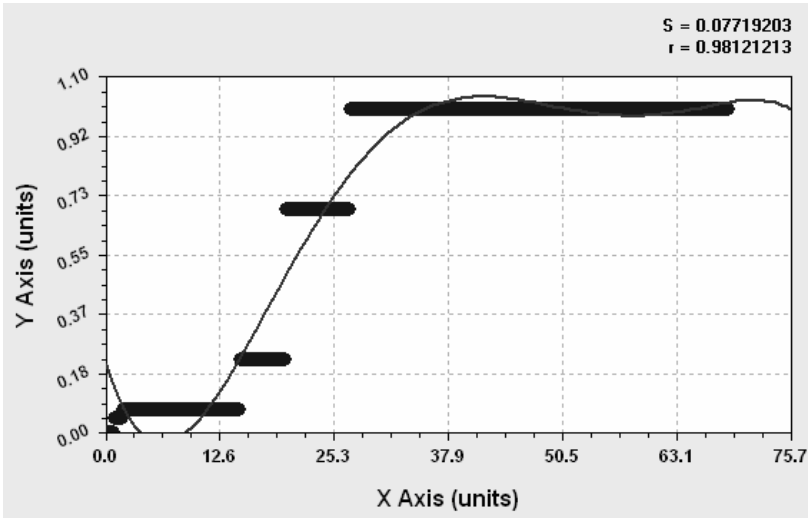
5. Análisis de variables V5 y V50.



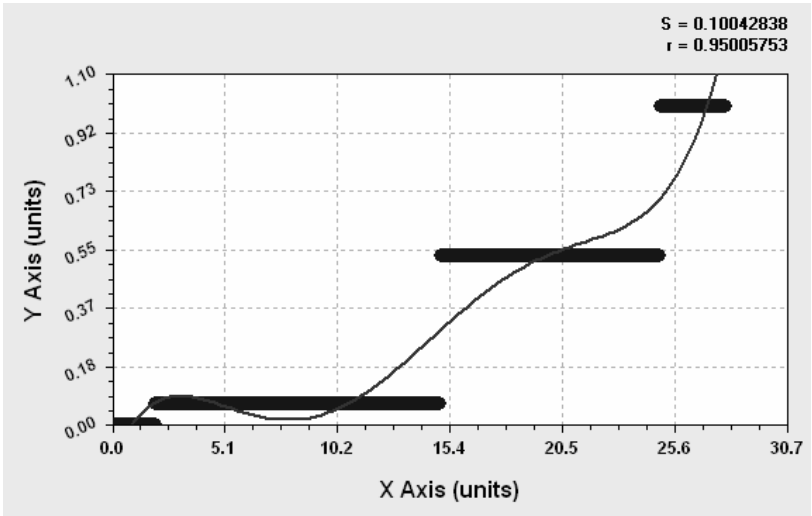
Muestra: 1
Modelo:
Polinomio de 5° grado
S = 0.136
r = 0.943



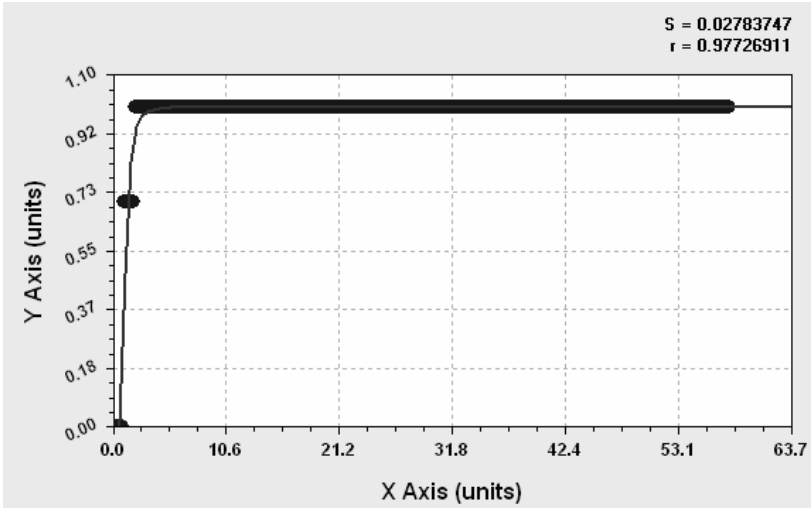
Muestra: 2
Modelo:
Polinomio de 5° grado
S = 0.103
r = 0.883



Muestra: 3
Modelo:
Polinomio de 5° grado
S = 0.077
r = 0.981



Muestra: 4
Modelo:
Polinomio de 5° grado
S = 0.100
r = 0.950



Muestra: 5
Modelo: MMF
S = 0.028
r = 0.977