Biblioteca Digital
F C E N - U B A

BIBLIOTECA CENTRAL LUIS F LELOIR
BIBLIOTECA CENTRAL LELOIR
FACULTAD DE CIENCIAS EXACTAS Y NATURALES UBA

Tesis de Maestría

Comparación interactiva de modelos de minería de datos utilizando técnicas de visualización

Padua, Luciana María

2014-07-24

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:
Padua, Luciana María. (2014-07-24). Comparación interactiva de modelos de minería de datos utilizando técnicas de visualización. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:
Padua, Luciana María. "Comparación interactiva de modelos de minería de datos utilizando técnicas de visualización". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2014-07-24.

EXACTAS UBA
Facultad de Ciencias Exactas y Naturales

UBA
Universidad de Buenos Aires

Dirección: Biblioteca Central Dr. Luis F. Leloir, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Intendente Güiraldes 2160 - C1428EGA - Tel. (++54 +11) 4789-9293

Contacto: digital@bl.fcen.uba.ar

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Facultad de Ingeniería

Comparación Interactiva de Modelos de Minería de Datos Utilizando Técnicas de Visualización

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires en Explotación de Datos y Descubrimiento del Conocimiento

Licenciada Luciana María Padua

Director de Tesis: Claudio Delrieux

Buenos Aires, 2014

Comparación Interactiva de Modelos de Minería de Datos Utilizando Técnicas de Visualización

Resumen

El siguiente trabajo presenta una metodología para la generación automática de múltiples árboles de decisión y la posterior evaluación de los resultados en forma visual e interactiva. El foco principal reside en el análisis de la incidencia de los parámetros de configuración de árboles de decisión en los resultados de los múltiples modelos generados. El objetivo del mismo consiste en contribuir a la creciente disciplina de Visual Analytics, la cual es en una combinación de técnicas de Visualización de la Información y Minería de Datos con el fin de involucrar a las personas en los procesos de aprendizaje automático. El enfoque presentado se basa en herramientas disponibles en el mercado, sugiriendo la adaptación de las mismas para lograr los objetivos de interacción del usuario. Con el fin de demostrar la potencialidad de la metodología se presenta al finalizar un caso de uso real de una empresa de Telecomunicaciones.

Palabras Clave: visualización, minería de datos, árboles de decisión, visual analytics, exploración espacio de parámetros.

Interactive Comparison of Data Mining Models using Visualization Techniques

Abstract

The following work presents a methodology to automatically generate multiple decision trees and to evaluate their results in an interactive and visual way. The focal point is the analysis of the influence of the configuration of decision trees' parameters in the several results generated. The goal is to contribute to the growing Visual Analytics discipline, which consists on a combination of Information Visualization and Data Mining techniques that aims to involve people in automatic processes. The developed approach is based on available software solutions in the market, suggesting their modifications to achieve human interaction. To demonstrate the application of the methodology we present a real use case from a Telecommunications company.

Key Words: Visualization, Data Mining, Decision Trees, Visual Analytics, Parameter Space Exploration.

Agradecimientos

Quiero agradecerle a mucha gente que directa o indirectamente me ayudaron a realizar esta tesis. Fue un camino largo, con tropezones y caídas, pero que finalmente llega a su fin y me llena de alegría.

A mi director Claudio Delrieux quien me ayudó a poner foco y llevar las ideas a la práctica, generando los espacios necesarios para la discusión a pesar de la distancia.

Al Doctor Krešimir Matković con quien tuve el agrado de publicar mi primer paper en una revista científica y que trabajó arduamente en la adaptación de su herramienta para aplicarla al análisis de árboles de decisión.

A las autoridades de la Maestría por generar el vínculo con Claudio y Krešimir.

A todos los que me apoyaron desde lo afectivo para lograrlo. A mi familia, especialmente mi mamá quien fue quien me introdujo al mundo de la Minería de Datos y sugirió que hiciera la maestría. A mis amigas quienes me perdonaron cada evento que me perdí por tener que dedicarle tiempo a esta tesis y tener la misma excusa durante años.

A todos, muchas gracias!

A mi mamá, Claudia Guzner.

Índice

I. Introducción 10
I.1 Motivación 10
I.2 Objetivos 11
I.3 Organización del Trabajo 11
II. Marco Teórico 13
II.1 Introducción 13
II.2 Visualización de la Información 13
II.3 Minería de Datos y Descubrimiento del Conocimiento 23
II.4 El Enfoque de Visual Analytics 28
II.5 Resumen 33
III. Metodologías Actuales para la Construcción y Evaluación de Árboles de Decisión 34
III.1 Introducción 34
III.2 Algoritmo CHAID 34
III.4 Metodología Actual para la Construcción y Evaluación de Árboles de Decisión de tipo CHAID 45
III.5 Resumen 53
IV. Análisis Interactivo de Árboles de Decisión 54
IV.1 Introducción 54
IV.2 Alcance de la Metodología Propuesta 54
IV.3 Nuestra Metodología 55
IV.4 Aplicación a un Caso Real 64
IV.5 Resumen 74
V. Conclusiones y Futuros Pasos 76
V.1 Introducción 76
V.2 Contribución al Proceso de Visual Analytics 77
V.3 Futuros Pasos 79
V.4 Conclusión 81
VI. Bibliografía 82

Listado de Figuras

- Figura II.1 Línea de tiempo, para enumerar los hitos en la historia de la cartografía temática, gráficos estadísticos, y visualización de datos (Fuente: Captura de pantalla de Friendly-Denis,2001) 15
- Figura II.2 William Playfair – Precio del Trigo y los Salarios (Playfair, 1786) 16
- Figura II.3 Mapa de Cólera – John Snow (1855). 17
- Figura II.4 Diagrama de Causas de Mortalidad en el Ejército en el Este. Florence Nightingale (1857). 18
- Figura II.5 gráfico que ilustra las pérdidas sucesivas de soldados franceses en la campaña a Rusia 1812-13 (Minard, 1861)) 19
- Figura II.6 Gráfico de Coordenadas Paralelas (Inselberg & Dimsdale, 1990). 20
- Figura II.7 Treemap (Plaisant, Shneiderman, Chintalapani, & Aleks, 2014) 20
- Figura II.8 Diagrama esquemático del Proceso de Visualización (adpatado de (Ware, 2004)) 21
- Figura II.9 Proceso de Visualización de Datos , adaptado al español de Ben Fry (2007). 22
- Figura II.10 Proceso de KDD (adaptado de (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)). 28
- Figura II.11 Interfaz de usuario de Baobview (van den Elzen & van Wijk, 2011) 30
- Figura II.12 El proceso de Visual Analytics (traducción de (Keim, Kohlhammer, Ellis, & Mansmann, 2010)). 32
- Figura III.1 Árbol de Análisis de los Comités Evaluadores de Sudáfrica (Adaptado de (Kass, 1980)) 41
- Figura III.2 Cinco clasificadores discretos plasmados en un gráfico ROC (adaptación (Fawcett, 2004)). 44
- Figura III.3 Caso Ejemplo Curva Lift y Lift Acumulado (adaptado de Coppock, (Coppock, 2002)) 46
- Figura III.4 Generación de un árbol de Decisión con SAS Enterprise Miner 47
- Figura III.5 Generación de varios árboles de decisión con diferentes combinaciones de parámetros 48
- Figura III.6 Múltiples nodos de Árbol de Decisión 49
- Figura III.7 Pantalla de Resultados del Nodo Árbol de Decisión de SAs Enterprise Miner 50
- Figura III.8 Nodo Comparación de Modelos SAS Enterprise Miner 50
- Figura III.9 Pantalla Resultados nodo “Compración de Modelos” SAS Enterprise Miner 51
- Figura III.10 Funcionalidad Construcción Árboles de Decisión Interactivamente 52
- Figura III.11 Selección de la variable a utilizar para dividir el nodo raíz en ramas 53

- Figura IV.1 Proceso de Modelado y Evaluación Interactiva de Modelos propuesto 55
- Figura IV.2 Funcionalidad Sugerida para la configuración de las combinaciones de parámetros de árbol de decisión en SAS Enterprise Miner 57
- Figura IV.3 Configuración sugerida para definir las combinaciones de parámetros en SAS Enterprise Miner (ejemplo para parámetro Profundidad Máxima – Maximum Depth). 57
- Figura IV.4 Ventana de Resultados Integrada para SAS Enterprise Miner 61
- Figura IV.5 ComVis para el análisis de resultados de múltiples configuraciones de árboles de decisión..... 62
- Figura IV.6 Análisis de Resultados mediante el uso de ComVis aplicado a múltiples árboles de decisión. 63
- Figura IV.7 Ejemplo para la selección visual del modelo final de minería de datos 64
- Figura IV.8 Pantallas de Configuración de Parámetros según mejora propuesta. 67
- Figura IV.9 Dos estrategias para el manejo de información de curvas. (a) Tiempo representado por una dimensión adicional. Los registros del 1 al 500 representan una única curva. (b) La curva es representada como un tipo atómico en la tabla. Las columnas de esta tabla representan familias de curvas. (extraído de (Konyha, Lez, & Matkovic, 2012)). 69
- Figura IV.10 Pantalla Inicial con información del archivo de entrada de ComVis. 70
- Figura IV.11 Resultados Generales sin utilizar filtros ni zoom 71
- Figura IV.12 Zoom en curvas interesantes y análisis de los parámetros asociados. 71
- Figura IV.13 Selección de curva más interesante y análisis de parámetros 72
- Figura IV.14 Análisis contrastante para parámetro Máxima Cantidad de Ramas 73
- Figura IV.15 Comparación de Configuración de Parámetros versus resultados reales.. 74
- Figura V.1 El proceso de Visual Analytics (traducción de (Keim, Kohlhammer, Ellis, & Mansmann, 2010)). 77
- Figura V.2 Visualización de árboles de decisión a través de icicle plots (izquierda) y gráficos de tipo nodo-vínculo (derecha)..... 80

Listado de Tablas

- Tabla III-1 Tabla de Contingencia Sexo versus Fumador 35
- Tabla III-2 Valores esperados de distribución en caso de independencia 36
- Tabla III-3 Tabla de Contingencia Año de Ingreso versus Comité Evaluador 38
- Tabla III-4 Frecuencias Esperadas para los valores Año de Ingreso y Comité Evaluador 39
- Tabla III-5 Tabla de Contingencia Año de Ingreso versus Comité Evaluador con los dos primeros valores unificados 39
- Tabla III-6 Tabla de Contingencia Año de Ingreso versus Comité Evaluador con los dos primeros valores unificados 40
- Tabla III-7 Matriz de Confusión 43
- Tabla IV-1 Configuración de parámetros del nodo de árboles de decisión de SAS Enterprise Miner 66

I. Introducción

I.1 Motivación

En el año 2012 la Universidad de Buenos Aires invita al Doctor Krešimir Matković a dar una charla en la carrera de Maestría en Explotación de Datos y Descubrimiento del Conocimiento sobre Visualización Interactiva de Información Compleja. Esta charla se centró Visualización de Información científica a través de la herramienta desarrollada por él mismo y un grupo de colaboradores, ComVis (Matkovic, Freiler, Gracanin, & Hauser, 2008). La funcionalidad principal de esta herramienta consiste en poder visualizar grandes colecciones de curvas en una misma interfaz en forma interactiva a través de múltiples filtros visuales. Si bien la aplicación fue desarrollada principalmente para la Visualización Científica, rápidamente pudimos observar su potencialidad de aplicación en problemas de Minería de Datos. A partir de ese momento surge el interés en trabajar en forma conjunta para adaptar la herramienta a problemas del entorno del aprendizaje automático.

De las ideas que surgieron, la que tuvo mayor peso fue la posibilidad de utilizar ComVis para visualizar interactivamente las curvas ROC[1], que son usualmente utilizadas para medir la efectividad de los modelos predictivos. La visualización de los resultados de modelos de Minería de Datos en forma interactiva no suele ser la característica más sobresaliente en la mayoría de los programas de Minería de Datos, por lo que se detectó una oportunidad de contribuir a la mejora en ese sentido.

Investigando sobre el tema y su estado del arte y saltando de vínculo a vínculo por páginas de internet, descubrimos una disciplina en auge: Visual Analytics. Esta disciplina surge como la necesidad de unir los esfuerzos de Visualización de la Información y Minería de Datos. El objetivo es poder involucrar a las personas en los procesos automáticos, en forma interactiva y visual. Trabajando sobre el marco teórico (capítulo II de esta tesis), descubrimos el trabajo de Bertini y Lalane (Bertini & Lalanne, 2009) quienes hacen un profundo estudio del estado del arte de Visual Analytics. Dentro de los posibles desarrollos que mencionan para la conjunción de Visualización de la Información y Minería de Datos, describen como necesaria la capacidad de visualizar los resultados de la configuración de diferentes parámetros de modelos de minería de datos en forma interactiva. Es esta necesidad la que nos sirve de disparador para el trabajo que aquí presentamos.

Una vez detectada esta necesidad, comenzamos a trabajar en la selección de los algoritmos de Minería de Datos que más se verían beneficiados con esta nueva herramienta. Dado que la misma se enfocaría en la incidencia de los parámetros sobre los resultados, seleccionamos los árboles de decisión. Este tipo de algoritmos dependen directamente de la configuración de ciertos parámetros y son los que más interacción generan por parte de los analistas.

Con el transcurso del tiempo la idea fue evolucionando y logramos resultados interesantes que fueron publicados en la sección especial sobre Análisis de Incertidumbre y Espacio de Parámetros en Visualización de la edición Computer & Graphics de Elsevier (Padua, Schulze, Matković, & Delrieux, 2014). Este trabajo se focalizó sobre el uso de ComVis para la evaluación

[1] ROC: Receiver Operating Characteristic Curve. Curva de evaluación para modelos de clasificación.

de la incidencia de los parámetros en los resultados de los árboles de decisión. En el presente trabajo, desarrollamos una metodología que abarca esta etapa, pero que además plantea la necesidad de modificar la forma en que hoy se realiza la configuración de los algoritmos en las herramientas de minería de datos.

Hoy en día la disciplina de Visual Analytics está en boca de todos gracias a que grandes vendedores de herramientas analíticas y bases de datos han incluido dentro de sus funcionalidades las promovidas por la disciplina. Contribuir con esta tendencia es la principal de nuestras motivaciones que nos llevan a presentar este trabajo.

I.2 Objetivos

Con el presente trabajo se pretende contribuir a la disciplina de Visual Analytics, particularmente en la visualización interactiva de parámetros de árboles de decisión. A diferencia de otras herramientas presentes en el mercado, no nos concentramos en la creación de un único árbol de decisión en forma interactiva. Por el contrario, pretendemos que el usuario pueda evaluar diferentes y múltiples resultados en forma interactiva y seleccionar el que más se ajuste a sus necesidades, comprendiendo el efecto de los parámetros de configuración durante este recorrido.

Las disciplinas de Visualización de la Información y Minería de Datos han evolucionado en forma separada. Sin embargo en los últimos años se ha visto un creciente interés por integrar ambas disciplinas. Desarrolladores del mundo de la Visualización comenzaron a interesarte por el “Big Data” e investigadores de Minería de Datos comprendieron la necesidad de incorporar elementos visuales en los procesos de explotación de datos. Es nuestro objetivo ser parte de esta comunidad al presentar una metodología que contiene elementos de ambas disciplinas.

El objetivo de esta investigación no consiste en desarrollar una herramienta única, sino presentar una forma de trabajo que pueda adaptarse a las diferentes soluciones existentes en el mercado. Para poder ejemplificar nuestra metodología seleccionamos dos herramientas a las cuales tenemos acceso y trabajamos en la adaptación de las mismas.

Nuestra contribución está orientada a usuarios principalmente de minería de datos, que se sienten más cómodos trabajando con herramientas visuales y no utilizando código de programación. Es nuestro objetivo mejorar su capacidad de análisis a través de la interacción.

A continuación presentamos el contenido de este trabajo para lograr los objetivos propuestos.

I.3 Organización del Trabajo

El trabajo comienza con la presentación del marco teórico de Visual Analytics. En este capítulo ahondamos sobre el estado del arte de las dos disciplinas principales que la componen: Visualización de la Información y Minería de Datos. A partir de ellas luego desarrollamos los contenidos específicos de Visual Analytics. Para las tres disciplinas planteamos la definición de las mismas, su evolución y los procesos que implican. Al final de este capítulo podrá obtenerse una idea del objetivo de Visual Analytics para luego evaluar cómo la metodología planteada contribuye a seguir construyendo la misma.

En el capítulo III explicamos cómo es desarrollado un modelo de árbol de decisión en la actualidad ejemplificando con una de las herramientas comerciales más utilizadas del mercado. Explicamos el algoritmo de árboles de decisión seleccionado, junto con los parámetros que es preciso configurar. Son los efectos de estos parámetros los que luego visualizaremos en forma interactiva en nuestra propuesta, siendo esta contribución la más novedosa. En esta misma sección también hacemos referencia a los principales métodos de evaluación de la efectividad de los modelos, concentrándonos principalmente en curvas ROC y Lift. Éstas, además de ser las formas visuales más representativas de los resultados, son las de mayor uso en la comunidad de Minería de Datos.

Una vez presentada la metodología actual, en el capítulo IV desarrollamos nuestra metodología en forma teórica, para luego aplicar la misma en un caso de uso real. Nuestra propuesta consiste en cinco pasos que explicamos en la sección teórica y que luego repetimos para el caso de uso real. Este caso de uso es de una empresa de telecomunicaciones argentina que actualmente realiza tareas de Minería de Datos de la forma tradicional y que podrá beneficiarse con la metodología.

Por último, en el capítulo V resumimos las contribuciones del trabajo tanto para Minería de Datos como Visualización de la Información. Finalizamos presentando algunos posibles futuros pasos para seguir desarrollando la metodología.

II. Marco Teórico

II.1 Introducción

Visual Analytics es una nueva área de investigación que combina técnicas de Visualización y Minería de Datos o Descubrimiento del Conocimiento. Esta disciplina surgió recientemente y se encuentra en pleno desarrollo. El objetivo principal de la misma es mejorar el análisis de grandes bases de datos y, por ende, la toma de decisiones mediante el uso de técnicas interactivas de visualización.

En este capítulo presentamos la evolución de cada uno de los pilares que componen Visual Analytics. Esto servirá para comprender cómo ambas disciplinas convergen en lo que hoy es conocido como Visual Analytics. La Visualización de la Información es el componente de mayor antigüedad, pudiendo rastrearse sus orígenes hasta los comienzos mismos de la civilización. La Minería de Datos, en cambio, es un dominio más reciente que no surge sino hasta la aparición de las computadoras. Sin embargo, a pesar de tener menor antigüedad, los trabajos en este campo son innumerables y su cantidad crece exponencialmente. El campo de Visual Analytics, por su parte, tiene tan sólo 10 años como concepto, pero gracias a la contribución de la tecnología, la investigación en esta disciplina también ha experimentado una gran evolución.

En la primera sección de este capítulo presentaremos el concepto de Visualización de la Información, junto con una breve reseña de su evolución y el delineamiento de un proceso típico de Visualización de la Información. A continuación, desarrollaremos el concepto de Minería de Datos o Knowledge Discovery (KDD), junto con un resumen de los avances en los algoritmos que forman parte de la disciplina y el proceso necesario para descubrir conocimiento. Por último evaluaremos la conjunción de ambas prácticas en lo que se conoce como Visual Analytics, desarrollando el concepto, su evolución y proceso.

II.2 Visualización de la Información

II.2.1 Definición

Para comenzar a desarrollar la noción de Visualización de la Información, es preciso definir con anterioridad los conceptos que la componen. En el diccionario de la Real Academia Española[2], el término visualización es definido como *"acción y efecto de visualizar"*. Visualizar, a su vez, tiene los siguientes significados:

1. Visibilizar (Hacer visible artificialmente lo que no puede verse a simple vista, como con los rayos X los cuerpos ocultos, o con el microscopio los microbios).

2. Representar mediante imágenes ópticas fenómenos de otro carácter; p. ej., el curso de la fiebre o los cambios de condiciones meteorológicas mediante gráficas, los cambios de corriente eléctrica o las oscilaciones sonoras con el oscilógrafo, etc.

3. Formar en la mente una imagen visual de un concepto abstracto.

[2] www.rae.es

4. Imaginar con rasgos visibles algo que no se tiene a la vista.

5. Hacer visible una imagen en un monitor.

Como puede observarse, en todas las definiciones se hace referencia a la capacidad de formar una imagen de algo que en realidad no tiene una forma específica o predeterminada. Tal es el caso de la información. La definición de este término es bastante amplia. Si consultamos el mismo diccionario para analizar su significado, encontramos que los conceptos propuestos son bastante diversos. De los sugeridos, podríamos adoptar el siguiente: la información es la *"*comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se *poseen sobre una materia determinada"*.

Podemos observar puntos de convergencia en ambas definiciones y al combinar estos conceptos podemos definir Visualización de la Información como:

"la comunicación de conocimiento a través de representaciones visuales"

Otros autores (Card, Mackinlay, & Shneiderman, 1999) prefieren hablar de percepción en lugar de visualización, para extender el concepto a otros sentidos tales como el auditivo y el táctil. Sin embargo, dado que la visión es el sentido que posee mayor capacidad de captar y procesar información en los humanos, se utilizan los conceptos de visualización y percepción como sinónimos.

En esta definición que hemos elaborado de Visualización de la Información, no hemos limitado la misma al uso de una herramienta específica, por ejemplo una computadora, para lograr el objetivo. Sin embargo, para limitar el alcance de este trabajo, podemos adoptar la definición de Card, MacKinlay y Schneiderman (Card, Mackinlay, & Shneiderman, 1999). Para los mismos, Visualización de la Información consiste en: *"el uso de representaciones visuales e interactivas* de datos abstractos *soportadas por el uso de computadoras, para amplificar el conocimiento"*. Los autores diferencian los conceptos de Visualización y Visualización de la Información, simplemente al distinguir el tipo de datos que se pretenden representar visualmente. En el primer caso, se refieren principalmente a los datos físicos (del cuerpo humano, geográficos, moléculas, etc.), que tienen una forma de representación visual inherente asociada a sus características espaciales. Al hablar de información en términos genéricos, se refieren a todo tipo de datos no físicos, tales como información financiera, de negocios, colecciones de documentos y otros conceptos abstractos.

II.2.2 Evolución de la Visualización de la Información

Si consideramos la primera definición de Visualización de la Información que propusimos en el apartado anterior, podemos encontrar los orígenes de la misma en épocas muy remotas de la historia de la humanidad. El hombre siempre ha buscado formas de comunicar su conocimiento en forma visual. Aún antes de la aparición de la escritura, nuestros antepasados utilizaban huesos de animales para representar las fases de la luna (Dürsteler, 2002). Los egipcios desarrollaron las escritura jeroglífica en el 3.300 a.C. que perduró hasta el año 394 con la última inscripción que figura en el templo de File (Wikipedia, 2014). Además de emplear signos consonánticos y determinantes, también utilizaban símbolos figurativos para representar algo tangible, fácil de reconocer. Los egipcios se inspiraron en su entorno cotidiano para el diseño de la escritura jeroglífica, resumiendo en un solo símbolo un significado más complejo. Esta esencia

es sostenida aún hoy en día, donde se desea utilizar técnicas de visualización para la simplificación de la comunicación de la información.

Investigar los orígenes de la Visualización de la Información como comunicación del conocimiento a través de representaciones visuales, nos llevaría a recorrer casi toda la historia de la humanidad. Por el contrario, limitándonos a la definición que considera el uso de computadoras como necesario para visualizar información, perderíamos de vista muchos de los grandes trabajos de los precursores de la disciplina, que se remontan hacia el siglo XVII inclusive.

En tal sentido, es interesante el límite propuesto por Friendly y Denis (Friendly & Denis, 2001) en su trabajo *"Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization"*. Para limitar el alcance de su investigación, se refieren a Visualización de Información como la representación de datos extraídos de forma esquemática, particularmente con dos focos: gráficos estadísticos y mapas cartográficos. En ambos casos se trata de caracterizar información cuantitativa y categórica, pero con diferentes objetivos de representación. Los autores utilizan a su vez una técnica de visualización muy conocida, un gráfico de línea de tiempo, para enumerar los hitos en la historia de la disciplina. Este tipo de gráfico fue utilizado por primera vez en 1765 por Joseph Priestley, tal como puede verse en el mismo compendio de los autores (Figura II.1 Línea de tiempo, para enumerar los hitos en la historia de la cartografía temática, gráficos estadísticos, y visualización de datos (Fuente: Captura de pantalla de Friendly-Denis,2001).

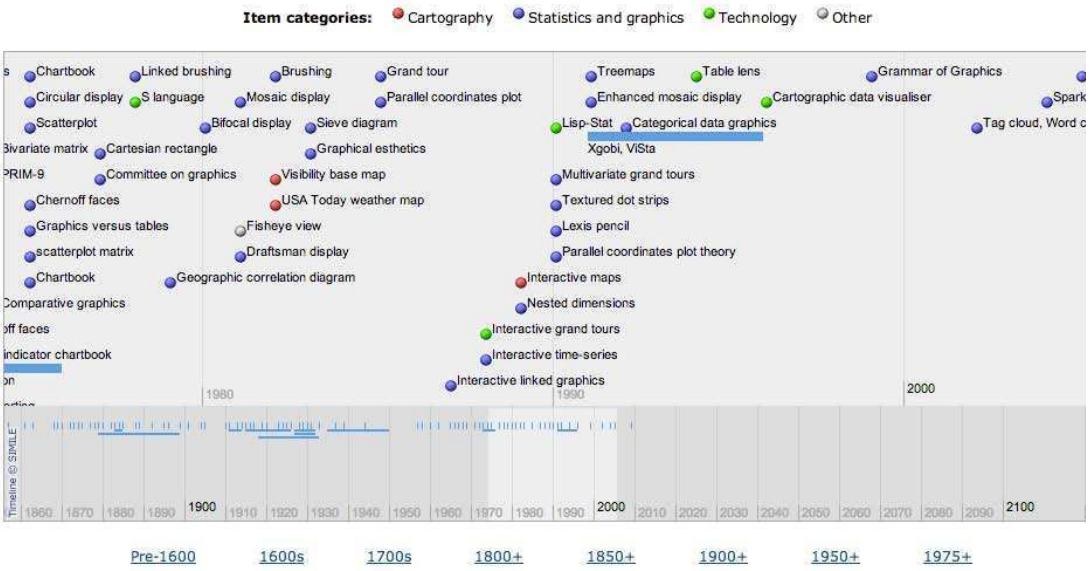

Figura II.1 Línea de tiempo, para enumerar los hitos en la historia de la cartografía temática, gráficos estadísticos, y visualización de datos (Fuente: Captura de pantalla de Friendly-Denis,2001)

La línea de tiempo comienza con los trabajos encontrados antes del año 1600 donde la mayoría de los gráficos y mapas eran utilizados en navegación y exploración. En el siglo XVI las técnicas e instrumentos de medición y observación estaban bien desarrolladas y es por ello que se empiezan a encontrar los primeros trabajos de visualización de datos.

Ya en 1600, los principales problemas consistían en la medición de tipo física (tiempo, distancia y espacio) para astronomía, encuestas, construcción de mapas, navegación y expansión territorial. Durante este siglo se observó un gran crecimiento en la teoría y práctica de la ciencia, con el surgimiento de la geometría analítica, la teoría de error en mediciones y estimaciones, el nacimiento de la teoría de la probabilidad y el comienzo de las estadísticas demográficas. Es en este período en donde René Descartes publica "Geometrie" (Descartes, 1937) en donde establece el sistema de coordenadas cartesianas que constituyen la base del dibujo científico y técnico desde entonces. Este sistema ya había sido utilizado anteriormente pero se relanza con el establecimiento de la relación entre la línea representada y la ecuación que la define. Hacia fines de este siglo entonces, ya existían algunas ideas firmes sobre la representación visual de los conceptos. Podría considerarse este período como el origen del pensamiento visual.

El siglo XVIII es el que la mayoría de los autores nombran al referirse a los comienzos de la Visualización de la Información. Durante este período tomaron forma las ideas propuestas en el siglo anterior. Hubo avances tanto en la visualización de información geográfica como estadística. A medida que nuevos datos comenzaron a ser recolectados (información política y económica), nuevas formas de representación fueron surgiendo. El trabajo más comentado en la bibliografía de Visualización de la Información como origen de esta disciplina, es el de William Playfair (1786) quien parece haber sido el primero en utilizar la línea y el área de un gráfico para representar datos visualmente (Figura II.2). Muchos de los reportes de negocios que vemos en la actualidad, constan de los gráficos desarrollados por Playfair en aquel entonces.

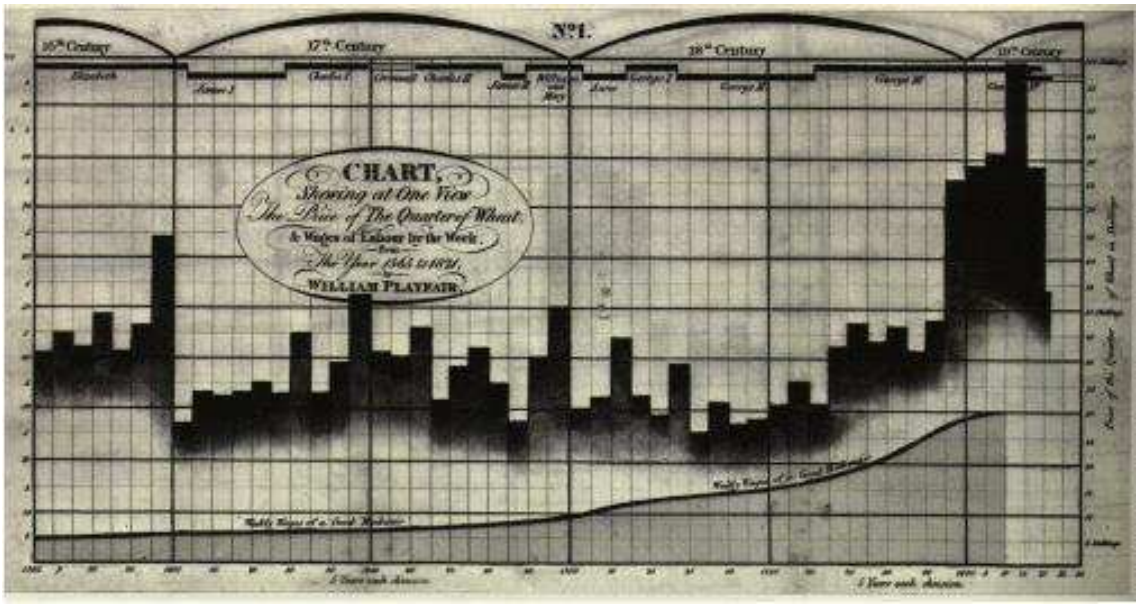

Figura II.2 William Playfair – Precio del Trigo y los Salarios (Playfair, 1786)

Durante la primera mitad del siglo XIX se vio una gran explosión de creatividad. En la visualización estadística, todas las formas de representación visual fueron creadas: gráficos de barras, tortas, líneas de series de tiempo, etc. En el ámbito de la cartografía, se evolucionó de simples mapas a atlas, y se introdujeron diversos simbolismos para los mismos. Fue Playfair quien

continuó con el desarrollo de gráficos para la representación de datos económicos y estadísticos, introduciendo por primera vez en 1801 el famoso gráfico de torta. En 1846, Adolphe Quetelet (1846) grafica la distribución normal, representado la campana que hoy es globalmente conocida.

La segunda mitad del 1800 es conocida como la Edad de Oro en la Visualización de la Información, por la cantidad de innovaciones y mejoras introducidas en la cartografía y gráficos estadísticos. Dentro de esta era encontramos el famoso gráfico de John Snow (1855), representando la epidemia de cólera que sufrió la ciudad de Londres en el año 1854 (Figura II.3).

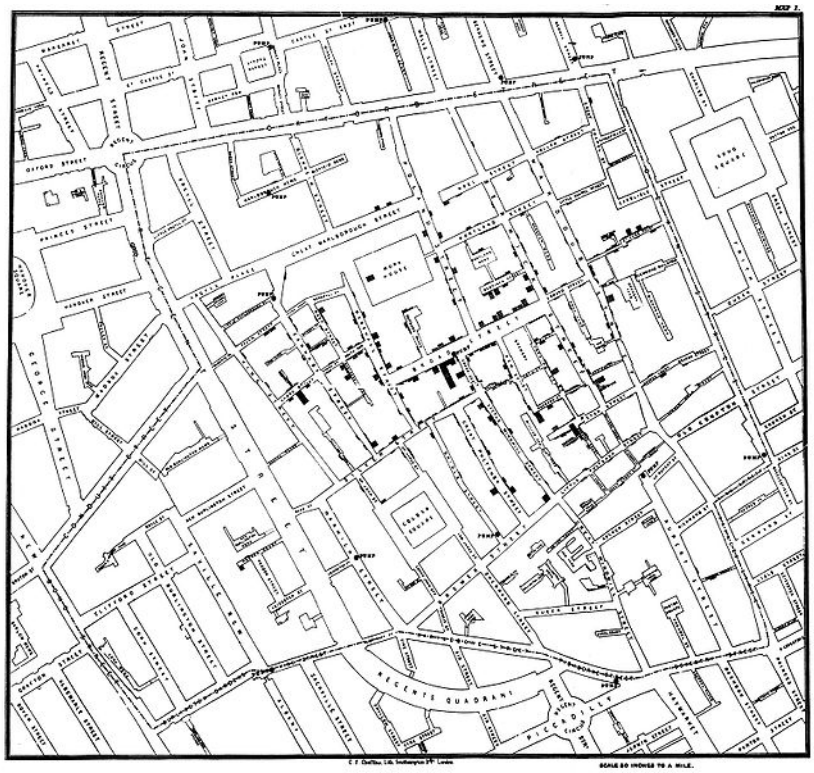

Figura II.3 Mapa de Cólera – John Snow (1855).

Snow combinó el mapa de la ciudad de Londres junto con gráficos de barras para identificar visualmente la causa de la epidemia. Así descubrió que la mayoría de los muertos por esta enfermedad se encontraban en las inmediaciones de la bomba de agua de la calle Broad, refutando la teoría de que el cólera se trasmitía por aire.

Florence Nightingale es otra de las grandes aclamadas de la Era de Oro de la Visualización. Se la conoce como la autora de los diagramas de área polar o “coxcombs” (Nightingale, 1857) aunque también los mismos podrían ser atribuidos a Andre-Michel Gerry (Friendly & Denis, 2001). Estos gráficos fueron utilizados para demostrar la diferencia entre las causas de mortalidad en hospitales británicos durante la guerra de Crimea (Figura II.4).

Edward Tufte (1983) clasifica el mapa que representa la Campaña del Ejército de Napoléon en Rusia de 1812-1813, de Charles J. Minard como el mejor gráfico estadístico que jamás se haya logrado (Figura II.5).

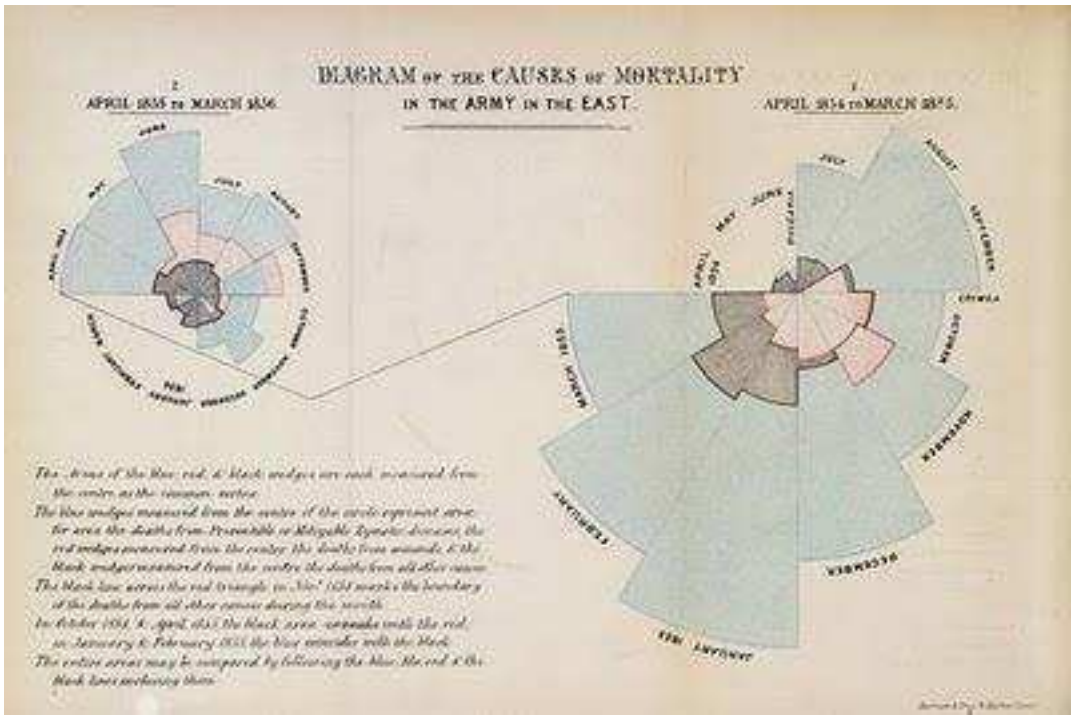

Figura II.4 Diagrama de Causas de Mortalidad en el Ejército en el Este. Florence Nightingale (1857).

Minard representa en un simple gráfico de dos dimensiones, cinco variables que explican la fatalidad del ejército de Napoleón. La banda más gruesa y en color marrón, representa el tamaño del ejército en cada uno de los lugares del mapa en el camino de ida hacia Moscú. Luego la banda más fina en color negro, representa la cantidad de efectivos durante la retirada. La misma está vinculada a la línea de temperatura en la parte inferior del gráfico, que también posee referencia a las fechas de la campaña. El gráfico demuestra cómo las inclemencias del tiempo junto con el cruce del río Berenzina destrozaron al ejército, que partió de Polonia con 442.000 soldados y regresó con tan solo 10.000.

El siglo XX, al igual que el XIX puede dividirse en dos partes para separar los avances de Visualización de la Información. Durante la primera mitad no se observaron grandes innovaciones en el área. Consistió en un período de adaptación a las nuevas formas de analizar la información que explotaron hacia fines del 1800. Los gráficos estadísticos se incorporaron en libros, currículos escolares, y se convirtieron en una práctica común en el gobierno, comercio y práctica científica (Friendly & Denis, 2001). De esta época podemos rescatar la creación del Gráfico de Gantt, en 1917, tan comúnmente utilizado hoy en la planificación de proyectos.

En la segunda mitad del siglo XX comienza otra era de creatividad en el área de visualización, especialmente de la mano de dos autores: Jacques Bertin y John Tukey. Jacques Bertin, cartógrafo francés, publica en 1967 “Semiologie Graphique” (Bertin, 1967) un libro cuyo contenido sigue vigente hasta el día de hoy. En su trabajo determina que toda visualización está formada por una serie de componentes que tienen distinto poder expresivo y que cada uno de

ellos funciona mejor dadas ciertas condiciones. Bertin sugiere seis variables básicas para la creación de gráficos: tamaño, valor, textura, color, orientación y forma. Para cada una señala en qué casos funcionan mejor y cómo utilizarlas.

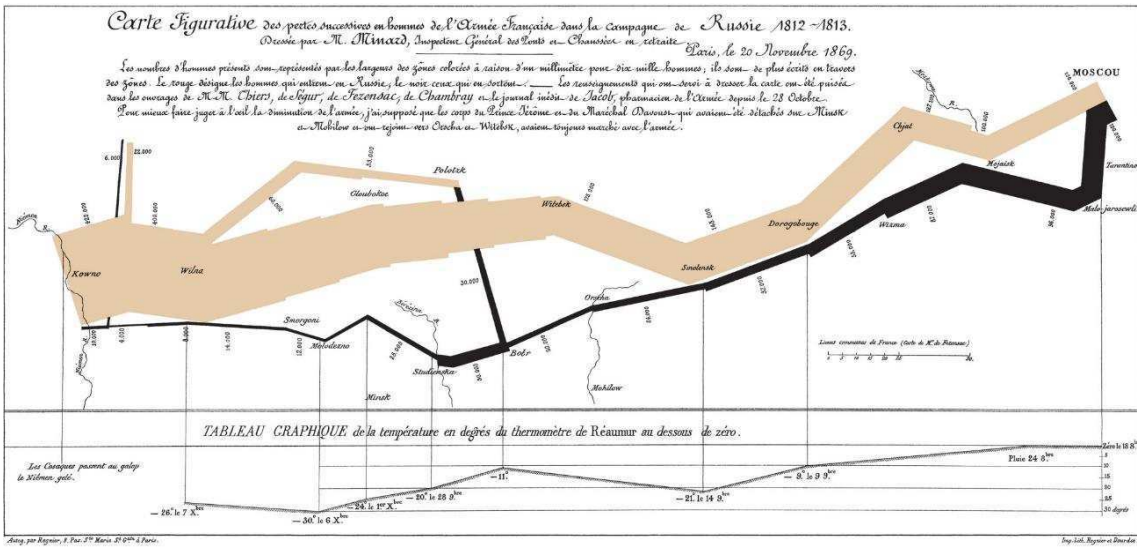

Figura II.5 gráfico que ilustra las pérdidas sucesivas de soldados franceses en la campaña a Rusia 1812-13 (Minard, 1861))

Si bien el trabajo de Tukey "Exploratory Data Analysis" (EDA) se publicó recién en 1977 (Tukey, Exploratory Data Analysis, 1977), ya en 1962 el autor proponía en su trabajo "The Future of Data Analysis" (Tukey, 1962), la legitimación del análisis de información estadística como una rama distinta a la estadística matemática. En EDA el énfasis está puesto en lograr rápidas compresiones de información estadística a través del uso de imágenes. En este trabajo introduce el ya famoso gráfico de *"box and whisker"* que le permite al analista comprender rápidamente los cuatro indicadores principales de la distribución de los datos (cuartiles, mediana, mínimo y máximo).

Con la aparición de las computadoras, se da otra era de explosión creativa en la disciplina y los desarrollos desde entonces son innumerables. Con el crecimiento de la informática, es cada vez más fácil generar y almacenar grandes cantidades de información, por lo que comienzan a ser necesarias nuevas técnicas de visualización específicamente diseñadas para grandes volúmenes de datos.

Un problema de particular importancia consistía en cómo visualizar conjuntos de datos con varias variables. Una de las contribuciones más importantes en este sentido fue la de Inselberg con sus sistema de Coordenadas Paralelas (Inselberg & Dimsdale, 1990). En este gráfico (Figura II.6) se representan cada dimensión como una escala vertical paralela a todas las demás. A cada elemento del conjunto de datos le corresponde una línea quebrada que une los valores que toman cada una de sus variables, el equivalente de un "punto" en representación bi- o tridimensional. Este trabajo fue presentado en 1990 en la primera conferencia de Visualización en el marco del Instituto de Ingenieros Eléctricos y Electrónicos, más conocido por sus siglas IEEE. Esta comunidad fue liderada en sus comienzos por científicos de recursos naturales, físicos y

computadores científicos. De los trabajos presentados en la misma, sólo algunos pocos se referían al uso de técnicas de Visualización de Información para datos no científicos. La mayoría de las investigaciones publicadas se enfocaron en técnicas de visualización científica.

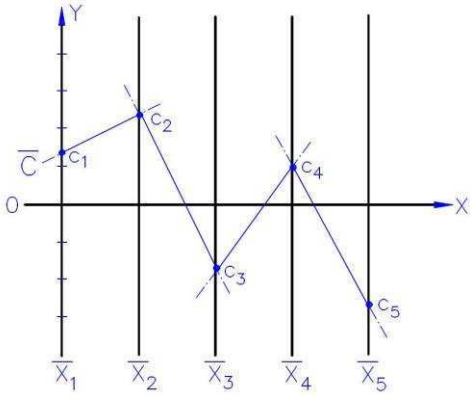

Figura II.6 Gráfico de Coordenadas Paralelas (Inselberg & Dimsdale, 1990).

Los “Treemaps” de Ben Schneiderman (1991) constituyen otros de los desarrollos más importantes de principios de los noventa. Surgieron como una herramienta de visualización de la estructura de discos duros, pero hoy son utilizados para explicar la composición de las bases de datos más diversas. El gráfico de Treemap (Figura II.7) consiste en un diseño de rectángulos anidados para visualizar estructuras jerárquicas. Combina tamaños y colores para representar diferentes atributos de la información.

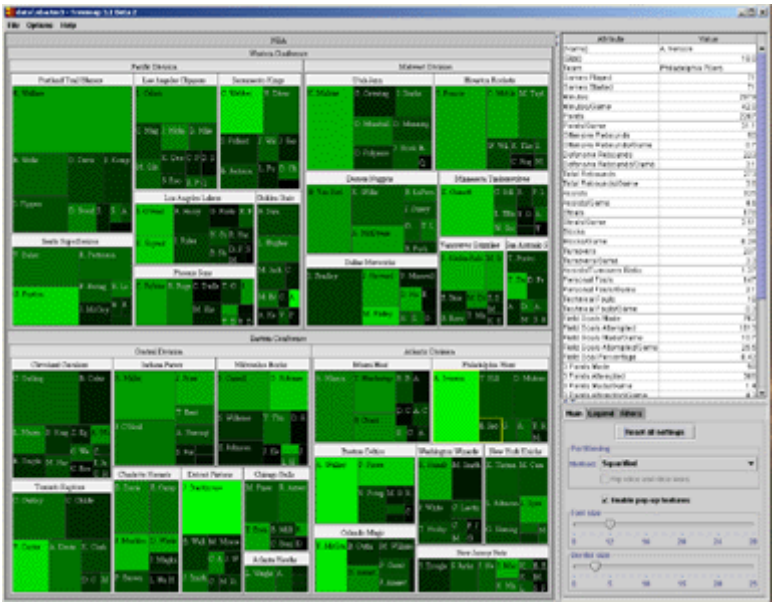

Figura II.7 Treemap (Plaisant, Shneiderman, Chintalapani, & Aleks, 2014)

Schneiderman es uno de los autores más referenciados dentro del mundo de la Visualización de la Información gracias al trabajo que realizó junto a Card y otros autores en el libro “Readings in Information Visualization: Using Vision to Think” (Card, Mackinlay, & Shneiderman, 1999).

A partir del año 2000, las técnicas de Visualización de la Información se expanden a la mayoría de las áreas del saber. Con la globalización del conocimiento a través del uso de internet, los trabajos en el área crecen a un ritmo exponencial. Los eventos y conferencias sobre Visualización también se multiplican y se subdividen cada vez más según las distintas ramas de aplicación. Tal es así que la conferencia de la IEEE, que comenzó en 1990, tiene hoy tres subsecciones: VAST, dedicada exclusivamente a los trabajos en el área de Visual Analytics (concepto que desarrollaremos más adelante y principal foco de este trabajo); InfoVis, para trabajos más genéricos de Visualización de la Información; y SciVis, la sección exclusiva para los trabajos de Visualización Científica.

Una de las áreas de conocimiento donde se introdujo la ciencia de la Visualización, es la Minería de Datos. Si bien ambas disciplinas han evolucionado en forma separada, comparten en la actualidad varios puntos en común. Muchos investigadores han trabajado en la integración de ambas y se encuentran aún haciéndolo. Dado que el foco de este trabajo es precisamente tal combinación, describiremos la evolución de la conjunción de ambas disciplinas en un apartado posterior.

II.2.3 Proceso de Visualización

Con la evolución de la disciplina, surgieron también varios enfoques sobre el proceso de construcción de visualizaciones. Diferentes referentes de Visualización de la Información enumeran diversas etapas e interacciones en el proceso de visualización de la información. Uno de los autores más mencionados es Colin Ware (2004) quien enumera las siguientes etapas en el proceso de Visualización (Figura II.8):

- Recolección y almacenamiento de la información
- Preprocesamiento prediseñado para transformar los datos en algo comprensible por la mente humana
- El hardware y algoritmos gráficos necesarios para producir imágenes en una pantalla
- El sistema de percepción y cognitivo del humano

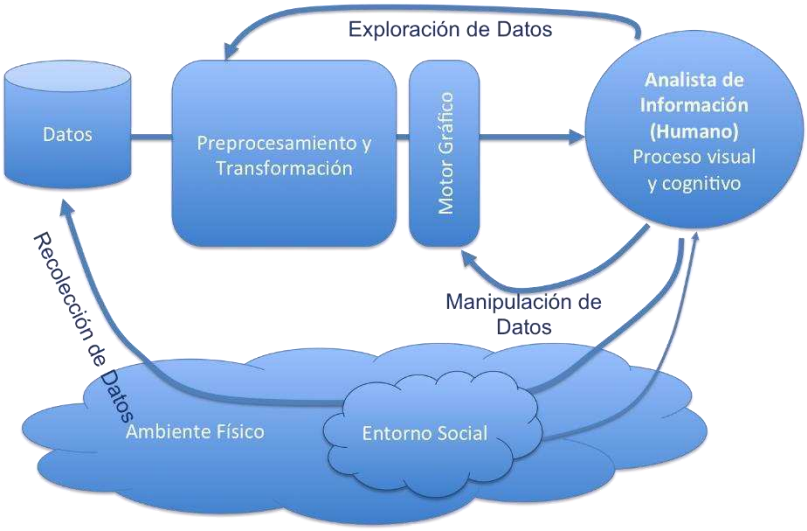

Figura II.8 Diagrama esquemático del Proceso de Visualización (adpatado de (Ware, 2004))

Como puede observarse, este proceso involucra el uso de un motor gráfico – entendiéndose una computadora – para la generación de la visualización, por lo que estamos considerando la definición acotada de Visualización de la Información que enunciamos en el apartado II.2.1.

En el esquema propuesto por Ware, la mayor parte del proceso está relacionada con la recolección de datos. Esta recolección está influenciada por un ambiente físico (fuentes de datos disponibles) y un entorno social que determina qué datos buscar en función de la percepción del analista. El proceso es altamente interactivo ya que es el humano quien determina cuándo se da fin al ciclo de visualización.

Ben Fry (2007) considera que el proceso de visualización de datos consiste en una serie de pasos para poder contestar una pregunta que da origen a la necesidad de visualización. Este proceso puede observarse en la Figura II.9.

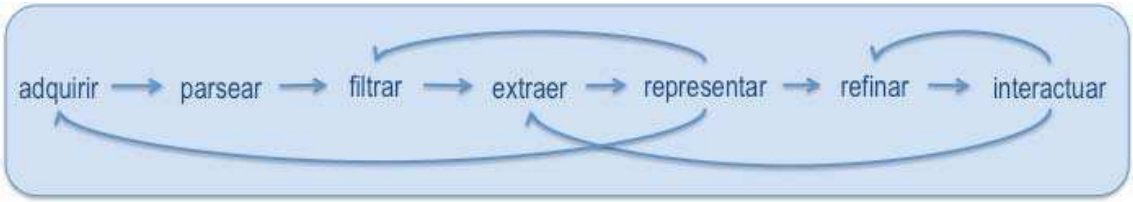

Figura II.9 Proceso de Visualización de Datos , adaptado al español de Ben Fry (2007).

Según Fry, estas etapas consisten en:

- **Adquirir**. Obtener los datos ya sea de un archivo, un disco o un sitio en la red.
- **Parsear**. Proveer cierta estructura a los datos y ordenarlos en categorías.
- **Filtrar**. Remover los datos que no sean de interés.
- **Extraer**. Aplicar métodos de estadística o minería de datos para encontrar patrones o ubicar la información en un contexto matemático.
- **Representar**. Elegir un modelo visual básico.
- **Refinar**. Mejorar el modelo básico para hacerlo más claro o más visualmente atractivo.
- **Interactuar**. Agregar métodos para manipular la información o determinar qué debe verse.

Se observan varios puntos en común con el proceso ilustrado por Ware. Sin embargo Fry ya menciona explícitamente dentro del proceso como tarea necesaria la extracción mediante técnicas estadísticas o de minería de datos. El autor ilustra las conexiones a lo largo del proceso para representar la importancia del trabajo en equipo. Por lo general se considera la práctica de visualización como un proceso en línea, donde los programadores se encargan de la adquisición de datos y parseo, mientras que los diseñadores se enfocan en las formas de representación. Es importante destacar que el libro de Fry (2007) está orientado a diseñadores, por ello pone especial atención a la integración de los especialistas en datos y los profesionales del área de diseño.

II.3 Minería de Datos y Descubrimiento del Conocimiento

II.3.1 Definición y Diferenciación

Históricamente la noción de descubrir patrones ocultos en los datos ha tenido muchas denominaciones (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996). Minería de Datos ha sido el más utilizado por estadísticos, analistas de datos y comunidades de gestión de la información. Adaptándonos a la definición propuesta por Fayyad y otros, podemos definir minería de datos como:

"la aplicación de algoritmos específicos para extraer patrones de los datos"

Según los autores, la Minería de Datos es sólo un paso en un proceso más amplio denominado descubrimiento del conocimiento (más conocido como KDD por sus siglas en inglés Knowledge Discovery in Databases). KDD se refiere al proceso general de descubrimiento de información útil en los datos. En este sentido, se define KDD como (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996):

"proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útil*es y entendibles en los datos"*.

Más allá de la diferenciación propuesta por los autores, en la actualidad ambos conceptos se tratan en forma indistinta la mayoría de las veces.

II.3.2 Evolución de Minería de Datos y KDD

A diferencia del concepto de Visualización de la Información, la noción de "descubrimiento del conocimiento" tiene una historia mucho más reciente. El primer taller de KDD fue llevado a cabo dentro del marco de la IJCAI (International Joint Conference on Artificial Intelligence) en el año 1989 (Piatetsky-Shapiro, 1991). Desde el año 1998, los talleres se convirtieron en conferencias internacionales sobre KDD y Minería de Datos y hoy en día han proliferado enormemente alrededor del mundo.

Existen dos enfoques para estudiar la evolución de la Minería de Datos. El primero parte de la evolución del concepto como resultado del desarrollo de los sistemas de bases de datos. Según un segundo enfoque, Minería de Datos es el resultado de la evolución de las tres disciplinas principales que la componen: la estadística, el aprendizaje automático y la inteligencia artificial.

Han (2006) describe los avances en los sistemas de bases de datos siendo la Minería de Datos un paso más del continuo. Para el autor la disciplina forma parte del gran crecimiento que tuvieron los sistemas de bases de datos desde los años sesenta. Desde esta década, las tecnologías de información y de bases de datos han estado evolucionando sistemáticamente desde primitivos sistemas de procesamientos de archivos a sistemas sofisticados y poderosos de bases de datos. En los años setenta, estos sistemas progresaron y se convirtieron en sistemas de bases de datos relacionales sobre los cuales los usuarios pueden acceder a los datos a través de lenguajes de extracción, interfaces de usuario y otras técnicas.

La información puede ser hoy en día almacenada en diversas bases de datos y repositorios de información. Uno de los más utilizados es el data warehouse. Esta tecnología incluye la

limpieza de datos, la integración de datos y el procesamiento en línea analítico (OLAP por sus siglas del inglés on-line analytical processing).

A pesar de que la tecnología OLAP permite el análisis multidimensional y la toma de decisiones, es preciso contar con herramientas analíticas más potentes que permitan llevar a cabo otro tipo de tareas, tales como clasificación, predicción o agrupamiento (clustering). Gracias a esta necesidad de análisis avanzado y la creciente disponibilidad de datos, surge lo que hoy se conoce como Minería de Datos.

El segundo enfoque para analizar la evolución de Minería de Datos, consiste en identificarla como la disciplina convergente de sus tres componentes principales: estadística, inteligencia artificial y aprendizaje automático (Kumar & Bhardwaj, 2011).

Los estadísticos han utilizado las computadoras por décadas como medio para demostrar o refutar hipótesis sobre los datos. La estadística es la base sobre la que han sido construidas la mayoría de las tecnologías de Minería de Datos. Sin embargo el uso de la estadística continúa asumiendo un conocimiento previo sobre los datos y posterior comprobación, mientras que la Minería de Datos se trata de todo lo contrario.

La Inteligencia Artificial es otro de los grandes componentes de Minería de Datos. Esta disciplina está construida sobre la heurística como oposición a la estadística y pretende aplicar procesamiento similar al pensamiento humano a los procesos estadísticos. Dado que este enfoque requiere gran capacidad de procesamiento computacional, no fue práctico hasta los años ochenta, cuando las computadoras se hicieron más poderosas. El Aprendizaje Automático puede ser considerado como una evolución de la Inteligencia Artificial (AI por sus siglas en inglés), dado que mezcla la heurística de AI con métodos estadísticos avanzados. Esta disciplina le permite a las computadoras aprender sobre los datos para luego aplicar el conocimiento aprendido sobre otros datos.

Tradicionalmente las técnicas de Minería de Datos pueden dividirse según el fin que tienen: extracción o identificación de patrones, agrupamiento de datos (clustering), clasificación o categorización predictiva. Cada uno de estos grupos ha experimentado una evolución diferente.

La técnica más conocida de extracción de patrones es la de Reglas de Asociación propuesta por Agrawal (Agrawal, Imielinski, & Swami, 1993) para la detección de patrones de compra en tickets de supermercados. Una regla de asociación es una regla probabilística que determina que si ciertos conjuntos de datos ocurren en forma simultánea, entonces otros conjuntos de atributos también tienen posibilidades de ocurrir. Esta técnica ha sido muy utilizada en la comunidad de Minería de Datos y a la vez referenciada en la mayor parte de los trabajos de la disciplina. Luego de este primer trabajo, el algoritmo se extendió a otras aplicaciones. Hoy el mismo es utilizado por muchos sistemas de recomendación para sugerir ítems de interés a las personas de acuerdo a su comportamiento de compra.

Otra de las grandes categorías de algoritmos utilizados en Minería de Datos es la de agrupamiento o clustering. Esta técnica consiste en agrupar los datos en grupos pero sin conocimiento previo de los mismos. A diferencia de la mayoría de los algoritmos de Minería de Datos, se trata de una técnica exploratoria. Berkhin (2009) presenta una extensa investigación sobre los diferentes tipos de clustering y su aplicación específica en Minería de Datos. La división

más común en clustering es separar los algoritmos entre métodos jerárquicos y métodos de particionamiento. Los primeros construyen los grupos gradualmente, mientras que los segundos aprenden los grupos directamente.

Las técnicas de agrupamiento han sido ampliamente utilizadas en aplicaciones de Minería de Datos y continúan desarrollándose para poder resolver las principales limitaciones de estos algoritmos. Entre ellas podemos mencionar el tipo de atributos que pueden manejar, la escalabilidad a grandes bases de datos, la habilidad para trabajar con datos con muchas dimensiones, la capacidad para encontrar grupos en formas irregulares (no basadas en elipses), el manejo de casos extremos (outliers), la complejidad del tiempo, la dependencia de orden de datos, el apoyo en el conocimiento a priori de los usuarios y la interpretabilidad de los resultados.

La investigación en el campo de clustering es muy amplia. Dado que nuestro trabajo no se centrará en este tipo de algoritmos, no profundizaremos en la evolución de los mismos.

El último grupo de algoritmos de Minería de Datos es el de clasificación o categorización. Estas técnicas tienen como objetivo construir clasificadores que pueden ser aplicados a nuevos datos para categorizar éstos en grupos. La diferencia con las técnicas de clustering, es que para clasificar necesitamos datos pre-categorizados que servirán para entrenar el algoritmo. Debido a esto, las técnicas de clasificación suelen ser llamadas algoritmos de aprendizaje supervisado mientras que las de agrupamiento son referenciadas como algoritmos de aprendizaje no supervisado. Dentro de este grupo se encuentran los árboles de decisión, técnicas de regresión y redes neuronales.

Los árboles de decisión son una de las técnicas de clasificación más utilizadas y es de hecho este tipo de algoritmo el que seleccionamos para el desarrollo de nuestro enfoque. Los algoritmos más conocidos son el ID3 (Quinlan J. R., 1986) y su sucesor C4.5 (Quinlan, 1993), CART (Breiman, Friedman, Stone, & Olshen, 1984) y CHAID (Kass, 1980).

ID3 es un algoritmo muy simple que utiliza la medida estadística de “ganancia de información” como criterio de división. El crecimiento del árbol termina cuando todas las instancias pertenecen a una solo clase o cuando el mejor indicador de ganancia de información no es mayor que cero. Esta técnica no aplica ningún procedimiento de poda y no puede manejar valores faltantes o atributos numéricos. C4.5 fue desarrollado por el mismo autor de ID3 (Quinlan, 1993). Las mejoras más relevantes que introdujo en el algoritmo son: aceptación de variables discretas y continuas, capacidad para manejar datos faltantes y la introducción de la técnica de poda para prevenir el sobreajuste.

CART proviene de las siglas en inglés de “Classification and Regression Trees”. Se caracteriza porque produce árboles binarios, es decir que cada nodo tiene sólo dos salidas posibles. La funcionalidad más importante de este algoritmo es la capacidad de generar árboles de regresión. Estos árboles predicen en sus hojas valores numéricos y no clases. Cuando se utiliza CART para generar un árbol de regresión, el criterio de división es buscar aquella partición que minimiza el error cuadrático. La predicción en cada hoja está basada en la media ponderada del nodo.

El árbol de decisión CHAID (CHi-squared Automatic Interaction Detection) proviene de desarrollos estadísticos. Para cada atributo de entrada, CHAID identifica el par de valores en el

conjunto de datos que es significativamente menos diferente respecto al valor objetivo. Esta diferencia significativa es medida a través del p-valor obtenido mediante pruebas estadísticas. El p-valor es utilizado en estadística para realizar contrastes de hipótesis. Este parámetro indica cuál es el menor nivel de significación que puede escogerse para el cual se aceptaría una hipótesis alternativa, siendo la hipótesis de nulidad la que se desea probar. El tipo de test a utilizar depende del atributo objetivo. Para cada par seleccionado, CHAID comprueba si el p-valor obtenido es mayor que cierto umbral de unión. Si la respuesta es positiva, junta los valores y busca por otros pares potencialmente unificables. Este proceso es repetido hasta que no hay más pares para unir. Se selecciona entonces el mejor atributo de entrada para ser utilizado para dividir el nodo actual, de tal manera que cada nodo secundario está hecho de un grupo de valores homogéneos del atributo seleccionado. Este algoritmo soporta valores faltantes y no emplea técnicas de poda.

Las técnicas de regresión también han sido utilizadas, y continúan siéndolo, para resolver problemas de clasificación en Minería de Datos. En casos simples, se puede utilizar la regresión lineal para predecir valores futuros de datos. Sin embargo, los problemas del mundo real suelen no ser proyecciones lineales de datos pasados, por lo que es necesario aplicar técnicas de regresión más avanzadas, como por ejemplo regresión logística.

La regresión logística es una generalización de la regresión lineal. Es utilizada principalmente para predecir variables binarias y ocasionalmente variables multi-clase. Esta metodología se enmarca dentro del conjunto de Modelos Lineales Generalizados (GLM). A diferencia de los árboles de decisión, la regresión logística devuelve la probabilidad de que un valor pertenezca a determinada clase. Dado que tiene sus orígenes en la rama de estadísticos de la Minería de Datos, muchas veces es preferida como técnica de clasificación por sobre los árboles de decisión o redes neuronales.

Este último método mencionado es de particular interés ya que las redes neuronales tienen la posibilidad de resolver complejos problemas. Los algoritmos de redes neuronales pueden ser utilizados para predecir variables objetivo de clase y numéricas.

La historia de las redes neuronales puede ser dividida en dos grandes épocas. Una primera etapa en donde tuvieron un corto éxito, de 1940 a 1970 y el resurgimiento a partir de los años 80 (Roberts, 2014) con su aplicación hasta el día de hoy.

El trabajo más antiguo en este ámbito data de 1940 cuando McCulloch y Pitts introdujeron el primer modelo de cómputo de redes neuronales. En 1950, Rosenblatt desarrolla el modelo más básico de redes neuronales, el perceptrón, que era capaz de aprender ciertas clasificaciones al ajustar los pesos de las conexiones. Dado que este modelo tenía muchas limitaciones, el interés en las redes neuronales decayó. En 1980 resurgen gracias al desarrollo de las redes multicapa, particularmente las redes “backpropagation”. Hoy en día el uso de algoritmos de redes neuronales por analistas de Minería de Datos, especialmente en ámbitos comerciales, no es tan común como el de árboles de decisión o regresión logística ya que los resultados de este tipo de algoritmo suelen ser muy difíciles de explicar para el usuario regular. A su vez, las redes neuronales aplicadas a grandes bases de datos suelen tener problemas de performance al necesitar de mucho tiempo para computar todos los pesos. La bibliografía sobre este tipo de algoritmos es muy vasta y no entraremos en detalle en este trabajo por estar fuera del alcance del mismo.

El campo de Minería de Datos experimentó y sigue experimentado un crecimiento muy rápido dada su amplia posibilidad de aplicación a diversas industrias. Las empresas de telecomunicaciones y de tarjetas de crédito, líderes en la aplicación de técnicas de Minería de Datos, utilizan algoritmos por ejemplo para detectar fraude en el uso de sus servicios. Las compañías de seguros y mercados de valores también están interesadas en la aplicación para el mismo fin. Las prácticas en medicina también son diversas. La estadística ha estado siempre presente en esta ciencia y la Minería de Datos logró hacerse un lugar en el campo. Se pueden utilizar diferentes técnicas predictivas para evaluar la efectividad de procedimientos quirúrgicos o medicamentos. La industria de supermercados utiliza la Minería de Datos para decidir qué productos tener en inventario como así también para mejorar sus promociones y ofertas al poder predecir la efectividad de las mismas. En general las áreas de marketing directo son las más beneficiadas por el uso de técnicas de Minería de Datos, ya que permiten detectar qué personas estarán más dispuestas a aceptar una promoción, logrando reducir los costos en las iniciativas de comercialización. Estos son sólo algunos de los ejemplos de campos de aplicación de KDD en la actualidad. Cada vez más disciplinas hacen uso de Minería de Datos para generar conocimiento sobre sus grandes bases de datos.

La complejidad creciente en varios de los campos de aplicación y los avances en la tecnología han impuesto nuevos desafíos para la disciplina; entre ellos diferentes formatos de datos, ubicaciones dispares, avances en redes, nuevos contextos de negocio, por mencionar sólo algunos. La integración con técnicas de Visualización de la Información también es un gran reto y es lo que pretende resolver la corriente de investigación de Visual Analytics que desarrollaremos en la sección siguiente.

II.3.3 Proceso de KDD

El proceso de KDD delineado por Fayyad y otros en 1996 (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996) sigue siendo hasta el día de hoy vigente. Este proceso es interactivo e iterativo y está compuesto por varios pasos que requieren de decisiones a ser tomadas por el usuario. En la Figura II.10 puede observarse un esquema del mismo.

Antes de realizar la selección de datos, es preciso comprender el dominio de aplicación y el objetivo para el cual se llevará adelante el proceso de KDD. Una vez comprendido esto, debe seleccionarse la información sobre la cual serán realizadas las tareas de descubrimiento de patrones. Estos datos deben ser pre-procesados, implicando esto la limpieza de los mismos, eliminación de ruidos y valores extremos, tomar decisiones sobre el tratamiento de datos faltantes y considerar los datos que cambian con el tiempo. Una vez procesados los datos, los mismos pueden ser transformados para reducir la dimensión del problema. El número de variables a considerar puede reducirse o se pueden buscar nuevas formas de representar los datos. Con los datos ya transformados, se debe decidir qué técnica de minería de datos se ajusta más al objetivo del proceso de KDD bajo estudio. Se podrá elegir por tanto entre modelos de clasificación, regresión, agrupamiento (clusters), redes neuronales, entre otros. El modelo producirá una serie de patrones que deberán ser interpretados y evaluados para finalmente convertirlos en conocimiento. Este nuevo conocimiento servirá para ejecutar acciones basadas en información.

Los resultados de estas acciones retroalimentarán todo el proceso de KDD que es altamente interactivo e iterativo.

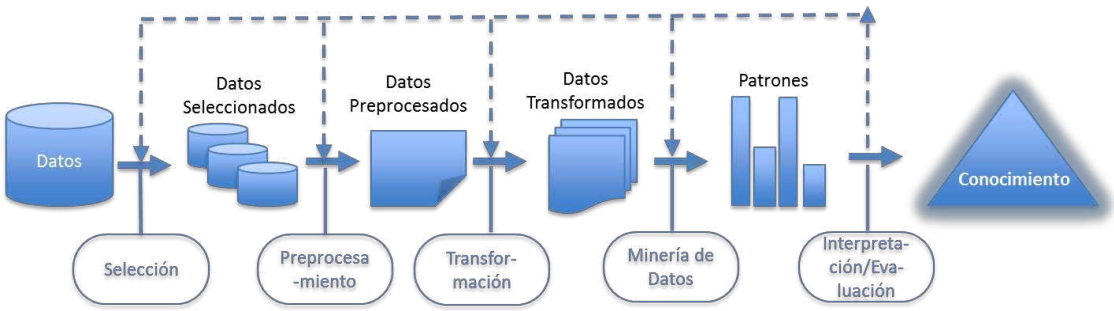

Figura II.10 Proceso de KDD (adaptado de (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996)).

II.4 El Enfoque de Visual Analytics

II.4.1 Definición

Como hemos podido comprobar en las secciones anteriores, las técnicas de análisis automático, como la Minería de Datos, se desarrollaron independientemente de las técnicas de Visualización. Sin embargo, desde hace menos de diez años existe una nueva comunidad científica que está investigando las ventajas y aplicaciones emergentes de la unificación de ambas disciplinas. Los investigadores se encuentran desarrollando el marco de lo que hoy se conoce bajo el concepto de "Visual Analytics"[3].

Según una de las primeras definiciones (Wong, 2004), Visual Analytics "es la ciencia de razonamiento analítico facilitado por interfaces humano-máquina interactivas". Sin embargo, rápidamente la misma fue mejorada, definiéndose hoy Visual Analytics como (Keim, Kohlhammer, Ellis, & Mansmann, 2010):

"la disciplina que combina técnicas de análisis automático con visualizaciones interactivas para un efectivo entendimiento, razonamiento y toma de decisiones basadas en conjuntos de *datos grandes y complejos"*

II.4.2 Evolución del Concepto

Tanto la Visualización de la Información como la Minería de Datos son métodos que sirven para transformar información en conocimiento. Sea un modelo, un patrón o una relación, la Minería de Datos puede ser vista como la presentación de un resumen de los datos en forma de información. Esta acepción está alineada con los objetivos principales de la Visualización de la

3 El término es conocido en inglés mundialmente por lo que en este trabajo no traduciremos el mismo al Español ya que se desconoce una traducción útil.

Información: encontrar una representación de los datos que reduzca la complejidad perdiendo la menor cantidad de información posible pero ganando foco en la información relevante.

Estas disciplinas han evolucionado a lo largo del tiempo en forma independiente. Por lo general la mayoría de los investigadores de KDD se han concentrado en el desarrollo de técnicas para la automatización del procesamiento de datos. Por el contrario, el objetivo de los investigadores de Visualización ha consistido en facilitar la interacción de las personas con la información mediante la reducción de datos a dimensiones procesables por la mente humana. En otras palabras, mientras que la Minería de Datos promueve la automatización, la Visualización impulsa la interacción de las personas. Esta diferencia de objetivos tiene dos causas principales (Fayyad, Grinstein, & Wierse, 2001). La primera es la creencia de que la intuición humana puede ser un estorbo más que una ayuda en el análisis de grandes bases de datos. La segunda, ha sido la fascinación por automatizar el modo en que la gente encuentra patrones en los datos.

En los últimos años, los investigadores han reconocido el potencial de la integración de procesos de Visualización y Minería de Datos. Uno de los primeros acercamientos a esta integración fue denominado "Visual Data Mining". En Agosto de 2001, se llevó a cabo el primer encuentro de investigadores de ambas áreas. El objetivo era explorar las posibilidades de involucrar a los humanos en el proceso de Minería de Datos. La idea principal consistía en que al combinar la flexibilidad, creatividad y conocimiento de las personas con el poder de las computadoras, se podría aumentar la efectividad de ambas.

La mayoría de los primeros trabajos en Visual Data Mining se enfocaron en dos instancias del proceso de KDD. Al comienzo, para la exploración visual de datos previo a las tareas de Minería de Datos, o al final para la representación de los resultados de los modelos. Sin embargo, ya desde entonces se planteaba la necesidad de un enfoque más promisorio: poder utilizar técnicas de visualización de datos e interacción en la fase de construcción de los modelos.

Estos primeros abordajes luego migraron a lo que hoy se conoce como "Visual Analytics". El término fue utilizado por primera vez en 2004 (Wong, 2004) y luego con el trabajo "Illuminating the Path: The Research and Development Agenda for Visual Analytics" (Thomas & Cook, 2005), el concepto fue ampliamente aceptado. Actualmente el término es utilizado en un contexto más amplio, describiendo nuevos campos que combinan áreas de investigación que incluyen visualización, interacción humano-computadora (HCI), análisis de datos, administración de datos, procesamiento de datos espaciales y temporales y estadística.

A pesar de que el término es relativamente nuevo, dado que se trata de una conjunción de varias disciplinas, varios trabajos anteriores a la aparición de este concepto ya contenían características de Visual Analytics. En este sentido es interesante la investigación realizada por Bertini y Lalanne (Bertini & Lalanne, 2009) en donde categorizan los trabajos en la disciplina según el peso de los componentes de visualización o minería de datos. De esta forma sugieren tres categorías: Visualización mejorada computacionalmente (V++), Minería mejorada visualmente (M++) e integración entre Visualización y Minería de Datos (VM). En la primera, agrupan todos los trabajos que son fundamentalmente de Visualización, pero que contienen algún tipo de procesamiento automático para soportar la misma. En el grupo M++ se engloban todas las investigaciones que si bien son principalmente sobre algoritmos de Minería de Datos, utilizan técnicas de Visualización para mejorar la interpretación de los resultados. Por último, separan

aquellos trabajos en donde las técnicas de Visualización y Minería de Datos están integradas de forma tal que no es posible diferenciar el rol de cada una de ellas. Este último grupo es el de nuestro interés, ya que concentra las investigaciones en lo que se conoce como "Visual Analytics".

La integración de Visualización y Minería de Datos, puede ser de tipo caja blanca o negra. En el primer caso la cooperación entre el humano y la máquina se da en el proceso de construcción de los modelos de Minería de Datos. Este tipo de integración es la que propusieron Ankerst y otros (Ankerst, Ester, & Kriegel, 2000) para la construcción cooperativa de árboles de decisión. Plantearon un sistema en donde el humano puede seleccionar los atributos de separación de los nodos y decidir qué clase asignar a cada rama del árbol. La herramienta permite incorporar el conocimiento del especialista reforzando su confianza en el modelo. Actualmente esta característica está implementada en varios programas comerciales de minería de datos, tales como SAS Enterprise Miner e IBM SPSS Modeler. En un trabajo más reciente, van den Elzen y Wijk (van den Elzen & van Wijk, 2011) presentan BaobabView. La solución permite, además de construir interactivamente árboles de decisión, comparar los diferentes resultados producidos por la selección de diferentes atributos de división y valores, características que no se encontraban presentes en los trabajos relacionados. En la Figura II.11 puede observarse la interfaz de usuario de la herramienta. Este trabajo se focaliza en la interacción del humano y en la configuración de los árboles por prueba y error. Más adelante compararemos este abordaje con el propuesto en nuestro enfoque.

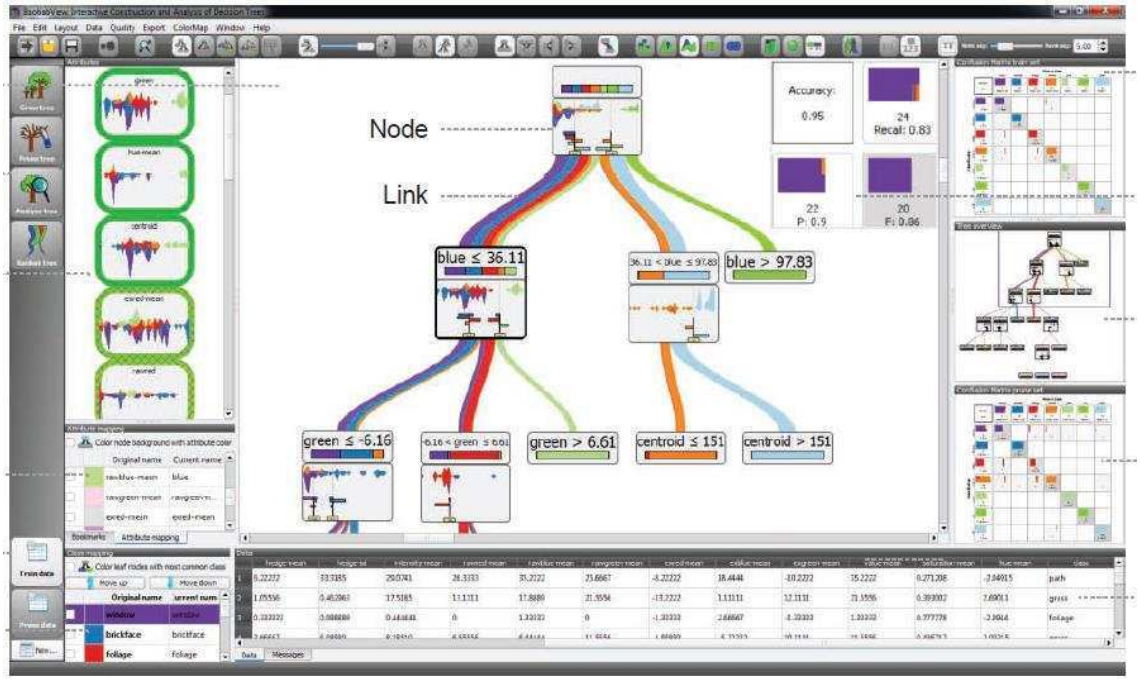

Figura II.11 Interfaz de usuario de Baobview (van den Elzen & van Wijk, 2011)

En el caso de la integración entre Visualización y Minería de datos de tipo caja negra, la interacción ocurre, pero indirectamente. Se utilizan los algoritmos de Minería de Datos en forma automática pero se le permite al usuario visualizar las diferentes configuraciones de los parámetros de los mismos. Una solución de este tipo es Morpheus (Müller, Assent, Krieger,

Jansen, & Seidl, 2008), que permite la exploración interactiva de clusters y la visualización de los resultados de las diferentes configuraciones de los parámetros de agrupamiento.

Desarrollos más recientes tienden a integrar ambas disciplinas en una forma intuitiva para el usuario, con funcionalidades exploratorias y analíticas. Tal es el caso por ejemplo de la herramienta SAS Visual Analytics[4], que casualmente lleva el nombre del área bajo estudio. Esta solución tiene funcionalidades para las dos primeras etapas del desarrollo de modelos analíticos, la exploración de los datos y la construcción de modelos de minería de datos. Sin embargo no posee aún características que le permitan al usuario comparar los resultados de los diferentes modelos. En su versión actual[5] el usuario puede generar árboles de decisión de tipo descriptivo, pero no puede realizar comparaciones visuales de los resultados de diversas configuraciones de parámetros ya que la herramienta no posee la capacidad para generar curvas ROC o matrices de confusión.

La mayoría de los desarrolladores de herramientas de Visualización tradicional han comenzado a incluir entre sus capacidades funcionalidades analíticas. Este es el caso por ejemplo de Tableau[6]. Esta herramienta nació como una solución para las tareas que se conocen como "Business Intelligence" y actualmente permite la integración con el software R para cubrir las necesidades analíticas de los usuarios[7]. La integración se basa principalmente en la visualización de modelos que ya han sido construidos en R pero no permite la modificación y visualización interactiva de cambios en los parámetros.

Por el momento, existen muy pocas soluciones de este tipo, en donde es posible interpretar visualmente las variaciones de los algoritmos de Minería de Datos respecto de cambios en los parámetros de configuración. Por tal motivo Bertini y Lalanne (Bertini & Lalanne, 2009) declaran de especial interés el estudio de técnicas de visualización eficientes que permitan realizar esta tarea. Si bien la necesidad de integración de técnicas de Visualización y Minería de Datos es evidente y continúa en constante crecimiento, no ha habido avances significativos en las herramientas de Minería de Datos para mejorar la interpretación de los parámetros visualmente. Este punto constituye el motivo principal de nuestra investigación.

En nuestro primer trabajo publicado en la sección especial sobre Análisis de Incertidumbre y Espacio de Parámetros en Visualización de la edición Computer & Graphics de Elsevier (Padua, Schulze, Matković, & Delrieux, 2014) presentamos una primera contribución a esta necesidad. En el mismo desarrollamos principalmente la visualización de los efectos de la configuración de los parámetros de árboles de decisión en los resultados medidos a través de curvas ROC y Lift principalmente. En el trabajo que aquí desarrollamos expandimos la metodología al incluir también el desarrollo de funcionalidades que faciliten la generación de múltiples árboles de decisión a los usuarios de herramientas de minería de datos.

4 http://www.sas.com/en_us/software/business-intelligence/visual-analytics.html
5 Versión Marzo 2014 – SAS Visual Analytics 6.4
6 http://www.tableausoftware.com/
7 http://www.tableausoftware.com/learn/whitepapers/tableau-and-r-faq

II.4.3 Proceso de Visual Analytics

El proceso de Visual Analytics presentado por Keim (Keim, Kohlhammer, Ellis, & Mansmann, 2010) combina métodos automáticos y visuales con la interacción humana para ganar conocimiento sobre los datos.

En la Figura II.12 puede observarse el modelo propuesto por Keim. El primer paso consiste en la preparación y transformación de los datos para su posterior exploración. Luego de esta tarea, el analista puede elegir entre el análisis visual o el automático. Si se utilizan primero métodos automáticos, se aplican algoritmos de Minería de Datos para el modelado de los mismos. Una vez que se crea el modelo, el analista debe evaluar y refinar el mismo, lo cual puede ser realizado de mejor forma a través de la interacción con los datos. Las visualizaciones permiten a los analistas interactuar con los métodos automáticos al modificar parámetros o seleccionar otros algoritmos. Luego se pueden visualizar los modelos para evaluar los conocimientos generados por los mismos.

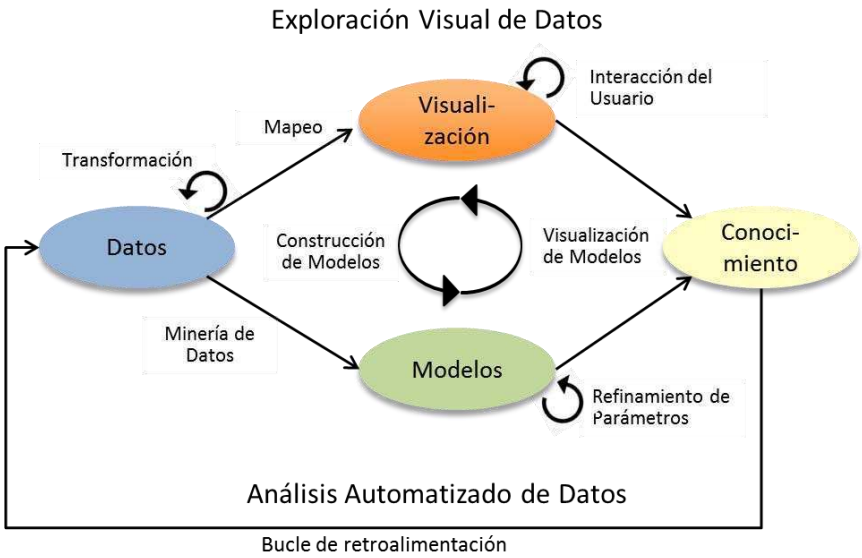

Figura II.12 El proceso de Visual Analytics (traducción de (Keim, Kohlhammer, Ellis, & Mansmann, 2010)).

Otra forma de análisis es comenzar por la exploración visual de los datos para confirmar una hipótesis generada por los métodos automáticos. La interacción es necesaria para revelar información útil que permita mejorar los procesos de análisis automático.

Este proceso tiene como objetivo integrar fuertemente los procesos automáticos con los de interacción visual. En nuestro trabajo nos enfocaremos en la visualización del refinamiento de parámetros.

II.5 Resumen

En el presente capítulo hemos presentado la visión de Visual Analytics como resultado de la conjunción de las disciplinas de Visualización de la Información y Minería de Datos. Repasamos los conceptos y evolución de cada una de las técnicas, junto con los procesos para generar visualizaciones y modelos de Minería de Datos.

Gracias al estudio del estado del arte en el área de Visual Analytics, pudimos detectar la necesidad analizar la influencia de la configuración de los parámetros de algoritmos de minería de datos mediante técnicas de visualización. Nuestro trabajo se centra en este punto, contribuyendo al desarrollo de la disciplina.

En los siguientes capítulos desarrollaremos nuestra idea y presentaremos una metodología para analizar visualmente los cambios en el poder predictivo de árboles de decisión debido a las diferentes configuraciones de los parámetros. Presentaremos un método interactivo de análisis en contraste a un método interactivo de construcción.

III. Metodologías Actuales para la Construcción y Evaluación de Árboles de Decisión

III.1 Introducción

En el capítulo anterior introdujimos la reciente disciplina de Visual Analytics a través de su historia y evolución. Mencionamos la necesidad de continuar desarrollando técnicas que combinen metodologías de Visualización de la Información y Minería de Datos. Cada algoritmo de Minería de Datos contiene parámetros específicos que precisan ser configurados para generar diferentes modelos y consecuentemente diversos resultados. A fin de acotar el alcance del presente trabajo, dentro de las técnicas de Minería de Datos disponibles, hemos seleccionado la metodología de árboles de decisión para desarrollar nuestro enfoque. Nos concentraremos en la configuración de los algoritmos de árboles de decisión del tipo CHAID (Kass, 1980). La elección de esta técnica en particular no ha sido al azar. Los árboles de decisión son ampliamente utilizados por analistas de Minería de Datos en ámbitos comerciales, debido a la facilidad de interpretación de los resultados generados por los mismos. A su vez estos algoritmos son muy flexibles dado que es posible cambiar varios parámetros para generar modelos alternativos y son menos sensibles a datos faltantes o a valores extremos (outliers). La elección del algoritmo CHAID específicamente se debe a que es el que se encuentra implementado en las herramientas de minería de datos más utilizadas por analistas corporativos, SAS Enterprise Miner e IBM SPSS Modeler (Rexer, 2012).

Para analizar la efectividad de los árboles de decisión, explicaremos el uso de las curvas ROC (Receiving Operating Characteristics Curves) y las curvas Lift. Estas curvas son ampliamente utilizadas por los analistas de Minería de Datos y constituyen un medio visual de interpretación de los resultados del modelado.

En este capítulo explicaremos la metodología de construcción y evaluación de árboles de decisión actual. Haremos foco en el estado de integración de Visualización de la Información y Minería de Datos en cada una de las etapas de construcción de un modelo predictivo. Este marco contribuirá a evaluar las mejoras que proponemos en nuestro aporte, el cual desarrollaremos en el capítulo siguiente.

III.2 Algoritmo CHAID

III.2.1 Método de Análisis

El acrónimo CHAID proviene de las siglas del inglés *"Chi*-Squared Automatic *Interaction Detection"*. Fue propuesto por Kass en 1980 (Kass, 1980) como una variación a las primeras técnicas presentadas en 1963 por Morgan y Sonquist (Morgan & Sonquist, 1963), denominadas *"Automatic Interaction Detection"* (AID). Estas técnicas son una familia de métodos utilizados para el análisis de conglomerados para variables categóricas no continuas. Estos algoritmos pueden ser utilizados tanto para describir conjuntos de datos como para predecir comportamientos.

El objetivo del procedimiento CHAID es dividir la población sucesivamente en dos o más grupos, a partir de una variable independiente que mejor explique la clase a la cual pertenece cada instancia evaluada. Esta evaluación del mejor predictor se hace utilizando el p-valor de la distribución Chi-Cuadrado, de ahí su nombre.

El p-valor es utilizado en Estadística para realizar contrastes de hipótesis. Este parámetro indica cuál es el menor nivel de significación que puede escogerse para el cual se aceptaría una hipótesis alternativa, siendo la hipótesis de nulidad la que se desea probar. El p-valor, aplicado al algoritmo CHAID, representa la probabilidad observada entre el predictor y la variable dependiente que se presentaría si fueran estadísticamente independientes. Por ejemplo, un p-valor de 0.05, significa que la relación observada entre el predictor y la variable dependiente se presentaría solamente 5% de las veces, si las variables fueran independientes.

CHAID utiliza el estadístico Chi-cuadrado para analizar si las diferencias entre la distribución de los valores observados y la distribución de los valores esperados son debidas al azar, o bien si no cumple la hipótesis establecida. En el caso de los árboles de decisión, la hipótesis nula es que las variables son independientes. El algoritmo calcula el estadístico Chi-cuadrado para cada combinación de la variable dependiente con respecto a la independiente y determina el valor del mismo según la tabla de contingencia.

Una tabla de contingencia es una tabla de doble entrada en donde en cada casilla figura la cantidad de casos o individuos que poseen las ambas características analizadas. Por ejemplo, en la Tabla III-1 se observa una tabla de contingencia de 2x2 en donde cada celda contiene la cantidad de individuos de determinado sexo (hombre o mujer) que cumplen con una condición de fumador (no fuma, fuma).

	Sexo		
	Hombre	**Mujer**	**Marginal**
Fumador	65	58	123
No Fumador	43	67	110
Marginal	108	125	233

Tabla III-1 Tabla de Contingencia Sexo versus Fumador

Con estos valores se puede calcular el estadístico Chi-Cuadrado para analizar la dependencia de las variables. Su cálculo permite conocer con un nivel de confianza estadístico determinado si los niveles de una variable influyen en los niveles de otra variable. Para el caso de ejemplo de la Tabla III-1, el estadístico determinará si el sexo de una persona es un factor determinante en su condición de fumador. Pearson planteó la utilización del estadístico Chi-Cuadrado (x^2) definido por:

$$x^2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(n_{ij} - E_{ij})^2}{E_{ij}}$$

donde:
n_{ij} número de observaciones (frecuencia) que tienen el atributo i y el atributo j
E_{ij} número de observaciones (frecuencia) esperadas en condiciones de independencia

Este valor es comparado contra el valor tabulado de la tabla Chi-Cuadrado para un nivel de confianza determinado y (n-1) (k -1) grados de libertad. Si el valor calculado es mayor que el valor de la tabla significará que las diferencias entre las frecuencias observadas y las esperadas son muy elevadas y por lo tanto se puede determinar con cierto nivel de confianza que existe dependencia entre los atributos analizados.

Para ejemplificar el cálculo de Chi-Cuadrado utilizamos los valores de la Tabla III-1 y realizamos los cálculos necesarios y los comparamos versus el valor del estadístico de la tabla Chi-cuadrado. En la Tabla III-2 se pueden observar las distribuciones esperadas en caso de independencia. Cada celda se calcula como el producto del total de casos para cada atributo dividido el total de casos bajo análisis.

	Sexo		
	Hombre	**Mujer**	**Marginal**
Fumador	123*108/233=57	123*125/233=66	123
No Fumador	110*108/233=51	110*125/233=59	110
Marginal	108	125	233

Tabla III-2 Valores esperados de distribución en caso de independencia

Luego para calcular el estadístico aplicamos la fórmula de Pearson:

$$x^2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(n_{ij} - E_{ij})^2}{E_{ij}} = \frac{(65-57)^2}{57} + \frac{(58-66)^2}{66} + \frac{(43-51)^2}{51} + \frac{(67-59)^2}{59}$$
$$= 4.42$$

Dado que el valor resultante es mayor que el de la tabla para un nivel de confianza del 95% con 1 grado de libertad (3.84), se rechaza la hipótesis nula de independencia entre los factores, es decir se considera que el sexo de una persona influye en su condición de fumador.

CHAID realiza varias pruebas Chi-cuadrado a lo largo de la ejecución del algoritmo. Esto tiene como consecuencia el aumento de la probabilidad de aceptar erróneamente la relación entre una variable dependiente y una independiente. Para disminuir este riesgo, Kass (1980) propuso el uso del “ajuste de Bonferroni”, que consiste en una técnica estadística que ajusta el nivel de significación en relación al número de pruebas estadísticas realizadas simultáneamente sobre un conjunto de datos. Este ajuste produce un p-valor ajustado que es el que utiliza el algoritmo para el análisis de la dependencia de las variables.

Si bien en su trabajo Kass propone el modelo sólo para variables independientes categóricas, la metodología también puede aplicarse con variables numéricas. Para lograrlo, estas variables se convierten a intervalos o rangos.

III.2.2 El Algoritmo

El algoritmo consiste en tres grandes pasos: fusión, división y detención.

Fusión

Para cada variable predictora X, unir las categorías no significativas. Cada categoría final de X resultará en un nodo hijo si X es utilizada para dividir el nodo. El paso de fusión también calcula el p-valor ajustado que será utilizado en el paso de división.

1. Si X tiene una sola categoría, parar y fijar el p-valor ajustado en 1.
2. Si X tiene dos categorías, ir al paso 8.
3. Si no se cumplen ninguna de las condiciones anteriores, encontrar el par de categorías de X que es significativamente menos diferente (más similar). El par más similar es aquel cuyo test estadístico tiene el mayor p-valor con respecto a la variable dependiente Y.
4. Para el par que tiene el p-valor mayor, verificar si el mismo es mayor que el definido por el usuario. Si lo es, este par es fusionado en una sola categoría compuesta. Luego un nuevo conjunto de categorías X es formado. Si no lo es, ir al paso 7.
5. Si la nueva categoría X formada consiste de 3 o más categorías originales, entonces encontrar la mejor división binaria dentro de la categoría compuesta en donde el p-valor sea el menor. Realizar esta división binaria si el p-valor no es mayor que un valor alfa de división.
6. Ir al paso 2.
7. Cualquier categoría que tenga muy pocas observaciones (comparado con el mínimo seleccionado por el usuario), es unida con la categoría más similar, medida por el p-valor más grande.
8. Se calcula el p-valor para las categorías fusionadas aplicando los ajustes de Bonferroni.

División

La mejor división para cada predictor es encontrada en el paso de fusión. El paso de división selecciona el mejor predictor para dividir el nodo. Esta selección es realizada comparando cada p-valor ajustado asociado a cada predictor. El p-valor ajustado es obtenido en el paso de fusión.

1. Seleccionar el predictor que tenga el menor p-valor ajustado, es decir, el más significativo.
2. Si el p-valor ajustado es menor o igual al valor definido por el usuario, dividir el nodo utilizando este predictor. Caso contrario, no dividir el nodo. Se convierte en un nodo terminal.

Detención

El paso de detención corrobora si el proceso de crecimiento del árbol debe continuar o si debe detenerse teniendo en cuenta las siguientes reglas.

1. Si el nodo es puro (todos los casos en el mismo tienen la clase de la variable dependiente), no se dividirá el mismo convirtiéndose en nodo terminal.
2. Si todos los casos en el nodo tienen valores idénticos para cada predictor, el nodo no será dividido.

3. Si la profundidad del árbol alcanza el máximo definido por el usuario, terminar el proceso de crecimiento.
4. Si el tamaño de un nodo es menor al límite definido por el usuario, el mismo no será dividido.
5. Si la división del nodo resulta en un nodo hijo cuyo tamaño es menor al mínimo definido por el usuario, este nodo hijo será fusionado con el nodo hijo más similar en términos de p-valor. Sin embargo, si el número de nodos resultantes es 1, el nodo padre no será dividido.

A continuación describimos los pasos del algoritmo mediante un pequeño ejemplo. El mismo ha sido adaptado del trabajo de Kass (1980). Los datos analizados corresponden un estudio realizado a 669 estudiantes de la Universidad de Witwatersrand, Sudáfrica. Los estudiantes son seleccionados a partir de los exámenes de matriculación. Se cree que los diferentes comités de evaluación no evalúan de la misma manera a los estudiantes y por ello los estudiantes tienen diferentes resultados dependiendo del comité que los examinó. Esta afirmación es producto de una simple matriz cruzada entre comité de evaluación y desempeño del estudiante en la universidad. Para poder refutar o afirmar la aseveración, se decidió realizar una investigación para poder asociar el desempeño de los estudiantes con otras variables explicativas.

Se clasificaron los comités evaluadores en 4 categorías: JMB (Joint Matriculation Board), NSC (National Senior Certificate), TUEC (Transvaal University Entrance Certificate) y Otros. Las variables independientes son:

- año de ingreso (predictor libre, es decir cualquier agrupación es permitida ya que no se trata como un número ordinal, sino como una categoría)
- número de materias comerciales incluidas en la matriculación (predictor libre, ídem año de ingreso)
- tipo de escuela a la que asistió
- calificación en matemáticas

Con el fin de reducir el ejemplo, se analizará solamente el primer paso del algoritmo para la primera variable predictora, año de ingreso. El objetivo en este paso es encontrar el par de categorías que mejor represente la variable independiente.

	Variable Dependiente				
Año	**Otros**	**JMB**	**TUEC**	**NSC**	**Total**
1960	7	39	107	17	**170**
1961	13	34	112	17	**176**
1962	11	30	93	26	**160**
1963	13	34	80	36	**163**
Total	**44**	**137**	**392**	**96**	**669**

Tabla III-3 Tabla de Contingencia Año de Ingreso versus Comité Evaluador

A partir de la tabla de contingencia (Tabla III-3) puede evaluarse el estadístico Chi-Cuadrado para saber si las variables son independientes. El cálculo del mismo se realiza con la fórmula propuesta por Pearson. Para ello es necesario calcular las frecuencias esperadas para cada combinación de atributos. Estos cálculos pueden observarse en la Tabla III-4.

	Variable Dependiente				
Año	**Otros**	**JMB**	**TUEC**	**NSC**	**Total**
1960	11	35	100	24	**170**
1961	12	36	103	25	**176**
1962	11	33	94	23	**160**
1963	11	33	96	23	**163**
Total	**44**	**137**	**392**	**96**	**669**

Tabla III-4 Frecuencias Esperadas para los valores Año de Ingreso y Comité Evaluador

Luego se calcula el x^2 con la fórmula antes expuesta:

$$x^2 = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(n_{ij} - E_{ij})^2}{E_{ij}} = 19.1$$

Dado que se busca disminuir la cantidad de categorías, el algoritmo evalúa las 6 combinaciones posibles para los valores de año de ingreso y calcula los valores Chi-Cuadrado para cada combinación. El valor año de ingreso es tomado en este caso como una variable libre, es decir que se toma como categórica y no ordinal, pudiendo realizar así combinaciones de años no consecutivas.

La Tabla III-5 muestra los valores para la combinación de los 2 primeros años en una sola categoría. Esta combinación otorga como resultado un estadístico Chi-Cuadrado de 2.2.

	Variable Dependiente				
Año	**Otros**	**JMB**	**TUEC**	**NSC**	**Total**
1960, 1961	20	73	219	34	**346**
1962	11	30	93	26	**160**
1963	13	34	80	36	**163**
Total	**44**	**137**	**392**	**96**	**669**

Tabla III-5 Tabla de Contingencia Año de Ingreso versus Comité Evaluador con los dos primeros valores unificados

El algoritmo evalúa a continuación las 3 fusiones posibles. La tabla 3.6 contiene 2 valores posibles para la categoría año. Esta fusión produce un valor Chi-Cuadrado de 3.

	Variable Dependiente				
Año	**Otros**	**JMB**	**TUEC**	**NSC**	**Total**
1960, 1961	20	73	219	34	**346**
1962, 1963	24	64	173	62	**323**
Total	**44**	**137**	**392**	**96**	**669**

Tabla III-6 Tabla de Contingencia Año de Ingreso versus Comité Evaluador con los dos primeros valores unificados

La última fusión posible (las 2 nuevas categorías), arroja un valor Chi-Cuadrado de 13.7, el cual es significativo con un nivel de confianza de 0.0035, por lo cual la unión no es realizada. Como no se puede fusionar más la categoría, se acepta esta tabla de 2x4 como la mejor y se estima su significación con el ajuste de Bonferroni en 0.024.

Este trabajo es realizado para cada una de las variables predictoras. Una vez que se tienen todas las tablas finales, se comparan los valores Chi-Cuadrado y se elige el más significativo para hacer la división del grupo en el primer nivel del árbol. La variable más significativa resultó ser el número de materias comerciales, agrupado en 4 categorías como muestra la Figura III.1 (Kass, 1980). Cada uno de los 4 grupos es luego utilizado para armar seguir dividiendo en grupos, excepto en el caso en donde se haya llegado a un grupo de 100 estudiantes (valor definido por el usuario). Para cada grupo (nodo del árbol) se procede de la misma forma que para el grupo total. En el caso de ejemplo, la variable tipo de escuela, resultó ser la mejor predictora para ambos subgrupos, generando 4 grupos en el segundo nivel del árbol. De estos grupos sólo 2 pueden ser subdivididos ya que tienen el número suficientes de observaciones. Sin embargo, para el grupo de 140 estudiantes, ninguna variable predictora superó el mínimo de nivel de significación configurado por el usuario, por lo que no pudo ser dividido.

III.3.3 Los Parámetros

En varios de los pasos del algoritmo descripto anteriormente, se compara el resultado obtenido versus un parámetro definido por el usuario. Estos parámetros determinan el modelo final, al fijar los límites de división de las ramas del árbol. A su vez estos parámetros determinan las reglas de parada del algoritmo.

P-Valor

Uno de los parámetros a definir por el usuario es el p-valor. Este valor será el que se compare contra el valor Chi-cuadrado en los pasos de fusión y división. Por lo general este estadístico suele fijarse en 0.05, aunque en algunos programas de minería de datos, el valor está predeterminado en 0.2. Recordemos que en el caso de los árboles de decisión, la hipótesis nula es que las variables son independientes, por lo tanto, mientras menor es el valor del estadístico, mayor es la dependencia entre las variables. Al fijar un valor pequeño, se le exige a las variables que sean más dependientes, por lo tanto más explicativas.

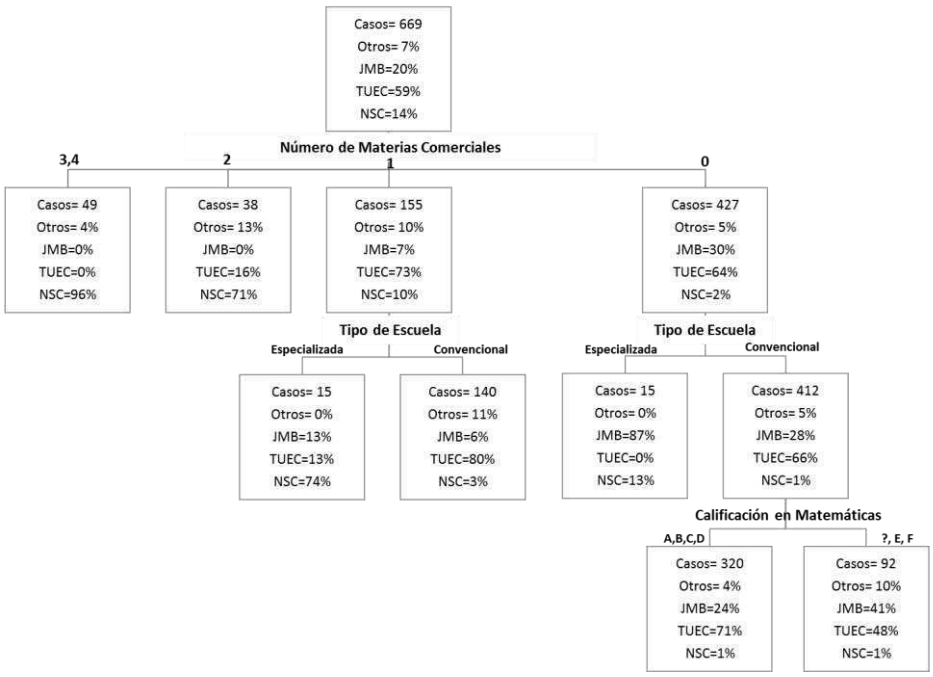

Figura III.1 Árbol de Análisis de los Comités Evaluadores de Sudáfrica (Adaptado de (Kass, 1980))

La fijación del p-valor en 0.05 es totalmente arbitraria. El uso de este valor fue impuesto por Fisher en su trabajo *"Statistical Methods for Research Workers"* (SMRW) (Dallal, 2012) y es utilizado desde entonces como medida de significancia estadística. Sin embargo, en otros trabajos de Fisher, la medida no es tan taxativa y se fija como máximo un p-valor de 0.1.

El valor que se desee fijar para este parámetro dependerá de cuán estricto se desee ser respecto de la prueba de independencia de variables y de la variable objetivo que se necesite predecir. Es decir, el nivel de error permitido podría ser distinto para predecir enfermedades o para predecir la respuesta a una campaña de marketing. En el primer caso se deseará ser más estricto ya que se trata de un problema que requiere mayor precisión en los resultados. En general se interpreta el p-valor de la siguiente manera:

- p-valor > 0.1 no puede encontrarse evidencia para rechazar la hipótesis nula.
- 0.05 < p-valor < 0.1 evidencia débil para rechazar la hipótesis nula
- 0.01 < p-valor < 0.05 evidencia relativamente fuerte para rechazar la hipótesis nula
- p-valor <0.01 evidencia muy fuerte para rechazar la hipótesis nula

Estos valores pueden ser encontrados en la mayoría de las aplicaciones estadísticas, sin embargo en minería de datos, el p-valor suele tener un máximo de 0.2 ya que hay otras combinaciones de parámetros que determinarán la relación entre las variables.

Máxima profundidad del árbol

El algoritmo seguirá distribuyendo los casos en ramas hasta tanto se cumpla alguna de las reglas de detención. Una de estas reglas es el parámetro fijado por el usuario como máxima profundidad. Establecer un límite a la generación de nodos hijos tiene como objetivo disminuir la complejidad de las reglas resultantes como así también el sobreajuste a los casos de entrenamiento.

El sobreajuste de un árbol implica que las reglas generadas pueden predecir muy bien el valor de los casos de entrenamiento pero no tienen la misma efectividad con los casos nuevos, es decir, no generaliza. Esto se mide analizando los resultados para el conjunto de entrenamiento y el de validación.

Mínima cantidad de casos por hoja

El paso 5 de la fase de detención del algoritmo determina que un nodo no será dividido si el resultado genera hojas de tamaño menor al límite definido por el usuario. Este parámetro suele ser conocido como "Tamaño de la hoja". Al igual que en el caso de máxima profundidad, se utiliza esta regla de parada para evitar el sobreajuste. No existe un criterio general para calcular el tamaño óptimo de una hoja. Algunos programas de minería de datos tienen como default 5 casos mínimos por hoja. Este número es arbitrario.

En el ejemplo analizado, este valor fue fijado en 100, por lo que las hojas con menor cantidad de casos no fueron evaluadas para continuar la división.

En la mayoría de los programas de minería de datos, además de los parámetros de parada y división propios del algoritmo CHAID, existe otra serie de parámetros que pueden ajustarse para obtener modelos. Uno de los más utilizados es el de *"Máxima Cantidad de Ramas".* Este parámetro determina en cuántas ramas puede dividirse un nodo. Por lo general el valor por defecto suele ser 2, para construir árboles binarios. Sin embargo este parámetro podría fijarse en otros valores. En el caso de CHAID la máxima cantidad de ramas que puede tener un árbol será igual al número máximo de categorías que existen en una variable categórica o intervalos en una variable numérica. La cantidad final dependerá del nivel de significación elegido. Para el caso del ejemplo, la variable "Número de Materias Comerciales" generó 4 ramas, 1 de las cuales es una conjunción de categorías, mientras que el resto está representada por un solo valor.

III.3.4 Medidas de Resultados

Matriz de Confusión

La matriz de confusión se crea analizando los 4 resultados posibles de un clasificador. Si la instancia es positiva y es clasificada como tal, se contabiliza como un verdadero positivo, mientras que si la misma se clasifica como negativa, sin serlo, se considera un falso negativo. Si el registro es negativo y es clasificado como tal, se considera verdadero negativo y de lo contrario, siendo clasificado como positivo, se considera un falso positivo. Estas métricas pueden visualizarse en lo que se denomina "Matriz de Confusión" (Tabla III-7). Esta matriz se ajusta a clasificadores binarios en donde cada caso puede ser considerado como positivo o negativo.

		Clase Verdadera	
		Positiva	Negativa
Clase Pronosticada	Positiva	Verdaderos Positivos (VP)	Falsos Positivos (FP)
	Negativa	Falsos Negativos (FN)	Verdaderos Negativos (VN)
	Totales	**Positivos (P)**	**Negativos (N)**

Tabla III-7 Matriz de Confusión

Los números en la diagonal principal representan las decisiones correctas, mientras que la diagonal inversa representa los errores, también llamados confusión, entre las clases. Esta matriz deriva en las métricas básicas de análisis de un clasificador:

$$tasa\ fp = \frac{FP}{N} \qquad tasa\ vp = \frac{VP}{P}$$

$$tasa\ vn = \frac{VN}{VN + FP} \qquad tasa\ acierto\ (recall) = \frac{VP}{P}$$

$$precisión = \frac{VP}{VP + FP} \qquad exactitud\ (accuracy) = \frac{VP+VN}{P+N}$$

La tasa de verdaderos positivos (tasa vp), suele llamarse también sensibilidad. Otra métrica que se utiliza es la de especificidad, que no es más que la tasa de verdaderos negativos (tasa vn).

Estas métricas suelen no ser suficiente en la mayoría de los casos en donde se necesita además de la clasificación, un ranking de posibles resultados positivos (Ling, Huang, & Zhang, 2003). Para solucionar esta limitación, surge el uso de las curvas ROC en el ámbito de la Minería de Datos. A continuación explicaremos las mismas.

Curva ROC

La curva ROC (del inglés Receiver Operating Characteristic Curve) es una técnica gráfica que sirve para visualizar, organizar y seleccionar clasificadores basados en su performance (Fawcett, 2004). El uso de curvas ROC se extiende a varios tipos de clasificadores, no sólo a árboles de decisión. Dado el alcance de este trabajo, sólo detallaremos su uso para el análisis de este tipo de clasificadores.

Los gráficos de curvas ROC se construyen utilizando los valores de la matriz de confusión explicadas anteriormente. La curva ROC se representa sobre un gráfico de dos ejes en donde en el eje horizontal se mapea la tasa de falsos positivos (tasa fp) y en el eje vertical la tasa de verdaderos positivos (tasa vp). Los ejes también suelen usarse en términos de sensibilidad y especificidad. La diferencia consiste en que en el eje de las x se grafica 1 – especificidad.

Los árboles de decisión producen una sola clasificación para cada instancia, por ende una única matriz de confusión, que corresponde a un único punto ROC. En la Figura III.2 podemos observar 5 puntos diferentes correspondientes a 5 árboles de decisión distintos.

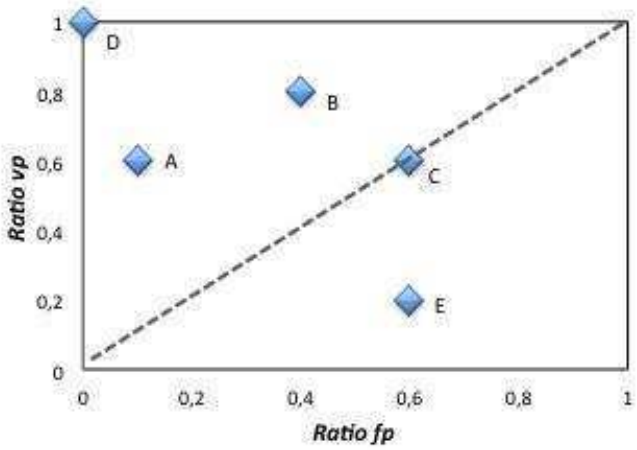

Figura III.2 Cinco clasificadores discretos plasmados en un gráfico ROC (adaptación (Fawcett, 2004)).

El punto D representa el caso de un clasificador perfecto, es decir, aquél que clasificó correctamente todas las instancias. La línea de la diagonal representa el azar, es decir que el clasificador califique el 50% de los casos positivos como tales y el 50% restante como negativos. Mientras más alejado hacia el extremo superior izquierdo se encuentre un modelo, mejor es el mismo. Pero es importante destacar que no hay una métrica universalmente aceptada al respecto (o sea, no alcanza con saber qué lejos está el punto de la diagonal). Por ejemplo el modelo correspondiente al punto A del gráfico es mejor que el C que está sobre la curva de azar y considerablemente superior al E, que se encuentra por debajo de la línea de referencia.

Una de las ventajas de la curva ROC es que es insensible a los cambios en las distribuciones de las clases. Si la proporción de casos positivos y negativos cambia en el conjunto de validación, la curva ROC no cambiará.

Para poder generar una curva ROC en lugar de puntos únicos en el espacio ROC, es preciso convertir la clasificación del árbol de decisión en una puntuación en lugar de una clase. Cada hoja final de un árbol clasificará las instancias contenidas en él según la proporción de casos positivos y negativos que contenga. Así, por ejemplo, en una hoja en donde el 90% de los casos sea positivo, el 10% restante también será clasificado como tal a pesar de no serlo. Para graficar la curva ROC para un árbol de decisión, cada hoja del árbol es convertida a un punto dentro del espacio ROC. Las hojas son ordenadas de menor a mayor en términos la tasa de falsos positivos (o 1 – especificidad) y luego los segmentos son unidos para obtener una curva. Los diferentes

sistemas de Minería de Datos tienen diferentes estrategias para graficar y suavizar la curva. Sin embargo en términos generales la mayoría de ellos funcionan de la forma anteriormente expuesta.

A diferencia de las curvas ROC para modelos que ordenan (rankeadores), cuando se trata de curvas ROC de modelos clasificadores, no se utiliza un nivel de corte (cutoff) para generar las mismas. Sin embargo algunas herramientas de Minería de Datos tienen la posibilidad de modificar el punto de corte de las hojas de los árboles. Es decir, se puede exigir que una determina hoja del árbol clasifique todos los casos de la misma en positivos, sólo si, por ejemplo, el 80% de los casos son positivos. La manera por defecto es clasificar la hoja según la mayoría, es decir, por encima del 50%.

Área bajo la Curva ROC (AUC)

Si bien las curvas ROC son fácilmente comparables visualmente, a veces es preciso contar con una sola medida escalar que determine qué modelo clasifica mejor. Una forma de realizar esto es mediante el cálculo del área bajo la curva ROC (AUC) (Bradley, 1997). Este valor será siempre variará entre 0.5 y 1, ya que ningún modelo puede dar una AUC menor que el azar, que es la diagonal que produce un área bajo la curva de 0.5.

Curva Lift

La curva Lift es muy utilizada por los analistas de minería de datos en áreas de marketing o ventas (Vuk, 2006). El lift es una medida de efectividad de un modelo predictivo calculada como la relación entre los resultados obtenidos con el modelo predictivo, versus los obtenidos sin el mismo. El lift se calcula por lo general dividiendo la base en deciles y ordenando los registros de mayor probabilidad de respuesta a menor probabilidad. Luego para cada decil se calcula la proporción de casos positivos que identificó el modelo versus el de la población en general. Estos datos se trasladan a un gráfico de dos ejes, en donde el eje de las x representa los deciles y el de las y la relación calculada (Figura III.3).

En la figura se observa además de la curva Lift, otra curva muy utilizada que es la curva Lift acumulada. Con esta curva es posible determinar qué nivel de ganancia (lift) se obtiene al realizar el corte de la base en determinado punto. Así por ejemplo para el caso de la Figura III.2 si seleccionamos el 40% de la base, obtendremos un lift de 2.05, es decir una proporción del doble de casos positivos que si hubiésemos seleccionado casos al azar. Esta curva es de gran utilidad en áreas de marketing directo. Si bien la curva es muy similar a la curva ROC, la misma difiere básicamente en que la curva lift depende de la cantidad de casos, mientras que la ROC no.

III.4 Metodología Actual para la Construcción y Evaluación de Árboles de Decisión de tipo CHAID

III.4.1 Proceso de Modelado Tradicional

En el capítulo II detallamos el proceso de KDD tradicional (Figura II.10), compuesto por 5 etapas: selección, pre-procesamiento, transformación, minería de datos y evaluación e interpretación. La etapa de minería de datos consiste en elegir el modelo que mejor se ajuste al objetivo propuesto. Por lo general en esta etapa son evaluadas diversas técnicas de minería de datos, tales como regresión, árboles de decisión, redes neuronales, entre otras. Cada uno de los

algoritmos debe ser configurado con diversos parámetros que luego producirán distintos resultados.

A continuación enumeramos los pasos que se realizan generalmente para construir un árbol de decisión basado en un algoritmo CHAID, ejemplificando este proceso con el uso de la herramienta de minería de datos SAS Enterprise Miner. Dado que el foco del presente trabajo es en las etapas de modelado y análisis de resultados, no detallaremos las tareas que involucran la selección, el procesamiento y transformación de datos.

		Métricas por Decil				Métricas Acumuladas			
Decil	Base	Nro de Casos por Decil	Cantidad Casos Positivos	% Casos Positivos	Lift	Nro de Casos Acum.	Cantidad Casos Positivos Acum.	% Acum Casos Positivos	Lift Acum.
1	1	100	16	16%	3,2	100	16	16%	3,20
2	1	100	12	12%	2,4	200	28	14%	2,80
3	1	100	8	8%	1,6	300	36	12%	2,40
4	1	100	5	5%	1,0	400	41	10%	2,05
5	1	100	3	3%	0,6	500	44	9%	1,76
6	1	100	2	2%	0,4	600	46	8%	1,53
7	1	100	1	1%	0,2	700	47	7%	1,34
8	1	100	1	1%	0,2	800	48	6%	1,20
9	1	100	1	1%	0,2	900	49	5%	1,09
10	1	100	1	1%	0,2	1000	50	5%	1,00
Total		**1000**	**50**	**5%**					

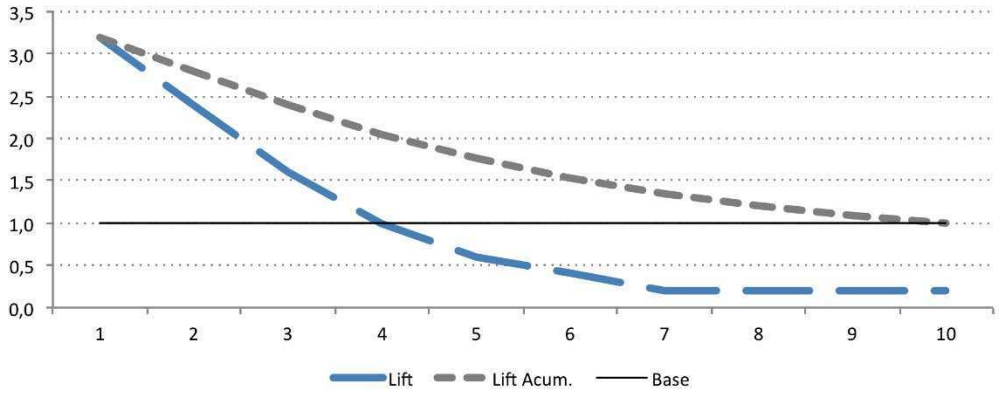

Figura III.3 Caso Ejemplo Curva Lift y Lift Acumulado (adaptado de Coppock, (Coppock, 2002))

III.4.1.1 Configuración inicial de los parámetros seleccionados

Una vez que el conjunto de datos y las variables a utilizar están seleccionadas, se procede a elegir el algoritmo a utilizar y configurar los parámetros del mismo. Por lo general el analista selecciona unos pocos parámetros a variar para realizar diversas configuraciones. Estos parámetros suelen ser los de parada del algoritmo: máxima profundidad y mínima cantidad de casos por hoja. También suele configurarse la apertura máxima en ramas que se desea. Este parámetro no constituye un criterio de parada pero es muy utilizado por los analistas.

SAS Enterprise Miner cuenta con un algoritmo de árboles de decisión propio, el cual puede ser aproximado a CHAID ajustando las opciones de configuración del mismo[8]. Este algoritmo es accedido mediante un nodo que contiene diversos parámetros que pueden ser ajustados (Figura III.4). El usuario de la herramienta debe ingresar los valores que desee en cada parámetro en el panel de la izquierda.

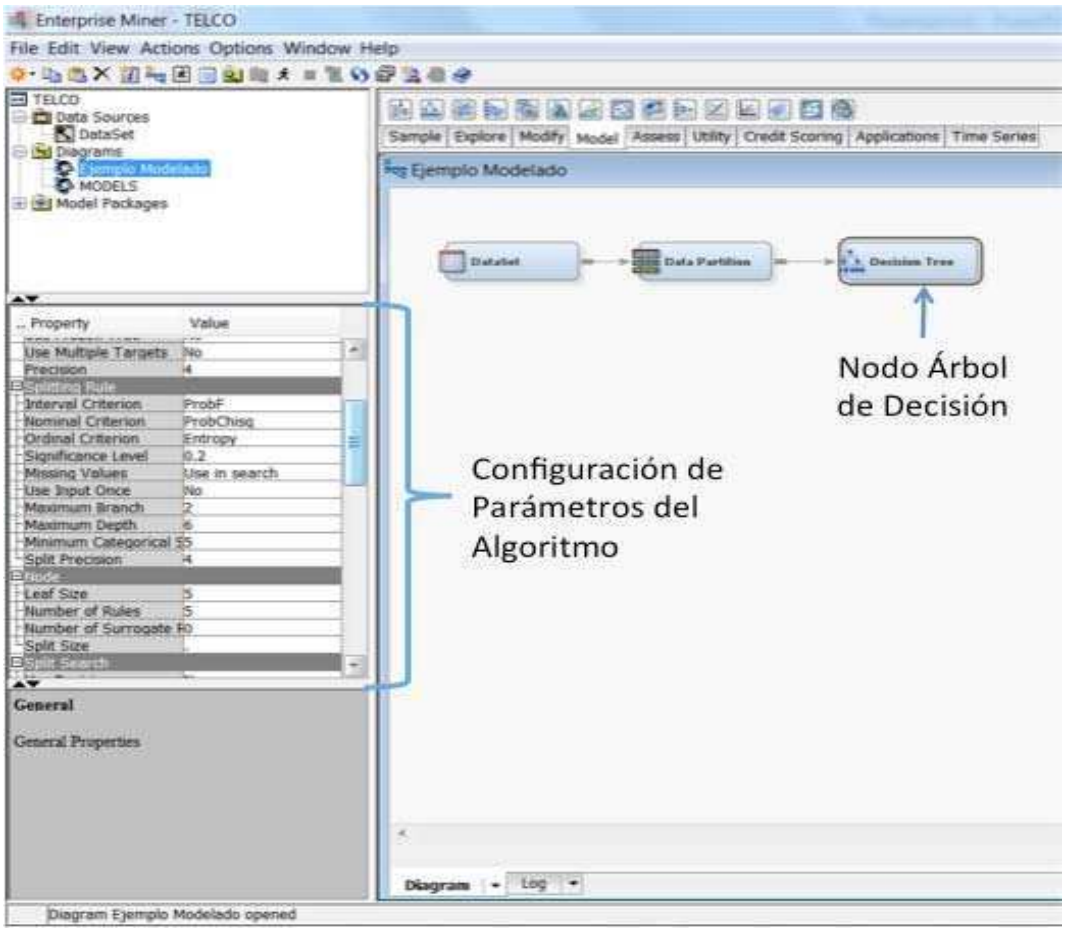

Figura III.4 Generación de un árbol de Decisión con SAS Enterprise Miner

III.4.1.2 Ejecución de las combinaciones seleccionadas

Por lo general estas combinaciones no son más de 3 o 4 ya que se prioriza el tiempo de ejecución. A su vez, la configuración de diferentes opciones suele ser un proceso tedioso en las herramientas de Minería de Datos ya que es preciso colocar un nodo por cada configuración que se desee hacer. Esta apreciación es válida para aquellos programas de Minería de Datos en donde no se utiliza código de programación sino una interfaz de usuario con flujos de trabajo, tales como SAS

8 Para aproximar el algoritmo a un CHAID, deben seguirse los pasos que se encuentran detallados en el manual de la herramienta.

Enterprise Miner[9] o IBM SPSS Modeler[10]. Ambas herramientas posibilitan al usuario generar código para poder programar la ejecución de varias combinaciones, sin embargo no todos los usuarios de estas herramientas tienen la capacidad para aprovechar esta funcionalidad.

En SAS Enterprise Miner, para cada nuevo árbol que se desee realizar, deberá colocarse un nuevo nodo y modificar los parámetros (Figura III.5). La herramienta nombra cada nodo automáticamente, para luego identificar cada árbol en la comparación de modelos. Este nombre no hace referencia a la configuración de parámetros seleccionada, por lo que para poder comparar cada combinación es preciso recordar los cambios que se hicieron en cada instancia.

Figura III.5 Generación de varios árboles de decisión con diferentes combinaciones de parámetros

El analista puede agregar cuántos nodos necesite según las combinaciones de parámetros que desea comparar. Cada uno de estos nodos generará diferentes modelos y consecuentemente diferentes resultados. Este procedimiento creará un diagrama muy difícil de visualizar, como puede apreciarse en la Figura III.6, en donde cada nueva combinación genera un nodo que es autonombrado. En esta figura alcanzan a verse 13 nodos y ya es muy incómodo para su uso.

III.4.1.3 Evaluación y comparación de resultados

Una vez ejecutados los algoritmos, se evalúan diversas métricas de resultados y se comparan las diferentes configuraciones seleccionadas. La evaluación puede ser de a una por vez, ejecutando un modelo por cada corrida, o una evaluación conjunta, ejecutando los algoritmos de una vez y

9 *Puede verse un ejemplo de configuración en el Capítulo 5 de la Guía online de la herramienta "Getting Started With SAS Enterprise Miner"*
http://support.sas.com/documentation/cdl/en/emgsj/66375/PDF/default/emgsj.pdf
10 Pueden verse las opciones de configuración del nodo en la ayuda en línea de la herramienta http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp

luego comparando los resultados. Cada herramienta de Minería de Datos posee diferentes funcionalidades para poder realizar esta comparación de modelos.

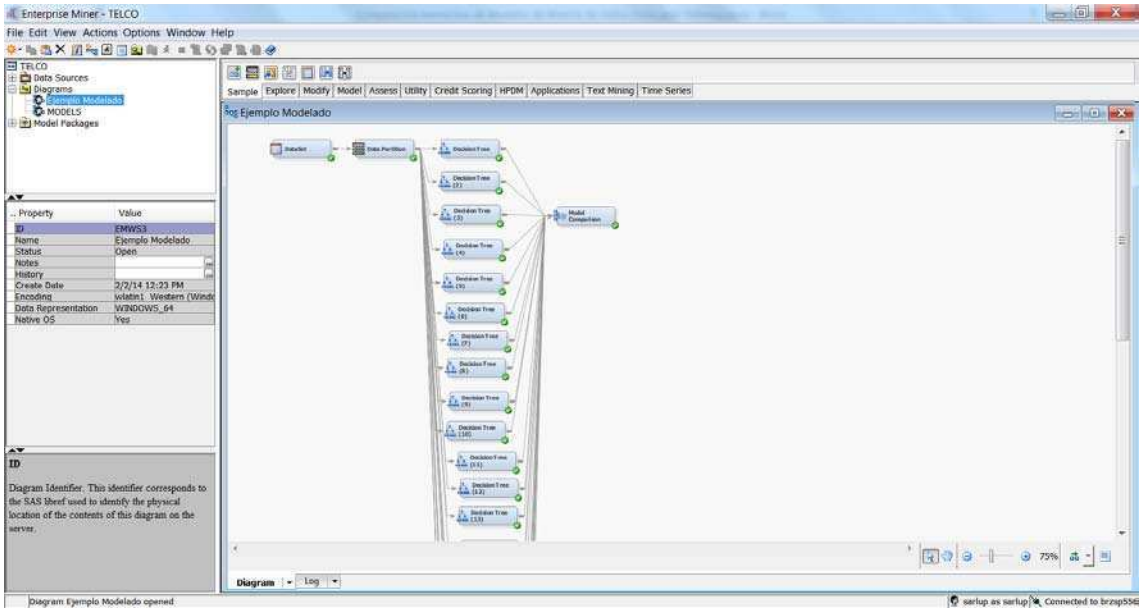

Figura III.6 Múltiples nodos de Árbol de Decisión

La evaluación de la efectividad de cada algoritmo puede realizarse comparando escalares como exactitud (accuracy) o AUC o analizando los gráficos de salida disponibles según la herramienta utilizada. Algunos programas de Minería de Datos no producen resultados visuales, por lo que el analista debe construir sus propios gráficos en función de la información tabular. Por lo general la interacción entre diferentes métricas es muy limitada, por lo que el analista tiene que evaluar uno por uno los resultados generados. En caso de no estar conforme con el resultado propuesto, vuelve al paso 1 para modificar los parámetros y seguir intentando obtener mejores resultados. Además de configurar los parámetros, el analista también puede volver a la etapa de selección de variables para descartar las que oscurecen el análisis o agregar aquellas que le parezca que puedan agregar valor.

Este proceso puede prolongarse cuanto el analista desee. Por lo general en cuanto se observa que ninguna variación produce mejores resultados, se finaliza el proceso de modelado y se publican los resultados obtenidos. Esta decisión también es arbitraria, porque no siempre se está seguro que "ninguna variación produce mejores resultados".

En SAS Enterprise Miner existen 2 instancias de evaluación de resultados. Cada nodo de árbol de decisión genera una pantalla de resultados en donde pueden evaluarse los estadísticos de resumen, curvas lift, el árbol en sí y otros resultados útiles para la evaluación individual de una configuración (Figura III.77).

En esta pantalla no se genera la curva ROC. Para poder generar la misma es necesario colocar el nodo de comparación de modelos (Figura III.88).

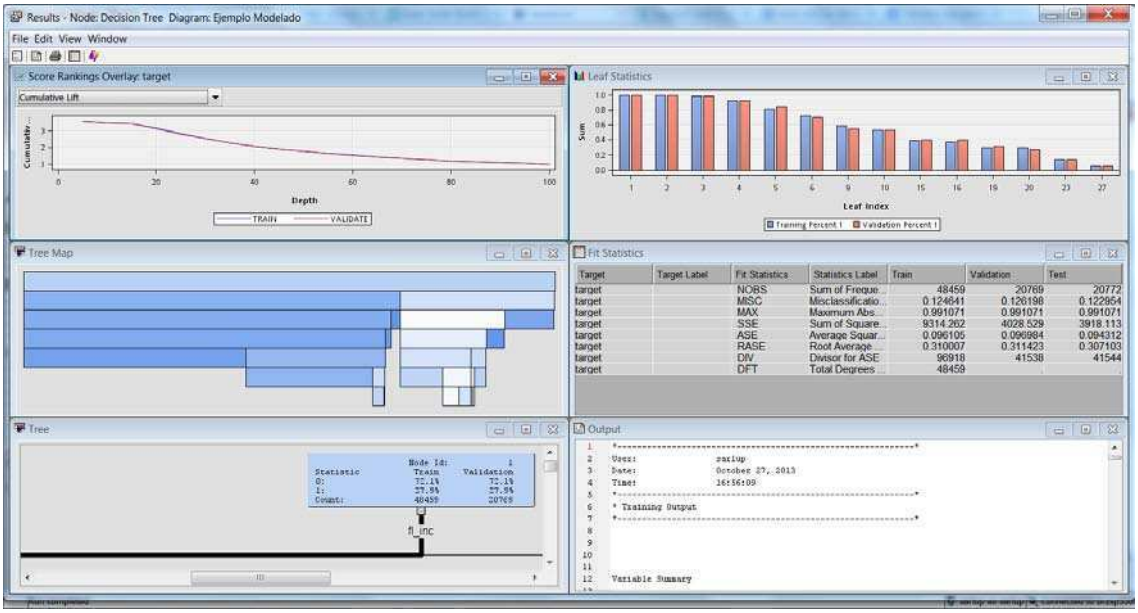

Figura III.7 Pantalla de Resultados del Nodo Árbol de Decisión de SAs Enterprise Miner

Este nodo permite comparar los resultados de 2 o más árboles de decisión, aunque también sirve para poder generar la curva ROC de un solo árbol. A diferencia de la pantalla de resultados de la Figura III.7, este nodo no muestra el árbol de decisión en sí (Figura III.99). El resto de las métricas si son repetidas (curvas lift, curvas de ganancia, métricas de accuracy, etc.).

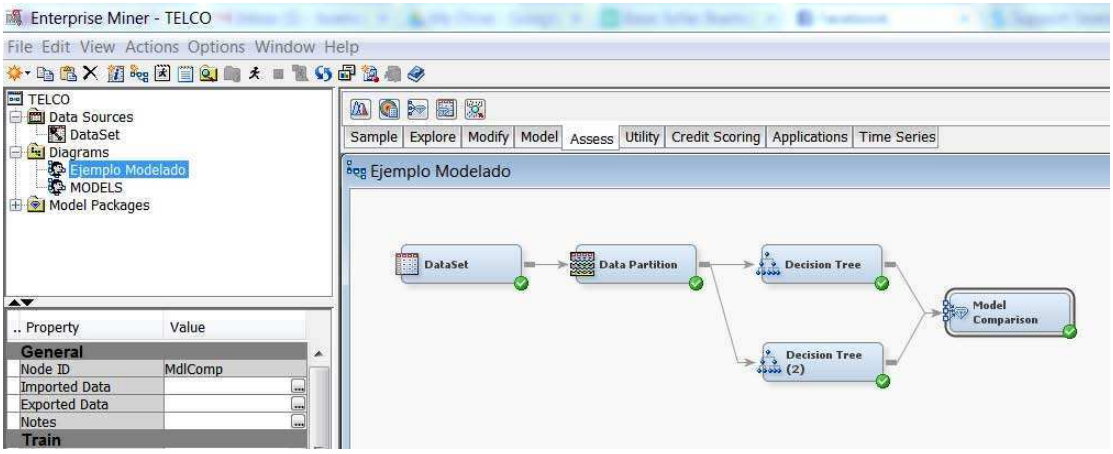

Figura III.8 Nodo Comparación de Modelos SAS Enterprise Miner

Como puede observarse, no hay forma de relacionar estos resultados con las diferentes configuraciones de parámetros seleccionadas. Para comprender la relación entre las diferentes configuraciones de parámetros y los resultados es preciso que el analista recuerde cómo nombró cada árbol.

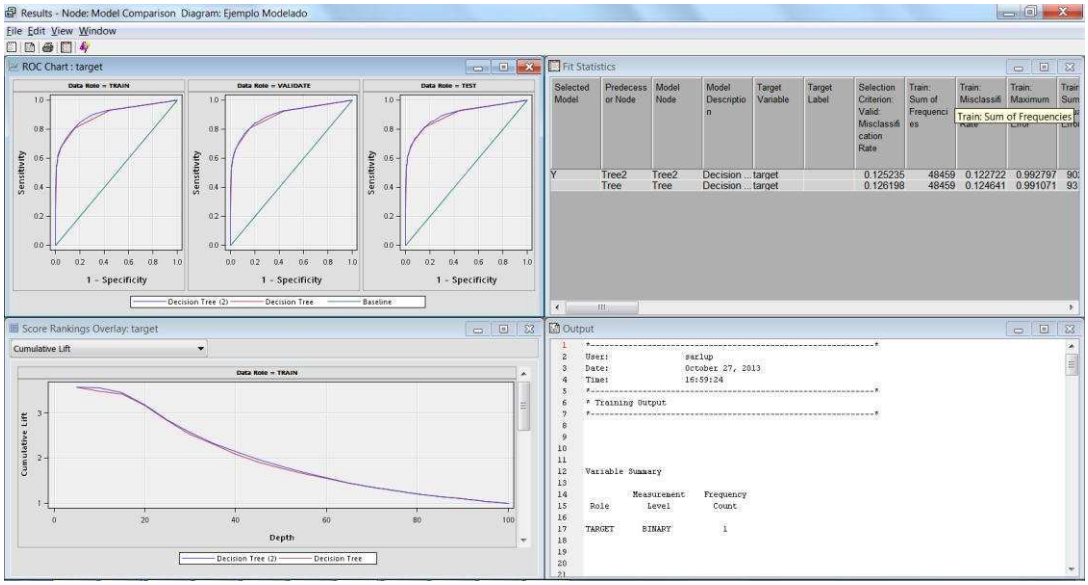

Figura III.9 *Pantalla Resultados nodo "Compración de Modelos" SAS Enterprise Miner*

En la Figura III.99 se observa la comparación de los modelos "Decision Tree" y "Decision Tree (2)". Ambos son muy similares en términos de curva ROC y estadísticos de precisión. La herramienta selecciona como mejor árbol "Decision Tree (2)". Sin embargo si el analista desea evaluar las diferencias entre las reglas producidas por cada árbol, debe retornar a la pantalla de resultados de cada nodo de árbol de decisión (Figura III.77), perdiendo de vista el comparativo general. A su vez si desea recordar qué parámetros configuró para cada nodo, debe pararse sobre el nodo en el diagrama de flujo que generó (Figura III.4) y evaluar los parámetros en el menú correspondiente.

La mayoría de las herramientas del mercado funciona en forma similar. No existe forma de evaluar la incidencia de los parámetros en los resultados en forma intuitiva, sino que el analista debe recordar para cada nombre de modelo, los parámetros que configuró. Esta limitación es la que deseamos resolver con nuestro enfoque.

III.4.2 Proceso de Modelado de Árboles de Decisión Interactivo

Algunas herramientas de Minería de Datos posibilitan la construcción de árboles de decisión en modo interactivo. Esta metodología propuesta en principio por Ankerst (Ankerst, Ester, & Kriegel, 2000) le permite al usuario tomar las decisiones de división y parada en forma interactiva, en contraposición a la forma automática por defecto. Los analistas que seleccionan esta metodología por lo general desean tener más incidencia en las variables que el árbol selecciona para la división ya que les permiten controlar los resultados que serán utilizados posteriormente.

Las herramientas que poseen esta capacidad funcionan de forma muy similar. Partiendo del nodo raíz del árbol, la herramienta propone las variables disponibles para abrir el árbol en ramas. Las mismas se encuentran ordenadas en función de su poder predictivo. El analista puede escoger la que le parezca mejor para distribuir los casos en las distintas ramas. Así por ejemplo si hay dos variables que empatan, puede elegir aquella que tenga mayor significado para la regla de

negocio que desea implementar. Una vez generados los nuevos nodos, se debe seleccionar cada uno de éstos para ir dividiendo en ramas sucesivas. El analista puede parar en donde mejor le parezca, ya sea porque llega al nivel de profundidad deseado o porque la importancia de las variables empieza a ser poco significativa. En algunas herramientas también existe la posibilidad de hacer un camino inverso, podando las ramas del árbol en forma interactiva. En esta modalidad, la herramienta construye el árbol máximo y luego el analista decide qué hojas y nodos podar.

En SAS Enterprise Miner, la funcionalidad de árboles interactivos es una opción dentro de las disponibles en la configuración del árbol de decisión (Figura III.1010).

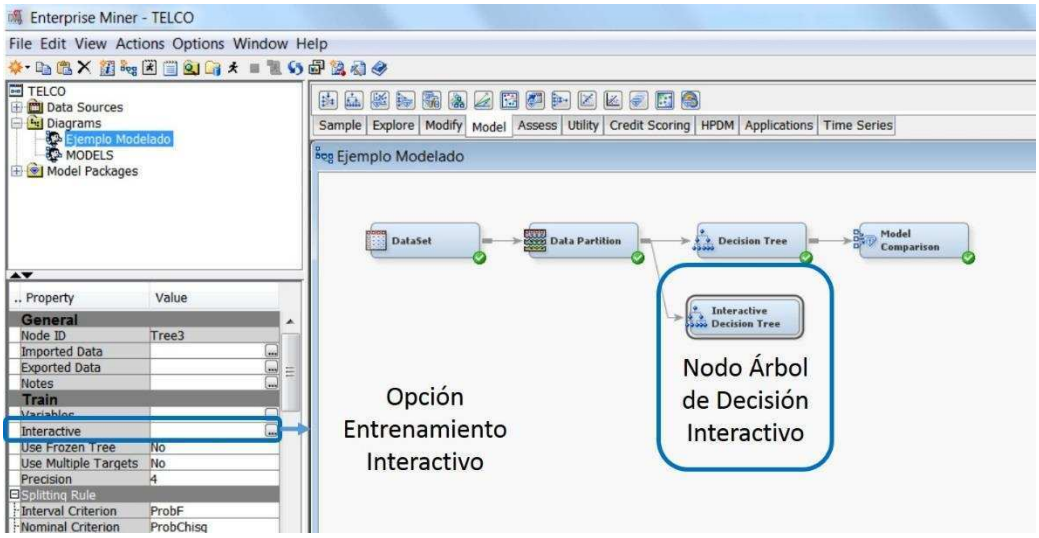

Figura III.10 Funcionalidad Construcción Árboles de Decisión Interactivamente

Una vez que se selecciona la opción correspondiente, puede comenzarse a construir el árbol de forma interactiva. Partiendo del nodo raíz, se selecciona la función "Split" y la herramienta sugiere las variables que pueden utilizarse para abrir el nodo en ramas (Figura III.111).

Como puede observarse en la Figura Figura III.111, el valor que dispone la herramienta es en realidad -Log(p). Este valor también conocido como Logworth es el opuesto del logartimo en base 10 del p-value. Mientras mayor sea este valor, mejor es la variable para abrir el árbol en ramas. Si para cada nodo se selecciona la variable con el mayor valor, entonces el árbol final obtenido será igual al generado automáticamente por la herramienta. Estos pasos pueden repetirse hasta llegar al árbol deseado por el analista.

A medida que se van seleccionando las particiones, se pueden conocer los resultados de la matriz de clasificación, sin embargo otras métricas de evaluación no pueden ser analizadas hasta tanto no se defina el árbol final. Esta limitación no permite hacer una comparación interactiva de los resultados de las diferentes variables que se van seleccionando en la construcción, sino que se debe esperar hasta finalizar la etapa de construcción para conocerlos.

En el trabajo presentado por van den Elzen y van Wijk (van den Elzen & van Wijk, 2011) se proponen muchas mejoras de visualización a las herramientas actuales, entre ellas la visualización de la matriz de confusión a medida que se va construyendo el árbol en forma

interactiva. Sin embargo los mismos autores plantean la necesidad de evaluar esta nueva propuesta en conjuntos de datos del mundo real, ya que el caso de estudio fue realizado con un conjunto de datos pequeño. A su vez tampoco poseen otro tipo de evaluaciones más consistentes como AUC o Lift, evaluaciones que ya hemos enunciado como necesarias.

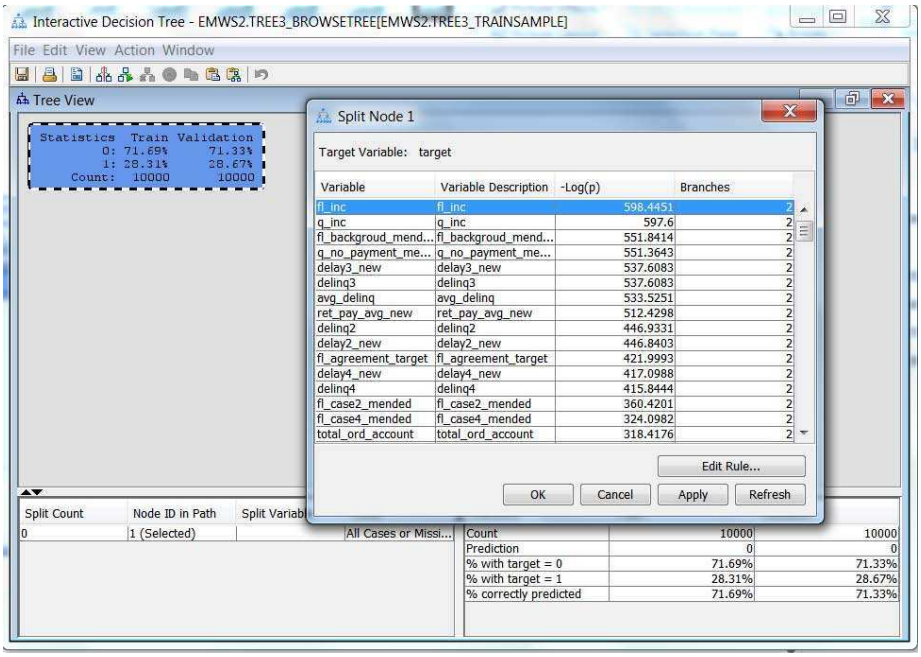

Figura III.11 Selección de la variable a utilizar para dividir el nodo raíz en ramas

III.5 Resumen

En el presente capítulo presentamos el algoritmo CHAID y lo explicamos mediante un simple ejemplo manual de un caso disponible en la bibliografía. Luego explicamos el proceso de modelado llevado a cabo actualmente por la mayoría de los analistas de Minería de Datos que utilizan las herramientas disponibles en el mercado para llevar a cabo sus tareas de análisis. Particularmente seleccionamos la herramienta SAS Enterprise Miner por ser la herramienta al alcance. Incluimos también la descripción de los métodos de análisis de resultados más usuales y la forma de visualizar los mismos, utilizando ejemplos genéricos.

En todos los casos resaltamos la falta de herramientas para evaluar la incidencia de la configuración de los parámetros en los resultados de los árboles de decisión generados. En el próximo capítulo presentaremos un nuevo enfoque mediante el cual pretendemos mejorar la metodología actual para poder analizar los resultados de las diferentes combinaciones de parámetros en forma visual.

IV. Análisis Interactivo de Árboles de Decisión

IV.1 Introducción

En los capítulos anteriores hemos realizado un recorrido por el estado actual del proceso de Minería de Datos y su intersección con Visualización de la Información. Detallamos el procedimiento actual llevado a cabo por analistas de Minería de Datos para construir árboles de decisión enumerando algunas limitaciones que podrían ser cubiertas mediante un enfoque de Visual Analytics.

Dentro de las limitaciones que se mencionaron en la metodología actual, es de especial interés la falta de integración entre el análisis de resultados y la evaluación de la incidencia de la configuración de los parámetros de árboles de decisión en los mismos. Desarrollar técnicas que permitan entender cómo un modelo cambia cuando uno o más parámetros son modificados fue declarado por Bertini (Bertini & Lalanne, 2009) como un área de especial interés en Visual Analytics. Según el autor, en las herramientas actuales es casi imposible lograr este nivel de interacción. No sólo porque la mayoría de los parámetros son difíciles de interpretar, sino también porque el usuario es forzado a realizar una serie de pasos de prueba y error, cambiando los parámetros, esperando la construcción de un nuevo modelo, evaluando los resultados e iterando hasta que el resultado sea lo suficientemente correcto. Este mismo punto fue el que comprobamos en el capítulo anterior al detallar los pasos en la creación de modelos de árboles de decisión en una de las herramientas de minería de datos más utilizadas del mercado.

En función de esta oportunidad de mejora detectada, presentaremos una metodología que consiste en el uso de técnicas de Visualización para la interpretación y comparación del efecto de la configuración de los parámetros en los resultados de árboles de decisión. La evaluación de resultados no sólo comprenderá el análisis de la precisión de los modelos (por ejemplo curvas ROC), sino también el estudio de las diferentes reglas generadas por las diferentes configuraciones de los algoritmos. Se pretende con este enfoque facilitar la tarea de análisis para los usuarios de aplicaciones comerciales. A diferencia de otros enfoques, proponemos la construcción de múltiples modelos en forma automática y la posterior interpretación a través de la visualización interactiva de los resultados.

En este capítulo desarrollaremos primero nuestro enfoque en términos conceptuales, introduciendo posibles desarrollos para la mejora de las herramientas actuales. Al finalizar demostraremos la aplicación en un problema de negocios.

IV.2 Alcance de la Metodología Propuesta

Tal como mencionamos en el capítulo anterior, explicaremos nuestro enfoque utilizando el algoritmo CHAID. Sin embargo, creemos que nuestro trabajo podrá expandirse fácilmente a otras técnicas de minería de datos que involucren la configuración de diversos parámetros para el entrenamiento de los algoritmos, como por ejemplos redes neuronales.

Nuestro trabajo consiste en plantear una metodología de análisis interactivo de árboles de decisión y la incidencia de los parámetros de configuración de los mismos. Para ello sugerimos el uso de diferentes herramientas disponibles. No desarrollamos una herramienta nueva, sino que

proponemos llevar nuestra metodología a las herramientas actuales del mercado. Para poder plasmar la idea en imágenes, presentamos posibles pantallas sugeridas que podrían verse implementadas en las diferentes soluciones.

IV.3 Nuestra Metodología

Teniendo en cuenta las limitaciones del proceso tradicional y del proceso interactivo, desarrolladas en el capítulo III, presentamos una alternativa para la construcción de árboles de decisión y su evaluación. Nuestra propuesta combina la generación automática de árboles de decisión junto con la evaluación interactiva de diferentes resultados, centrándonos en la incidencia de los parámetros en la efectividad de los modelos. Presentaremos primero el desarrollo teórico o conceptual y luego aplicaremos el mismo al modelado de un caso de negocios de la vida real.

Nuestra metodología consiste en los pasos enumerados en la Figura IV.1. Explicaremos cada uno de ellos en las secciones siguientes haciendo referencia a las diferencias con el procedimiento actual explicado en el capítulo anterior.

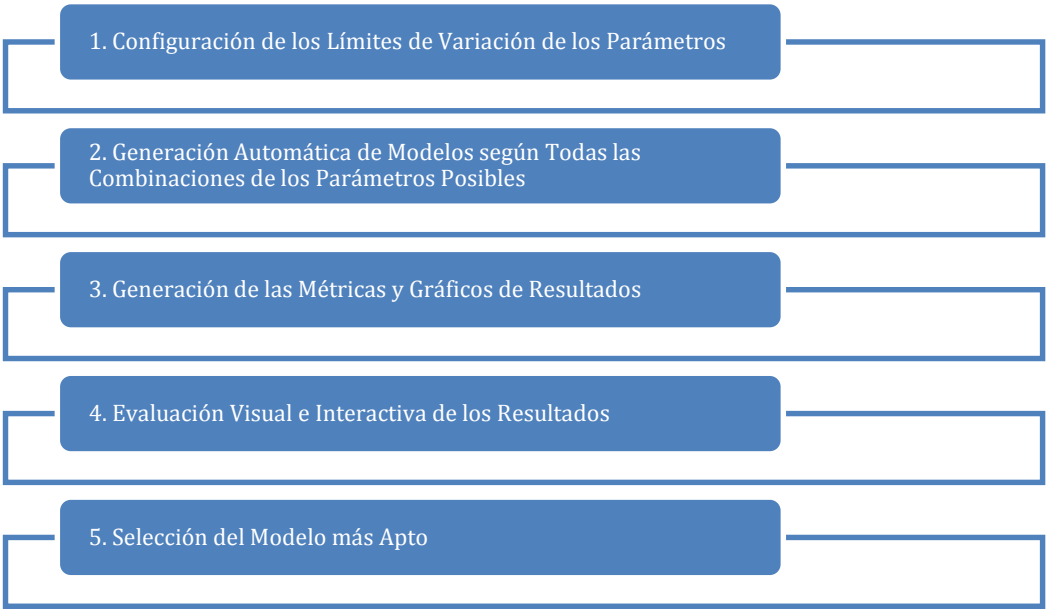

Figura IV.1 Proceso de Modelado y Evaluación Interactiva de Modelos propuesto

Los pasos propuestos son posteriores a la definición del conjunto de datos de entrenamiento y la definición de las variables a utilizar. Dentro del proceso de KDD (Figura II.10), nuestra metodología se aplicaría a partir del paso de minería de datos, involucrando también la evaluación e interpretación de los resultados.

IV.3.1 Configuración de los Límites de Variación de los Parámetros

Actualmente para realizar pruebas de diferentes parámetros de CHAID, el analista debe seleccionar los valores de cada uno de ellos y ejecutar cada combinación de una por vez (sección III.4.1.1).

En nuestra metodología, proponemos que el analista configure para cada parámetro del algoritmo que desee variar el límite superior e inferior, así como el incremental que desea para cada uno de estos parámetros. Otra opción sería la configuración de lista de valores a utilizar para el caso en donde los intervalos de variación no sean regulares.

Por ejemplo si desea evaluar cómo afecta el máximo de profundidad del árbol al resultado predictivo del mismo, podría fijar para el parámetro máxima profundidad un mínimo 5 niveles y un máximo 10 y que este a su vez se incremente de a una unidad. Es decir que la herramienta debería generar automáticamente árboles de 5, 6, 7, 8, 9 y 10 niveles de profundidad. Para el parámetro mínima cantidad de casos por hoja podría resultar poco significativo incrementar el mismo en una unidad, por lo que podría configurarse para este parámetro una variación, por ejemplo, de 10 en 10 o un valor porcentual del total de casos de entrenamiento. Para el parámetro mínima cantidad de casos por hoja, tal vez sea mejor colocar una lista de valores posibles en lugar de intervalos ya que es un parámetro que no genera grandes variaciones en los resultados según diversas pruebas realizadas. Otro tipo de configuración para este parámetro, podría una fórmula. Sin embargo, sería preciso analizar qué tipo de fórmula podría colocarse ya que no en todos los casos sería aplicable por el tamaño del conjunto de datos o por la precisión que se espera en el modelo.

Este tipo de pruebas de combinación de parámetros puede ser realizado actualmente en la mayoría de las herramientas disponibles a través del uso de código de programación. Nuestro enfoque pretende facilitar la tarea de analistas no familiarizados con lenguajes de programación, por lo que proponemos que las herramientas de Minería de Datos dispongan de una pantalla amigable para el usuario en donde el mismo pueda configurar las combinaciones deseadas.

Tomando como ejemplo el nodo de árboles de decisión de SAS Enterprise Miner, podría incluirse en el mismo la opción de configuración de los límites en el menú de configuración de parámetros. En la Figura IV.2 comparamos la pantalla de configuración actual con una sugerida. En esta última, incluimos la posibilidad de expandir las opciones de configuración en los parámetros de Nivel de Significación (Significance Level), Máxima Cantidad de Ramas (Maximum Branch), Máxima Profundidad (Maximum Depth) y Tamaño de la Hoja (Leaf Size). Utilizamos el ícono de puntos suspensivos al ser esta la forma en que SAS Enterprise Miner indica que existen más opciones de configuración.

En las opciones de configuración actual, el analista debe ingresar un único valor para cada opción de configuración. En nuestro enfoque sugerimos la posibilidad de ingresar intervalos de valores o listas, en contraposición al enfoque actual. Al hacer click en los puntos suspensivos en cada parámetro a configurar, el analista podría establecer los límites de los parámetros para generar las combinaciones deseadas de los mismos o la lista de valores a utilizar. En la Figura IV.3 puede observarse un ejemplo de interfaz de usuario sugerida creada para la configuración del parámetro Máxima Profundidad.

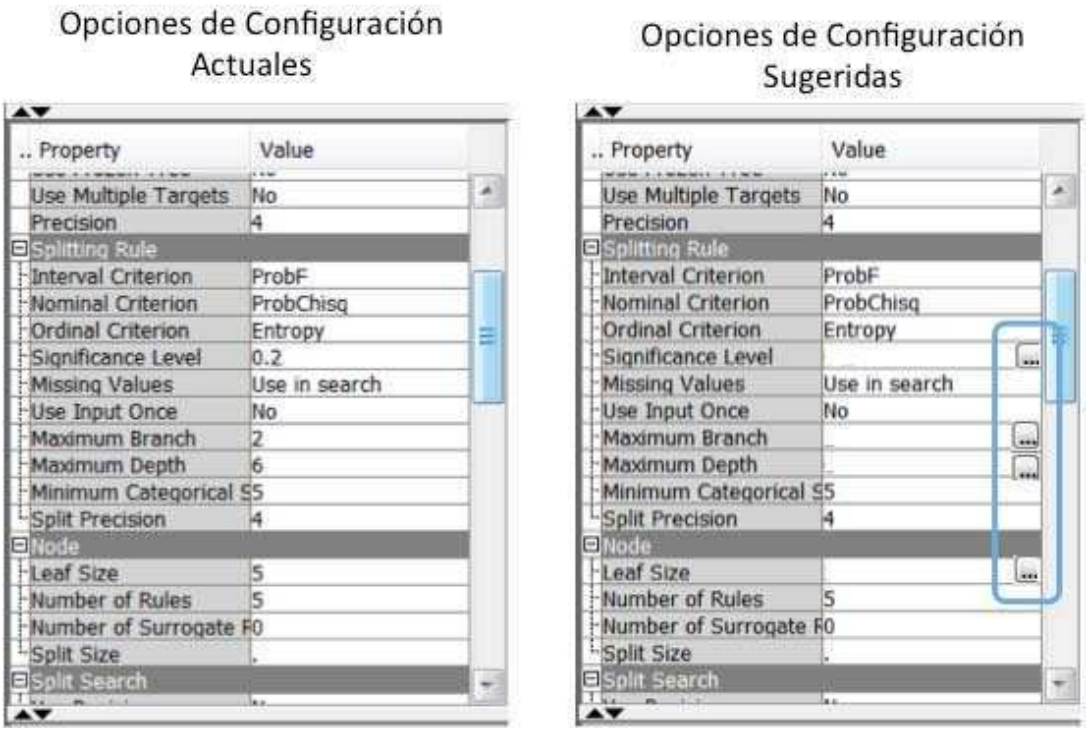

Figura IV.2 Funcionalidad Sugerida para la configuración de las combinaciones de parámetros de árbol de decisión en SAS Enterprise Miner

Si el analista deseara configurar diferentes valores mediante una lista, debería ingresar los valores separados por comas (por ejemplo Depth 1, 3, 6). De esta forma sólo se formarían combinaciones con estos valores. En el caso de rangos, se debería ingresar los valores separados por un guion (por ejemplo Depth 5-8) y la unidad incremental para este intervalo. Para configuraciones únicas, bastaría con colocar un solo valor, como se coloca actualmente.

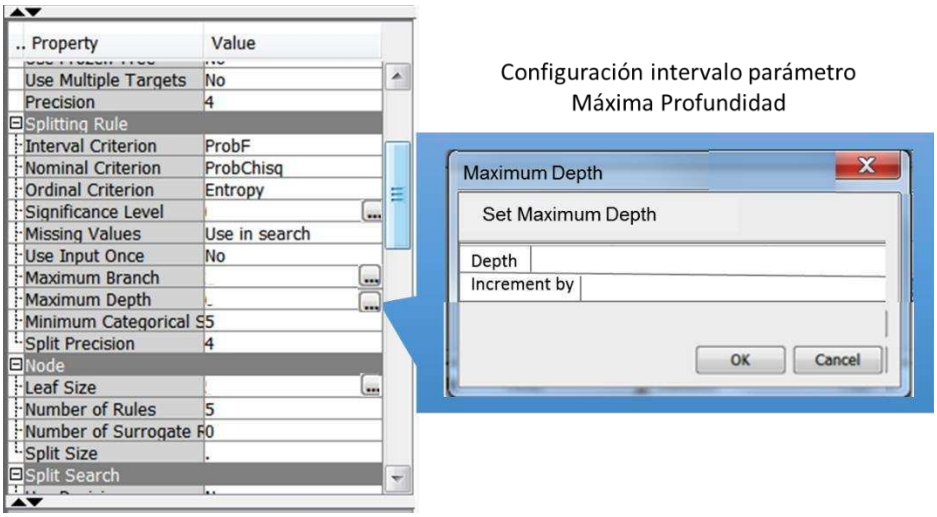

Figura IV.3 Configuración sugerida para definir las combinaciones de parámetros en SAS Enterprise Miner (ejemplo para parámetro Profundidad Máxima – Maximum Depth)

Con una funcionalidad de este tipo, el analista evitaría tener que seguir los pasos de prueba y error que usualmente se practican ya que la herramienta generará todos los árboles posibles y luego el esfuerzo del analista estará enfocado en el análisis visual e interactivo de los resultados.

A su vez al disponer de una interfaz de usuario fácil de utilizar, no debería conocer código de programación para realizar las iteraciones, siendo esto una característica de alto valor para usuarios de perfiles menos técnicos.

IV.3.2 Generación Automática de Modelos según Todas las Combinaciones de los Parámetros Posibles

Una vez configurados los parámetros y sus valores posibles (listas o rangos), la herramienta generará todos los árboles posibles y sus consecuentes resultados. Esto funcionará mediante una programación de tipo "do while" en donde las tareas se ejecutan mientras que se cumplan las condiciones fijadas. En este caso se generarán tantos árboles como combinaciones de parámetros posibles haya, finalizando la ejecución cuando no haya más combinaciones de parámetros a procesar.

A diferencia de la metodología de construcción de árboles de decisión en forma interactiva (sección III.4.2), nuestro enfoque implica que las tareas de construcción de los diferentes árboles se realicen en forma automática. Es decir que la herramienta procese el algoritmo sin involucrar al analista en el mismo. Todos los pasos del algoritmo CHAID serán ejecutados en forma automática y secuencial para todas las combinaciones de parámetros. La interacción del analista se reserva para la etapa de evaluación de los resultados y análisis de incidencia de la configuración de los parámetros en los mismos.

IV.3.3 Generación de las Métricas y Gráficos de Resultados

Una vez que se han procesado todas las combinaciones posibles, la herramienta mostrará todos los resultados en forma gráfica. Cuando hablamos de resultados nos referimos tanto a los resultados del modelo, como ser matrices de confusión, Curvas ROC, Curvas Lift, Esquema del árbol y reglas, como así también a los gráficos de parámetros.

Estos gráficos deberán estar relacionados, de forma tal que sea posible interactuar con todos ellos y observar la incidencia de la configuración del algoritmo no sólo en los resultados de precisión, sino también en las reglas del árbol que se generen.

Para lograr esta interacción, se sugiere el uso de técnicas de visualización de múltiples vistas coordinadas, MVC (también conocidas como CMV por sus siglas en inglés, Coordinated Multiple Views).

IV.3.3.1 Sistemas de Múltiples Vistas Coordinadas

Un sistema de vistas múltiples utiliza dos o más vistas diferentes para soportar la investigación de una entidad conceptual única (Wang Baldonado, Kuchinsky, & Woodruff, 2000). Estos autores proponen una serie de guías o reglas para definir cuándo un sistema de este

tipo puede resultar de utilidad. A continuación detallamos las mismas y la aplicación al problema bajo análisis.

- **Regla de la Diversidad:** se deben utilizar vistas múltiples cuando existe una diversidad de atributos, modelos, perfiles de usuarios, niveles de abstracción o géneros. La presencia de diversidad es uno de los motivos principales para decidir la construcción de un sistema de MVC.

  En nuestro caso, la presencia de diversidad es evidente, ya que se precisan analizar diferentes entidades, tales como curvas ROC, Lift y parámetros de configuración de árboles de decisión, entre otras.

- **Regla de Complementariedad:** se deben utilizar múltiples vistas cuando las mismas pueden demostrar correlaciones o disparidades en los datos. Cuando se poseen vistas únicas, el usuario debe recordar lo analizado en una imagen para luego analizar la siguiente y construir la relación mentalmente. Dado que recordar es más demandante mentalmente que comparar, se sugiere el uso de MVC para facilitar la tarea.

  En el caso de análisis de múltiples árboles de decisión, tener vistas complementarias puede producir información nueva, que no se conocía de antemano. Por ejemplo, poder relacionar los valores de los parámetros con los resultados de los modelos visualmente, libera al usuario de tener que recordar cómo configuró cada árbol de decisión para luego hacer el comparativo con los nuevos árboles generados por otras configuraciones.

- **Regla de Descomposición:** fraccionar información compleja en múltiples vistas para crear porciones manejables y proveer información mediante la interacción en diferentes dimensiones. Esta regla se basa en el principio de "divide y reinarás", al permitir al usuario analizar pequeños fragmentos para luego construir la visión general.

  En nuestro problema de múltiples árboles y múltiples parámetros, el usuario también podría beneficiarse con esta regla, ya que puede descubrir nueva información al analizar pequeñas secciones de cada uno de los gráficos de resultados, cuando así se precise.

- **Regla de Economización:** usar la mínima cantidad de vistas múltiples. Si bien las vistas múltiples pueden contribuir al análisis de información, también pueden ser muy confusas y abrumadoras. Se debe usar una cantidad limitada de ellas.

  Esta regla es muy importante para tener en cuenta a la hora de analizar toda la información propuesta en nuestro enfoque. El analista podría perder el foco de análisis si se encontrara ante demasiadas posibilidades.

Una vez evaluado si el problema que se desea analizar es apto para un desarrollo de un sistema de MVC, es preciso seguir ciertas reglas de diseño para asegurar la usabilidad y navegabilidad de la herramienta.

- **Regla de Optimización de Recursos y Tiempo:** es preciso realizar un balance entre los costos y beneficios espacio-temporales de presentar vistas múltiples. Debe decidirse si se precisan vistas simultáneas o secuenciales. Las vistas simultáneas pueden ser difíciles de asimilar por el usuario, ya que no podría visualizar todos los gráficos de una vez. Mostrar todos los resultados de una vez podría causar una sobrecarga de información.

  En este sentido puede ser también relevante darle la posibilidad al usuario de seleccionar la cantidad de vistas que precisa para poder analizar el problema bajo estudio. Algunos analistas podrán sentirse más cómodos visualizando algunos pocos gráficos mientras que otros pueden preferir ver todo de una vez.

- **Regla de Obviedad:** utilizar pautas perceptuales para hacer las relaciones entre las múltiples vistas más visibles para el usuario. Distinguir las relaciones en un sistema de MVC puede ser una tarea muy difícil para el usuario. Es por ello que es preciso utilizar diferentes técnicas para mostrarle al usuario ciertas relaciones. Entre estas técnicas se encuentran el resaltado, la organización del espacio y las interacciones entre diferentes componentes de las vistas a través de filtros.

  Esta regla implica en nuestro enfoque proporcionarle información al usuario sobre las mejores combinaciones de parámetros de antemano para que el analista no deba analizar todo el espacio de soluciones, sino que podría comenzar por un reducido conjunto de resultados a explorar.

- **Regla de Consistencia:** hacer las interfaces para múltiples vistas consistentes, y hacer que los estados de múltiples vistas sean consistentes. La consistencia de interfaces es un tema ampliamente desarrollado en la literatura de Interfaz Hombre-Máquina (HCI por sus siglas en inglés). Schneiderman (Shneiderman & Plaisant, 2004) plantea 8 reglas para el diseño de interfaces, siendo la primera de ellas “Esforzarse por la Consistencia”. Esta regla implica que en situaciones similares se deben requerir secuencias de acciones consistentes; terminología idéntica en el uso de prompts, menús y pantallas de ayuda.

  En nuestro ejemplo, lograr la consistencia entre las diferentes vistas implica que el usuario pueda profundizar o hacer comparaciones de los resultados mediante gráficos y menús que sean coherentes en toda la aplicación. Esto garantiza un aprendizaje más simple de la herramienta.

- **Regla de Gestión de la Atención:** se deben usar técnicas perceptuales para enfocar la atención del usuario en la vista correcta en el momento correcto. Para lograr esto técnicas tales como animaciones, sonidos, resaltado y movimiento son muy útiles.

  Para analizar diferentes árboles de decisión, la técnica más valiosa es la de resaltado, ya que otras técnicas tales como sonido y movimiento no aplican en nuestra problemática.

En la mayoría de las herramientas de minería de datos, el uso de sistemas de MVC para el análisis de resultados es limitado. En SAS Enterprise Miner se encuentra implementado en cierta forma un sistema de múltiples vistas tanto para el análisis de resultados del nodo de árboles de decisión como para el resultado de comparación de modelos (Figura III.7, Figura III.9). Sin embargo, la coordinación entre las vistas es bastante limitada. Sólo algunos gráficos pueden ser analizados en conjunto mediante el uso de relaciones de tipo selección y resaltado. A su vez no es posible analizar la curva ROC de un árbol junto con la visualización del árbol en sí, ya que ambos gráficos se encuentran en ventanas diferentes y no guardan ningún tipo de relación.

Esta limitación se observa en la mayoría de los sistemas de Minería de Datos, como así también en otras aplicaciones que requieren el uso de sistemas de tipo MVC (Andrienko & Andrienko, 2007).

Siguiendo con el ejemplo de SAS Enterprise Miner, presentamos en la Figura IV.4 una forma para poder convertir las ventanas de resultados actuales en un sistema de tipo MVC.

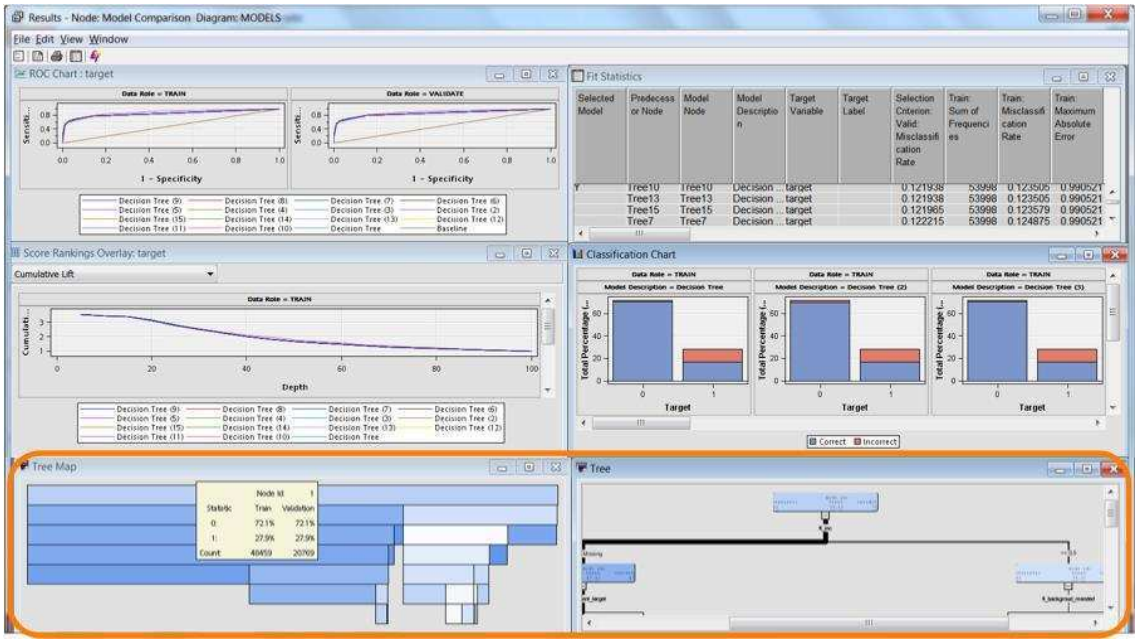

Figura IV.4 Ventana de Resultados Integrada para SAS Enterprise Miner

Los 4 gráficos superiores son producidos por el nodo de Comparación de Resultados de la herramienta, mientras que los dos inferiores son generados sólo en cada nodo de árbol de decisión. En nuestro enfoque, proponemos la integración de estos gráficos. El analista debería poder observar el árbol subyacente a la curva ROC seleccionada como así también analizar la relación entre la matriz de confusión (representada gráficamente por el gráfico de clasificación, Classification Chart). También es necesario contar con una visión general de los diferentes árboles para poder comparar los mismos. Siguiendo con los principios de MVC, el usuario debería poder seleccionar un elemento de un gráfico y ver cómo el resto cambia en función de la selección.

En cuanto a la visualización de los parámetros, puede observarse que ninguna de las salidas de la herramienta permite evaluar los mismos. El analista debe recordar qué nombre le puso a cada configuración y así comparar los resultados. Desde ya esta modalidad es poco práctica y factible con determinado volumen de árboles a analizar. Por este motivo es preciso desarrollar otro tipo de vistas para los parámetros.

Para ejemplificar mejor nuestro enfoque seleccionamos la herramienta de MVC ComVis (Matkovic, Freiler, Gracanin, & Hauser, 2008). Este sistema permite configurar varias vistas y ser adaptado en función de cada necesidad de análisis. El sistema permite subir un archivo de tipo csv (comma separated values) y utilizar diferentes gráficos para representar la información. A su vez, mediante el uso de los pinceles simples y múltiples, pueden seleccionarse diferentes porciones de la información, logrando el tipo de interacción que describimos anteriormente como necesaria. La Figura IV.5 muestra un ejemplo de ComVis aplicado para la evaluación de los resultados de modelado de árboles de decisión.

En esta vista, el usuario seleccionó 4 paneles para la interpretación de los resultados: curvas ROC, curvas Lift, vista de parámetros y vista de árboles. Si bien en este caso existen 4 gráficos preseleccionados, el usuario podría configurar más vistas utilizando los menús de la herramienta. Esta capacidad respeta las reglas de diseño de sistemas de MVC.

La vista de parámetros fue especialmente desarrollada en ComVis (Padua, Schulze, Matković, & Delrieux, 2014) para poder analizar la incidencia de las combinaciones de los parámetros en las curvas ROC y Lift, como así también en la estructura general de los árboles de decisión. En esta vista se evalúan todas las combinaciones en forma de histogramas anidados. Cada parámetro tiene su distribución y se muestra la relación con el resto. En la Figura IV.5 no hay ninguna selección realizada, por lo que podemos observar todo el espacio de configuraciones.

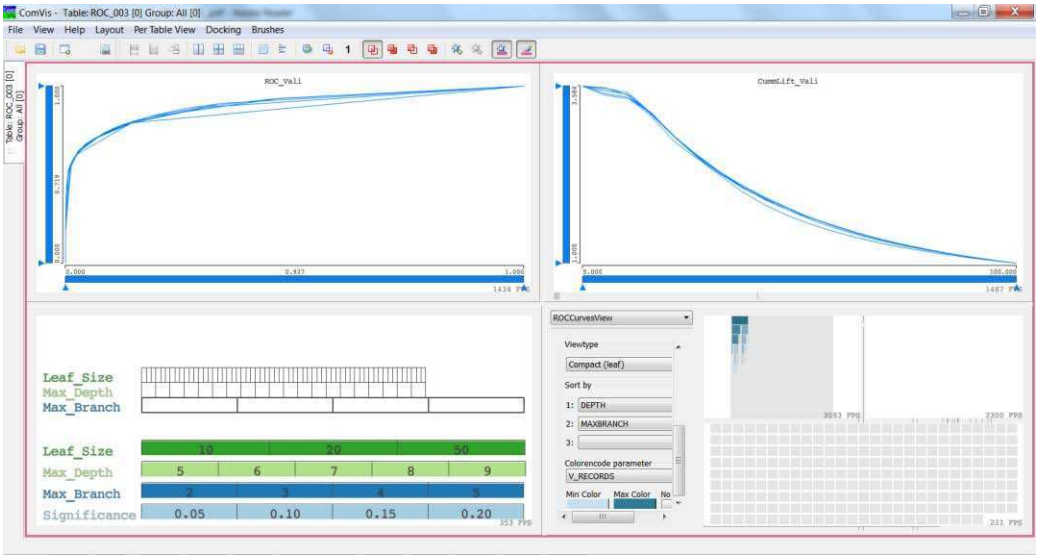

Figura IV.5 ComVis para el análisis de resultados de múltiples configuraciones de árboles de decisión

IV.3.4 Evaluación Visual e Interactiva de los Resultados

Una vez que la herramienta generó todos los árboles y los gráficos necesarios para analizar los múltiples árboles de decisión, el analista podrá dedicarse a la tarea de análisis interactivo de los resultados. Las MVC servirán al analista para poder evaluar la incidencia de los parámetros, observar diferentes resultados y seleccionar aquél que más útil le sea, no sólo en términos de efectividad, sino también como herramienta de negocio. Esta información le servirá para limitar en un futuro la configuración de los parámetros y así poder disminuir la cantidad de combinaciones a analizar.

En un sistema como ComVis estas funciones son muy simples de realizar, ya que existe la posibilidad de relacionar varios gráficos y así sacar conclusiones a través de la interacción.

La técnica más usual de interacción es la de resaltado. Esta técnica consiste en seleccionar datos interactivamente en una representación visual. La intención es resaltar estos datos en diferentes visualizaciones del conjunto de datos bajo análisis.

ComVis soporta dos tipos de pinceles, simple y compuesto. El compuesto permite al usuario utilizar operaciones de tipo booleanas (OR, AND, SUB) de forma interactiva. Esto permite al analista ampliar (operación de tipo OR) o acotar (AND, SUB) su selección para descubrir características interesantes en los datos.

En la Figura IV.5 observábamos la vista de la configuración inicial del sistema. En la figura 4.6 el usuario ha realizado una selección sobre las curvas ROC más interesantes utilizando un pincel de tipo simple. Esta selección se refleja en el resto de los gráficos en donde se resaltan sólo los resultados que corresponden a tales curvas.

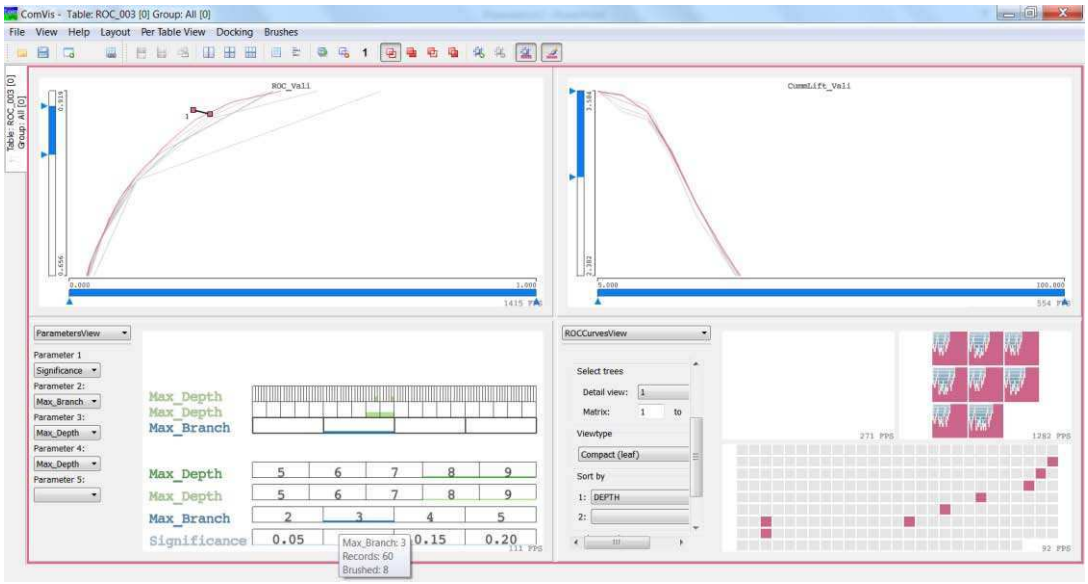

Figura IV.6 Análisis de Resultados mediante el uso de ComVis aplicado a múltiples árboles de decisión.

Este tipo de interacción le permite al usuario descubrir relaciones entre los datos de forma visual y simple, que de otra forma hubiese implicado varias consultas a una base de datos. En el ejemplo de aplicación al caso de negocio explicaremos esta funcionalidad con mayor detalle.

IV.3.5 Selección del Modelo más Apto

Luego de interactuar con todos los resultados visualmente, el analista habrá seleccionado aquel modelo que le resulte más adecuado para su problemática. Esta selección no necesariamente será en términos de poder predictivo del árbol de decisión, sino también por las reglas que éste generó, influenciadas por los parámetros de configuración.

En el gráfico de parámetros se podrá conocer la configuración necesaria para arribar a los resultados seleccionados. El analista podría utilizar estos datos para volver a configurar la herramienta sólo con estos parámetros y así generar el modelo para los nuevos datos empleando menos tiempo.

Lo ideal, sin embargo, sería tener la posibilidad de seleccionar el modelo más adecuado a través de las vistas generadas. Así por ejemplo el analista podría pararse sobre cualquiera de los gráficos generados, hacer click derecho e indicarle a la herramienta que ese resultado es el seleccionado como modelo final. Podemos ver un ejemplo de aplicación de esta modalidad en la Figura IV.7. Esta funcionalidad no está disponible en la herramienta ComVis, sino que es una sugerencia de mejora.

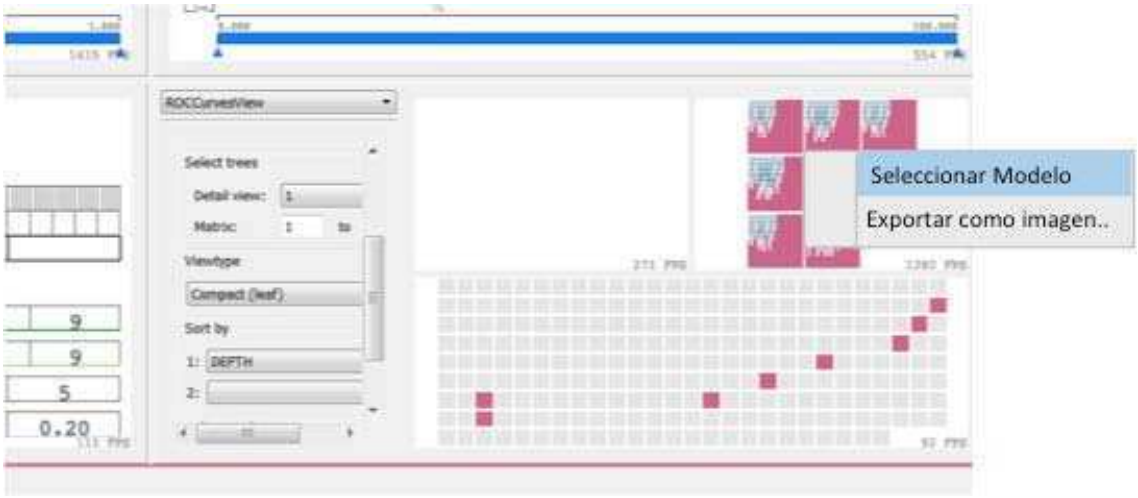

Figura IV.7 Ejemplo para la selección visual del modelo final de minería de datos

Con este último paso, el analista finalizaría su tarea de análisis de múltiples árboles de decisión. A través del proceso, el usuario pudo descubrir diferentes características que no habrían sido posibles sin herramientas interactivas. Esta metodología podría ser aplicada a diversos problemas de Minería de Datos.

IV.4 Aplicación a un Caso Real

Con el fin de demostrar nuestra metodología en la práctica, solicitamos la colaboración de una empresa de telecomunicaciones de Argentina que realiza tareas de minería de datos para generar modelos predictivos para diferentes problemáticas. En esta oportunidad trabajamos en la

generación de un modelo predictivo para la detección de aquellos clientes que tienen mayor probabilidad de falta de pago de su factura telefónica en los próximos 60 días.

Dado que nuestro enfoque comienza desde la etapa de generación de modelos, sin tener en cuenta la preparación de los datos, obtuvimos un conjunto de datos listo para modelar. El conjunto de datos contiene 90.000 filas y 144 columnas que sirven para poder identificar los clientes que tendrán mayor propensión a no pagar su factura telefónica en los próximos 60 días. De las 144 variables, 132 son variables numéricas, incluyendo la variable target y el identificador de cliente, y 12 categóricas. Entre las variables numéricas podemos encontrar variables de facturación, uso de los servicios, variables de tendencias tales como variación en el uso y consecuente variación en la facturación, y las propias variables de mora. Las variables categóricas son de identificación de segmento de cliente, productos que posee y demográficas.

El conjunto de datos es una muestra del total de líneas de la compañía y a su vez se encuentran balanceadas las clases. El balanceo es una técnica que se utiliza para poder incrementar la proporción de casos positivos versus los negativos con el fin de identificarlos más fácilmente. En el conjunto de datos proporcionado el 25% de los casos no pagarán su factura 60 días luego del vencimiento. El porcentaje real de casos no se informa por temas de confidencialidad.

El proceso de modelado en la empresa es del tipo tradicional. Los analistas utilizan SAS Enterprise Miner para generar árboles de decisión para calcular la propensión de los clientes a no pagar sus facturas. Una vez generado el conjunto de datos correspondiente el mismo es utilizado como base para entrenamiento. Se realiza la partición en conjunto de entrenamiento y validación y se selecciona el nodo de árboles de decisión de la herramienta. Este tipo de algoritmo es preferido a otros por la facilidad para explicar los resultados al área comercial. Las reglas generadas pueden ser fácilmente traducidas a acciones de negocio. Los analistas utilizan la pantalla de configuración de SAS Enterprise Miner para generar diversos árboles de decisión. Por lo general los analistas comienzan la tarea de análisis con la configuración por defecto de la herramienta y luego van ajustando los parámetros en función de los resultados obtenidos. Dado que la herramienta presenta ciertas limitaciones a la hora de visualizar los resultados, aquellas combinaciones de parámetros que no producen buenos resultados son eliminadas del análisis sin ser incluidas en el nodo de comparación de modelos. Los analistas realizan algunas corridas y luego de algunos intentos se quedan con el modelo que mejores resultados produjo para la problemática bajo análisis. Este proceso es de duración variable, ya que depende del poder de cómputo de los sistemas y de la cantidad de pruebas que se realicen. Una vez seleccionado el modelo más conveniente, se pone en práctica el mismo, generando todos los meses la puntuación correspondiente para cada cliente.

A continuación demostraremos la aplicación de la metodología propuesta para el caso de uso planteado. Detallaremos para cada paso de nuestro enfoque las herramientas utilizadas y los procedimientos llevados a cabo. Si bien en este caso hemos utilizado ciertas herramientas, creemos que el enfoque propuesto puede ser logrado con otras que posean capacidades similares. Haremos hincapié en las diferencias versus en enfoque tradicional de prueba y error que suele llevarse a cabo en la mayoría de los casos.

IV.4.1 Configuración de los diferentes parámetros

Para generar los diferentes árboles de decisión utilizamos SAS Enterprise Miner, particularmente el nodo de árboles de decisión. La herramienta permite generar árboles de decisión mediante la programación de código SAS[11]. Sin embargo existen ciertos resultados necesarios para generar métricas de medición y comparación de los modelos que no son generados por este procedimiento. Por este motivo se decidió realizar un procedimiento manual para poder generar todos los resultados necesarios.

Si bien el nodo de árboles de decisión de la herramienta cuenta con 24 parámetros configurables, para nuestro ejercicio seleccionamos 4 de ellos. Estos parámetros son los más utilizados por analistas de minería de datos. En la Tabla III-1 se detallan los parámetros utilizados y las variaciones de los mismos.

Parámetros	Valores					Valores Posibles
Nivel de Significación	0.05	0.1	0.15	0.2		4
Máxima Cantidad de Ramas	2	3	4	5		4
Profundidad Máxima	5	6	7	8	9	5
Tamaño mínimo de la hoja	10	20	50			3
Combinaciones Totales						**240**

Tabla IV-1 Configuración de parámetros del nodo de árboles de decisión de SAS Enterprise Miner

La configuración por defecto de estos parámetros en el nodo de árboles de decisión de SAS Enterprise Miner es: nivel de significación 0.2, máxima cantidad de ramas 2, profundidad máxima 6, tamaño de la hoja 5. Para nuestro ejemplo fijamos el nivel de significación de 0.05 a 0.2, con un incremento de 0.05. Como indicamos en el capítulo 3, este parámetro permite controlar la precisión del cálculo de dependencia de variables. Mientras menor es el nivel de significación mayor es la dependencia entre las variables, es decir son más explicativas. Para la apertura en ramas, seleccionamos un intervalo de 2 a 5, con incrementos de una unidad. Esta apertura permitirá crear mayores intervalos para las variables numéricas, así como también asignar menor cantidad de categorías a cada rama en caso de ser necesario. Como profundidad máxima, fijamos un intervalo de 5 a 9, con incrementos de una unidad. Si bien por lo general no se recomienda generar árboles de gran profundidad por el riesgo de sobreajuste, se desea comparar los resultados de las diferentes combinaciones. Un árbol más profundo puede también ser más preciso sin necesidad de estar sobreajustado. Cada resultado podrá ser evaluado en función de las necesidades del negocio. Por último, fijamos la cantidad mínima de hojas en 10, 20 y 50. En este caso en un lugar de un rango de valores, se trata de una lista. Seleccionamos estas variaciones para el último parámetro ya que en pruebas iniciales comprobamos que pequeñas variaciones no producían diferencias significativas en los resultados.

11 http://support.sas.com/documentation/onlinedoc/miner/em43/allproc.pdf

Esta tarea que fue realizada manualmente, podría ser mejorada significativamente contando con la posibilidad de configurar los parámetros del árbol de la forma en que planteamos en la sección IV.3.1. De contar con tales características en la herramienta, el usuario sólo debería ingresar los rangos o listas en las pantallas correspondientes y así no debería invertir tiempo en la creación de cada combinación de parámetros, sino que la herramienta lo haría automáticamente. En la Figura IV.8 mostramos las pantallas de configuración para el caso propuesto según las mejoras que se proponen sobre la herramienta SAS Enterprise Miner.

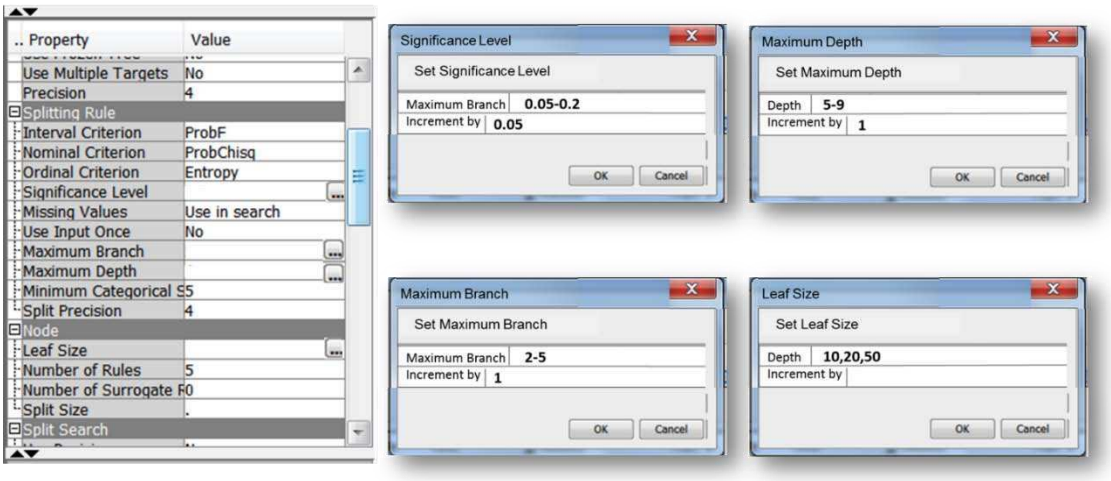

Figura IV.8 Pantallas de Configuración de Parámetros según mejora propuesta.

Como se puede observar en la Figura IV.8, para cada parámetro seleccionado, se configuran los valores que se desean probar. En este caso tres parámetros están configurados en modo rango y uno en modo lista. De esta forma el usuario sólo configura una vez el nodo de árboles de decisión, ahorrando tiempo y a su vez, utilizando una interfaz amigable para el usuario.

IV.4.2 Generación Automática de Modelos según Todas las Combinaciones de los Parámetros Posibles

Luego de configurar los diferentes nodos de árboles de decisión, se ejecutan los mismos. El tiempo de procesamiento dependerá de la capacidad de la máquina con la que se trabaje. En nuestro caso trabajamos con una computadora de procesador Intel Core i5-2540M CPU 2.60GHz con 8GB de memoria RAM con sistema operativo Windows 7 de 64bits. La corrida de punta a punta (desde la partición del data set hasta el nodo de comparación de modelos) tardó en total 2.5 horas. El promedio fue de 40 segundos por cada árbol de decisión. Este tiempo de procesamiento podría mejorarse utilizando una máquina más potente o un enfoque de tipo “high-performance”.

Hoy en día la mayoría de los proveedores de herramientas de Minería de Datos están desarrollando lo que se conoce como “High Performance Analytics”. Esta metodología consiste en utilizar varios procesadores para realizar una misma tarea, de forma tal que la carga de trabajo pueda distribuirse acelerando los procesos. En el caso de SAS existen tres formas diferentes de procesamiento de tareas tipo high-performance: en memoria (in memory analytics), directamente

en las bases de datos (in data base) y por grilla de servidores (grid computing)[12]. El desarrollo de estos temas está fuera del alcance de este trabajo.

IV.4.3 Generación de las Métricas y Gráficos de Resultados

SAS Enterprise Miner genera automáticamente gráficos y métricas de resultados para cada árbol de decisión como así también para cada nodo de comparación de modelos que se utilice. Estos gráficos se alimentan de tablas de datos que pueden ser utilizadas en otros programas para otro tipo de análisis.

En nuestro enfoque proponemos el uso de una herramienta de Múltiples Vistas Coordinadas para la visualización interactiva de resultados. Dado que SAS Enterprise Miner no posee esta funcionalidad, tomamos las tablas de datos generadas por la herramienta para visualizarlas en ComVis (Matkovic, Freiler, Gracanin, & Hauser, 2008). Esta herramienta permite generar gráficos interactivos con una amplia cantidad de opciones de personalización las cuales explicaremos en esta sección.

De todos los resultados generados por SAS Enterprise Miner tomamos los datos necesarios para graficar las curvas ROC, Lift, Lift Acumulada y Respuesta Acumulada. A su vez, tomamos los datos de diseño de los árboles finales para la representación de los mismos en ComVis.

Ninguna de las tablas generadas por SAS Enterprise Miner contiene información sobre la configuración de los parámetros del algoritmo que produjo cada resultado. Por tal motivo fue necesaria la construcción de una tabla específica con estos datos. Cada árbol fue identificado con un nombre y 4 columnas, una por cada parámetro configurado: Nivel de Significación, Máxima Profundidad, Máxima Cantidad de Ramas y Mínima Cantidad de Casos por Hoja. Plasmar esta información en un formato tabular para la explotación de los datos a través de visualizaciones, es uno de los principales aportes de nuestra metodología. Al poder unir la información de los parámetros con las tablas de resultados, es posible visualizar toda la información en forma interactiva.

ComVis fue desarrollado con el objetivo de ser lo suficientemente flexible para poder ser adaptado a diferentes casos de uso (Matkovic, Freiler, Gracanin, & Hauser, 2008), lo cual permitió adaptarlo nuestra problemática. Para poder implementar nuestra metodología, se trabajó especialmente en la generación del archivo de entrada de la herramienta. Todas las tablas de resultados generadas por SAS Enterprise Miner y la creada manualmente con la configuración de los parámetros, fueron unidas en una sola tabla que permite relacionar luego las diferentes vistas.

Una de las principales características de ComVis es la forma en que maneja la información de curvas para poder graficar y analizar las mismas. La herramienta se centra en la representación de familias de curvas a través de un único registro en una tabla de datos (Figura IV.9).

La principal aplicación de este concepto fue pensada para gráficos de series de tiempo (Konyha, Lez, & Matkovic, 2012). En la Figura IV.9 se muestran dos formas de almacenar información sobre diferentes simulaciones de temperaturas para una región en particular. La

12 http://www.sas.com/en_us/software/high-performance-analytics.html

Figura IV.9(a) muestra la forma convencional de almacenar este tipo de información. Para cada simulación se genera una nueva línea en donde los escalares (Diff_H y Diff_V) deben repetirse para cada uno de estos registros, duplicando la información. En el caso de la Figura IV.9(b) en lugar de duplicar las filas, se genera una nueva columna que contiene la curva en sí, que en este caso representa la variación de la temperatura simulada en las diferentes corridas.

a.

Record #	Time step	Diff_H	Diff_V	Tropical temp.
1	1	1250	7.5e-5	24.50
2	2	1250	7.5e-5	24.50
3	3	1250	7.5e-5	24.50
...	...	...	...	
500	500	1250	7.5e-5	24.30
501	1	1250	7.7e-5	24.20
...	...			
1001	1	1250	7.9e-5	24.01
...	...			
50000	100	2500	9.5e-5	24.25

b.

Record #	Diff_H	Diff_V	Tropical temp. (t)
1	1250	7.5e-5	
2	1250	7.7e-5	
3	1250	7.9e-5	
...	...	...	
100	2500	9.5e-5	

A Family of Curves

Figura IV.9 Dos estrategias para el manejo de información de curvas. (a) Tiempo representado por una dimensión adicional. Los registros del 1 al 500 representan una única curva. (b) La curva es representada como un tipo atómico en la tabla. Las columnas de esta tabla representan familias de curvas. (extraído de (Konyha, Lez, & Matkovic, 2012)).

Si bien esta técnica de tratamiento de datos de curvas fue pensada para series de tiempo, rápidamente pudo observarse su aplicación para otro tipo de gráficos bidimensionales. En nuestro caso esta forma de manejar los datos de curvas puede aplicarse a los gráficos de curvas ROC, Lift, Lift Acumulado, Respuesta y Respuesta Acumulada.

El archivo de entrada de ComVis contiene una única línea para cada árbol generado, es decir 240 registros, y 58 columnas representando las diferentes características de estos árboles que serán analizadas visualmente. De estas 58 columnas, 12 corresponden a las familias de curvas, mientras que el resto son escalares tales como KS, error cuadrático, etc, datos que también obtuvimos de las salidas por defecto de SAS Enterpirse Miner.

IV.4.4 Evaluación Visual e Interactiva de los Resultados

Una vez que se generaron los archivos y se cargaron en ComVis, el analista puede comenzar con las tareas de análisis que lo guiarán en la selección del modelo final que más se ajuste a su necesidad.

Acorde a nuestra metodología, el analista sólo debería interactuar en la etapa de configuración de parámetros, en la evaluación y selección final. Los dos pasos que describimos anteriormente (generación de modelos y generación de métricas y gráficos) deberían ser ejecutados por la herramienta en forma automática. Es decir, una vez que la herramienta genera los modelos y sus resultados, debería mostrarlos directamente en una vista como la de la Figura IV.10 sin necesidad que el analista tenga que cargar el archivo. Esto se podría lograr en una herramienta totalmente integrada.

Una vez que el archivo plano es cargado en ComVis, puede personalizarse la visualización de diferentes gráficos que luego servirán para realizar la evaluación visual e interactiva de los diferentes árboles de decisión. En la Figura IV.10 se muestran la mayoría de los gráficos disponibles.

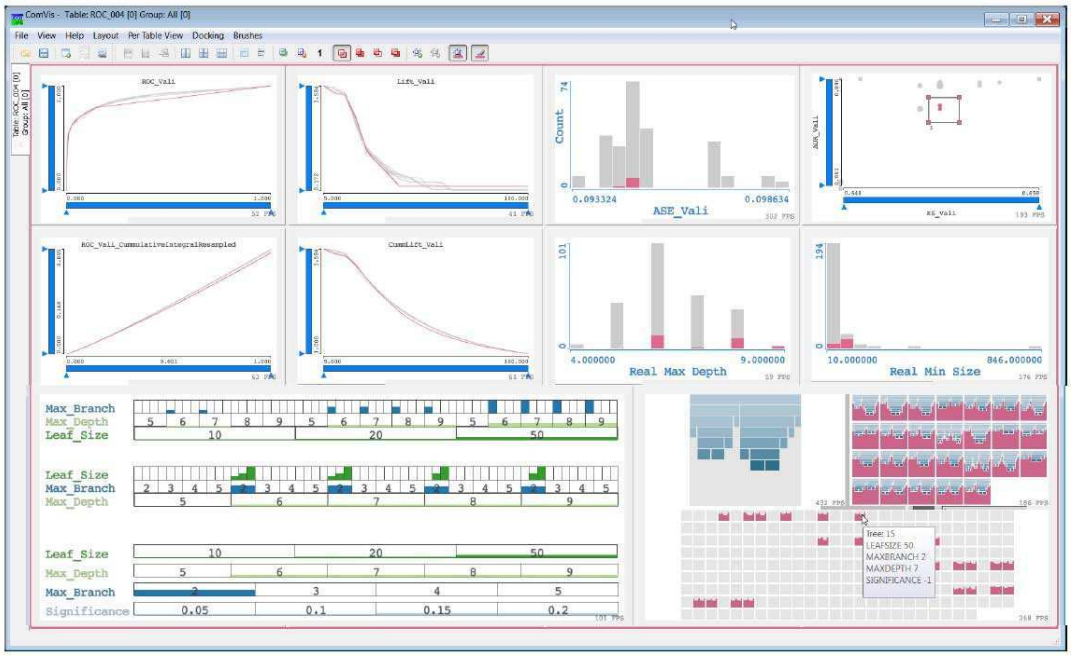

Figura IV.10 Pantalla Inicial con información del archivo de entrada de ComVis.

Para el caso que estamos analizando, deseamos conocer el modelo que mejor identifique aquellos clientes que son más propensos a no pagar su factura en un lapso de 60 días. Los modelos cuya curva ROC esté más cercana al eje y son los más interesantes, ya que son los que mejor capacidad de predicción tienen. También podemos utilizar la curva Lift para evaluar la bondad del modelo. En este caso la curva más alta será la que represente el modelo con mayor ganancia.

En la Figura IV.11 podemos observar a grandes rasgos que varios modelos tienen la misma curva ROC, muy pocas sobresalen sobre otras. Sin embargo al hacer zoom en el eje de las y (Figura IV.12) podemos observar que algunos de ellos son más interesantes que otros ya que sus curvas ROCs se separan de las otras. Para poder analizar los parámetros que generaron cada curva, seleccionamos la curva de interés mediante la funcionalidad de resaltado, la cual permite señalar una curva en este gráfico y observar la selección en los gráficos restantes. El hacer zoom en el eje y nos permite seleccionar las curvas con mayor precisión.

Al hacer esta primera selección es interesante descubrir que una única curva representa 8 árboles diferentes. Esto se visualiza fácilmente en la vista de Múltiples árboles (extremo inferior de la Figura IV.13) en donde hay 8 cuadros seleccionados. Esto quiere decir que diferentes combinaciones de parámetros produjeron los mismos resultados en términos de curva ROC.

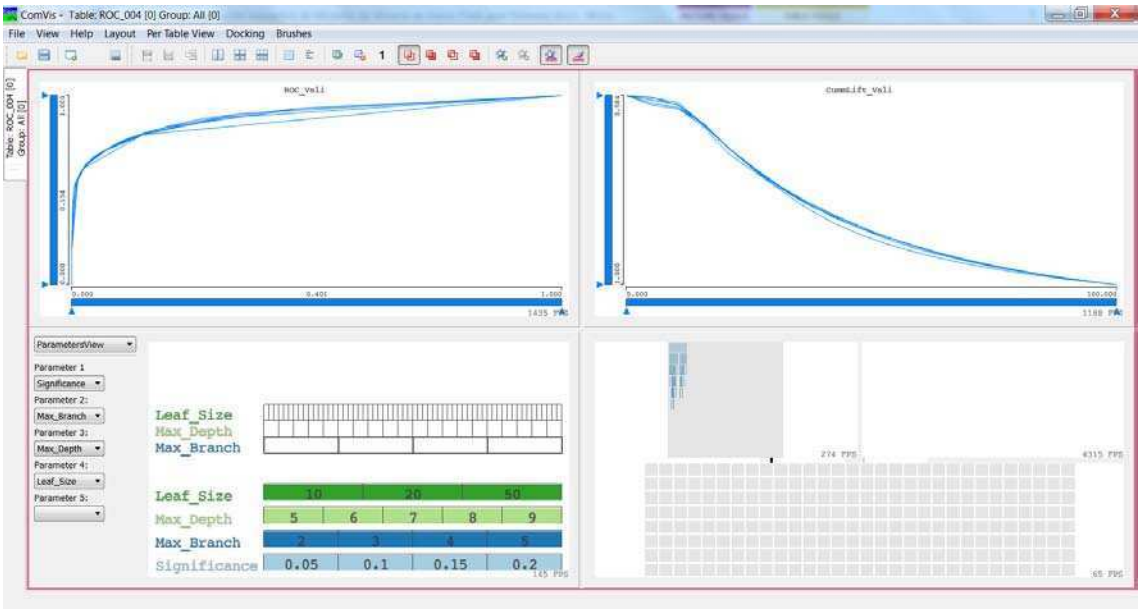

Figura IV.11 Resultados Generales sin utilizar filtros ni zoom

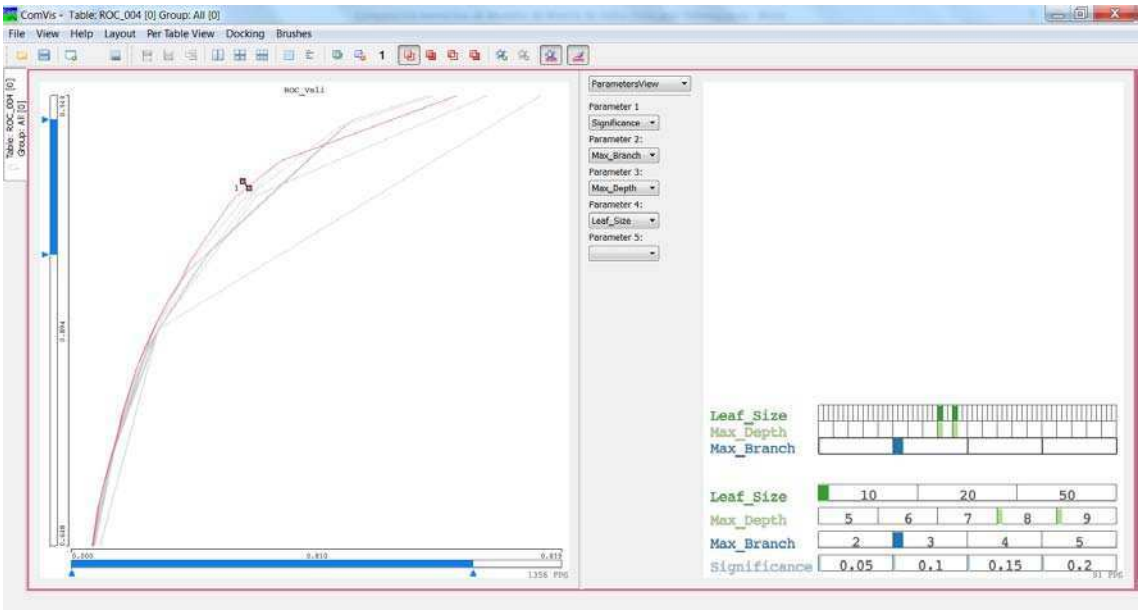

Figura IV.12 Zoom en curvas interesantes y análisis de los parámetros asociados.

En la vista de parámetros cruzados de la Figura IV.13 también podemos observar fácilmente que la curva resaltada se obtiene configurando la máxima cantidad de ramas (Max_Branch) en 3, la máxima profundidad (Max_Depth) en 8 ó 9, y la mínima cantidad de casos por hoja (Leaf_Size) en 10. El nivel de significación (Significance_Level) no parece tener incidencia ya que todos los valores posibles se encuentran resaltados. En este primer análisis podríamos concluir que el mejor modelo se obtiene al configurar la máxima cantidad de ramas en

3, y la profundidad en 8 ó 9, sin embargo esto no sería válido ya que sólo hemos seleccionado 8 de los 60 árboles que tienen este nivel de apertura. Para poder validar el punto, es preciso realizar otros análisis contrastantes.

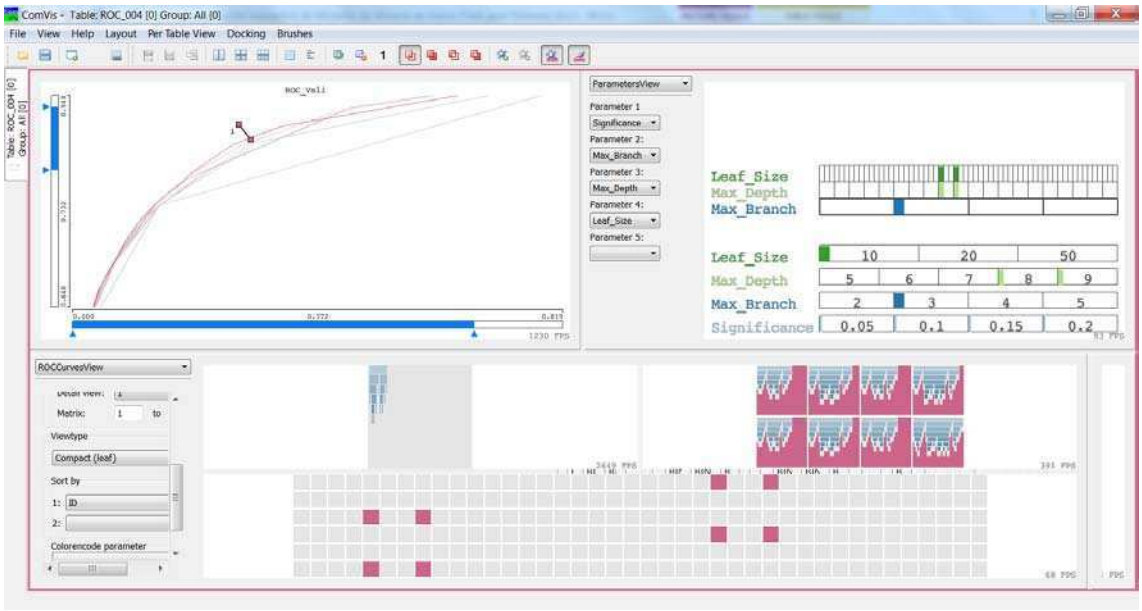

Figura IV.13 Selección de curva más interesante y análisis de parámetros

En la Figura IV.14 se puede observar un nuevo gráfico en la visualización en el panel inferior derecho. Seleccionamos un gráfico de dispersión para representar la relación entre el parámetro máxima cantidad de ramas y el área bajo la curva ROC[13]. En el eje de las x representamos el parámetro Máxima Cantidad de Ramas (Max_Branch) y en el eje de las y el área bajo la curva ROC. De esta forma podemos seleccionar fácilmente aquellos puntos que están por debajo del óptimo para el parámetro Max_Branch=3 y observar las diferencias versus la curva seleccionada en la Figura IV.12.

En el gráfico de la curva ROC podemos ver que se trata de una de las curvas medias. En la vista de parámetros podemos observar que hemos seleccionado 44 árboles de los 60 posibles y que la mayoría corresponden a árboles de poca profundidad, aunque también hay una menor cantidad de árboles de 8 y 9 de profundidad. Por lo tanto no podemos todavía asegurar que el mejor resultado se obtenga con una configuración de máxima apertura de ramas en 3 y máxima profundidad en 8 y 9 niveles.

Para realizar un análisis más profundo de los resultados obtenidos, seleccionamos otras variables y las representamos en nuevos gráficos. En la Figura IV.155 se puede observar el panel dividido en 6. En la parte superior encontramos tres gráficos de dispersión. El primero (de izquierda a derecha) es el mismo de la Figura IV.144. Los dos adicionales representan la relación

13 SAS Enterprise Miner utiliza el acrónimo AUR para referirse al área bajo la curva ROC en lugar de utilizar AUC, que es el acrónimo más utilizado en general.

entre el parámetro configurado y la real dimensión obtenida en cada modelo. En el gráfico del medio observamos la relación entre el parámetro Mínima Cantidad de Casos por Hoja (Leaf_Size) y el real menor tamaño de hoja en cada árbol. En el gráfico de la derecha podemos observar la relación entre el parámetro Máxima Profundidad y la real profundidad del árbol. Estas métricas fueron calculadas para poder representarlas ya que SAS Enterprise Miner no las genera.

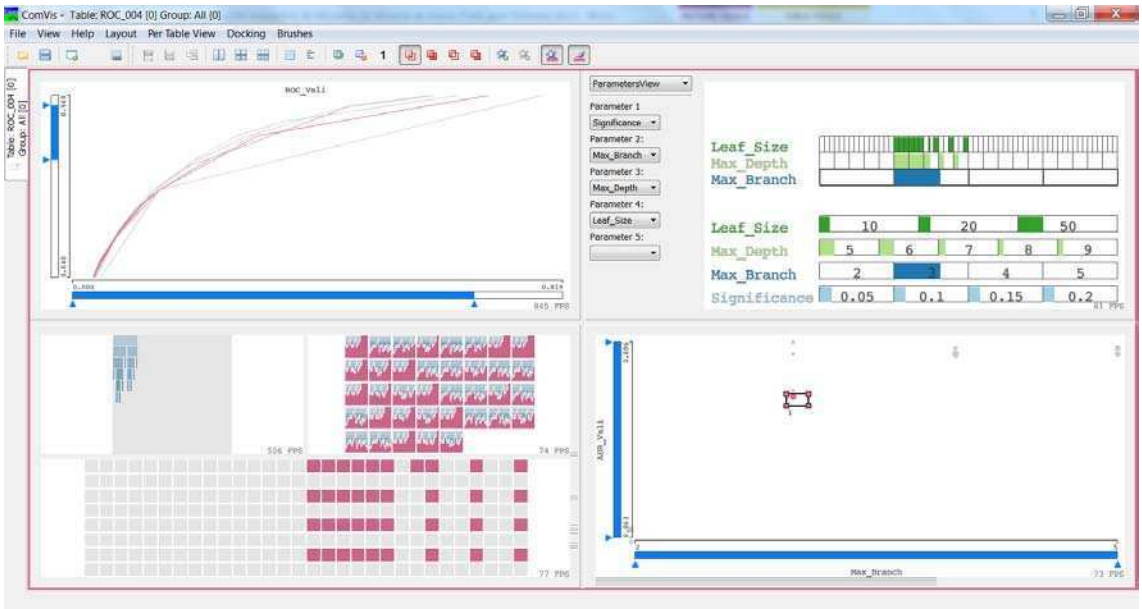

Figura IV.14 Análisis contrastante para parámetro Máxima Cantidad de Ramas

En la Figura IV.155 se encuentran seleccionados los árboles con mejor área bajo la curva ROC. Tanto árboles de 2 ramas como de 3 producen los mejores resultados. La máxima profundidad configurada en ambos casos es de 8 y 9 niveles. Sin embargo, si observamos el gráfico del extremo superior derecho, podemos observar que en realidad ningún árbol pudo llegar a tener 9 niveles de profundidad (el par 9,9 no existe). Esto quiere decir que las reglas de parada del algoritmo podaron el árbol antes de llegar a esta profundidad. La mínima cantidad de casos por hoja, en cambio, sí fue alcanzada.

El usuario puede ahora seleccionar aquel modelo que más se ajusta a sus necesidades. Tiene información de todas las pruebas generadas y conoce los resultados que puede obtener de sus datos. La selección del modelo más conveniente involucrará también decisiones basadas en datos cualitativos, como por ejemplo qué modelo resultará más fácil de comprender por aquellos que tienen que luego implementarlo.

IV.4.5 Selección del Modelo más Apto

Las evaluaciones realizadas en el paso anterior concluirán con la selección del modelo más apto para la problemática bajo análisis. En este ejemplo en particular el usuario puede elegir entre un árbol de 2 ramas y uno de 3. El árbol de 2 ramas produce reglas más simples por lo que por lo general se preferirá antes que un árbol más ancho.

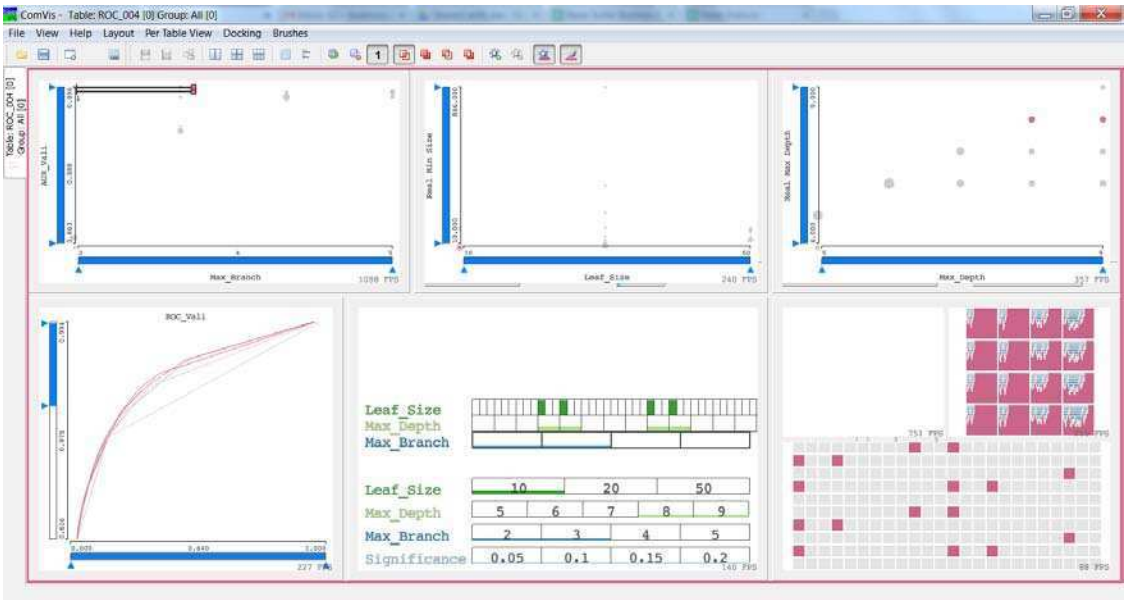

Figura IV.15 Comparación de Configuración de Parámetros versus resultados reales

Para seleccionar el modelo, el usuario debería hacer click en cualquiera de los gráficos e indicar que es su modelo elegido. Este modelo quedará implementado de forma tal que mes a mes puedan generarse las puntuaciones que reflejan la probabilidad de falta de pago de la factura telefónica. Para lograr este tipo de integración, es preciso que toda la plataforma sea integrada, reduciendo al mínimo la cantidad de interacciones del usuario para tareas operativas y no de análisis.

IV.4.6 Conclusiones

Mediante el caso desarrollado demostramos la forma en que un analista podría evaluar diferentes resultados posibles de forma visual e interactiva sin necesidad de tener que invertir su tiempo en la configuración por prueba y error de los árboles de decisión, como debe hacer mediante la metodología tradicional. Al poder evaluar múltiples combinaciones el usuario puede decidir cuál es la más conveniente, asegurándose que analizó una cantidad suficiente de posibilidades.

A su vez demostramos que para el caso analizado, diferentes combinaciones produjeron los mismos resultados, lo cual le sirve al usuario para limitar las combinaciones posibles en el futuro y sólo seleccionar los rangos de parámetros que pueden producir resultados más dispares. Al ejecutar una gran cantidad de modelos sin necesidad de hacerlo en forma iterativa, el usuario puede analizar los resultados de múltiples combinaciones de parámetros, encontrando mejores resultados o resultados más adecuados a sus necesidades.

IV.5 Resumen

En este capítulo presentamos nuestro enfoque para el análisis interactivo de árboles de decisión junto con un caso de aplicación. Nuestra metodología tiene como objetivo contribuir a la

disciplina de Visual Analytics en un área que hasta el momento no ha sido lo suficientemente explorada (Bertini & Lalanne, 2009). El principal diferencial de nuestro enfoque consiste en incluir como parte de la visualización de resultados, la visualización de los parámetros de configuración de árboles de decisión. El objetivo de ello es poder contribuir al aprendizaje del analista sobre la incidencia de los parámetros del algoritmo de árboles de decisión en la efectividad del modelo. Para ello es preciso contar con sistemas de visualización que soporten el análisis a través de múltiples vistas. La metodología aquí presentada no está presente en ninguna de las herramientas de minería de datos del mercado. El enfoque se basa en los principios de la disciplina de Visual Analytics, en donde el usuario tiene mayor incidencia en las etapas de análisis, interactuando de manera ágil e intuitiva con la interfaz. La Minería de Datos suele tener un enfoque contrario, orientado a la mayor automatización de las tareas. Nuestra propuesta consiste en utilizar la automatización de la minería de datos en las etapas de ejecución de los algoritmos y generación de resultados e involucrar al humano en las tareas que requieran análisis y conocimiento del dominio. Combinando ambas disciplinas se pueden producir resultados no sólo óptimos en términos matemáticos, sino también útiles para el negocio.

En el capítulo siguiente presentaremos las conclusiones del trabajo y las mejoras propuestas que constituirán la base de trabajos futuros.

V. Conclusiones y Futuros Pasos

V.1 Introducción

En este trabajo hemos desarrollado una metodología para la evaluación visual e interactiva de múltiples árboles de decisión. El objetivo de este desarrollo fue contribuir a la creciente disciplina de Visual Analytics, cubriendo falencias de las herramientas actuales en lo que respecta a la visualización del espacio de parámetros.

En la actualidad existen múltiples desarrollos tendientes a acortar la brecha entre Visualización de la Información y Minería de Datos. Muchos de ellos se concentran específicamente en la exploración visual de los datos y en los resultados de los modelos, pero ninguno hace referencia a las configuraciones de esos modelos.

Nuestra propuesta es un enfoque metodológico que bien podría implementarse en cualquiera de las herramientas de Minería de Datos, Visualización de la Información o las conocidas como herramientas de Business Intelligence. Si se deseara implementar esta metodología en una herramienta de Minería de Datos, sería necesario desarrollar los componentes de visualización de forma tal de lograr la interacción deseada entre todos los componentes. Si en cambio nuestro enfoque fuese adoptado por una herramienta de visualización, se debería desarrollar el algoritmo de árboles de decisión, junto con los componentes necesarios para generar las combinaciones de parámetros.

Actualmente las herramientas de Visual Analytics están proliferando a un ritmo muy acelerado. Este desarrollo está siendo liderado principalmente por los proveedores de herramientas de Visualización que comienzan a ofrecer dentro de sus capacidades funcionalidades analíticas. Los líderes de herramientas de Minería de Datos todavía no presentan un claro camino hacia la integración de Minería de Datos y Visualización.

Empresas tales como Tableau, líder en Business Intelligence, en lugar de desarrollar capacidades propias, se integra con la plataforma R para poder cubrir las necesidades analíticas de sus usuarios. Microstrategy también opta por esta alternativa otorgando la capacidad de integrar código desarrollado en R para mejorar sus capacidades analíticas. En ambos casos la integración se realiza a nivel de resultados, es decir que se utilizan las capacidades gráficas para evaluar las salidas de los modelos de R, pero no para configurar los mismos visualmente.

SAS, empresa líder en Minería de Datos, aprovecha sus modelos analíticos para incluirlos en sus nuevos desarrollos visuales, tales como la herramienta SAS Visual Analytics. Esta herramienta en constante cambio pretende llevar el poder de Minería de Datos a los usuarios de negocios en forma visual. Tal es así que una de sus funcionalidades es la construcción de árboles de decisión. Esta capacidad sigue siendo limitada a la construcción de un árbol por vez. IBM, otro de los grandes proveedores de herramientas de Minería de Datos, no presenta una estrategia tan clara en la integración de sus modelos de Minería de Datos (adquiridos mediante la compra de SPSS) y su plataforma de Business Intelligence.

Si bien nuestro enfoque está desarrollado sólo para visualizar árboles de decisión, nuestra metodología podría aplicarse a otros algoritmos de Minería de Datos que involucren tareas de configuración de parámetros, como por ejemplo Redes Neuronales. En este sentido sería

interesante complementar el software Matlab, especialista en este tipo de modelos, con las capacidades de visualización interactiva de ComVis propuestas.

En el este capítulo repasaremos las contribuciones de nuestro enfoque al proceso de Visual Analytics y presentaremos algunas ideas sobre las cuales podrían trabajarse para que la metodología cubra más etapas del ciclo.

V.2 Contribución al Proceso de Visual Analytics

Nuestra propuesta tiene como objetivo mejorar alguna de las tareas del proceso de Visual Analytics presentado por Keim (Keim, Kohlhammer, Ellis, & Mansmann, 2010). Como puede observarse en la Figura V.1[14] con nuestro aporte cubriríamos las tareas de construcción de modelos, visualización de los mismos y refinamiento de los parámetros. En el trabajo elaborado no hicimos foco en las etapas iniciales del proceso relacionadas con los datos de entrada, sino que nos concentramos en la visualización de los modelos en sí y en el análisis de sus características y resultados. A continuación detallamos cada una de las contribuciones para luego explorar y proponer áreas de mejora.

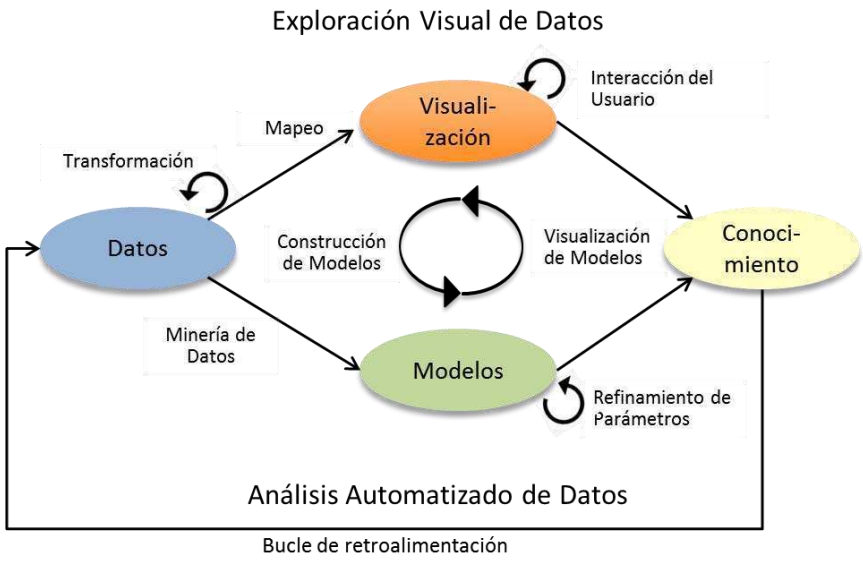

Figura V.1 El proceso de Visual Analytics (traducción de (Keim, Kohlhammer, Ellis, & Mansmann, 2010)).

V.2.1 Construcción de Modelos y Minería de Datos

La metodología que hemos desarrollado comienza desde la etapa del modelado, por lo que supone la preexistencia de un conjunto de datos ya transformados y listos para modelar. Por lo general la mayoría de las herramientas de Minería de Datos precisa de datos en formato tabular, o también llamados datos estructurados. En el caso de nuestro desarrollo, las condiciones sobre

[14] Esta figura es la misma que la Figura II.12. Repetimos nuevamente la misma para facilitar la lectura.

los datos no difieren. Los árboles de decisión soportan datos extremos y faltantes, por lo que no es necesario trabajar con los mismos. Si nuestro enfoque fuese aplicado para otros modelos que no puedan manejar este tipo de problemas en los datos, sería preciso una etapa adicional de transformación de los mismos, ya sea mediante técnicas de rellenado o filtrado de los mismos.

Una vez que se tienen los datos en formato tabular, el proceso continúa con la etapa de construcción de modelos. La mejora propuesta para esta tarea, consiste en realizar la construcción de múltiples modelos en un solo paso en contraposición al enfoque interactivo propuesto por otros autores (Ankerst, Ester, & Kriegel, 2000) (van den Elzen & van Wijk, 2011) para la generación de árboles de decisión.

Si bien la interacción es el eje central de la Visualización de la Información, en nuestro trabajo creemos que es más relevante la interacción al momento de evaluar los resultados de diferentes configuraciones y no tanto en la construcción de los modelos paso a paso. Sostenemos que la construcción interactiva no permite evaluar configuraciones alternativas, siendo necesarios más pasos para arribar a la misma conclusión. El trabajo de van den Elzen presenta varias mejoras respecto al de Ankerst en este sentido, sin embargo sigue estando limitado a la construcción de un modelo por vez.

Un enfoque intermedio podría ser la generación de múltiples combinaciones en varios pasos. Si bien en el caso desarrollado la construcción de modelos se realizó en una sola vez, esta tarea también podría ser interactiva al seleccionar diferentes valores para los parámetros que luego derivarían en diversas combinaciones de parámetros. Es decir, en lugar de generar 240 combinaciones de una vez (como hicimos en el caso planteado) podría generarse una menor cantidad e ir refinando los modelos en sucesivas iteraciones. Nuestra metodología, tal como está planteada, permite realizar la tarea de esta forma progresiva. Podrían seguirse diferentes estrategias en la generación de combinaciones de parámetros. El usuario podría dejar fijo un parámetro y luego jugar con el resto para ir evaluando los diferentes resultados. De cada tanda de pruebas podría seleccionar aquellas configuraciones de parámetros que mejor se ajustaran a la problemática e incluirlas, junto con unas nuevas, en la siguiente tanda de combinaciones.

Esta interacción sería muy fácil de lograr con los menús propuestos en la sección IV.3.1 de este trabajo. Mediante el uso de la función de listas de valores para los parámetros, el usuario podría introducir como valores posibles sólo los que haya seleccionado en cada evaluación.

Dado que la herramienta de visualización sugerida, ComVis, facilita la visualización de múltiples colecciones de curvas junto con la vista cruzada de parámetros, la aproximación iterativa no sería necesaria. Sin embargo algunos usuarios podrían preferir un enfoque del tipo “divide y conquistarás” el cual también es posible de la forma en que hemos propuesto la configuración de parámetros.

V.2.2 Visualización de Modelos y Refinamiento de Parámetros

En nuestro trabajo entendemos visualización de modelos en dos sentidos: visualización de los resultados y visualización de la configuración de parámetros, estando éstos directamente relacionados. La mejora que proponemos con nuestra metodología está ligada mayormente a la visualización de la configuración del modelo. Actualmente ninguna herramienta del mercado

permite visualizar las relaciones entre parámetros y resultados de la forma en que proponemos en nuestro desarrollo metodológico.

Observando la incidencia de la configuración de parámetros en los resultados, es posible tomar decisiones respecto a los mismos, contribuyendo a la tarea de refinamiento de parámetros propuesta por Keim para el proceso de Visual Analytics. Así por ejemplo para el caso de estudio se pudo observar que varias configuraciones produjeron los mismos resultados, lo cual sirve para limitar en un futuro la cantidad de combinaciones a generar.

Se observó también que el nivel de significación no influyó en la mayoría de los resultados, por lo que este parámetro podría no ser de interés para futuras pruebas. Es decir que en corridas sucesivas podría seleccionarse un único valor para este parámetro, disminuyendo así la cantidad de modelos generados y por ende el tiempo de cómputo de los modelos. Este punto, sin embargo, debería validarse con otros conjuntos de datos. Arribar a estas conclusiones mediante la metodología disponible en herramientas de Minería de Datos es más difícil ya que el analista debe ir evaluando una a una, por prueba y error, las diferentes pruebas que realiza.

Otra característica de las propuestas que resulta de gran valor para el refinamiento de parámetros, es la posibilidad de comparar la configuración del modelo versus los niveles alcanzados. En el caso de ejemplo demostramos que ciertas configuraciones de profundidad y cantidad de casos por hoja mínimo no fueron logradas en los árboles de decisión resultantes. Es decir que si bien el usuario configuró la herramienta para obtener un árbol de determinada profundidad, la misma no fue alcanzada en algunos casos debido a la interacción con el resto de los parámetros. Este tipo de información es útil a la hora de generar nuevas combinaciones ya que limita la cantidad de pruebas a realizar. El dato sobre los reales niveles de profundidad y el tamaño de las hojas fue calculado a partir de los datos de los árboles de decisión generados por la herramienta de Minería de Datos, siendo esto una innovación ya que no es un dato disponible en tal herramienta. Si bien para este trabajo sólo desarrollamos estas dos métricas, también podría generarse el mismo cálculo para la cantidad de ramas del árbol.

V.3 Futuros Pasos

En la metodología propuesta hemos cubierto sólo algunas de las etapas del proceso de Visual Analytics, concentrándonos específicamente en la visualización de árboles de decisión y generación de conocimiento sobre la configuración de parámetros de los mismos. Nos enfocamos en estas tareas ya que eran las que presentaban menor progreso en la bibliografía estudiada.

Sin embargo a lo largo del desarrollo de nuestro enfoque fuimos detectando áreas de mejora que podrían ser desarrolladas en trabajos futuros. En esta sección presentamos algunas de estas ideas.

V.3.1 Visualización de Reglas de Decisión

En nuestra metodología no planteamos mejoras en la visualización de reglas de decisión. Por lo general las reglas de decisión generadas por los modelos se representan a través de una

estructura de árbol de tipo nodo-vínculo (Figura V.2, derecha). En la adaptación realizada a la herramienta ComVis, se seleccionó la visualización de árboles de decisión a través de "gráficos de estalactita" (icicle plots) (Figura V.2, izquierda). Estos gráficos son muy útiles para resumir información sobre la estructura general de un árbol (ancho y profundidad) y ocupan menos espacio. Sin embargo, los mismos no contienen información sobre las reglas de decisión generadas por los modelos.

Conocer las diferentes reglas para cada árbol resultante es de especial interés para los analistas, ya que los mismos deben tomar decisiones sobre cuáles utilizar en la práctica. Por ejemplo ante dos modelos que poseen la misma área bajo la curva ROC (AUC) o una muy similar, producto de diferentes configuraciones de parámetros, el analista podría seleccionar un árbol en lugar de otro dependiendo de las reglas que se hayan generado.

Dado que el analista podría seleccionar un modelo por sobre otros en función de las reglas generadas, es preciso trabajar sobre la forma de visualizar las mismas como así también en la comparación de ellas. Visualizar más de un árbol a la vez podría ser complejo para el usuario, por lo que deberían evaluarse visualizaciones alternativas.

Schulz (2011) agrupa en su proyecto treevis.net[15] cientos de posibilidades para la representación de árboles de decisión. Los diferentes diseños pueden ser filtrados según sus características de dimensión, representación y alineación. Seleccionar el más adecuado para comparar una gran cantidad de árboles podría consistir en un trabajo de investigación en sí. Este punto es de sustancial interés para desarrollos futuros de la metodología.

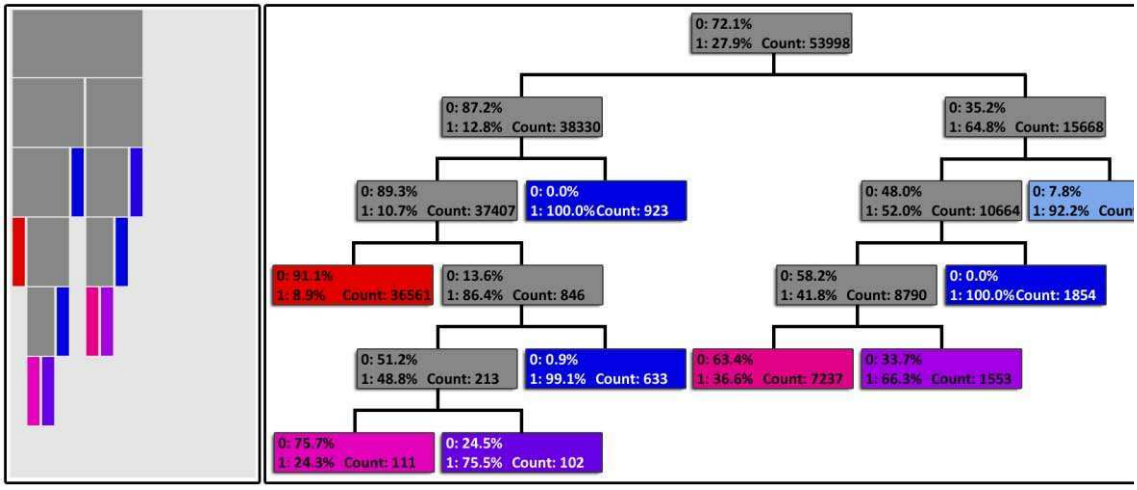

Figura V.2 Visualización de árboles de decisión a través de icicle plots (izquierda) y gráficos de tipo nodo-vínculo (derecha).

V.3.2 Preselección de Resultados Relevantes

En el caso que utilizamos como ejemplo, descubrimos que varias configuraciones de parámetros produjeron los mismos resultados. A pesar de ello todos estos resultados fueron

15 Treevis.net

representados visualmente. En desarrollos futuros, podría trabajarse en la posibilidad de mostrar sólo aquellos modelos que presenten diferencias sustanciales y resumir en una única vista las características de aquellos que no difieren de los otros. Si bien podría decidirse eliminar estos árboles del análisis, consideramos que no sería correcto ya que descubrir que diferentes combinaciones de parámetros producen los mismos resultados, puede ser también de interés para el analista. En el caso de existir resultados repetidos, deberían tomarse decisiones sobre cuál de todos los modelos mostrar, existiendo múltiples opciones para ello.

Otra forma de limitar la cantidad de resultados a visualizar podría ser dándole la posibilidad al usuario de predefinir filtros para ciertas métricas. Así por ejemplo el usuario podría determinar que sólo desea visualizar los árboles que tienen como mínimo una cierta precisión o AUC. En nuestra implementación este filtro puede ser realizado una vez generadas todas las visualizaciones a través de los pinceles de selección. Este filtro podría realizarse antes de la producción de las visualizaciones, introduciéndolo como parámetro en la generación. Sin embargo la información sobre las configuraciones que no alcanzaron el mínimo deseado debería estar también disponible como resumen para conocimiento del analista.

V.4 Conclusión

La disciplina de Visual Analytics sigue creciendo y cada vez son más numerosos los trabajos y desarrollos que se realizan en esta área. El punto central es acercar al humano al análisis de datos a través de herramientas visuales e interactivas. Si bien estas características se contradicen con el objetivo de automatización de Minería de Datos, creemos que nuestro enfoque toma los puntos más sobresalientes de cada ciencia para plantear una forma de trabajo que se ajuste a la definición de Visual Analytics.

A lo largo de este trabajo demostramos cómo es posible unir los mundos de Visualización de la Información y Minería de Datos mediante la adaptación de herramientas y formas de trabajo actuales. Creemos que nuestra metodología presenta características de gran valor y esperamos en un futuro poder contar con una solución integral que tenga todos los componentes necesarios para llevar a cabo las tareas aquí planteadas.

VI. Bibliografía

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference of Management of Data, (págs. 207--216). Washingtong DC.

Andrienko, G., & Andrienko, N. (2007). Coordinated Multiple Views: A Critical View. Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (págs. 72--74). Washington, DC, USA: IEEE Computer Society.

Ankerst, M., Ester, M., & Kriegel, H.-P. (2000). Towards an effective cooperation of the user and the computer for classification. Proc. 6th Intl. Conf. on Knowledge Discovery and Data *Mining (KDD '00)* (págs. 179--188). ACM Press.

Berkhin, P. (2009). Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data, 25--71.

Bertin, J. (1967). Sémiologie graphique : les diagrammes, les réseaux, les cartes. Mouton.

Bertini, E., & Lalanne, D. (2009). Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. SigKDD Explorations Journal, 9--18.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1145--1159.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). Readings in Information Visualization: Using Vision to Thin. Morgan Kaufmann Publishers Inc.

Coppock, D. (21 de 06 de 2002). Why Lift? Obtenido de Information Management: http://www.information-management.com/news/5329-1.html

Dallal, G. E. (23 de 05 de 2012). Jerry Dallal's Home Page. Obtenido de Why P=0.05?: http://www.jerrydallal.com/LHSP/p05.htm

Descartes, R. (1937). La Géométrie. Paris.

Dürsteler, J. (30 de 12 de 2002). InfoVis.net. Obtenido de http://www.infovis.net/printMag.php?num=110&lang=1

Dürsteler, J. (09 de 01 de 2003). InfoVis.net. Obtenido de http://www.infovis.net/printMag.php?num=111&lang=1

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers.

Fayyad, U., Grinstein, G., & Wierse, A. (2001). Information Visualization in Data Mining and Knowledge Discovery. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 37--54.

Friendly, M., & Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Obtenido de http://www.datavis.ca/milestones/

Fry, B. (2007). Visualizing Data, Exploring and Explaining data with the processing environment. O'Reilly Media.

Han, J. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Inselberg, A., & Dimsdale, B. (1990). Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. Proceedings of the 1st Conference on Visualization '90 (págs. 361--378). San Francisco, California: IEEE Computer Society Press.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of the Royal Statistical Society. Series C (Applied Statistics), 119--127.

Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). Mastering The Information Age - Solving Problems with Visual Analytics. Goslar, Germany: Eurographics.

Konyha, Z., Lez, A., & Matkovic, K. (2012). Interactive visual analysis of families of curves using data aggregation. Proceedings of the 12th International Conference on Knowledge Management (págs. 24:1--24:8). Graz, Austria: ACM.

Kumar, D., & Bhardwaj, D. (2011). Rise of Data Mining: Current and Future Application Areas. IJCSI International Journal of Computer Science.

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: a statistically consistent and more discriminating measure than accuracy. PROCEEDINGS OF 18TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI-2003), (págs. 519--524).

Matkovic, K., Freiler, W., Gracanin, D., & Hauser, H. (2008). ComVis: a Coordinated Multiple Views System for Prototyping New Visualization Technology. Proceedings of the 12th Intern. Conference Information Visualization (IV 2008), 215-220.

Minard, C. J. (1861). Des Tableaux Graphiques et des Cartes Figuratives. Paris: Thunot.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Asso-, 415--434.

Müller, E., Assent, I., Krieger, R., Jansen, T., & Seidl, T. (2008). Morpheus: interactive exploration of subspace clustering. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (págs. 1089--1092). Las Vegas, Nevada, USA: ACM.

Nightingale, F. (1857). Mortality of the British Army. London: Harrison and Sons.

Padua, L., Schulze, H., Matković, K., & Delrieux, C. (2014). Interactive exploration of parameter space in data mining: Comprehending the predictive quality of large decision tree collections. Computers & Graphics, 99--113.

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine, 68--70.

Plaisant, C., Shneiderman, B., Chintalapani, G., & Aleks, A. (2014). Treemap. Obtenido de University of Maryland, Human-Computer Interaction Lab: http://www.cs.umd.edu/hcil/treemap/

Playfair, W. (1786). Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century. London.

Quetelet, A. (1846). Lettres sur la théorie des probabilités, appliquée aux sciences morales et politiques. Hayez.

Quinlan. (1993). C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Quinlan, J. R. (1986). Induction of Decision Tree. Machine Learning, 81--106.

Rexer, K. (2012). 2011 Data Miner Survey. Obtenido de Rexer Analytics: http://rexeranalytics.com/Data-Miner-Survey-Results-2011.html

Roberts, E. (17 de 05 de 2014). Neural Networks. Obtenido de http://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/index.html

Schulz, H.-J. (2011). Treevis.net: A Tree Visualization Reference. IEEE Computer Graphics and Applications, 11--15.

Shneiderman, B. (1991). Tree visualization with Tree-maps: A 2-d space-filling approach. ACM Transactions on Graphics, 92--99.

Shneiderman, B., & Plaisant, C. (2004). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Pearson Addison Wesley.

Snow, J. (1855). On the Mode of Communication of Cholera. John Churchill.

Thomas, J., & Cook, K. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Computer Society Press.

Tufte, E. (1983). The Visual Display of Quantitative Information. Graphic Press.

Tukey, J. W. (1962). The future of data analysis. j-ANN-MATH-STAT, 1--67.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

van den Elzen, S., & van Wijk, J. J. (2011). BaobabView: Interactive construction and analysis of decision trees. IEEE VAST 2011 (págs. 151--160). IEEE.

Vuk, M. (2006). ROC Curve, Lift Chart and Calibration Plot. Advances in methodology and Statistics.

Wang Baldonado, M. Q., Kuchinsky, A., & Woodruff, A. (2000). Guidelines for Using Multiple Views in Information Visualization. Proceedings of the Working Conference on Advanced Visual Interfaces (págs. 110--119). New York, NY, USA: ACM.

Ware, C. (2004). Information Visualization. Perception for design. Morgan Kaufman.

Wikipedia. (2 de Mayo de 2014). Jeroglífico. Obtenido de Wikipedia, La enciclopedia libre: http://es.wikipedia.org/w/index.php?title=Jerogl%C3%ADfico&oldid=74151706

Wong, P. C. (2004). Visual analytics. IEEE Computer Graphics, 20--21.