UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

MINERÍA DE DATOS EN LA FACULTAD DE CIENCIAS PARA DETECTAR PATRONES QUE INCIDEN EN LA CONCLUSIÓN DE LA CARRERA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

SONIA YOSELIN OVANDO RETIZ

DIRECTOR DE TESIS:

DRA. AMPARO LÓPEZ GAONA

2014

1. Datos del alumno
Ovando
Retiz
Sonia Yoselin
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
408056120

2. Datos del tutor
Dra.
Amparo
López Gaona

3. Datos del sinodal 1
Dr.
Javier
García García

4. Datos del sinodal 2
Dra.
Ruth Selene
Fuentes García

5. Datos del sinodal 3
Dra.
Sofía Natalia
Galicia Haro

6. Datos del sinodal 4
Act.
Jaime
Vázquez Alamilla

7. Datos del trabajo escrito
Minería de datos en la Facultad de Ciencias para detectar patrones que inciden en la conclusión de la carrera.
155 p.
2014

Con todo mi cariño y mi amor para las personas que hicieron todo en la vida para que yo pudiera lograr mis sueños, por motivarme y darme la mano cuando sentía ya no tener más fuerza, a ustedes por siempre mi corazón y mi agradecimiento.

Papá y Mamá
Sandra Marylin Ovando Retiz
Raquel Dorantes Ortiz
Miguel Ángel Saavedra Bello
Gabriel Cabeza Esquivel

A mi amada Universidad, a la Facultad de Ciencias por abrirme las puertas de sus instalaciones, enseñarme tantas cosas y otorgarme las facilidades para la realización de este trabajo.
A mis maestros, tutora y sinodales que influyeron con sus vastos conocimientos y experiencias en mí para lograr la visión y cumplimiento de ésta meta.
A amigos, compañeros de aventuras por estar siempre al pie del cañón a lo largo de todos estos años.
Y finalmente a la persona que inspiró mi camino. Por compartirme su pasión por las Matemáticas y la Actuaria y sembrar en mí esa semilla que ahora comienza a rendir frutos. Anita Balboa

A todos y cada uno de ellos les dedico cada una de estas páginas de mi tesis.

Sonia Yoselin Ovando Retiz

" POR MI RAZA HABLARÁ EL ESPÍRITU"
~ José Vasconcelos ~

ÍNDICE

INTRODUCCIÓN

En la vida moderna, tratar de encontrar patrones, tendencias y anomalías es uno de los grandes retos a los que se enfrentan las organizaciones en el ámbito de los negocios y de solución y detección de problemas. La fuerte competencia, los cambios constantes y la automatización de procesos han llevado a estas asociaciones a buscar nuevas tecnologías en almacenamiento de la información y que minimicen los costos y generen respuesta a sus requerimientos institucionales.

La mayoría de los empleados toman decisiones en base a su experiencia, conocimiento, capacidad de análisis. Pero este tipo de decisiones pueden no ser óptimas. Para ello se aplican modelos matemáticos y metodologías de análisis que explotan las bases de datos de las empresas para generar información y conocimiento útil y dinámico para complejos procesos de toma de decisiones.

Como parte de la Inteligencia de Negocios, aplicada a empresas, organizaciones y asociaciones, es que la Minería de Datos tiene lugar para utilizar eficazmente los datos almacenados y obtener información útil sobre los hábitos y preferencias de los individuos objeto de estudio, para después convertirlos en conocimiento que lleve a elaborar estrategias que impulsen a la organización a adquirir o conservar una fuerte presencia en su ramo.

Las áreas de aplicación de la Minería de Datos se dividen principalmente en dos grandes grupos, el primero es el que está enfocado a detectar patrones de comportamiento para generar reglas, agrupar datos según ciertas características y a partir de ahí tomar decisiones y el segundo es el que está enfocado al análisis de la información para generar predicciones de ciertas variables para canalizar las acciones hacia esos resultados.

El presente trabajo tiene como objetivo general la aplicación de técnicas de Minería de Datos que ayuden a detectar patrones que inciden en que un alumno de la carrera de Ciencias de la Computación o Actuaría de la Facultad de Ciencias de la UNAM, concluya su carrera. Tomando en cuenta tanto los factores que lo impulsan a terminarla como los que dificultan el proceso o que lleven al alumno a desertar de la Facultad.

Para poder llegar a la parte aplicada del trabajo es importante describir de manera más explícita qué es la Minería de Datos, cuáles son sus orígenes, sus ventajas y desventajas, la diferencia que existe con otro tipo de análisis de la información, la metodología que sigue y sus aplicaciones. Estos temas serán presentados en el Capítulo 1.

Así mismo es indispensable conocer la teoría y parte técnica de las herramientas de Minería de Datos que se aplican en la resolución de problemas. El Capítulo 2

tiene por objetivo cubrir estos temas para que el lector pueda tener una mayor retroalimentación en conocimiento y entendimiento del proceso de aplicación.

Una vez que se presenta la estructura anterior, se lleva a cabo un caso práctico. Se trabaja con una base de datos de la Facultad de Ciencias para realizar el análisis de deserción y culminación de las carreras antes mencionadas; iniciando en el Capítulo 3 con la preparación de los datos, haciendo correcciones a algunos registros, generando nuevas variables y realizando transformaciones matemáticas en algunas de ellas.

Este análisis se realizará en Microsoft Excel 2010, herramienta que permite trabajar con bases de datos grandes, analizarlas y modificarlas de manera dinámica y sencilla, a pesar de que existen muchas otras herramientas probablemente más potentes, se decidió utilizar Microsoft Excel debido a que se tiene un conocimiento alto del software. Se utilizará tanto la parte interactiva del software como la parte de su programador Microsoft Visual Basic.

Finalmente en el Capítulo 4, se generan los modelos, mostrando paso a paso la forma en que se elaboraron y las razones que llevaron a la selección de los mismos. Resultando así algunas reglas y patrones que ayudan a tener una visión más específica sobre el comportamiento de los alumnos que sirva para en un futuro realizar proyectos de mejora o afianzamiento de los planes de estudio o de las bases de datos.

En esta última etapa de la tesis se seleccionó el programa estadístico “R” ya que es un software de uso libre y gratuito, que permite hacer estudios exploratorios de bases de datos y así mismo tiene la capacidad de realizar minería de datos usando código en la consola del programador como con una interfaz gráfica llamada Rattle que permite al usuario realizar minería sin introducir el código de cada modelo.

CAÍTULO 1. MINERÍA DE DATOS

1.1 Antecedentes

En un entorno globalizado, el éxito de una organización depende altamente del conocimiento y habilidades para hacer negocio, además de la capacidad para actualizar la información y adaptarse a los nuevos desafíos que se presenten.

Actualmente las actividades de muchas organizaciones se centran en el almacenamiento, transferencia e interpretación de la información utilizando para realizar estas tareas tecnologías cada vez más avanzadas.

"En los años 60's se generaron los procesos Batch (por lotes) que consisten en determinar trabajos comunes y realizarlos todos a la vez".[10] Los sistemas eran usados principalmente para transacciones de negocios y sus capacidades de realizar reportes se limitaban a un número determinado de ellos. Dichos sistemas, se sobrecargaban de información y los usuarios tenían que esperar por días o semanas para obtener sus reportes en caso que requirieran reportes distintos a los estándares disponibles. En resumen, la información era difícil de capturar y analizar, poco flexible para modificar y se necesitaba programar cada actividad que se requería utilizar.

Al verse en la necesidad de tener una ventaja competitiva en los mercados, en los años 70's surgen los primeros DSS (Decision Support Systems) que son sistemas de soporte para toma de decisiones muy flexibles y adaptables, proporcionan un enfoque específico hacia el análisis de los datos de una organización.[27]

En esta misma época surge otra herramienta de software de inteligencia empresarial muy poderosa conocida como EIS (Executive Information Systems), que pretendía mostrar informes y listados de información relevante y consolidada de diversas áreas del negocio mediante el acceso a una base de datos y de esta forma permitir a los altos ejecutivos optimizar la labor de resumir y presentar los datos de la manera más sencilla posible.[27]

Posteriormente a mediados de los 80's y principios de los 90's surgirán una gran cantidad de herramientas para el análisis, consulta, informes y acceso a datos debido al crecimiento de la información como por ejemplo hojas de cálculo, interfaces gráficas, reportes operacionales, modelación estadística, etc.[10]

Precisamente en esta época se puede destacar el surgimiento del Data Warehouse, que de acuerdo con W. H. lnmon es "Un conjunto de datos integrados orientados a un tema, que varían con el tiempo y no son transitorios, los cuales soportan un proceso de toma de decisiones en una administración" [11] y surge con el objetivo de hacer que la información de la organización sea accesible, consistente, adaptable y elástica y así tener un sistema de soporte a las

decisiones; los datos existentes y las tecnologías no cambian ni se corrompen y se protegen los valores de la información.

Teniendo ya un sistema de control de las bases de datos, se comienzan a buscar mejores formas de encontrar información a partir de ellas. Fue entonces cuando Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro entre otros, empezaron a consolidar el término de Minería de Datos.[25]

Comienzan a existir cambios que incluyen el crecimiento de las redes de computadoras, técnicas de redes neuronales y algoritmos avanzados, se propagan las aplicaciones cliente/servidor por lo que para las organizaciones el conocimiento del cliente y sus patrones de comportamiento se convierte en una necesidad fundamental.

1.2 ¿Qué es la Minería de Datos?

La minería de datos (MD) o Data Mining, es un proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, el aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.[11]

La tarea de MD es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos. Estos patrones pueden ser vistos como una especie de resumen de datos de entrada y pueden ser utilizados en un análisis adicional.

A menudo la MD es vista como una parte del descubrimiento de conocimiento en las bases de datos (KDD Knowlegde Discovery in Databases), el cual consta de varias etapas: elegir los datos apropiados, pre-procesarlos, transformarlos si es necesario, realizar la minería de datos, encontrar patrones y relaciones y luego interpretar las estructuras descubiertas.

Existen varias fases de un proyecto de Minería de Datos que son siempre las mismas, independientemente de la técnica utilizada para la extracción del conocimiento. En la siguiente figura se pueden ver, de forma esquemática, los pasos de un proyecto de minería de datos.[23]

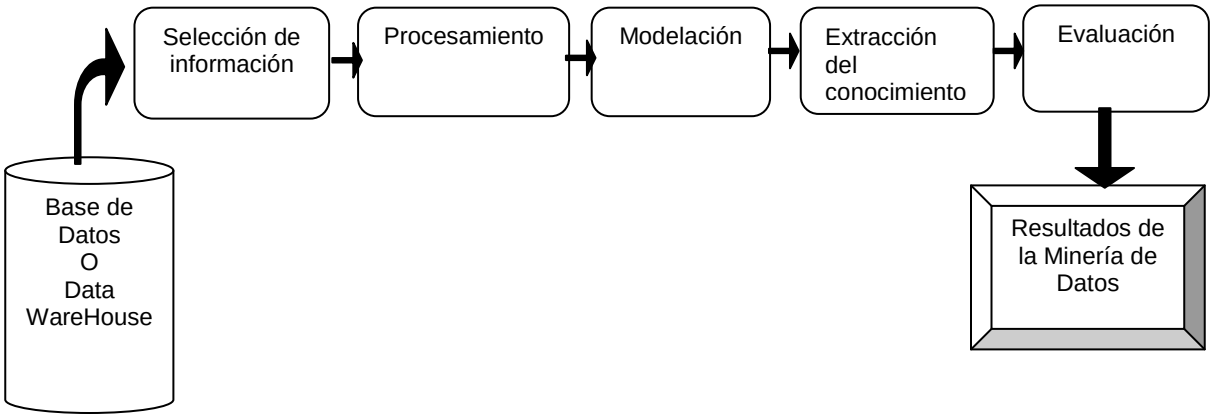


Figura 1.1 Fases de un proyecto de Minería de Datos

1. Selección de la información.- Los datos contenidos en una base de datos pocas veces son idóneos para trabajar directamente sobre ellos por lo que se hace un filtro de forma que se eliminan o modifican valores incorrectos, no válidos, desconocidos y atípicos, etc., según convenga al proyecto.

Uno de los procesos para organizar el flujo de datos entre diversos sistemas de una organización, moviéndolos, reformateándolos, limpiándolos y cargándolos es el llamado proceso ETL conocido así por sus siglas en inglés extract, transform, load; éstas son las etapas de las que se compone, en español son: extracción, transformación y carga. En la primera se obtiene la información de diversas fuentes tanto internas como externas de la organización; en la segunda se realiza un filtrado, limpieza, depuración, homogeneización y agrupación de la información; y en la tercera se organiza y actualiza los datos y los metadatos en una la base de datos correspondiente. [23][23]

La idea del ETL es que después de leer los datos primarios de una base, realizarle alguna transformación, validación y filtrado y al final escriba datos en el almacén o base de datos nueva para que estén disponibles para ser analizados por los usuarios.

2. Procesamiento.- En esta fase se analiza la información seleccionada mediante técnicas estadísticas para obtener una mejor perspectiva del comportamiento de los datos y poder saber qué tipo de modelo conviene más utilizar para el cumplimiento del objetivo principal.

3. Modelación.- De acuerdo a las características se seleccionan y aplican las técnicas adecuadas de modelado (las cuales se abordarán en el Capítulo 2). Se debe calibrar la configuración del modelo para optimizar los resultados. Pueden

usarse varias técnicas a la vez para generar distintos modelos y compararlos entre sí al finalizar el proceso.

4. Extracción del conocimiento.- Después de aplicar algún modelo de minería de datos se identifican los patrones de comportamiento entre las variables observadas.

5. Evaluación.- Debido a que los modelos y el conocimiento adquirido a través de ellos son determinantes para la toma de decisiones, cada modelo realizado debe ser validado de manera que las conclusiones que arroje sean satisfactorias y suficientes para el objetivo establecido. La información obtenida debe ser coherente y se selecciona el modelo que mejor se ajuste a resolver el problema establecido como objetivo.

6. Resultados.- Las conclusiones obtenidas a través del proceso de minería de datos son presentadas de manera resumida y concreta para que los ejecutivos puedan entenderlas sin necesitar altos conocimientos matemáticos ni estadísticos, y así tomar decisiones convenientes para la organización.

1.3 Tipos de repositorios de datos.

Un proceso de minería de Datos debería poder aplicarse a cualquier repositorio de datos. A continuación se mencionan brevemente las características, importancia y aplicación de algunas de éstas:

a) *Base de datos relacionales:* Consiste en un conjunto de datos interrelacionados y un software para manejar y acceder a dichos datos. Está representada por un conjunto de tablas con una serie de atributos y almacenan una gran cantidad de elementos que a su vez contienen distintos atributos. Cuando se aplica la minería de datos en este tipo de bases de datos, es posible buscar tendencias o patrones entre los datos basándose en los atributos.[8]

b) *Base de datos temporales (Series de Tiempo):* Son las que contienen muchos atributos relacionados con el tiempo, pueden referirse a distintos instantes o intervalos. Una base de datos de series de tiempo almacena secuencias de valores que cambian con el tiempo. Este tipo de base de datos es muy utilizada en la estadística y minería de datos para encontrar las características en evolución y cambio.[8]

c) *Base de datos orientada a objetos:* Son un tipo de base de datos que se basa en la programación orientada a objetos (empleados, clientes o productos). A cada objeto se le asocia un conjunto de variables que lo describe, un conjunto de mensajes que utiliza para comunicarse y un conjunto de métodos que contienen el código para implementar un mensaje. La función de minería de datos en este tipo de bases de datos puede ser clasificar, subclasificar y ordenar en jerarquías.[8]

d) *Base de datos espaciales:* Son las que contienen información relacionada con el espacio físico como zona geográfica, una ciudad, redes de transporte, información del tráfico, imágenes satelitales etc. Sus aplicaciones pueden ser proveer información de los servicios públicos, descubrir patrones que describan características de casas localizadas en algún área geográfica específica, describir el clima en distintas latitudes o ver patrones económicos en la población de alguna ciudad.[8]

e) *Base de datos de texto:* También llamadas bases de datos documentales. Contienen descripciones escritas que normalmente son oraciones largas o párrafos. Estas bases de datos pueden contener información estructurada, semi-estructurada o no estructurada. En este tipo de bases de datos la minería es ocupada para obtener asociaciones entre los contenidos, agrupar y clasificar el texto a través de métodos lingüísticos y matemáticos.[8]

f) *Base de datos multimedia:* Son aquellas que almacenan imágenes, audio y video. Soportan objetos de gran tamaño y se utilizan en aplicaciones como la World Wide Web e interfaces basadas en comandos de voz. Para la minería de datos es importante integrar técnicas de almacenamiento y búsqueda, en el caso de la información en la Web, la tarea de la minería de datos no es sencilla debido a que muchos datos no están estructurados y es difícil determinar la información útil para extraer el conocimiento.[8]

g) *Almacenes de datos o Data Warehouse:* es una base de datos y metadatos (datos sobre los datos) de una organización que integra y depura información de diversas fuentes para posteriormente procesarla, utiliza diversas estructuras para almacenar la información (modelos de tablas en estrella, en copo de nieve, bubos relacionales, etc). [27]

Según Bill Inmon, quien formalizó el término *data warehouse,* un almacén de datos se caracteriza por ser integrado (en una estructura consistente), temático (con datos específicos y necesarios), histórico y no volátil (para ser leído, no modificado). [23]

El Data Warehouse organiza la información mediante múltiples dimensiones (colección de miembros o entidades del mismo tipo). Así mismo, los miembros de una dimensión pueden estar organizados en una o más jerarquías (conjunto de miembros en una dimensión que se definen por su posición relativa con respecto a otros miembros de la misma dimensión). [23]

Las tres principales estructuras dimensionales son[13]:

- Esquema de estrella: es un esquema que consta de una sola tabla de hechos central rodeada de tablas de otras dimensiones. En la tabla de hechos se encuentran los atributos destinados a cuantificar el hecho y en las tablas de dimensión, los atributos se destinan a elementos de nivel y su descripción. Las tablas de dimensión se encuentran desnormalizadas (toda la información referente a una dimensión se almacena en una misma tabla).

- Esquema de copo de nieve: consta de una tabla de hechos conectada a muchas tablas de dimensiones conectadas a otras tablas de dimensiones. Cada nivel puede tener varias dimensiones y cada dimensión varios niveles, así cada nivel de dimensiones representa un nivel en una jerarquía. Las tablas de dimensión están normalizadas en múltiples tablas.

- Esquema de constelación: este esquema resulta de una combinación de los dos anteriores. Su objetivo es aprovechar las ventajas de cada uno.

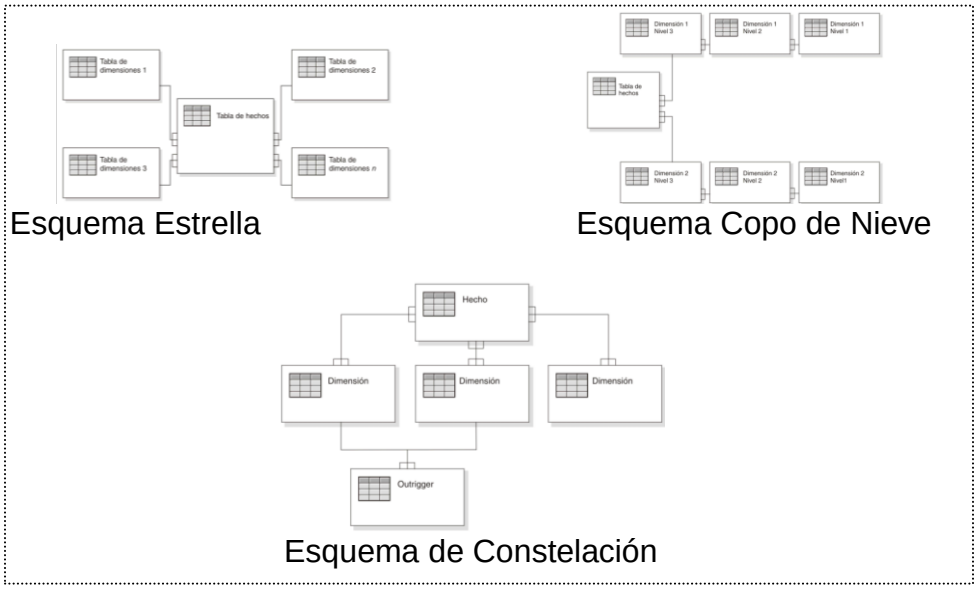


Figura 1.2 Esquemas bidimensionales de un almacen de datos [13]

1.4 Ventajas de la Minería de Datos

La Minería de Datos ha atraído la atención en la industria en los últimos años debido a que la información y conocimientos obtenidos pueden ser utilizados para aplicaciones que van desde la gestión de un negocio, el control de la producción y el análisis del mercado hasta el diseño en ingeniería y exploración científica.

Entre algunos de los beneficios y ventajas de la MD se destacan:

- Se puede extraer conocimiento enfocado a grupos de clientes con objetivos comunes gracias a que las bases de datos se exploran mediante varios modelos que comparados entre sí arrojan patrones acerca del comportamiento de los clientes.

- Los modelos son fáciles de entender por personas que no tienen gran conocimiento estadístico o matemático, de modo que al presentarse, los ejecutivos podrán aplicar sus resultados de la manera más conveniente para la organización.

- Los modelos son confiables gracias a que antes de aplicarlos son validados mediante otras técnicas estadísticas y matemáticas. Así mismo, los modelos se construyen rápidamente gracias a la aplicación de algún software especializado.

- Enormes bases de datos pueden ser analizadas sin ningún problema y descubrir patrones ocultos sin pérdida de información en el camino.

- Ayuda en el análisis de outliers que son elementos que no cumplen con el comportamiento general o el modelo de datos. La mayoría de los métodos de minería de datos descartan este tipo de datos como excepciones pero también pueden ser justo los patrones buscados.

- Describe y modela regularidades y tendencias para objetos cuyo comportamiento cambia a lo largo del tiempo

Por supuesto la minería de datos también tiene limitantes ya que depende completamente de la pureza y veracidad de las bases de datos, si los datos no son correctos el modelo creado estará perfectamente hecho pero no arrojará conclusiones válidas para la solución del problema. Otra limitación es que a pesar de descubrir patrones, conexiones y relaciones entre las variables, no necesariamente se sabe la causa de esta relación.

1.5 Tareas que se resuelven con Minería de Datos[16]

Es conveniente categorizar las tareas que se realizan en minería de datos y saber con exactitud la diferencia entre ellas para poder saber cuál aplicar dependiendo de los objetivos de la persona u organización que analizará los datos. La siguiente lista muestra algunas de las tareas más comunes:

- Descripción
- Clasificación
- Estimación
- Pronóstico
- Agrupación
- Asociación

Descripción: Es la representación o explicación detallada de las cualidades, características o circunstancias de algún objeto o persona.

Los modelos de minería de datos deben ser lo más transparente posible, es decir, los resultados del modelo deben describir patrones claros que sean susceptibles de interpretación intuitiva y explicación. Algunos métodos son más adecuados que otros para la interpretación transparente.

Podemos obtener una descripción de alta calidad por medio de un análisis exploratorio de datos o un método gráfico en busca de patrones y tendencias.

Un ejemplo de la aplicación de esta tarea puede ser etiquetar tres tipos de especies de plantas de acuerdo a características dadas, siendo el resultado final el conocimiento de la cantidad de plantas que se tienen de cada especie.

Clasificación: El proceso consiste en asignar un conjunto de datos a grupos determinados. El modelo de minería de datos examina un gran número de registros o variables que contienen a su vez información sobre el registro que se quiere clasificar (variable objetivo). Después de examinar todos los registros, el algoritmo aprenderá cuáles combinaciones de registros están asociadas a la variable objetivo, obteniendo así los datos base para realizar la clasificación de otros datos futuros.

Ejemplos de aplicaciones de algoritmos de clasificación pueden ser:

Determinar si una transacción particular de tarjeta de crédito es fraudulenta.

Evaluar si una solicitud de hipoteca es un riesgo de crédito bueno o malo.

Dividir una base de datos de un banco en grupos que sean lo más homogéneos posibles para reconocer la las personas y su capacidad crediticia como buena o mala.

Estimación: La estimación es similar a la clasificación, excepto que la *variable objetivo* es numérica más que categórica. Los modelos se construyen utilizando registros completos que proporcionan el valor de la variable objetivo y valores predictores. Entonces, las nuevas observaciones y estimaciones son hechas con base a los valores predictores.

Ejemplos de estimación en los negocios son:

Estimación de la cantidad de dinero que una familia elegida al azar gasta al ir de compras de regreso del trabajo en una estación del año.

Estimación del crecimiento de la población adulta en un país en los próximos 5 años.

Pronóstico: Según el Diccionario Ilustrado de la Lengua[1], pronóstico se define como la “acción y efecto de conocer lo futuro por ciertos indicios”.

Para la minería de datos, el pronóstico es similar a la clasificación y la estimación con la diferencia de que para la predicción, los resultados, como se ve en la definición, se encuentran en el futuro. Ejemplos:

Pronosticar el precio de una acción tres meses en el futuro.

[1] Diccionario Ilustrado de la Lengua. TEIDE. Barcelona 1986. PP. 789

Pronosticar el porcentaje de aumento de las muertes de tráfico el próximo año si el límite de velocidad aumenta.

Agrupación: Un algoritmo de agrupación (en inglés, clustering), es un procedimiento de concentración de registros, observaciones o casos en clases de objetos similares. Un cluster es un conjunto de registros que son similares entre sí y diferentes a los registros de otros grupos.

Clustering difiere de la clasificación en el hecho de que no existe una variable objetivo, en su lugar, los algoritmos de agrupación buscan segmentar los datos de todo el conjunto en subgrupos relativamente homogéneos, donde se maximiza la similitud de los registros del interior del grupo y la similitud con los registros fuera del grupo se reduce al mínimo.

Asociación: En minería de datos, es el trabajo de encontrar cuáles atributos "van juntos". En el mundo de los negocios se le conoce como análisis de afinidad o análisis de *market basket*.

La asociación trata de descubrir las reglas para cuantificar la relación ente dos o más atributos. Las reglas de asociación son de la forma "Si una condicionante, entonces una consecuencia.", junto con una medida del apoyo y la confianza asociada a la regla.

Un ejemplo de su aplicación puede ser encontrar la mejor forma de asociar comunidades autónomas de México que refleje posibles similitudes entre subconjuntos de ellas.

1.5 Estadística Vs Minería de Datos

La Estadística y la Minería de datos se refieren a dos disciplinas que están fuertemente relacionadas ya que conducen al mismo objetivo, el de efectuar "modelos" compactos y comprensibles sobre relaciones establecidas de un evento, pero también presentan fuertes contrastes y diferencias que hacen precisamente que existan áreas del conocimiento distintas.

La diferencia fundamental reside en las técnicas utilizadas para obtener estos modelos.

Entonces es importante describir brevemente algunas de las técnicas más significativas para realizar y utilizadas de cada una de las dos disciplinas, claro está que existen muchas más técnicas estadísticas que las listadas aquí. En el capítulo 2 se profundizarán a detalle las técnicas de Minería.

1.5.1 Técnicas estadísticas

Regresión lineal [2]

Es una técnica utilizada para estudiar la relación entre las variables. Puede utilizarse para explorar y cuantificar la relación entre la variable llamada dependiente y una o más variables independientes o predictivas, así como para desarrollar la ecuación lineal con fines predictivos. [2]

La fórmula general de regresión es:

$$y = a_0x_{0+}a_1x_{1+}a_2x_{2+}\ldots+a_nx_n + e$$

Donde y indica la variable dependiente, x_i indica las variables independientes, a_i son los coeficientes o parámetros del modelo y e suponen errores aleatorios con distribución normal, media cero y varianza σ^2.

El modelo de regresión lineal sería la línea que minimiza la tasa de error entre el valor real y el valor predicho. Se trata de encontrar los parámetros mediante diversos métodos como la técnica de mínimos cuadrados por ejemplo para posteriormente obtener una predicción.

Ajuste de curva univariable

El Ajuste de curva univariable obtiene una función que describe exactamente la distribución de los datos a través del tiempo. Entre los tipos de curva que pueden seleccionar se encuentran la exponencial, Hiperbólica, Lineal, Potencia, Poisson, Normal, el Logaritmo.[18]

Análisis de componentes principales[18]

Es una técnica de reducción de información o de la dimensión de las características (número de variables). El objetivo será reducir las variables a un menor número de ellas pero perdiendo la menor cantidad de información posible.

Los componentes principales serán una combinación lineal de las variables originales y serán independientes entre sí.

Existen 2 tipos de componentes principales:

Normalizado que se basa en la correlación de los datos de entrada. La correlación puede estar entre 1 y -1.

Centrado que se basa en la covarianza de los datos de entrada, la cual mide la tendencia de dos variables a cambiar juntas.

Análisis Factorial[19]

Es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto de muchas variables.

Cada grupo se forma con las variables que correlacionan mucho entre sí y procurando que los grupos sean independientes entre sí.

Su propósito consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos.

Consta de cuatro fases características: el cálculo de una matriz capaz de expresar la variabilidad conjunta de todas las variables, la extracción del número óptimo de factores, la rotación de la solución para facilitar su interpretación y la estimación de las puntuaciones de las variables en las nuevas dimensiones.

1.5.2 Técnicas de Minería de Datos

La minería de datos ha dado lugar a una sustitución progresiva del análisis de los datos dirigido a la verificación y descubrimiento, mediante un enfoque de análisis, de conocimiento nuevo y oculto a simple vista.

La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, lo cual es más eficiente que un análisis estadístico clásico. Estos algoritmos o técnicas se encuentran en continua evolución como resultado de la colaboración entre campos de investigación como bases de datos, inteligencia artificial, sistemas expertos, estadística, recuperación de la información y visualización.

Las técnicas de minería de datos se clasifican en dos categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento[11].

Supervisados o predictivos: encuentran el valor de un atributo (etiqueta) de un conjunto de datos en un tiempo futuro a partir de otros atributos llamados descriptivos, induciendo una relación entre los atributos que sirven para realizar la predicción. A esta forma de trabajo se le conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento y prueba.

En el entrenamiento se hace una construcción del modelo usando un subconjunto de datos con etiqueta conocida; En la prueba se aplica el modelo sobre el resto de los datos.

No supervisado o de descubrimiento del conocimiento: descubren patrones y tendencias en los datos sin utilizar datos históricos. Sirve para llevar a cabo acciones y obtener un beneficio científico o de negocio a partir de los patrones descubiertos.

La siguiente tabla muestra las técnicas más utilizadas en minería de datos, algunas de ellas también utilizadas en estadística, éstas técnicas se analizarán a fondo en el capítulo 2.

SUPERVISADO	NO SUPERVISADO
Árboles de decisión	Detección de desviaciones
Redes neuronales	Segmentación
Regresión	Agrupamiento (Clustering)
Series de tiempo	Técnicas de Asociación
	Patrones Secuenciales

Tabla 1.1 Clasificación de las técnicas de Minería de datos

1.5.3 Comparación entre Estadística y Minería de Datos.

La esencia de la minería de datos se encuentra en la posibilidad de descubrimiento de información oculta y sumamente valiosa. Esto significa que su naturaleza es exploratoria entendiéndose con esto que en un principio no se tienen hipótesis claras, mientras que la naturaleza de muchas áreas de la estadística es confirmar alguna hipótesis.

La MD aventaja a la Estadística en los siguientes supuestos:

- Cuando el objetivo es meramente exploratorio para concretar un problema o definir cuáles son las variables más interesantes de un sistema de información es necesario utilizar MD ya que no se parte de ningún supuesto y pretendemos buscar algún conocimiento nuevo.

- A mayor dimensionalidad del problema la minería de datos ofrece mejores soluciones.

- Las técnicas de minería de datos son menos restrictivas que las estadísticas y permiten ser utilizadas con los mínimos supuestos posibles.

Cuando los datos de la organización son muy dinámicos, las técnicas de MD inciden sobre la inversión y actualización del conocimiento del negocio. Un almacén de datos poco dinámico permite una inversión en un análisis estadístico que da respuestas a preguntas muy concretas.

Existen otros contextos en los que es más adecuando el análisis estadístico que las técnicas de minería de datos como son:

- El objetivo de la investigación es encontrar causalidad. Las relaciones complejas que se encuentran en la MD muchas veces impiden una interpretación certera de causa-efecto.

- Cuando se pretende generalizar sobre poblaciones desconocidas, es decir si las conclusiones son aplicadas a otros elementos de población similares, convendría más utilizar técnicas de inferencia estadística.

A manera de resumen se pueden ver en la Tabla 1.2 las características de la Estadística y la MD de forma comparativa.

Se debe destacar que pese a las diferencias que existen entre Minería y Estadística, ambas disciplinas constituyen una sinergia excelente para la toma de decisiones y que no son excluyentes la una de la otra. Por ejemplo, dentro de la metodología de un proyecto de minería de datos debemos incluir un análisis estadístico para la preparación de los datos y entendimiento del negocio.

En conclusión la Minería de Datos y la Estadística, son técnicas complementarias que permiten obtener conocimiento inédito de nuestros almacenes de datos o dar respuestas a cuestiones concretas de un negocio.

ESTADÍSTICA	MINERÍA DE DATOS
El objetivo es obtener un modelo estadístico que ayude al entendimiento del negocio y proyección de ciertos patrones y características.	El objetivo es disponer de un modelo que describa y realice predicciones sobre un negocio
Trabaja sobre muestras definidas y acotadas.	Trabaja sobre grandes bases de datos
Para aplicar sus métodos, los datos deben estar depurados y sin anomalías. Solucionando estos problemas con otras técnicas especializadas.	Por si solas estas técnicas contendrían multitud de valores no informados e inconsistentes. Varios de sus métodos no requieren que se realice una limpieza previa de los datos.
Las variables siguen distribuciones conocidas.	Las variables pueden provenir de distribuciones desconocidas.
Se pueden hacer manipulaciones en los datos de manera directa	No se pueden hacer manipulaciones en los datos de manera directa

Tabla 1.2 Diferencias entre Minería de Datos y Estadística

1.6 Metodologías de la Minería de Datos

En cualquier organización, para implementar una tecnología determinada, se requiere de una metodología apropiada según los proyectos que se aborden.

La sistematización del proceso de minería de datos es un punto importante para la planificación y ejecución de un proyecto. Esta sistematización está dada por dos metodologías específicas para la minería de datos principalmente: CRISP-DM y SEMMA (específica para software SAS), pero muchas organizaciones implementan el proceso de KDD (Knowledge Discovery in Databases) y en la actualidad surgió la metodología CATAYST o P3TQ que va ganando popularidad debido a su completitud y flexibilidad para adaptarse a distintos escenarios.[11]

1.6.1 Metodología CRISP-DM[4][16][11][16][19]

Es una metodología creada en Europa por un grupo de empresas SPSS, NCR y Daimler Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en proyectos de MD gracias a la versatilidad de su proceso.[16]

Trata de desarrollar los proyectos de minería de datos bajo un proceso estandarizado de definición y validación de tal forma que se desarrollen proyectos minimizando los costos y obteniendo resultados de alto impacto en el negocio.

El estándar incluye un modelo y una guía estructurados en seis fases, algunas de estas fases son bidireccionales, lo que significa que alguna de las fases permitirá revisar parcial o totalmente las fases anteriores.

Podemos observar de manera esquemática las seis fases de esta metodología en la siguiente figura.

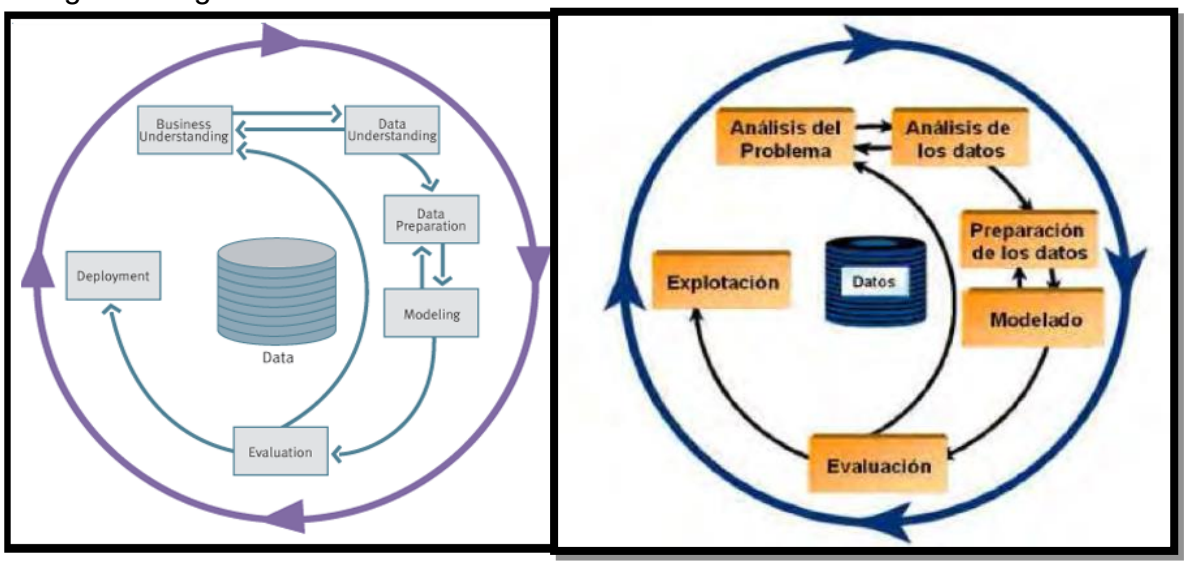


Figura 1.3 Fases de la metodología CRISP-DFM [4][24]

La primera fase es **entendimiento del negocio** o business understanding. Debe enunciar los objetivos del proyecto y los requisitos del negocio, traducir estos objetivos y restricciones en la formulación de un problema de minería de datos definido y preparar una estrategia preliminar para el logro de los objetivos. Esta fase es una de las más importantes ya que si no se comprende bien la meta a cumplir, entonces se podrían realizar todas las demás fases de manera correcta pero llegando a resultados que no servirían para el problema planteado.[4][16][11][16][19]

La segunda fase **es análisis de los datos** o data understanding. Comprende la recopilación de los datos necesarios para el logro de los objetivos planteados; la realización de un análisis exploratorio de los datos para familiarizarse con ellos y descubrir ideas iniciales; la evaluación de la calidad de los datos y la selección opcional de subconjuntos interesantes de datos que pueden contener patrones interesantes y procesables.

La tercera fase es **preparación de los datos** o data preparation. En esta fase, como su nombre lo dice se preparan, a partir de los datos brutos iniciales, el conjunto final de datos que se va a utilizar para todas las fases subsecuentes. Se deben seleccionar las variables que se desean analizar y ver que sean apropiadas para el análisis, de lo contrario se deben realizar transformaciones convenientes en esas variables. [22]

Se necesita limpiar los datos sin procesar, corregir valores “missing (faltantes”, “outliers (atípicos)” y los datos mal capturados mediante técnicas estadísticas especiales para cada caso por lo que esta fase suele ser muy laboriosa.

La cuarta fase es **modelado** o modeling. Aquí se selecciona y aplica la técnica más adecuada para el proyecto de minería según el objetivo; se debe calibrar la configuración del modelo para optimizar los resultados. Una descripción de las tareas de esta fase es la siguiente:

Selección de la técnica de modelado. Para esta selección se debe considerar el objetivo principal del problema a resolver y la relación con las herramientas de MD existentes.

Generación del plan de prueba. Después de seleccionada la técnica, se ejecuta, sobre los datos previamente preparados, para generar uno o más modelos. La selección de los mejores parámetros en cada técnica es un proceso iterativo y se basa en los resultados generados que serán interpretados posteriormente.

Evaluación del modelo. Las personas encargadas del proyecto de MD interpretan los modelos de acuerdo a los criterios de éxito preestablecidos (seguridad del conjunto de prueba, total de pérdida permitida, cantidad de ganancia, etc.…).

La quinta fase es **evaluación de resultados** o evaluation. Se debe realizar una valoración de la calidad y efectividad del modelo desde el punto de vista del cumplimiento de los objetivos establecidos en la primera fase.

Si es un modelo bueno entonces se debe llegar a una decisión sobre el uso de los resultados de minería de datos; si no es un modelo bueno entonces se debe regresar a cualquiera de las fases para volver a elaborar uno nuevo.

La última fase es **explotación** o deployment. En esta fase y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones basadas en la observación del modelo y los resultados.

En la fase de explotación se debe asegurar el mantenimiento de la aplicación y la difusión de los resultados.

Se realiza un informe final el cual contiene la conclusión del proyecto de minería de datos realizado a través de un resumen de los resultados, las relaciones y patrones encontrados, para que éstos sean utilizados en elaboración de estrategias y toma de decisiones de la organización.

1.6.2 Metodología SEMMA[11][22]

Esa metodología fue desarrollada por SAS Institute, quien la define como “el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocios desconocidos” [12]. Se caracteriza por priorizar sus fases desde un punto de vista técnico, es decir, dando prioridad a las prácticas usadas para la implementación y obtención de resultados.

Cada una de las iniciales de esta metodología hace referencia a una de las fases de un proyecto de minería de datos.

En la siguiente figura se muestran las fases de la metodología SEMMA:

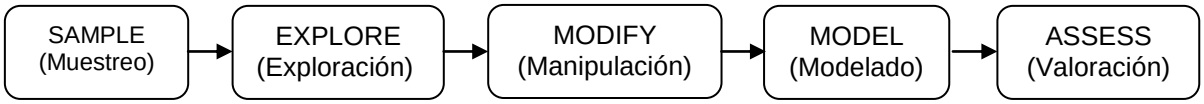


Figura 1.4 Fases de la metodología SEMMA

Fase de **muestreo** o sample**:** Se preparan los datos para su posterior exploración, seleccionando una muestra representativa del problema en estudio. Dicha muestra es elegida mediante un proceso de muestreo aleatorio simple y se le asocia un nivel de confianza.

Fase de **exploración** o explore: Se trata de analizar los datos con el fin de simplificar el problema y optimizar la eficiencia del modelo. Normalmente esta fase se ayuda de un análisis exploratorio estadístico que para evidenciar posibles relaciones entre variables y establecer las variables explicativas que se pondrán como entradas del modelo.

Fase de **manipulación de los datos** o modify: Cuando se llega a esta parte, las actividades se centran en la selección y transformación de variables y datos que servirán para la construcción de los modelos con la finalidad de tener una base de datos correcta y útil para el cumplimiento de los objetivos. Aquí se realiza la reducción de la dimensión, imputación de valores “missing”, “outliers”, etc.

Fase de **modelado de los datos** o model: La selección el modelo va a depender esencialmente de los datos y tipos de variables que se tengan. Se sugiere aplicar más de uno a la vez y luego comparar los resultados obtenidos.

Fase **valoración de los resultados** o assess: Consiste en la evaluación de los resultados mediante un análisis de bondad o un análisis de diagrama ROC que enfrenta dos variables: sensibilidad y especificidad y busca que ambas categorías sean altas, u otras herramientas estadísticas clásicas.

Como se puedo observar ambas metodologías comparten la misma esencia, estructurar el proyecto de minería de datos en fases que se encuentran altamente relacionadas entre sí. Por esta razón es fácil adaptar cualquiera de las dos metodologías a cualquier proyecto de minería de datos.

1.7 Aplicaciones de la minería de datos

Las aplicaciones de las diversas técnicas de la MD tienen un extenso campo de acción y principalmente se pueden aplicar dentro de los siguientes aspectos: existencia de sistemas parcialmente desconocidos, enormes cantidades de datos y presencia de un potente hardware y software con capacidad de realizar tareas de minería de datos.

Entre las áreas de aplicación de la minería de datos se encuentran la Astronomía (clasificar cuerpos celestes), Medicina (asociación de síntomas), Biología Molecular (predicción de enfermedades y causas), Climatología (predicción de desastres), Mercadotecnia (selección de clientes y productos), Economía (detección de fraudes, inversiones, seguros), Internet (comportamiento de redes sociales), etc.

En la siguiente figura se puede ver de manera esquemática varios campos de acción de la minería de datos.

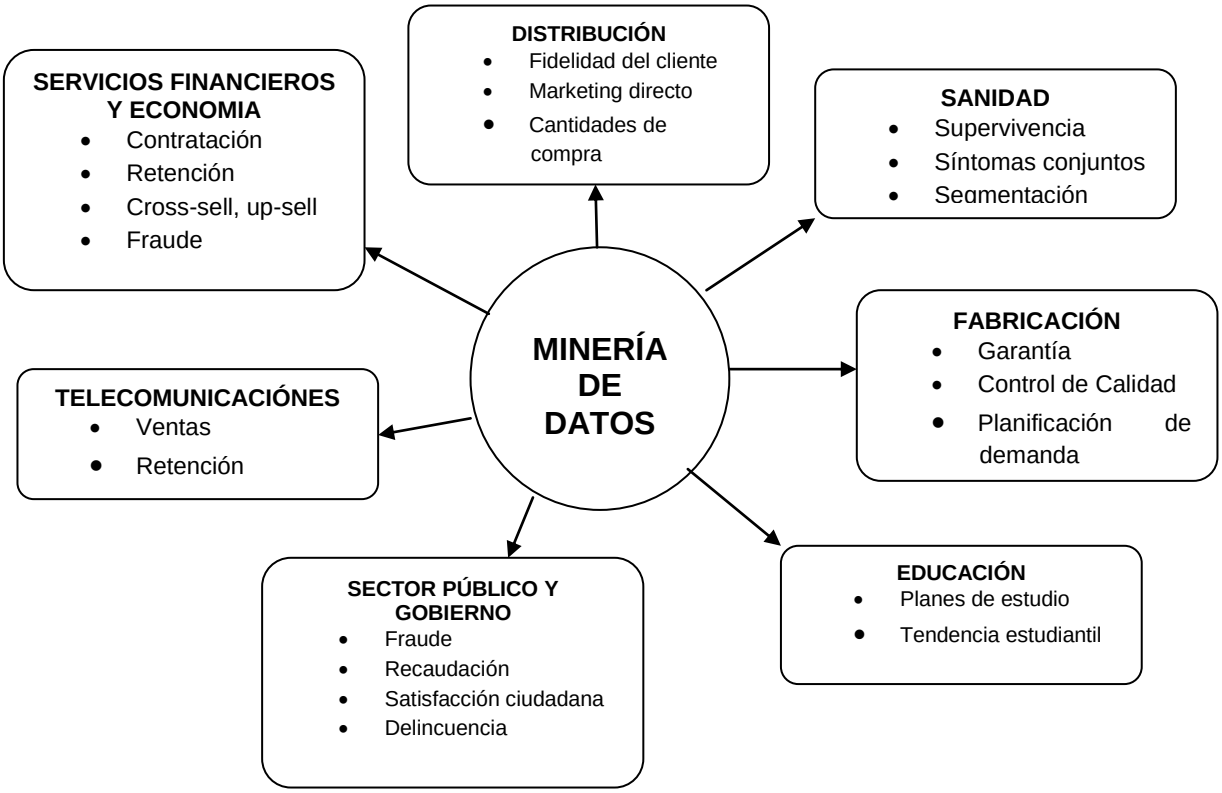


Figura 1.5 Áreas de aplicación de la Minería de Datos.

Actualmente se han implementado acciones muy exitosas en empresas y organizaciones utilizando minería de datos, entre las más populares podemos destacar[11][28]:

- A principios del mes de julio de 2002, el director del FBI en Estados Unidos, anunció que el Departamento de Justicia implementaría técnicas de minería de datos para analizar las bases de datos comerciales con la finalidad de detectar terroristas.

- En 2001, las instituciones financieras a nivel mundial perdieron alrededor de 2.000 millones de dólares en fraudes cometidos con tarjetas de crédito por lo que implementaron el *Falcon Fraud Manager (*sistema inteligente de minería de datos que examina transacciones, propietarios de tarjeta y datos financieros para intentar detectar fraudes y tratar de disminuirlos).
  El sistema Falcon ha permitido ahorrar más de seiscientos millones de dólares al año y monitorea alrededor de 450 millones de cuentas distribuidas en 6 continentes.

- La BBC (British Broadcasting Corporation) del Reino Unido emplea un sistema para predecir el tamaño de las audiencias televisivas de un programa determinado, así como la hora óptima de emisión.

- Hace algunos años una de las sucursales de Wal-mart se hizo la pregunta sobre qué productos se venderían con mayor frecuencia junto con los pañales; realizaron un estudio de minería de datos y encontraron que en asociación con los pañales se vendían muy frecuentemente cervezas. Además notaron que se vendían principalmente los viernes en la tarde y que eran comprados por hombres con edades entre 25 y 35 años.

  Como consecuencia de esto el supermercado colocó la cerveza junto con los pañales, logrando que el patrón de comportamiento observado aumentara y con esto las ganancias de esa sucursal.

- En el Instituto Tecnológico de Chihuahua, México, se realizó un estudio sobre los recién titulados de la carrera de Ingeniería en Sistemas Computacionales. Se quería observar si los recién titulados comenzaban a trabajar en actividades profesionales relacionadas con sus estudios.

  Mediante la aplicación de minería de datos se descubrió que existían cuatro variables que determinaban la adecuada inserción laboral: zona económica, colegio de procedencia, nota al ingresar y promedio final.

  A partir de estos resultados la universidad podía plantearse nuevas soluciones de tipo socioeconómico como por ejemplo el otorgamiento de becas, ya que la mayoría de las variables determinantes no dependían de la universidad.

- El club AC. de Milán utiliza un sistema inteligente para prevenir lesiones y optimizar el acondicionamiento de cada atleta. Con ello el club intenta ahorrar dinero evitando comprar jugadores que presenten una alta probabilidad de lesión.

Como ya se dijo, el campo de acción de la minería de datos es muy extenso y variado pero no siempre puede ser aplicado con éxito, esto se debe a que no siempre se cuenta con los suficientes datos que permitan obtener información de interés.

Las herramientas comerciales de Minería de Datos que existen actualmente en el mercado son muy variadas y excelentes en diversas aplicaciones. Su correcta elección depende de la necesidad de la organización y de sus objetivos a corto y largo plazo.

Para implementar una solución de minería de datos es necesario consultar expertos en el área para poder seleccionar la técnica más adecuada al problema de la empresa y no tener un desgaste económico en implementaciones erróneas
En resumen, la Minería de Datos surge como una tecnología emergente con varias ventajas para los investigadores y personas de negocios ya que ahorra grandes cantidades de dinero y de tiempo. No cabe duda que trabajar con esta tecnología es cada vez más importante para la correcta toma de decisiones.

CAPITULO 2. MÉTODOS DE MINERÍA DE DATOS

2.1 Introducción

Las técnicas de minería de datos aparecen como una especie de colador de bases de datos que produce una serie de patrones claros e interpretables para el minero.

Se tiene que el mismo método o técnica puede resolver una gran cantidad de tareas, esto se debe a que la mayoría de las tareas que se vieron en el Capítulo 1 son de aprendizaje inductivo.[2]

En la Sección 1.5.2 se mencionó la principal clasificación que existe entre las técnicas de minería de datos (supervisado y no supervisado), pero no es la única clasificación existente ya que hay muchas otras que dividen las tareas de acuerdo al tipo de aprendizaje, la rama matemática de la que provienen o inclusive haciendo una mezcla de tareas con técnicas.

Se mencionará a continuación otra de las clasificaciones existentes[11]:

Técnicas algebraicas y estadísticas: Se basan en expresar modelos y patrones mediante fórmulas algebraicas, funciones lineales y no lineales, valores estadísticos (medias, varianzas y correlaciones), etc.

Los patrones obtenidos mediante éstas técnicas se obtienen a partir de un modelo ya predeterminado del cual, se estiman unos coeficientes o parámetros. Algunos de los algoritmos más conocidos dentro de este grupo se encuentran la regresión lineal, regresión logarítmica, regresión logística, análisis de discriminante y técnicas de estadística no paramétrica.

Técnicas Bayesianas: Se basan en estimar la probabilidad de pertenencia a un grupo o clase mediante la estimación de probabilidades condicionales inversas, utilizando el teorema de Bayes, generalizando topologías entre las interacciones probabilísticas de las variables y representando dichas interacciones gráficamente. Los algoritmos más populares son: Clasificador bayesiano naive, máxima verosimilitud y el algoritmo EM.

---

[2] " El aprendizaje inductivo es una de las formas de aprendizaje, en la que el aprendiente realiza un proceso que parte de la observación y el análisis de una característica de la lengua, hasta la formulación de una regla que explique dicha característica**.**"- Castañeda, A. (1997). *Aspectos cognitivos en el aprendizaje de una lengua extranjera*. Granada: Lingüística y Método.

Técnicas relacionales, declarativas y estructurales: Representan los modelos mediante lenguajes declarativos, lógicos, funcionales y lógico-funcionales. Aquí se encuentran las llamadas técnicas de minería de datos relacionales como las ILP (programación lógica inductiva).

Técnicas estocásticas y difusas: Son técnicas en las que se utilizan funciones de preferencia difusas (Fuzzy) o componentes fundamentales (métodos evolutivos y genéticos). Forman lo que se denomina computación flexible.

Técnicas basadas en caos, densidad o distancia: Se basan en la distancia de un elemento de la base de datos al resto de elementos. Entre los algoritmos de esta categoría se encuentran: vecinos más próximos, estimación de funciones de densidad, Two-Step o COBWEB y K-medias.

Técnicas basadas en árboles de decisión: Son quizá el método más fácil de utilizar y entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica de tal manera que al final se puede detectar algún tipo de condición o característica entre los elementos de un nodo.

Técnicas basadas en redes neuronales artificiales: Son técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. Existen innumerables variantes como el preceptrón simple, redes multicapa, redes de base radial, redes de Kohonen, etc.

Ambas formas de clasificación de las técnicas de minería se podrían resumir en la tabla 2.1 que, además de una clasificación, contiene una comparación entre las tareas que se pueden desarrollar con cada técnica.

Un aspecto fundamental a considerar en el aprendizaje de minería es el esfuerzo computacional que se requiere. Entre dos métodos se preferirá el que obtenga patrones de una manera más rápida. La eficiencia de los métodos depende generalmente del número de atributos, complejidad del espacio de hipótesis que se está considerando, del conocimiento previo existente, del nivel de error permitido y de lo innovador que se deseen los patrones.

Para solucionar este problema se ha diseñado el algoritmo *anytime* (utilizado dentro de otros algoritmos) que produce mejores soluciones entre más tiempo se otorgue al algoritmo, puede ser interrumpido en cualquier momento y mostrar la mejor solución obtenida hasta el momento, esto facilita mucho el aprendizaje en minería de datos.

Ahora bien, existe una característica diferenciadora fundamental entre todos los métodos de aprendizaje; la manera en que se expresan los patrones aprendidos, que es la razón por la que unos métodos van mejor para unos problemas que para otros.

NOMBRE	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones y Factorizaciones
Redes neuronales	*	*	*		
Árboles de decisión ID3, C4.5, C5.0	*				
Árboles de decisión CART	*	*			
Otros árboles de decisión	*	*	*	*	
Redes de Kohonen			*		
Regresión lineal y logarítmica		*			*
Regresión logística	*			*	
K-Means			*		
Naive Bayes	*				
Vecinos más próximos	*	*	*		
Análisis factorial y de Componentes Principales					*
Two step, Cobweb			*		
Algoritmos genéticos y evolutivos	*	*	*	*	*
Análisis discriminante multivariante	*				

Tabla 2.1 Técnicas de minería de datos y tareas que se realizan con ellas.

Los patrones que no dependen de la proximidad, similitud espacial o geométrica en individuos no pueden ser analizados por los métodos de “cerca y rellena”, para ellos existen otros conceptos llamados “relacionales” que hacen referencia a las relaciones que se pueden establecer entre atributos.

Cuando el modelo no es capaz de ajustarse a los datos de ninguna manera se obtiene un efecto de **subajuste** (underfitting), caso contrario es el **sobreajuste** (overfitting). Si la resolución es buena (y la densidad de datos es alta) podemos tener una buena idea del patrón, pero si hay pocos datos, la idea del patrón es muy difusa. [16]

La mayoría de los métodos que son capaces de obtener separaciones, lo hacen componiendo trozos, es decir ajustándose localmente pero no globalmente.

Existen tres medidas que permiten determinar el grado de continuidad o separabilidad de un problema considerando una medida de distancia [11].

Separabilidad lineal: Un objeto es separable linealmente si existe una función lineal (una recta, un plano o un hiperplano) que separa las clases nítidamente.[11]

Separabilidad geométrica: Ajusta patrones que aunque no sean lineales si pueden tener un gran componente geométrico. Se define de la siguiente manera:

$$SG(f) = \left(\frac{\sum_{i=1}^{n} eq(f(ei), f(m(ei)))}{n} \right)$$

Donde *f* es la función definida por los datos, *n* es el número de ejemplos, *m* es el vecino más cercano y $eq(a, b)=1$ si a=b y 0 en otro caso.[11]

Separabilidad geométrica radial[11]: El objetivo es establecer un radio adecuado, estimado a partir de la densidad de los datos. Se define como:

$$SGR(f) = \frac{\sum_{i=1}^{n} \frac{\sum_{ej:dist(ei,ej)<r} eq(f(ei), f(ej))}{|ej:dist(ei,ej)<r|}}{n}$$

Estas medidas pueden utilizarse antes de abordar un problema de clasificación para decidir qué método servirá mejor. Sin embargo, si la separabilidad es baja se debe intentar con métodos no basados en distancias que ayudarán a aumentar la expresividad de los datos, dichos métodos pueden ser: Aumento en la dimensionalidad y/o creación de atributos, localidad (para espacios de hipótesis iguales) o combinación de modelos y métodos.

En minería de datos se definen métricas de distancias para medir la similitud. La función de distancia más común es, la distancia Euclidiana ($d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$.

Cuando se mide la distancia, ciertos atributos que tienen valores grandes, pueden abrumar la influencia de otros atributos que se miden en una escala más pequeña. Para evitar esto, se debe asegurar la normalización de los valores de los atributos. Para las variables continuas, la normalización min-max (diferencia entre la observación y el mínimo valor, escalado por su rango $x^* = \frac{X - \min(muestra)}{\max(muestra) - \min(muestra)}$) o normalización Z-score (tomando la diferencia entre el valor observado y el valor medio, dividido por la desviación estándar $x^* = \frac{x - medi(muestra)}{sd(muestra)}$). Para las variables categóricas, la métrica de distancia euclidiana no es la adecuada. En su lugar, se puede definir una función de indicadoras (dummys), utilizada para comparar los valores del atributo i-ésimo de

un par de registros de la siguiente forma: $dif(x_i, y_i) = \begin{cases} 0 & si\ x_i = y_i \\ 1 & c.o.c \end{cases}$. Entonces, se sustituye $dif(x_i, y_i)$ en la métrica Euclidiana.

Si bien se han mencionado ya algunas técnicas utilizadas para la detección de patrones en la minería de datos, aún no se ha profundizado en cada una de ellas, por tanto en las siguientes secciones de este capítulo se abordarán las principales metodologías, sus características, importancia, utilidad y desventajas.

2.2 Reglas de asociación

Las reglas de asociación consisten en encontrar patrones interesantes en forma de relaciones de implicación entre los valores de los atributos (ítems) y los objetos de un conjunto de datos. Estas reglas son utilizadas para analizar atributos **nominales**, pueden servir para conocer el comportamiento general de los datos y asistir en la toma de decisiones. [16]

Una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos ítems de una base de datos. Puede ser vista como reglas de la forma **SI α ENTONCES β** donde α y β son dos conjuntos de ítems disjuntos que forman una transacción. Se suele trabajar con dos medidas para conocer la calidad de la regla: **cobertura** (soporte o support) y **confianza** (confidence). La cobertura es el porcentaje de instancias que la regla predice correctamente del total de transacciones. La confianza mide el porcentaje de veces que la regla se cumple dado que ya se cumplió el evento **α**. [11]

En minería de datos, las reglas de asociación se restringieron inicialmente a colecciones de datos binarios. Posteriormente estas técnicas se desarrollaron también para colecciones de datos mezclados. A continuación se presentan las técnicas para ambos tipos de colecciones.

Reglas de Asociación en colecciones de datos binarios[23]

El objetivo de la asociación en colecciones binarias consiste en encontrar todas las reglas interesantes a partir de un conjunto de transacciones.

Este tipo de reglas formalmente se conceptualiza de la siguiente manera: sea $I = \{i_1, \ldots, i_m\}$ un conjunto de ítems, $T = \{t_1, \ldots, t_n\}$ un conjunto de transacciones tal que $t_i \subseteq I$.

Una regla de asociación es una implicación de la forma $X \Rightarrow Y$, donde $X \subseteq I$, $Y \subseteq I$ y $X \cap Y = \varnothing$.

Se considera que una regla es interesante si su soporte es mayor que el umbral mínimo de soporte (min Supp) y su confianza es mayor que un umbral de mínima confianza (min Conf), ambas medidas basadas en el soporte de un conjunto de ítems. Es decir X U Y debe ser un conjunto frecuente de ítems.

El soporte de un conjunto de ítems X en un conjunto de transacciones (atributos), t es la fracción de atributos de T que contienen los ítems e X.

$$Supp\,(\mathrm{X}) = \frac{|\{t \in T \mid X \subseteq t\}|}{|T|}$$

Si el soporte de X es mayor o igual que el mini Supp, se dice que el conjunto de ítems X es frecuente.

El soporte de una regla $\mathrm{X} \Rightarrow \mathrm{Y}$ en un conjunto de transacciones.

$$Supp\ (\mathrm{X} \Rightarrow \mathrm{Y}) = Supp\,(\mathrm{X\ U\ Y})$$

La confianza de una regla es la fracción de atributos de T que conteniendo a X, también contienen a Y.

$$Conf\ (\mathrm{X} \Rightarrow \mathrm{Y}) = \frac{supp\ (X\ U\ Y)}{\mathrm{supp\ (X)}}$$

Al realizar el minado de datos se requiere hacer dos pasos fundamentales. Primero se buscan los conjuntos frecuentes de ítems y a partir de ellos, son generadas las reglas de asociación interesantes.

Conjuntos frecuentes: El espacio de los posibles conjuntos frecuentes de ítems depende exponencialmente del tamaño del conjunto de ítems $\mathrm{I}: (\,|\,\{\mathrm{X}|\mathrm{X} \subseteq \mathrm{I}, \mathrm{X} \neq \varnothing\}| = 2|\mathrm{I}| - 1)$. Sin embargo la propiedad de Clausura Descendente del soporte permite que no se recorra todo el espacio de búsqueda para encontrar los conjuntos frecuentes de ítems. Esta propiedad dice que todo subconjunto de un conjunto frecuente de ítems es frecuente, mientras que todo superconjunto de un conjunto no frecuentes de ítems tampoco es frecuente.

Debido a esta propiedad se puede crear una red de ítems I= { i_1,...i_m} que es dividida por una frontera en dos subespacios, uno solo contiene conjuntos frecuentes de ítems y el otro solo contiene conjuntos no frecuentes de ítems y recursivamente se generan los conjuntos de ítems por cada rama de la estructura.

La siguiente figura muestra un ejemplo del tipo de red que se pueden crear.

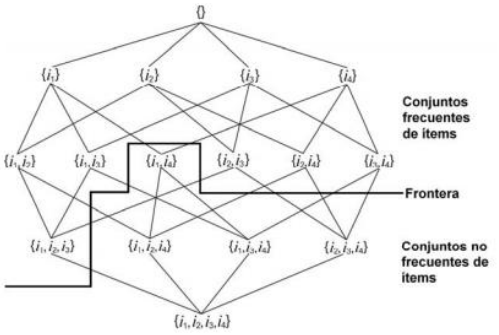


Figura 2.1 Red formada por el conjunto de ítems I = {i_1, i_2, i_3, i_4} [23]

El algoritmo más común y utilizado que ocupa la propiedad *Clausura Descendente del Soporte* se conoce como **Apriori** (ver la Algoritmo 2.1)[11].

En este algoritmo, L_k contiene los conjuntos frecuentes de ítems de tamaño k y C_k los conjuntos de ítems candidatos a frecuentes de tamaño k. Como primer paso el algoritmo almacena en L_1 los conjuntos frecuentes de ítems de tamaño 1, posteriormente en cada iteración se generan los conjuntos de ítems de tamaño k, candidatos a frecuentes, combinando los conjuntos frecuentes de ítems de tamaño k-1. Este proceso se realiza hasta que al comienzo de una iteración del conjunto de conjuntos frecuentes de ítems, el tamaño de k-1 sea vacío. Para generar los conjuntos de ítems candidatos a frecuentes de utilizan las operaciones de *Join* y *Prune*.

Join: consiste en tomar los pares de conjuntos de ítems de tamaño k-1 que coincidan en sus k-2 primeros ítems y generar conjuntos de ítems de tamaño k manteniendo los k-2 ítems comunes y adicionando los (k-1)-ésimos ítems de los conjuntos que se unen. Su fórmula es:

$$Join\ (\{i_1,\ldots, i_k\}, L_{k-1}) \equiv (\{i_1,\ldots, i_{k-2}, i_k\} \in L_{k-1})$$

Prune: consiste en aplicar la propiedad de Clausura Descendente del soporte para podar los conjuntos de ítems de tamaño k que tengan al menos un subconjunto no frecuente de ítems de tamaño k-1. Su fórmula es:

$$Prune(c, L_{k-1}) \equiv (\forall s[s \subset c \wedge |s| = k-1 \rightarrow s \in L_{k-1}])$$

Existen también otros algoritmos como el ECLAT, FP-Growth, Patricia Trie-Mine, CT-ITL, CT-PRO y Apriori-TFP que sirven para encontrar los conjuntos frecuentes pero no se abordarán en éste trabajo.

Input D: Set of transactions
$minSupp$: minimum support threshold
Output F: Frequent itemsets
$L_1 \leftarrow \{i | |\{t \in D | i \in t\}| \geq minSupp\}$
for $(k = 2; L_{k-1} \neq \emptyset; k++)$ **do**
$C_k \leftarrow \{c | Join(c, L_{k-1}) \wedge Prune(c, L_{k-1})\}$
forall $transaction \quad t \in D$ **do**
$C_t \leftarrow \{c \in C_k | c \in t\}$
forall $candidates \quad c \in C_t$ **do**
$c.Support \leftarrow c.Support + 1$
$L_k \leftarrow \{c \in c_k | c.Support \geq minSupp\}$
$F \leftarrow \cup_k L_k$

Algoritmo 2.1 Algoritmo Apriori. [11]

Reglas de asociación: El conjunto de reglas representativas (RR) se define como el conjunto de reglas de asociación interesantes (con soporte y confianza mayores a los umbrales de mínimo soporte y mínima confianza) que no están cubiertas por otras reglas de asociación interesantes (AR), cumpliéndose lo siguiente:

$$RR = \{r \in AR | \nexists r' \in AR, r' \neq r \wedge r \in C(r')\}$$

$$Si\ r' \in C(r) entonces\ supp(r') \geq supp(r)\ y\ conf(r') \geq conf(r)$$

$$Si\ r \in AR\ y\ r' \in C(r) entonces\ r' \in AR$$

El método más común de generación de reglas de asociación interesantes es el llamado ***Gen Rules*** (ver algoritmo 2.2) que consiste en, por cada conjunto frecuente de ítems generar todas las reglas posibles separando el conjunto de ítems en dos subconjuntos disjuntos.

El número de reglas generadas para un conjunto de transacciones y umbrales de mínimo soporte y confianza de datos, puede ser muy grande por lo que se crea solo un subconjunto de reglas interesantes a partir del cual pueden ser generadas el resto de las reglas interesantes mediante el método de cubrimiento de una regla que se define como: $C(X \rightarrow Y) = \{X \cup Z \rightarrow V | Z, V \subseteq Y \wedge Z \cap V = \emptyset \wedge V \neq \emptyset\}$.

Input F: Set of frequent itemsets
Output RA: Set of association rules
$L_1 \leftarrow \{i | |\{t \in D | i \in t\}| \geq minSupp\}$
forall $itemset \quad Z \in F$ **do**
forall $itemset \quad X \subset Z \quad and \quad X \neq \emptyset$ **do**
if $(Z.support / X.support) \geq minConf$
$RA = RA \cup \{X \Rightarrow (Z/X)\}$

Algoritmo 2.2 Algoritmo Gen Rules [11].

Reglas de Asociación Multinivel[11]

Cuando existe una gran cantidad de atributos con respecto al número de ítems presentes en cada registro se agrupan los atributos en categorías y es más fácil encontrar reglas con niveles adecuados de confianza y cobertura.

Se debe proporcionar una jerarquía de conceptos que contiene un árbol de relaciones entre los atributos. Esta jerarquía define una secuencia de relaciones entre conceptos que va de los específicos a los generales.

El nivel de agrupamiento debe ser flexible, de manera que permita a un usuario seleccionar el grado de asociación que considere adecuado. Se debe ser cuidadoso ya que un nivel excesivo de abstracción puede conducir a que se aprendan reglas poco informativas.

El algoritmo de reglas de asociación multinivel es muy parecido al de datos binarios. Se utiliza una aproximación basada en la búsqueda de reglas con un mínimo de confianza, realizando esta búsqueda por niveles de conceptos más generales a conceptos más específicos, hasta que no se encuentran conjuntos de ítems que no cumplan los requisitos mínimos. Para cada nivel, se puede utilizar cualquier algoritmo de búsqueda de conjuntos de ítems frecuentes como los mencionados en las asociaciones binarias.

Se utilizan dos tipos de soportes en la asociación multinivel:

- **Soporte uniforme:** Utiliza el mismo límite de cobertura en todos los niveles y permite optimizar la búsqueda de los conjuntos de ítems debido a la propiedad de que un concepto de un nivel que no cumpla con el requisito de cobertura no contiene ningún concepto más específico que cumpla el requisito.

- **Soporte reducido:** En cada nivel se establece un límite de cobertura específico que normalmente se va incrementando conforme desciende en el nivel de conceptos, es decir, a conceptos más específicos se colocan límites más altos.

A diferencia de la asociación en colecciones de datos binarios, las reglas del tipo Si ___entonces___ combinan ítems o conjuntos de ítems a diferente nivel. A este tipo de reglas se les denominan reglas de asociación de niveles cruzados.

Reglas de asociación secuenciales[11]

Este tipo de reglas expresan patrones de comportamiento que, como su nombre lo dice, son secuenciales o que se dan en instantes distintos pero cercanos en el

tiempo. Normalmente son aplicadas para mejorar la estructura de los servidores web, en marketing y distribución.

Se trabaja con una base de datos cuyos registros están formados por transacciones que contienen un identificador de usuario o cliente, un identificador de tiempo y un conjunto de ítems de la transacción.

Se basa en encontrar una lista de conjuntos de ítems más comunes de un mismo cliente, ordenada con respecto al tiempo.

Uno de los algoritmos más populares para la aplicación de reglas de asociación secuenciales es el algoritmo **A priori All** que se divide en cinco fases:

1. *Ordenación*: Tomando como clave el ID del cliente como clave primaria y el tiempo como clave secundaria, se ordena la base de datos y se construye una secuencia de conjuntos de ítems por cada cliente[11]

2. *Selección de conjuntos e ítems*: Se eligen los conjuntos de ítems que cumplan con una mínima cobertura, calculando la cobertura con respecto a cada cliente.

3. *Transformación y renombramiento*: A cada conjunto de ítems se le asigna un identificador, se transforma cada secuencia de manera que sólo contenga ítems frecuentes y se renombra cada conjunto por su identificador en cada una de las secuencias.

4. *Construcción de secuencias frecuentes*: A partir del conjunto de ítems frecuente se construye el conjunto de secuencias que cumplan con el criterio de cobertura.

5. *Selección de secuencias máximas:* Filtra el conjunto de secuencias frecuentes de manera que no haya subsecuencias.

Existen otros algoritmos más sofisticados como el **Generalized Sequential Patterns (GSP)** que permite considerar restricciones temporales, es decir secuencias de una determinada duración, así como ítems con diferentes niveles de abstracción o secuencias multinivel.

2.3 Técnicas de agrupación (clustering)

El clustering es una colección de métodos estadísticos que permiten clasificar un conjunto de elementos de una muestra en un determinado número de grupos, basándose en las semejanzas y diferencias existentes entre los componentes de

la muestra. En pocas palabras se trata de encontrar relaciones entre los objetos, a diferencia de la asociación que busca una relación entre los atributos.[1][17][20]

El proceso de agrupación depende de una serie de criterios habitualmente de distancia, ya que individuos similares (cercanos) deberán ir al mismo grupo. El objetivo es que los elementos de un mismo grupo sean lo más parecidos posible, mientras que elementos de distintos grupos sean muy diferentes.

Estas técnicas son ocupadas para la exploración de datos científicos, recuperación de información y minería de texto, aplicaciones sobre bases de datos espaciales, aplicaciones Web, marketing, diagnóstico médico, etc.

El procedimiento para realizar una técnica de clustering consta de 4 etapas principales[17]:

- *Elección de las variables:* Pueden ser variables cualitativas, ordinales, nominales, cuantitativas, discretas o continuas

- *Elección de media de asociación o similitud:* Las observaciones se agrupan según la similitud expresada en términos de una distancia, entre las distancias más utilizadas se encuentran: distancia euclidiana, distancia de Manhattan y distancia de Mahalanobis.

- *Elección de técnicas de minado:* Se emplea la técnica que más se ajuste para alcanzar los objetivos deseados. Estas Técnicas pueden ser Mapas auto-organizativos, K-medias, Jerárquicos, Cobweb, EM, ISODATA, GRASP, etc.

- *Validación de los resultados:* Se debe evaluar que tan bueno es el ajuste. (Ver 2.8)

A continuación se describirán dos de las técnicas más comunes de agrupamiento utilizadas en minería de datos:

K-Medias (K-Means)[9][19]

El algoritmo K-Means, creado por McQueen en 1967, es el algoritmo de clustering más conocido y utilizado ya que sigue una forma fácil y simple de clasificación de un conjunto de objetos en un determinado número de K grupos determinados a priori.

Se representa a cada uno de los clusters (prototipo) por la media de sus puntos, es decir por su centroide. Tiene la ventaja de que se observa un significado gráfico y estadístico inmediato.

La técnica K-Medias puede seguir dos enfoques distintos: K medias por lotes (*batch)* y K medias en línea (*on line)*. El primero se aplica cuando los datos de entrada se conocen desde un principio, mientras que el segundo se aplica cuando no se conocen los datos en un comienzo.

Para realizar el algoritmo se define un conjunto de objetos $D_n=(x_1, x_2,...x_n)$, para todo i en los reales y k, v_1 los centros de los K cluster. El procedimiento puede dividirse en 4 etapas:

Etapa 1. Se toman al azar k clusters iniciales.

Se eligen aleatoriamente *k* objetos que formarán los *k* clusters iniciales. Para cada objeto x_i se calcula el cluster más próximo A_g y se incluye en la lista de objetos de dicho cluster (habitualmente se utiliza la distancia euclidiana).

$$A_g = \arg\min_{A_j} \{ d(x_i, A_j) \} \quad \forall k = 1, \ldots, n$$

Etapa 2. Se asignan los objetos al cluster.

Cada prototipo A_k tendrá un conjunto de objetos a los que representa (D_n).

Etapa 3. Se reasignan los centroides.

Se desplaza el prototipo hacia el centro de masa de su conjunto de objetos.

$$A_k = \frac{\sum_{i=1}^{n} xi}{m}$$

Etapa 4. Se repite el procedimiento hasta que ya no se desplazan los prototipos.

Una vez terminado este algoritmo el espacio de objetos quedará dividido en k clases o regiones y el prototipo de cada clase estará en el centro de la misma. Dichos centros se determinan para minimizar las distancias cuadráticas euclídeas entre los patrones de entrada y el centro más cercano, es decir minimizando el valor de J:

$$J = \sum_{i=1}^{k} \sum_{n=1}^{m} M_{i,n} * d_{eucl} (x_n - A_i)^2$$

Donde m es el conjunto de clusters, x_n es el objeto de entrada n, A_i es el prototipo de la clase y $M_{i,n}$ es la función de pertenencia del ejemplo n a la región i que vale 1 si el prototipo A_i es más cercano al objeto x_n y 0 en otro caso.

El problema del empleo de estos esquemas es la necesidad de inicializar el número de prototipos al principio de la ejecución. Esto perjudica la eficacia del algoritmo ya que en la práctica no se conoce el número de cluster final.

Ejemplo:

Supongamos dos variables x_1 y x_2 y 4 elementos: A, B, C, D. Con la siguiente tabla de valores:

	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Se requiere dividir estos elementos en dos grupos (k =2).

De modo arbitrario, se dividen los elementos en dos clusters (AB) y (CD) y se calculan los centroides de los dos clusters.

Cluster (AB):

$\overline{x_1}$	$\overline{x_2}$
$\frac{5-1}{2}=2$	$\frac{3+1}{2}=2$

Cluster (CD):

$\overline{x_1}$	$\overline{x_2}$
$\frac{1-3}{2}=-1$	$\frac{-2-2}{2}=-2$

Se calculan las distancias euclídeas de cada observación al grupo de centroides y se reasignan cada una al grupo más próximo. Si alguna observación se mueve del grupo, hay que volver a calcular los centroides de los grupos. Así las distancias son:

$$d^2(A, (AB)) = (5-2)^2 + (3-2)^2 = 10$$

$$d^2(A, (CD)) = (5+1)^2 + (3+2)^2 = 61$$

Como A está más próximo al cluster (AB) que al cluster (CD), no se reasigna.

Se hace lo mismo para el elemento B:

$$d^2(B, (AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B, (CD)) = (-1+1)^2 + (1+2)^2 = 9$$

Por lo cual, el elemento B se reasigna al cluster (CD) dando lugar al cluster (BCD). A continuación se vuelven a calcular los centroides:

Cluster	$\overline{x_1}$	$\overline{x_2}$
A	5	3
(BCD)	-1	-1

Nuevamente, se vuelven a calcular las distancias para cada observación para ver si se producen cambios con respecto a los nuevos centroides.

	A	(BCD)
A	0	52
B	40	4
C	41	5
D	89	5

Como n se producen cambios, entonces la solución para *k* = 2 clusters es: A y (BCD).

Si se requiere comprobar la estabilidad de los grupos, es recomendable volver a hacer el algoritmo con una *k* distinta.

Una vez considerados los clusters finales, se deben interpretar; para realizar esto, se pueden cruzar con otras variables categóricas o se pueden ordenar de modo que los objetos del primer cluster aparezcan al principio y los del último cluster al final.

Mapas auto-organizativos de Kohonen[19]

En 1982 Teuvo Kohonen presentó un modelo de red denominado mapas auto-organizativos o SOM (Self-Organizing Maps) basado en ciertas evidencias descubiertas a nivel cerebral.

La red auto-organizada debe descubrir rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada e incorporarlos a una estructura interna de conexiones.

El objetivo de esta técnica es categorizar los datos que se introducen en la red. Se clasifican valores similares en la misma categoría y, por tanto, deben activar el mismo prototipo de salida.

Se elige un gran número de clusters y se colocan en forma de una red bidimensional. La idea es que los representantes de cada grupo (o pesos, según la notación de Kohonen) estén correlacionados espacialmente, de modo que los puntos más próximos en la rejilla sean más parecidos entre sí que los que estén muy separados.

Consta de dos capas, una capa de entrada (formada por *N* neuronas, una por cada variable de entrada), donde se introducen los objetos y otra de competición o salida (formada por *M* neuronas), en la que se procesa la información y se forma el mapa de rasgos. Cada neurona representa a un prototipo, la idea es que los prototipos capturen objetos similares. Normalmente, las neuronas de cada capa de salida se organizan en forma de mapa bidimensional como se muestra en la figura 2.2.

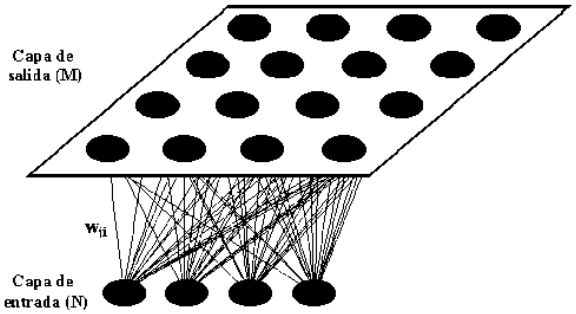


Figura 2.2 Mapa bidimensional de Kohonen[19]

Entre las conexiones de las dos capas que forman la red la información se propaga desde la capa de entrada hacia la capa de salida. Cada neurona de entrada *i* está conectada con cada una de las neuronas de salida *j* mediante un peso w_{ij}. De esta forma, las neuronas de salida tienen asociado un vector de peso W_j llamado vector de referencia (codebook), debido a que constituyen el vector prototipo (vector promedio) de la categoría representada por la neurona de salida *j.*

También existen conexiones laterales entre las neuronas de la capa de salida pues cada una de estas neuronas va a tener cierta influencia sobre las que estén a su alrededor. Esto se consigue a través de un proceso de competición entre las neuronas y la aplicación de una función denominada de vecindad (N_j), que produce la topología o estructura del mapa. Las topologías más frecuentes son la rectangular y la hexagonal.

El proceso de aprendizaje de los mapas auto-organizativos se puede resumir en tres pasos:

Paso 1. Se establece el número de ciclos, llamados de entrenamiento, que se repetirán hasta concluir el algoritmo para calcular la tasa de convergencia de la función de vecindad y la tasa de aprendizaje.

Se selecciona al azar un vector $\mathbf{x}$ del conjunto de datos y se calcula su distancia (similitud) a los vectores de referencia, usando por ejemplo la distancia euclídea. La magnitud calculada estará regida por la tasa de aprendizaje.

Las coordenadas de cada vector referencia quedan determinadas por los pesos de las neuronas de la capa de salida.

Paso 2. Se selecciona el vector más próximo o BMU (best matching unit) y el resto de vectores próximos es actualizado.

Mientras se va realizando el proceso de actualización y nuevos vectores se van asignando al mapa, la tasa de aprendizaje decrece gradualmente hasta llegar a cero.

La regla de actualización para el vector de referencia dado *i* es la siguiente:

$$m_j(t+1) = \begin{cases} m_j(t) + \propto (t)\left(x(t) - m_j(t)\right) & j \in N_c(t) \\ m_j(t) & j \notin N_c(t) \end{cases}$$

Una vez terminado el entrenamiento, el mapa se ordena en sentido topológico: *n* vectores topológicamente próximos se aplican en *n* neuronas adyacentes o incluso en la misma neurona.

Paso 3. Se corrobora que el mapa se ha adaptado adecuadamente a los datos. Como medidas de calidad de los mapas se considera la precisión de la proyección y la preservación de la topología.

La medida de precisión de la proyección describe cómo se adaptan las neuronas a los datos. Normalmente e número de datos es mayor que el número de neuronas y el error de precisión es siempre diferente de cero.
La medida de preservación de la topología describe la manera en la que el mapa auto-organizativo preserva la topología del conjunto de datos. Si el error de preservación es grande, entonces el mapa esta realizado incorrectamente.

Las fórmulas para las dos medidas mencionadas son las siguientes:

Precisión de la proyección

$$\varepsilon_q = \frac{1}{N}\sum_{i=1}^{N} ||x_i - m_c||$$

Preservación de la topografía

$$\varepsilon_t = \frac{1}{N}\sum_{k=1}^{N} u(x_k) \qquad u(x_k)\begin{cases} 1 & si\ el\ primer\ y\ segundo\ BMUs \\ & de\ x_k\ no\ estan\ proximos \\ 0 & c.o.c. \end{cases}.$$

Para estudiar un mapa auto-organizativo se requiere de herramientas gráficas como la matriz unificada de distancias o matriz U o histogramas de datos.

La matriz U representa el mapa como una rejilla regular de neuronas, el tamaño y topología pueden ser observados en el gráfico, donde cada elemento representa una neurona. Los colores en la rejilla se seleccionan de modo que cuanto más oscuro es el color entre dos neuronas, menor es la distancia entre ellas (ver fig. 2.3).

En el histograma de datos muestra cuántos vectores pertenecen a un cluster definido por cada neurona (ver fig. 2.4).

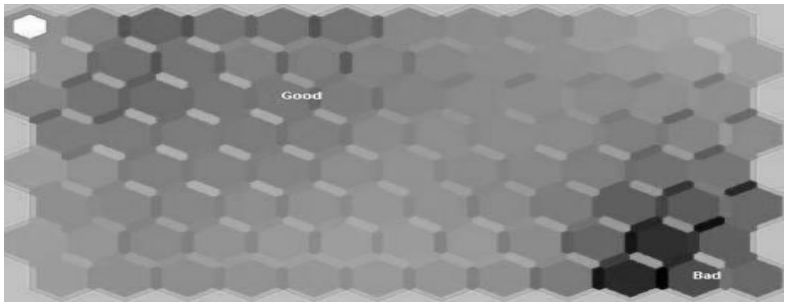


Figura 2.3 Ejemplo de representación gráfica de una matriz unificada de distancias [19].

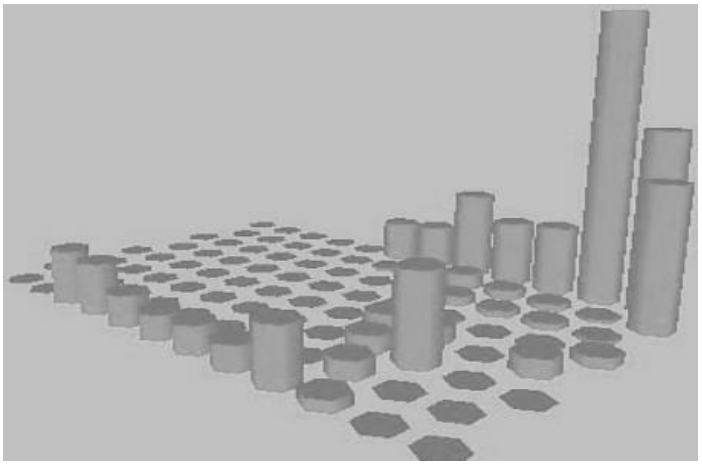


Figura 2.4 Ejemplo de visualización de histograma de datos [19].

2.4 Métodos de clasificación

La clasificación pretende construir un modelo que permita predecir la categoría de las instancias en función de una serie de atributos de entrada.

En esta sección se abordará el problema de la clasificación, que es el más frecuente en la práctica. En ocasiones, el problema de clasificación se formula como un refinamiento en el análisis, una vez que se han aplicado algoritmos no

supervisados de agrupamiento y asociación para describir relaciones de interés en los datos.

Existen varios algoritmos que sirven para realizar la tarea da clasificación, entre los más representativos se encuentran el algoritmo del vecino *k* más cercano, el algoritmos IBk, algoritmo LVQ (adaptación del método Kohonen a la tarea de clasificar), árboles de decisión (ver 2.6), etc. Se describirá a continuación únicamente los dos primeros ya que son los algoritmos más utilizados.

Algoritmo IBk[6][11]

Este algoritmo se encuentra dentro de los algoritmos utilizados en las redes neuronales (ver 2.7), el aprendizaje lo deja al final por lo que necesita menos tiempo para terminar el proceso de análisis pero más tiempo para realiza las predicciones y clasificaciones, pero las realiza con alta exactitud.

El algoritmo consta de cuatro fases que se pueden resumir de la siguiente manera:

1. Calcula la distancia euclidiana del prototipo deseado a aquellos que tengan atributos seleccionados por el usuario, es decir que fueron muestreados.

2. Ordena las muestras tomando en cuenta las distancias calculadas.

3. Escoge a priori la *k* más óptima con ayuda de una validación cruzada.

4. Calcula un promedio ponderado de distancia inversa con los vecinos multivariados *k* más cercanos. Y utilizando una función de aprendizaje, obtener los resultados.

Este método utiliza una medida de distancia para calcular el prototipo ganador así como para la función de aprendizaje, tanto para aproximar como para alejar los prototipos de los atributos base.

En general el número de prototipos suele ser mayor que el número de clases, ya que una misma clase puede estar distribuida en zonas distantes del espacio y necesitar más de un “grupo” para capturar un prototipo.

Algoritmo k- vecinos más cercanos[11][29]

El algoritmo *k* vecinos más cercanos asigna la característica (clase) del prototipo más próximo, utilizando una función de distancia. La regla utilizada se conoce como k-NN (k nearest neighbor). Es de las técnicas más utilizadas para clasificar, pero también se puede utilizar para estimación y pronóstico.

Esta técnica almacena datos de entrenamiento, para que una clasificación de un registro (prototipo) nuevo no clasificado, pueda encontrarse simplemente por comparación con los registros más similares en la base de datos.

Se denominará al conjunto de referencia (R), al conjunto de prototipos sobre el que se buscarán los vecinos más cercanos.

La regla de clasificación k-NN más simple es cuando el valor de k es uno y se basa en la suposición de que la clase del patrón a etiquetar, **X,** es la del prototipo más cercano en R (conjunto de referencia) al que denotaremos X_{NN}. Esta regla puede expresarse como:

$$d(X) = min_{i-1-N} d(X, X_i)$$

Por ejemplo, en la figura 2.5 se observan prototipos con la característica de ser círculos o cuadrados, entonces al prototipo interrogante, el algoritmo le asignaría la característica círculo, ya que el individuo más cercano posee dicha característica.

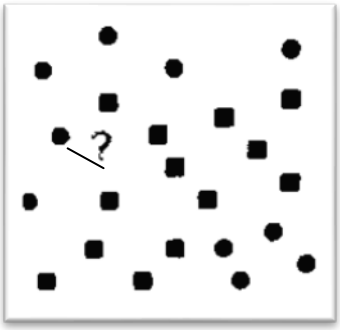


Figura 2.5 Imagen de 1-NN [11].

El problema de éste algoritmo es que ignora la densidad o la región donde se encuentra el prototipo que se desea clasificar. Por ello es más recomendable utilizar diversas k, estos algoritmos se basan en la suposición de que los prototipos más cercanos tienen una probabilidad similar.

Los vecinos se toman de un conjunto de objetos para los cuales la correcta clasificación es conocida. Esto se puede considerar como el conjunto de entrenamiento para el algoritmo. Para identificar a los vecinos, los objetos son representados por vectores de posición en un espacio característico multidimensional.

Dado un método o métrica para determinar qué registros son más similares al nuevo registro sin clasificar, se tiene que establecer cómo estos registros similares se combinan para proporcionar una decisión de clasificación para el nuevo registro. Es decir se necesita una **función combinación.** Las funciones de combinación más básicas son:

Votación no ponderada simple: Primeramente se decide el valor de *k* (número de registros que tendrán influencia en la clasificación del nuevo registro). A continuación, se compara el nuevo registro a los *k* vecinos más cercanos, es decir, a los *k* registros que son de distancia mínima desde el nuevo registro en términos de la distancia euclidiana o cualquier otra métrica; cada registro tendrá un voto y la característica con más votos es la que se le asignará al nuevo registro.

Votación ponderada: Supone que los vecinos que están más cerca al nuevo registro tienen mayor peso que los vecinos más distantes. La influencia de un determinado registro es inversamente proporcional a la distancia del registro nuevo a ser clasificado, es decir $voto(característica\ i) = \frac{1}{d(X,Yi)^2}$ donde X es el nuevo registro y Yi es un vecino cercano con la característica i. Así el prototipo nuevo tendrá la característica con mayor votación.

Si se quisiera utilizar el algoritmo *k*-vecino más cercano para estimación y pronóstico o encontrar valores continuos de una variable objetivo determinada, se debe utilizar el método **Locally weighted averaging.**
Este método estima la variable objetivo como la media ponderada de los registros históricos de esa variable para los *k* vecinos más cercanos, utilizando la inversa cuadrada de las distancias de los pesos (votos) obtenidos de igual manera que para clasificar. Es decir: $\widehat{y_{est}} = \frac{\sum_i w_i y_i}{\sum_i w_i}$ donde $\widehat{y_{est}}$ es la variable a estimar, w_i son los pesos con la característica i y y_i son los valores conocidos de la variable con característica i.

2.5 Métodos Bayesianos[11][19]

Para muchas técnicas de minería de datos el hecho de trabajar con incertidumbre es un gran problema pero no lo es en absoluto para los métodos y técnicas bayesianas debido a que una de sus principales características es el uso explícito de la teoría de la probabilidad para cuantificar la incertidumbre.

La teoría de la probabilidad y los métodos bayesianos son una de las técnicas más utilizadas en minería de datos, su principal herramienta son las redes bayesianas que son un modelo que permite un doble uso: predictivo y descriptivo. En el modelo descriptivo, los algoritmos de redes bayesianas se centran en el descubrimiento de las relaciones de independencia y/o relevancia entre sus variables, por lo que el modelo resultante refleja de forma explícita muchas relaciones de interés. En el modelo predictivo las técnicas bayesianas se ocupan para tareas de clasificación.

Redes Bayesianas (RBs)

Las RBs proporcionan una representación gráfica para un conjunto de variables aleatorias y para las relaciones existentes entre ellas.

Representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico. Este conocimiento se articula en la definición de relaciones de independencia o dependencia entre las variables que componen el modelo con el criterio conocido como *d-separación o separación dirigida*.

La tarea de las redes bayesianas es encontrar el gráfico dirigido acíclico (G) que mejor represente el conjunto de dependencias o independencia presentes, dado un conjunto de datos D y realizar predicciones sobre el comportamiento de cualquier variable a partir de otras variables conocidas.

Para realizar esta tarea es necesario calcular la probabilidad de una red bayesiana en concreto, dado el conjunto de datos conocidos (*P (G|D)*); una vez que se sabe cómo calcular la probabilidad, se tiene una medida de adecuación de cada red bayesiana a los datos de partida y como consecuencia se puede comparar entre distintas redes para elegir la mejor. La idea es que teniendo una métrica definida para calcular la adaptación de los datos a una red bayesiana, se buscará encontrar una solución que maximice esta métrica.

Las RBs trabajan con el teorema de Bayes el cual es derivado de la fórmula de probabilidad condicional y permite establecer la probabilidad a posteriori de una variable Y, dado un conjunto de eventos X.

$$P^*(Y) = \frac{P(X|Y)P(Y)}{P(X)}$$

Formalmente, una red bayesiana es una dupla B= (G, θ), donde G es el grafo y θ es el conjunto de distribuciones de probabilidad $P(X_i|Pa(X_i))$ con $i \in \{1, \dots, n\}$ y $Pa(X_i)$ representa los padres (nodos anteriores) de la variable X_i en el grafo G.

La codificación de las relaciones de independencia condicional en una RB hace que la distribución de probabilidad conjunta se pueda almacenar de una manera mucho más eficiente y local a cada una de las variables del modelo, es decir:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$$

Un ejemplo de red bayesiana se muestre en la figura 2.6 y la función de probabilidad conjunta representada en esta red es $P(x_1, \dots, x_5) = P(x_1|x_2, x_5)P(x_2)P(x_3|x_5)P(x_4|x_3, x_5)P(x_5)$.

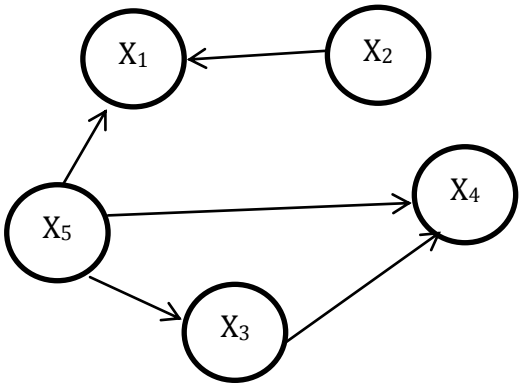


Figura 2.6 Red Bayesiana

El modelo más simple de clasificación con redes bayesianas es el modelo **Naïve Bayes** (NB) que se basa en que la estructura de la red es fija y solo necesitamos aprender los parámetros (probabilidades). El fundamento principal es la suposición de que todos los atributos son independientes y que es conocido el valor de la variable objetivo. Sus resultados son competitivos con otras técnicas e incluso las superan en algunas tareas.

La hipótesis de independencia asumida por el clasificador NB da lugar a un modelo gráfico probabilístico en el que existe un único nodo raíz (la clase) y todos los demás nodos tienen un único nodo antecesor.

Para realizar las redes bayesianas las personas se ayudan de numerosos paquetes de software que están dedicados al manejo y creación de estas redes. Dichos paquetes utilizan varios algoritmos especializados como:

Algoritmo K2

Se basa en la búsqueda u optimización de una métrica bayesiana, parte de que las variables de entrada están ordenadas, de forma que los posibles padres de una variable aparecen en orden antes que ella misma.

Algoritmo B

Está basado en la optimización de una métrica de calidad de redes bayesianas, este algoritmo no impone la restricción de proporcionarle como entrada un orden específico entre las variables. Trata de introducir el arco que mayor ganancia representa con respecto a la red anterior y a la métrica utilizada. El algoritmo se detiene cuando la inclusión de un arco no representa ninguna ganancia.

Algoritmo HC (Hill Climbing)

Parte de una solución inicial, como podría ser una red vacía de enlaces. A partir de esta solución se calcula el nuevo valor de la métrica utilizada de todos los nodos vecinos a la solución actual y se queda con el vecino que tenga mejor valor de la métrica. El algoritmo parará cuando no exista ningún nodo que pueda mejorar la solución actual.

Algoritmo TAN (Tree Augmented Naïve Bayes)

Constituye una extensión del clasificador NB. Se construye una red bayesiana donde se da un tratamiento especial a la variable objetivo. Se pretende mantener la simplicidad computacional del clasificador NB pero intentando mejorar la tasa de acierto durante la clasificación.

Algoritmo EM (Expectation Maximization)

Se parte de la hipótesis de que una red bayesiana está perfectamente definida y que sólo se tiene que estimar las tablas de distribución condicional asociadas a cada nodo o variable. El algoritmo consta de dos etapas iterativas: en la primera de ellas se calculan las esperanzas de los estadísticos necesarios (frecuencias esperadas) y la segunda etapa es la de maximización que consiste en estimar los parámetros a partir de las frecuencias esperadas.

La aplicación de las redes bayesianas para el análisis de minería de datos es más que evidente. La capacidad descriptiva y explicativa que poseen las hacen muy adecuadas para analizar las relaciones entre los atributos de una manera gráfica y mucho más sofisticada que las reglas de asociación o estudios correlaciónales convencionales.

2.6 Árboles de decisión [19]

Los métodos basados en árboles de decisión son bastante populares en minería de datos, se pueden utilizar para tareas de clasificación y estimación. Son un conjunto de condiciones organizadas en una estructura jerárquica, es decir, una colección de nodos de decisión, conectados por ramas, que se extienden hacia abajo desde el nodo raíz hasta que terminan en los nodos hoja.

Comenzando en el nodo raíz, que por convención se coloca en la parte superior del diagrama de árbol de decisión, los atributos se prueban en los nodos de decisión con cada resultado posible en una rama. Entonces cada rama conduce o bien a otro nodo decisión o a un nodo hoja de terminación. (Ver figura 2.7)

La decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

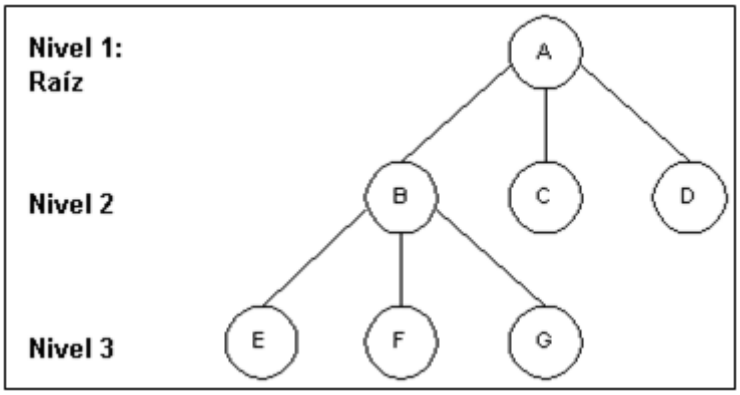


Figura 2.7 Árbol de decisión

Los árboles de decisión son útiles para la exploración inicial de datos y apropiados cuando hay un número elevado de datos. La variable objetivo en ésta técnica debe ser una variable discreta.

Los problemas donde se pueden utilizar árboles de decisión son: Estimación de una variable dependiente continua, regresión binaria, problemas de clasificación con categorías múltiples ordinales y problemas de clasificación con categorías nominales múltiples.

Para el conocimiento de la metodología se deben conocer los siguientes conceptos:

Entropía: es la medida de la incertidumbre que hay en un sistema. Es decir, ante una determinada situación, la probabilidad de que ocurra cada uno de los posibles resultados. [16]

La función de entropía más usada es la denominada binaria, descrita por la fórmula: $H_2\,(p,1-p) = plog_2\left(\frac{1}{p}\right) + (1-p)log_2\left(\frac{1}{1-p}\right)$. En general su fórmula es $H(X) = -\sum_j p_j log_2(p_j)$.

Ganancia de información: es la diferencia entre la entropía de un nodo y la de uno de sus descendientes. En el fondo no es más que una nueva entropía que servirá para la elección del mejor atributo en cada nodo.[19]

Para el nodo con el conjunto de entrenamiento S (datos usados para ese nodo S) y el atributo A. *Ganancia(S)= H(A) - H_S(A)*.

Profundidad: Cantidad de niveles o capas del árbol. Depende de la cantidad de datos y se debe procurar no quedar corto en la cantidad de divisiones. Algunos algoritmos manejan la profundidad como condición de paro.

Poda: Es eliminar condiciones en las ramas del árbol o quitar algunas reglas de división mediante un límite de poda. Los nodos que están por debajo del límite de

poda se eliminan. Se puede realizar este proceso antes (prepoda) o después (postpoda) de la elaboración de un árbol, esto dependerá de cada usuario. El objetivo es evitar un sobreajuste en la técnica.[11]

2.6.1 Clasificación de árboles de decisión

Las categorías existentes para esta técnica de minería, dependiendo de la tarea que realizan. Se destacan tres:

Árboles de decisión para clasificación:[29]

La característica más importante del problema de clasificación es que se asume que las clases son disjuntas, es decir, una variable es de clase A o de clase B, pero no puede ser al mismo tiempo de ambas. Debido a esto, los árboles de clasificación siempre conducirán un registro hasta un solo nodo hoja.

Los datos de entrenamiento se van dividiendo de arriba abajo, utilizando cada vez un conjunto de condiciones excluyentes y exhaustivas. Estos algoritmos se llaman algoritmos de partición. Dependen de un número específico de profundidad y de un criterio para la partición.

El aprendizaje de este tipo de árboles de decisión es el denominado criterio de partición ya que una mala elección de la partición generará un mal árbol. La única condición que se debe exigir es que las particiones al menos separen los datos en distintos nodos hijos.

Los algoritmos utilizados para clasificar son el CART, ID3 y C4.5 entre otros.

Árboles de decisión para regresión, agrupamiento o estimación:[1][16][19]

Este tipo de árboles se construyen de manera muy similar a los de clasificación pero la función de clasificación tiene dominio en los números reales y los nodos de las hojas del árbol se etiquetan con valores también reales.

Por ejemplo para realizar una regresión, se considera una función lineal en los nodos; para agrupamiento la idea es modificar el criterio de partición y de evaluación para que considere particiones que separen entre zonas densas y poco densas; para estimación los registros tienen una etiqueta discreta, denominada clase, se trata de determinar para cada nuevo registro cuál es la probabilidad de que pertenezca a cada uno de los nodos.

Los algoritmos utilizados para estas tareas son el CART, CHAID, C4.5 e ID3.

Árboles de decisión relacionales:[20][11]

Pueden definirse como un conjunto ordenado de reglas que se llaman listas de decisión de primer orden.

Tienen la misma estructura que los otros dos tipos de árbol, los nodos de decisión contienen condiciones mientras que los nodos hoja indican el valor predicho para la clase. Se trata de árboles binarios, realizan un test como criterio de división; la rama izquierda indica que el test es cierto y la rama derecha representa que el test es falso.

2.6.2 Algoritmos de aprendizaje

Existen varios algoritmos de aprendizaje en árboles de decisión, las más importantes son:

Random Forest – Bosque Aleatorio:[11]

Un bosque aleatorio es un clasificador desarrollado por Leo Breiman, que consiste en realizar muchos árboles de decisión y despliega como resultado la clase que es la moda de las clases de salida de árboles individuales.

El algoritmo tiene un parámetro $1 \leq v \leq d$ entero positivo fijado por el usuario. En cada paso de construcción del árbol, una hoja es elegida uniformemente aleatoria. v variables son seleccionadas aleatoriamente de las d candidatas (x_i). Entre las v variables, se selecciona la partición que minimiza el número de puntos mal clasificados de la base de datos con el criterio de votación ponderada (ver 2.4 b). Este criterio es repetido hasta que cada nodo contenga un punto de entrenamiento x_i.

Este tipo de algoritmo incluye un buen método para estimar los datos faltantes, provee una manera experimental para detectar interacciones entre las variables y calcula proximidades entre casos, lo cual es útil para el agrupamiento, detección de casos extraños, visualización de los datos y su aprendizaje es rápido.

Por otro lado los bosques aleatorios son propensos al sobreajuste en tareas de regresión o clasificación con mucho ruido (datos irrelevantes).

CART (Método de árboles de clasificación y regresión):[29][19][11]

Este método fue sugerido por Breiman et al. en 1984. Se basa en el lema “divide y vencerás” y realiza únicamente árboles binarios, conteniendo exactamente dos ramas en cada nodo de decisión. El algoritmo recursivamente particiona los

registros contenidos en los datos de entrenamiento en subconjuntos de valores similares para observar su comportamiento en función a la variable objetivo.

El criterio para una división óptima es de acuerdo a las siguientes características:

Sea $\Phi(s|t)$ una medida de "bondad" de una división candidata s en el nodo t, y cuya fórmula es

$$\Phi(s|t) = 2P_l P_r \sum_{j=1}^{\#class} |P(j|t_l) - P(j|t_r)|$$

t_l= *nodo hijo izquierdo del nodo t*
t_r = *nodo hijo derecho del nodo t*

$$P_i = \frac{\#\ de\ obs\ en\ t_l}{\#\ de\ obs\ en\ datos\ de\ entrenamiento} \quad i \in \{l, r\}$$

$$P_r(j|t_i = \frac{\#\ de\ obs\ de\ la\ clase\ j\ en\ t_i}{\#\ de\ obs\ en\ datos\ de\ entrenamiento}\ i \in \{l, r\}$$

Así, la división óptima es la que maximice esta medida $\Phi(s|t)$, utilizando todas las posibles divisiones en el nodo t.

El proceso general de esta metodología es:

Paso 1. El nodo raíz es dividido en subgrupos determinados por la partición de una variable (elegida mediante $\Phi(s|t)$), generando nodos hijos.

Paso 2. Los nodos hijos son divididos usando la partición de una nueva variable. El proceso recursivo se repite para los nuevos nodos hijos sucesivamente hasta que se cumpla alguna condición de paro.

Paso 3. Algunos nodos resultantes son terminales (nodos hoja puros), mientras que otros nodos podrían seguir dividiéndose (nodos hoja impuros).

Paso 4. En cada árbol se cumple la propiedad de tener un camino único entre el nodo raíz y cada uno de los demás nodos del árbol. Aquí se calcula la ganancia de la información.

C4.5:[29][16][11]

Este algoritmo fue propuesto por Quinlan y es la evolución del algoritmo ID3, creado por él mismo en 1986 el cual usa la ganancia de información como criterio de separación; el árbol crece hasta encontrar un nodo terminal y no emplea procedimientos de poda ni manejo de valores perdidos.

Al igual que el algoritmo CART, el algoritmo C4.5 recursivamente visita cada nodo de decisión selecciona la división óptima hasta que ya no haya más divisiones posibles, sin embargo se diferencian en que el algoritmo C4.5 no se limita a divisiones binarias y para los atributos categóricos por defecto produce una rama separada para cada uno.

El algoritmo C4.5 al ser evolución del algoritmo ID3, también utiliza el concepto de ganancia de información para seleccionar la división más óptima.

El concepto de entropía es utilizado de la siguiente manera. Sea una división candidata S, que divide los datos en T subgrupos, T_1, T_2,…, T_k (se tomarán en cuenta todas las posibles divisiones existentes).

Se calcula la suma ponderada de las entropías para los subconjuntos individuales mediante la fórmula $H_s(T) = -\sum_{i=1}^{k} P_i H_S(T_i)$ donde P_i representa la porción de observaciones en el subgrupo i.

Por último se calcula la ganancia de la división S y la división candidata que resulte con mayor ganancia es elegida y se vuelve a repetir el proceso hasta llegar a la condición de paro.

2.7 Redes Neuronales Artificiales (RNA)[19]

Las RNA son sistemas conexionistas dentro del campo de la Inteligencia Artificial, pueden utilizarse para varias tareas como: el reconocimiento de patrones la reducción de la dimensionalidad, la clasificación, la visualización, pronóstico, etc.[19]

El primer modelo de red neuronal fue propuesto en 1943 por Mc Culoch y Pittis en términos de un modelo computacional. Este modelo era un modelo binario, donde cada neurona tenía un escalón o umbral prefijado, y sirvió de base para modelos posteriores.[15]

La idea básica de las redes neuronales es imitar ciertos aspectos de la arquitectura del cerebro, para que ésta resuelva los problemas modelados.

Las redes neuronales biológicas tienen como objetivo principal desarrollar operaciones de síntesis y procesamiento de información, relacionadas con los sistemas biológicos. Las neuronas y las conexiones entre ellas *(sinapsis*) constituyen la clave para el procesado de la información.

El proceso biológico que se realiza puede describirse de la siguiente manera, aunque la forma completa en que dos neuronas se relacionan no es totalmente conocida: Una transmisión electro-química tiene lugar en la *sinapsis,* las *dentritas* reciben las señales de entrada procedentes de otras neuronas a través de la

sinapsis, finalmente, una vez recibida y procesada la información, el *axón* lleva la información de salida a las *dentritas* de otras neuronas (ver Figura 2.8).[11]

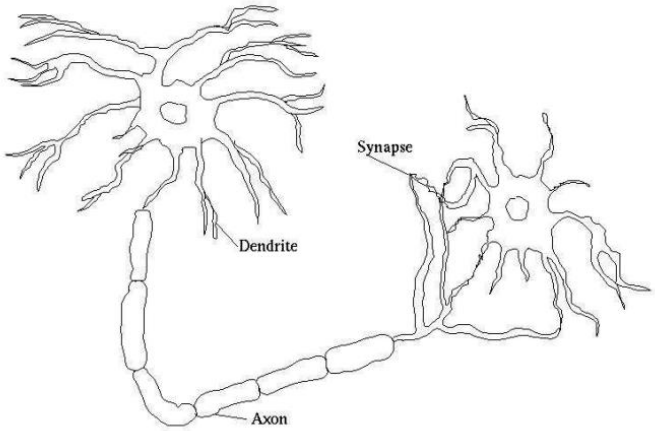


Figura. 2.8 Red neuronal biológica [11]

Las redes neuronales artificiales están inspiradas en las redes neuronales biológicas pero poseen otras funcionalidades y estructuras de conexión, así mismo tienen las siguientes características que las diferencian de las biológicas: Auto-Organización, adaptabilidad, procesado no lineal y procesado paralelo (usa un gran número de nodos de procesado).

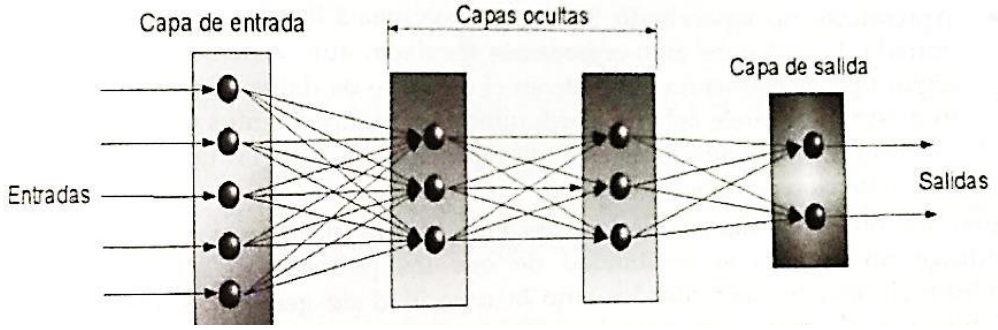


Figura 2.9 Ejemplo de red neuronal artificial [11].

Una RNA se determina por las neuronas y la matriz de pesos, en su estructura contiene una capa de entrada (cada nodo corresponde a una variable independiente), una o varias capas ocultas (nodos internos) y una capa de salida (posibles soluciones al problema) (Fig. 2.9). Entre dos capas existe una red de pesos de conexión, que pueden ser hacia delante, hacia atrás, lateral y de retardo (Fig. 2.10). [15]

Conexiones hacia delante: los valores de las neuronas de una capa inferior son propagados hacia las neuronas de la capa superior.

Conexiones hacia atrás: estas conexiones llevan los valores de las neuronas de una capa superior a otras de una capa inferior.

Conexiones laterales: la neurona de salida que da el valor más alto es la que tendrá el peso total (por ejemplo 1, mientras todas las demás tendrán un valor 0).

Conexiones de retardo: los elementos de retardo se incorporan en las conexiones para implementar modelos dinámicos y temporales (modelos que requieren de memoria).

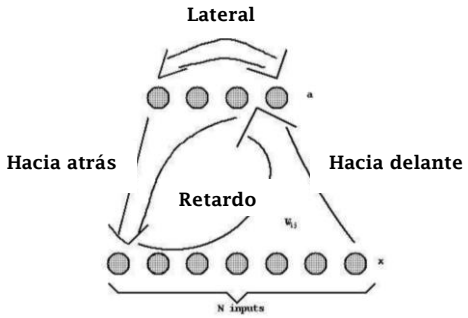


Figura 2.10 Tipos de conexiones en una RNA [23]

El tamaño de las redes neuronales depende del número de capas y del número de neuronas ocultas por capa. El número de capas ocultas está directamente relacionado con las capacidades de la red, para que el comportamiento de la red sea correcto, se tiene que determinar apropiadamente el número de neuronas de la capa oculta.

El proceso de trabajo de una red neuronal se puede sintetizar de la siguiente forma: Cada nodo inyecta un valor de y_i a su salida, este valor se propaga por la red mediante conexiones unidireccionales hacia otros nodos de la red, asociada a cada conexión hay un peso sináptico denominado w_{ij} (los pesos iniciales son generados aleatoriamente en un rango de -0.5 y 0.5), que determina el efecto del nodo j-ésimo sobre el nodo i-ésimo.[15]

Las entradas al nodo i-ésimo que provienen de los otros nodos son acumulados junto con el valor umbral θ_i, y se aplica la función base *f*, obteniendo u_i. La salida final y_i se obtiene aplicando la función de activación sobre de u_i.

Función base f[16]: Esta función puede ser una función lineal de tipo hiperplano, en la cual el valor de red es una combinación lineal de las entradas $u_i(\mathrm{w},\mathrm{x}) = \sum_{j=1}^{n} w_{ij}x_j$ o una función radial de tipo hiperesférico, la cual es una función base de segundo orden no lineal y donde el valor de red representa la distancia a un determinado patrón de referencia $u_i\,(w,x) = \sqrt{\sum_{j=1}^{n}(x_j - w_{ij})^2}$

Función de activación[16]*:* El valor de red, obtenido por la función base $u(w,x)$, es transformada mediante una función no lineal que puede ser una función sigmoidal $f(u_i) = \frac{1}{1+\exp\{-\frac{u_i}{\sigma^2}\}}$ o una función gausiana $f(u_i) = c * \exp\{-\frac{u_i^2}{\sigma^2}\}$.

En las tareas de clasificación normalmente se toma la función de activación sigmoidal y en las tareas de predicción se utiliza la función de activación lineal.

El proceso de modelización tiene dos fases: Fase de entrenamiento y fase de prueba, en la primera se usa un conjunto de datos o patrones de entrada para determinar los pesos que definen el modelo de RNA de manera iterativa y con el objetivo de minimizar el error, en la segunda se corrige un posible subreajuse utilizando un segundo grupo de datos de entrada (grupo de validación) que permita controlar el proceso de aprendizaje.

Siendo que todas las RNA siguen en general el mismo proceso de modelización y trabajo, existe una clasificación que las agrupa en términos de sus algoritmos o métodos e entrenamiento en redes neuronales supervisadas y no supervisadas.

RNA supervisadas:[11][15]

Este tipo de redes neuronales es el más popular, los datos para el entrenamiento están formados por pares de patrones de entrada y salida, antes de producir una salida, los datos pasan por un proceso de supervisión. En la etapa (m+1)-ésima, los pesos se adaptan de la siguiente forma $w_{ij}^{m+1} = w_{ij}^{m} + \Delta w_{ij}^{m}$.

El método más utilizado en esta clasificación es el **Perceptrón multicapa (MLP por sus siglas en ingles)** que es capaz de actuar como un aproximador universal de funciones, convirtiéndolo en una herramienta de propósito general, flexible y no lineal.

Un preceptrón multicapa está compuesto por una capa de entrada, una capa de salida y una o más capas ocultas; las conexiones entre neuronas son siempre hacia delante por lo tanto la información siempre se transmite desde la capa de entrada hacia la capa de salida.

Intenta resolver problemas de predicción y de clasificación, para el primer problema, estima una variable continua de salida a partir de las variables predictoras o de entrada, para el segundo problema asigna la categoría de pertenencia de un determinado patrón a partir de un conjunto de variables predictoras.

El MLP utiliza el algoritmo de retropropagación. Cabe señalar que antes de poder meter los datos a nuestro algoritmo se les debe realizar una transformación mediante el proceso min-max o con funciones dummys con el objetivo de contener valores únicamente entre cero y uno.

El algoritmo utiliza una tasa de aprendizaje que controla el tamaño del cambio de los pesos en cada iteración, en general el valor de esta tasa suele estar comprendida entre 0.05 y 0.5, así mismo utiliza un factor de momento que acelera la convergencia de los pesos, a éste se le suele asignar un valor próximo a 1.

RNA no supervisadas:[11][15]

Para estos modelos, el conjunto de datos de entrada consiste solo en patrones de entrada. Por tanto, la entrada a cada capa de la red se produce sin una supervisión de los datos de salida. La red se adapta con respecto a las experiencias recogidas de los patrones de entrenamiento anteriores.

El método más utilizado en RNA no supervisadas es el **Aprendizaje de Hebb**, denominado así debido gracias a Donald Hebb [Hebb 1949] que se basa en el supuesto de que si dos neuronas en cualquier lado de la sinápsis son activadas simultáneamente, la longitud de la sinápsis se incrementará.

Consiste básicamente en el ajuste de los pesos de las conexiones de acuerdo con la correlación de los valores de activación (salidas) de las neuronas conectadas.

Este algoritmo pretende medir la familiaridad o extraer las características de los datos de entra.

Ahora bien con la finalidad de llegar al entendimiento global de una RNA, se suele adoptar la perspectiva *top-down* que empieza por la aplicación, después pasa al algoritmo y de ahí a la arquitectura de la red (Fig.2.11).

La ventaja de las RNA reside en el procesado paralelo, adaptativo y no lineal, además tienen muchas aplicaciones en el procesado de señales e imágenes, reconocimiento del habla y caracteres, sistemas expertos, análisis de imágenes médicas, control de robos e inspección industrial.

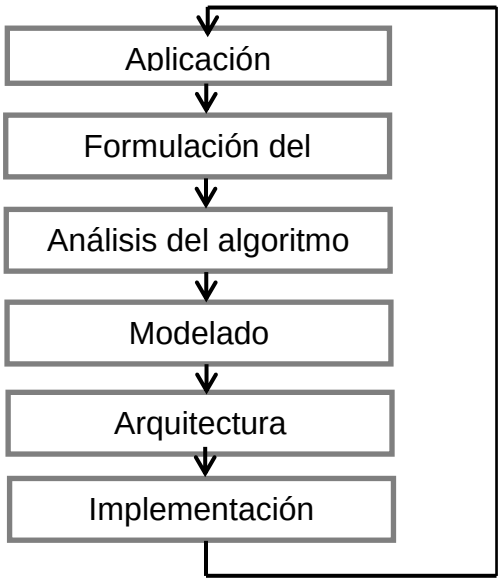


Figura 2.11 Perspectiva top-down para RNA

2.8 Técnicas de evaluación

Dentro del proceso de minería de datos, la etapa de evaluación es de suma importancia ya que al llevarla a cabo, el usuario sabrá si sus modelos son funcionales y apropiados para el cumplimiento del objetivo inicial.

Existen varias técnicas que sirven para evaluar la calidad de un modelo. De manera global se pueden utilizar criterios de evaluación como lo son el interés, la novedad, la comprensibilidad, la simplicidad y la aplicabilidad del modelo. De forma más específica las técnicas de evaluación dependerán de la tarea de minería que se realice.

Evaluación de clasificadores:

Las técnicas de clasificación contienen una función objetivo, considerando un espacio de posibles hipótesis. Las medidas de calidad para evaluar la hipótesis deben reflejar similitud semántica de la hipótesis con respecto a la función objetivo. En el caso de los clasificadores, la función tiene como dominio una categoría o clase.

Existen técnicas de evaluación basadas en la precisión de la hipótesis (*h)* o dicho de otro modo en el porcentaje de error de la hipótesis con respecto a la función (*f).* Al porcentaje de error se le llama error de la muestra, lo que se espera es que sea pequeño, se obtiene de la siguiente manera:[11]

$$error_s(h) = \frac{1}{n}\sum_{x \in S} \delta(f(x) \neq h(x))$$

Donde *n* es el número de componentes de la base de datos S, $\delta = \begin{cases} 1 & verdadero \\ 0 & falso \end{cases}$.

Un mecanismo más correcto para calcular el error de la muestra es separar el conjunto de datos en dos subconjuntos disjuntos. El primero subconjunto se llama de **entrenamiento** y el segundo se llama de **prueba** y solo se emplea para calcular el error de muestra de las hipótesis construidas con los datos de entrenamiento. Normalmente la partición se realiza aleatoriamente dejando 50% de los datos a cada grupo.[16]

Otra técnica de validación es la llamada **validación cruzada**, que permite reducir la dependencia del resultado del experimento en el modo en el que se realiza la partición. Consiste en dividir el conjunto de datos en *k* subconjuntos disjuntos de tamaño similar, se establece una hipótesis utilizando el conjunto formado por la unión de *k-1* subconjuntos y el subconjunto restante se emplea para calcular un error de muestra parcial. Este procedimiento se repite *k* veces utilizando siempre un subconjunto diferente para estimar el error. El error final será la media aritmética de los *k* errores de muestra parciales. [11]

Los **intervalos de confianza** [3] son muy comunes para evaluar este tipo de técnicas ya que dada una muestra *S* de *n* registros tomada a partir de una función objetivo *f* con una distribución *D*, es posible establecer un intervalo de una hipótesis *h* a partir del error de muestra utilizando, en grupos grandes, la siguiente fórmula para calcular el intervalo:

$$error_s(h) \pm Z_{\%}\sqrt{\frac{error_s(h)(1-error_s(h))}{n}}$$

Evaluación de modelos de regresión (h)[2]:

Para evaluar estas técnicas de minería, se suele estimar la calidad de un modelo calculando la diferencia entre las predicciones del modelo y las de la función objetivo. Dos de los métodos utilizados son el cálculo de la raíz del error cuadrático medio y mucho más efectivo el error absoluto medio. El primero está dado por la fórmula $RECM = \sqrt{\frac{1}{n}\sum_{x \in S}(h(x) - f(x))^2}$ y el segundo está dado por la fórmula $EAM = \frac{1}{n}\sum_{x \in S}|h(x) - f(x)|$.

Evaluación de modelos de agrupamiento:

Una de las opciones de evaluación de estos modelos es utilizar el método de máxima verosimilitud que significa quedarse con la hipótesis con la cual los datos sean más verosímiles. Sea p(x) la probabilidad estimada de observar un punto en la posición x utilizando el modelo a evaluar, si el modelo es bueno, se deben esperar probabilidades altas para los puntos que son observados. Podemos tomar p(x) como función de evaluación para el punto x.[11]

Otra opción es utilizar la distancia entre los grupos como medida de calidad del modelo, entre mayor sea la distancia entre los grupos, mejor se habrá efectuado la separación. Las distancias utilizadas son la distancia media (media de las distancias entre todos los componentes de ambos grupos), la distancia simple (distancia entre los vecinos más próximos de los dos grupos) y la distancia completa (utiliza los vecinos más lejanos).

Otra alternativa es realizar varios modelos con diversas o inclusive la misma técnica de minería y comparar entre sí.

Evaluación de reglas de asociación:

Como las reglas de asociación expresan patrones de comportamiento entre los datos (combinaciones de los atributos que suceden más frecuentemente) no hay un atributo definido como clase sobre el cual evaluar la regla, no podemos utilizar las técnicas de evaluación mencionadas anteriormente.

Entonces se suele trabajar con la medida de cobertura y la medida de confianza. La primera se define como el número de instancias en las que la regla se puede aplicar, la segunda mide el porcentaje de veces que la regla se cumple cuando se puede aplicar.

Para cualquier tarea de minería, en ocasiones se desea comparar dos técnicas de aprendizaje distintas, para esto se utiliza el llamado *t-test (test de t-Student)*[3] el cual certifica si las medidas de dos grupos son estadísticamente diferentes y se basa en calcular el valor *t* que está definido por la fórmula $t = \frac{\hat{d}}{S_{\hat{d}}}$, donde $\hat{d}$ representa la diferencia entre la media de los dos grupos y $S_{\hat{d}}$ la variabilidad entre dos grupos.

CAPITULO 3. "APLICACIÓN DE MINERÍA DE DATOS. Fases: Entendimiento del Negocio, Preparación y Análisis de los Datos"

3.1 Introducción y Fase de entendimiento del negocio

En este Capítulo se realiza la primera parte de la aplicación de Minería de Datos en la Facultad de Ciencias[33], en las carreras de Actuaría y Ciencias de la Computación, con el objetivo de encontrar patrones que inciden en la culminación de la carrera.

Se utiliza para su realización el método CRISP-MD.

En el siguiente estudio únicamente se busca encontrar patrones basados en historiales académicos e información académica, dejando fuera los factores económicos y sociales.

Cabe resaltar que se delimitará a estudiar los criterios de culminación en las generaciones de 2002 a 2013 para ambas carreras, teniendo así 4 planes de estudio, en Actuaría se encuentran los planes 2000 y 2006, mientras que en Ciencias de la Computación se encuentran los planes 1994 y 2013.

3.2 Fase entendimiento de los datos

Una vez establecido el objetivo de estudio, se procede a analizar la base de datos proporcionada para conocer las tablas que servirán y las que harán falta para el proceso.

Los datos contenidos en la base de datos inicial fueron proporcionados en un archivo Excel y se encuentran de forma transaccional por lo que habrá que modificarlos a un esquema horizontal. Las variables contenidas son:

ID: De forma encriptada ya que la información personal de los alumnos es confidencial

Generación: Año de generación de cada alumno, de la generación 1958 a la generación 2013

Carrera: Si el alumno es de Ciencias de la Computación o Actuaría.

Plan: Para los alumnos de Actuaría el plan 2000 corresponde a la clave 119, el plan 2006 corresponde a la clave 1176 y para los alumnos de Ciencias de la computación el plan 1994 corresponde a la clave 218 y el plan 2013 corresponde a la clave 1556.

Ingreso: Contiene la forma en que ingresó el alumno a la Facultad de Ciencias, puede ser de 10 formas: Años posteriores al primero AC, años posteriores al primero RE, Cambio de Carrera y/o Plantel (concurso), cambio de plantel, cambio interno de carrera, carrera simultánea, concurso de selección, examen de selección, pase reglamentario o segunda carrera.

Avance: Contiene el porcentaje de avance académico del alumno expresado del 0 al 1.

Promedio: Contiene el promedio general del alumno hasta el momento.

Aprobadas: Contiene 0 si el alumno reprobó la materia o 1 si el alumno aprobó la materia.

Clave: Contiene el número asignado a la materia inscrita por el alumno.

Materia: Contiene el nombre de la materia inscrita.

Grupo: Contiene la clave del grupo en el que se inscribió el alumno.

Semestre: Contiene la clave del semestre en el que el alumno inscribió la materia.

Tipo: Si la materia es inscrita de forma ordinaria el valor será “ORD”, si la materia fue inscrita en extraordinario largo el valor será “EXTL” y si la materia fue inscrita en extraordinario el valor será “EXT”.

Cal: Contiene la calificación obtenida por el alumno, ésta puede ser de 5 a 10, NP o NA.

Hora_ini: Contiene la hora de inicio de la materia inscrita.

Vg: Contiene el total de inscripciones en la materia.

Con esta información, lo que se pretende hacer es, eliminar los registros y columnas innecesarias para el estudio.

Debido a que la información se encuentra en forma transaccional, se crearán algunas columnas con información relevante usándola tal cual está y posteriormente se cambiará el formato de la base a horizontal de manera que las filas sean únicas y con información de un solo alumno.

Se dividirá la información en varias tablas que podríamos clasificar en 2 tipos, el primer tipo contendrá la información general de los alumnos, el segundo tipo contendrá los historiales académicos de forma detallada. Debido a que los alumnos se encuentran en diferente carrera y plan de estudios, es necesario crear una tabla de historiales académicos para cada plan.

Una de las columnas o variables a crear será la variable objetivo, definida como **situación escolar** y dividida en tres categorías: Estudia, terminó y desertó.

Los criterios tomados para poner la información en situación escolar son los siguientes:

- Un alumno desertó si el número de créditos en su registro es menor al número de créditos requeridos para terminar sus estudios y que tenga más de cuatro semestres sin inscribirse, es decir que su último semestre inscrito sea menor a 2011-2.
- Si el alumno tiene menos créditos de los requeridos en su plan de estudios, pero su último semestre inscrito fue 200131, entonces el alumno sigue estudiando y se le asignará la categoría Estudia.
- Si el alumno tiene un avance académico mayor o igual a uno, entonces el alumno terminó la carrera.

La información necesaria par a crear las columnas nuevas se obtiene de dos fuentes principales, la primera es la base de datos proporcionada y la segunda es la página http://www.fciencias.unam.mx.

3.3 Fase de preparación de los datos

La primera fase de preparación de los datos, se llevó a cabo en Excel 2010, aprovechando el formato ya dado, utilizando las funciones de Excel y creando algunas macros en su programador (Ver Fig. 3.1). En esta fase se pretende organizar la información para poder ser analizada, al finalizar, dicha las bases de datos será pasada a formato txt (delimitado por tabulaciones) para poder utilizar las tablas en el programa R en las siguientes fases del estudio.

La segunda fase de preparación de datos se llevará a cabo a la par del análisis exploratorio. En esta segunda fase se realizará la modificación, transformación y selección de variables de acuerdo con los resultados del análisis.

B	C	D	E	F	G	H	I	J	K
	plan	ingreso	generación	carrera	clave	materia	grupo	semest	tipo
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	399	Estadística II	6096	20071	ord
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	1705	Productos Financieros Derivados I	6157	20071	ord
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	621	Programación Lineal	6159	20071	ord
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	77	Análisis de Redes	6144	20082	exl
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	9	Análisis Matemático I	4111	20082	exl
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	1605	Legislación en Seguro Privado y Social	6212	20082	ord
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	1050	Administración Financiera	6124	20101	exl
z6R7oYUXkvPcy3ZdQ=	119	Cambio de Carrera y/	2003	Actuaría	1707	Análisis Numérico	6101	20101	exl

Figura 3.1 Ejemplo de las 1^{eras} Columnas de la base original en forma transaccional. Primeras Columnas.

El primer paso que se realizó fue quitar aquellos registros con generaciones anteriores a 2002.

Posteriormente se procedió a crear las nuevas columnas utilizando aún la forma transaccional de la base otorgada, estas columnas ayudarán para la creación de la vista minable (tabla a la que se le aplicará la minería de datos). Dichas columnas y su proceso de elaboración son:

Turno: Contiene una M si es matutino y una V si es vespertino. El criterio tomado para asignar los valores fue: si Hora_ini es mayor a 13 hrs, entonces se asigna V, de lo contrario se asigna M.

SemCorrecto: Contiene “si”, si el alumno inscribió la materia de acuerdo al plan de estudios en el semestre que le corresponde, o “no” si el alumno inscribió la materia en un semestre que no toca. La asignación de su valor requirió la elaboración de 3 columnas auxiliares Aux1, Aux2, Aux3.

Aux1: Con la función *consultav(),* se realizó una búsqueda de la materia inscrita en los historiales académicos obtenidos de la página de la facultad de ciencias para cada plan, dichos historiales se encuentran en otra hoja de Excel, su primera columna contiene el nombre de la materia y la segunda columna contiene si la materia es par o impar. Los valores que puede contener Aux1 son 1 si la materia es impar, 2 si la materia es par o #N/A si no se encontró la materia.

Aux2: Contiene valores 1, 2 o 3, uno si aux1 es igual a 1, 2 si aux1 es igual a 2 y 3 si aux1 es igual a #N/A. El número 3 significa que la materia es optativa.
Aux3: Se extrajo con la función *derecha()* aplicada a la columna “Semestre” y extrayendo su último dígito, ya que éste contiene la información de si el alumno inscribió en semestre par o impar. Toma valores 1 o 2.
Finalmente se asigna el valor a la columna Semestre correcto con el siguiente criterio, si Aux3=Aux2 o Aux3=3, entonces se asigna “si”, de lo contrario se asigna “no”.

UltimoSemInsc y 1erSem: Dos columnas que contienen la información del primer y último semestre inscritos por el alumno respectivamente. Para el llenado de ésta columna se requirió elaborar una macro en Visual Basic de Excel, ésta función extrae el máximo y el mínimo del total de semestres inscritos por cada alumno y rellena cada columna con el dato correspondiente y en las filas asignadas a ese alumno. Dicha función se encuentra en la Fig.3.2

TotalSem: Contiene el total de los semestres inscritos por el alumno. Se requirió hacer una macro para ésta columna, dicha macro se encuentra en la Fig. 3.3.

TotalAprobadas *: Contiene el total de materias aprobadas que tiene el alumno.

AprobadasYrecursadas *: Contiene el total de materias aprobadas, pero que fueron recursadas.

AprobadasNoRecursadas *: Contiene el número de materias que el alumno acredito en su primera inscripción.

TotalReprobadas*: Contiene el total de materias reprobadas del alumno.

ReprobadasYrecursadas *: Contiene el total de materias que el alumno tiene reprobadas y que ha recursado.

ReprobadasNoRecursadas *: Contiene el total de materias que el alumno reprobó pero que no se recursaron.

```
Sub primer_semestre()
        Dim alumno, alumnoaux As String
        Dim cont, i, cont2, k, cont3 As Long
        Dim tamaño As Long
        '_____________________________________CUENTA ALUMNOS
        cont3 = 1
        Range("B2").Select
        tamaño = Range(Selection, Selection.End(xlDown)).Count
        alumno = Range("B2").Value
        alumnoaux = Range("B3").Value
        For i = 2 To tamaño
             If alumno <> alumnoaux Then cont3 = cont3 + 1
             alumno = Cells(i + 1, 2).Value
             alumnoaux = Cells(i + 2, 2).Value
        Next i
        '_____________________________________-
        i = 3
        cont = 0
        cont2 = 1
        alumno = Range("B2").Value
        alumnoaux = Range("B2").Value
        For j = 1 To cont3
            cont = 0
                '_____________________________________Cuenta las veces que aparece el alumno
                While alumno = alumnoaux
                    alumno = Cells(i, 2).Value
                    i = i + 1
                    cont = cont + 1
                Wend
            cont2 = cont2 + cont
            alumnoaux = Cells(cont2 + 1, 2).Value
               alumno = Cells(cont2 + 1. 2).Value
              '_____________________________Rellena con el valor maximo
              For k = 0 To cont - 1
                  Cells(k + cont2 - cont + 1, 25).Value = WorksheetFunction.Max(Range(Cells(cont2 - cont + 1, 8), Cells(cont2 - 1, 8)))
                  Cells(k + cont2 - cont + 1, 27).Value = WorksheetFunction.Min(Range(Cells(cont2 - cont + 1, 10), Cells(cont2 - 1, 10)))
              Next k
        Next j
 End Sub
```

Figura. 3.2 Código para rellenar último semestre y primer semestre

TotalRecursadas *: Contiene el total de materias que el alumno ha recursado a lo largo de su trayectoria escolar.

Recursamientos *: Contiene el total de los recursamientos que el alumno ha tenido a lo largo de su trayectoria escolar

Matutino *: Contiene el número total de materias que el alumno inscribió en el horario matutino.

Vespertino*: Contiene el total de materias que el alumno inscribió en el horario vespertino.

InscritasBien *: Contiene el total de materias que el alumno inscribió en el semestre correcto (par o impar) de acuerdo al mapa curricular de su carrera y plan de estudios.

InscritasMal *: Contiene el total de materias que el alumno inscribió en el semestre incorrecto (par o impar) de acuerdo al mapa curricular de su carrera y plan de estudios.

```
Sub total_semestres()

    Dim alumno, alumnoaux As String
    Dim cont, i, cont2, k, cont3 As Long
    Dim tamaño As Long
    i = 3
    cont = 0
    cont2 = 1
    alumno = Range("B2").Value
    alumnoaux = Range("B2").Value
        For j = 1 To 5393 'total alumnos
            cont = 0
            sem = 0
            trecu = 0

        '_____________________________________________Cuenta las veces que aparece el alumno
            While alumno = alumnoaux
                If Cells(i - 1, 10) <> Cells(i, 10) Then sem = sem + 1
                If Cells(i - 1, 16) > 1 Then trecu = trecu + Cells(i - 1, 16).Value - 1

                alumno = Cells(i, 2).Value
                i = i + 1
                cont = cont + 1

            Wend
            cont2 = cont2 + cont
            alumnoaux = Cells(cont2 + 1, 2).Value
            alumno = Cells(cont2 + 1, 2).Value
        '_____________________________________________Rellena con el valor maximo
            For k = 0 To cont - 1
                  Cells(k + cont2 - cont + 1, 28).Value = sem
                  Cells(k + cont2 - cont + 1, 36).Value = trecu


            Next k
        Next j
```

Figura 3.3 Código para rellenar columna total semestres

Años: Representa los años que el alumno ha estado en la Facultad de Ciencias desde la primera hasta la última inscripción sin tomar en cuenta el tiempo efectivo en que el alumno se inscribió, bajas temporales o interrupción de estudios. Cada semestre tiene un valor de 0.5 años.

SitEscolar: Los valores que toma son “Terminó”, “Desertó” o “Estudia. Un alumno terminó si la columna avance es mayor o igual a uno. Un alumno desertó si la columna avance es menor a uno y la columna último semestre es menor que 20131. Un alumno estudia si la columna avance es menor a uno y el último semestre inscrito es igual a 20131.

Termino, Deserto y Estudia: Son el resultado de descomponer a la variable SitEscolar. Toman valores Si o No, esto con el objetivo de crear variables objetivo dummy, formato requerido en algunos modelos de minería.
Para las variables con * se elaboraron las macros descritas en la Fig. 3.4a y 3.4b.
Una vez que se terminaron todas estas columnas, se procedió a cambiar la presentación de los datos, de transaccional a horizontal con la finalidad de poder utilizarla para el minado.

El proceso requirió nuevamente la programación de una macro que identificara a cada alumno, haciendo una lista con registros únicos, terminada esta tarea se procedió a buscar la información de cada uno de los alumnos con la función *consulatv ( )* aplicada a cada columna. Llamaremos a esta base “Datos Alumnos” y contiene las columnas Id, plan, ingreso, carrera, avance, promedio, último sem, 1er sem, total sem, aprobadas, aprobadas recursadas, aprobadas no recursadas, reprobadas, reprobadas recursadas, reprobadas no recursadas, total recursadas, recursamientos, matutino, vespertino, bien inscritas, mal inscritas y situación escolar.

```
Sub total_materias()

    Dim alumno, alumnoaux As String
    Dim cont, i, cont2, k, cont3 As Long
    Dim tamaño As Long
    i = 3
    cont = 0
    cont2 = 1
    alumno = Range("B2").Value
    alumnoaux = Range("B2").Value
        For j = 1 To 5393 'total alumnos
            cont = 0
            rec = 0
            apro = 0
            aproR = 0
            aproNr = 0
            repro = 0
            reproR = 0
            reproNr = 0
            mat = 0
            vesp = 0
            correcto = 0
            incorrecto = 0
            '_____________________________________Cuenta las veces que aparece el alumno
            While alumno = alumnoaux
                If Cells(i - 1, 19) = 1 Then apro = apro + 1
                If Cells(i - 1, 19) = 1 And Cells(i - 1, 16) > 1 Then aproR = aproR + 1
                If Cells(i - 1, 19) = 1 And Cells(i - 1, 16) = 1 Then aproNr = aproNr + 1
                If Cells(i - 1, 19) = 0 Then repro = repro + 1
                If Cells(i - 1, 19) = 0 And Cells(i - 1, 16) > 1 Then reproR = reproR + 1
                If Cells(i - 1, 19) = 0 And Cells(i - 1, 16) = 1 Then reproNr = reproNr + 1
                If Cells(i - 1, 16) > 1 Then rec = rec + 1
                If Cells(i - 1, 20) = "M" Then mat = mat + 1
                If Cells(i - 1, 20) = "V" Then vesp = vesp + 1
                If Cells(i - 1, 24) = "si" Then correcto = correcto + 1
```

Figura 3.4a Código de macros para columnas tipo * .

```
                If Cells(i - 1, 24) = "no" Then incorrecto = incorrecto + 1
                alumno = Cells(i, 2).Value
                i = i + 1
                cont = cont + 1
            Wend
            cont2 = cont2 + cont
            alumnoaux = Cells(cont2 + 1, 2).Value
            alumno = Cells(cont2 + 1, 2).Value
        '_____________________________________Rellena con el valor maximo
            For k = 0 To cont - 1
                    Cells(k + cont2 - cont + 1, 29).Value = apro
                    Cells(k + cont2 - cont + 1, 30).Value = aproR
                    Cells(k + cont2 - cont + 1, 31).Value = aproNr
                    Cells(k + cont2 - cont + 1, 32).Value = repro
                    Cells(k + cont2 - cont + 1, 33).Value = reproR
                    Cells(k + cont2 - cont + 1, 34).Value = reproNr
                    Cells(k + cont2 - cont + 1, 35).Value = rec
                    Cells(k + cont2 - cont + 1, 37).Value = mat
                    Cells(k + cont2 - cont + 1, 38).Value = vesp
                    Cells(k + cont2 - cont + 1, 39).Value = correcto
                    Cells(k + cont2 - cont + 1, 40).Value = incorrecto
            Next k
        Next j

End Sub
```

Figura 3.4b Código de macros para columnas tipo * .

id	plan	ingreso	generac	carrera	avance	promed	Ultimo s	1er Sem	Total se	Total ap
02KcqHH5ouWKianEnr6WOQSu1TY=	119	Concurso	2002	Actuaría	0.977169	8.45	20121	20021	12	39
05BcJZGx3GynpHNFIPALc/aUZZE=	119	Pase Regl	2002	Actuaría	1	8.5	20071	20021	9	40
0A35NJwVxSk+DH4sPtfxlOsJq54=	119	Concurso	2002	Actuaría	1	8.525	20052	20021	8	40
0AsHXEfl6kodG9/eUC3AbFvsJKE=	119	Concurso	2002	Actuaría	1	8.675	20052	20021	8	40
0AU4sg9AeAuhB5B7BNaJl9LtFL4=	119	Cambio de	2002	Actuaría	0.392694	8	20072	20021	11	11
0K0HySzBiB/Lk2vx4A5Tr5cAIRk=	119	Carrera Sir	2002	Actuaría	0.205479	7.333333	20071	20032	6	4
0SE5lNfpTOk0cUQ05Q/2Km5gXv0=	119	Pase Regl	2002	Actuaría	1.013699	8.804878	20111	20021	18	41
0SLrMJY/vSkLdq0GS4TC6KCp3Ql=	119	Pase Regl	2002	Actuaría	1	8	20062	20021	10	40
1as5eNbCd6ucAeM22nRPXFvFV0l=	1176	Cambio Int	2002	Actuaría	0.39726	9.583333	20131	20021	6	12
1EHEyXZz9WUHfWxDkGl0Y+9gjqM=	119	Pase Regl	2002	Actuaría	0.954338	7.25641	20131	20021	19	38
1mzyw3Ytt7aCkcr4DRcJDoE2iPc=	119	Pase Regl	2002	Actuaría	1	9.325	20052	20021	8	40
1ZpuCl72ZleD62R+GdVR5dYSICw=	119	Pase Regl	2002	Actuaría	0.136986	0	20021	20021	1	2
27xQAU0bR3LspCzWUcRlDe4lVDU=	119	Pase Regl	2002	Actuaría	0.643836	8.227273	20052	20021	8	20
/2CoFHqoW5Lty4UElDlFb17rsE0=	119	Pase Regl	2002	Actuaría	1.045662	8.547619	20061	20021	9	42

Figura 3.5 Ejemplo de la tabla Datos Alumnos en formato horizontal

Después de crear todas estas columnas, se procede a dividir la información en distintas tablas de acuerdo con ciertas características, en la primera tabla, que llamaremos “DatosAlumnos” se encuentran las siguientes columnas:

Id, Plan, Ingreso, Gen, Carrera, Avance, Promedio, UltimoSemInscrito, 1erSem, TotalSem, TotalAprobadas, AprobadasYrecursadas, AprobadasNoRecursadas, TotalReprobadas, ReprobadasYrecursadas, ReprobadasNoRecursadas, TotalRecursadas, Recursamientos, Matutino, Vespertino, InscritasBien, InscritasMal, Años y SitEscolar dividida en Estudia, Termino y Desertó.

Como se puede observar, toda la información sobre las materias se dejó fuera de la tala “Datos Alumnos”, la razón de hacerlo fue que los alumnos están inscritos en distintos planes por lo que al cambiar al formato horizontal se requirió hacer otras 4 tablas, una por cada tipo de plan. Cada tabla contiene las columnas id, SitEscolar y las correspondientes columnas de cada nombre de materia (contienen 0 si está reprobada, 1 si se aprobó y 2 si no se ha inscrito), así mismo contiene la información sobre el grupo en que se inscribió cada materia (se pone NI si no está inscrito), su tipo de inscripción (se pondrá NI si el alumno no ha inscrito esa materia) y las veces que se ha cursado la materia.

Cabe señalar que no todas las calificaciones son numéricas, algunas contenían valores como NA, NP, AC, B, MB y S, dichas calificaciones se cambiaron a calificaciones numéricas con base a la información encontrada en el reglamento de la UNAM encontrado en la página de internet https://www.dgae.unam.mx/normativ/legislacion/regexa78.html, quedando de la siguiente manera:

> NP=NA=5 S=6 B=8 MB=10
> AC: Representa una materia que no cuenta para promedio pero que debe acreditarse. Para efectos de este trabajo se le asignará el número 1.

El nombre de cada tabla llevará el nombre de la siguiente manera: Act2000, Act2006, Comp1994, Comp20013. Se dejan fuera las materias optativas ya que el estudio se concentrará en las materias obligatorias que afectan directamente el desempeño de los alumnos.

Una vez organizada la información de esta manera, se procedió a cambiar el formato de todas las tablas guardadas como libros de Excel habilitadas para macros, a formato txt delimitado por tabulaciones.

3.3.1 Análisis Exploratorio

Previo a la realización de la fase de minería de datos, es importante efectuar un análisis estadístico exploratorio, con la finalidad de obtener una idea general del comportamiento de los datos y así tomar decisiones sobre qué variables se incluirán en los modelos de minería y cuáles no dependiendo de su correlación y valor explicativo. Así mismo, saber qué variables requieren un tratamiento especial como por ejemplo transformaciones, tratamiento de missing y outliers, etc.
Cabe señalar que éste análisis exploratorio es elaborado únicamente con estadística descriptiva, ya que no es el objetivo de este trabajo llegar a resultados y conclusiones usando esta herramienta si no únicamente lograr tener un mayor entendimiento de los datos.

Los datos serán importados desde archivos con extensión .txt al programa estadístico R, se analizarán las bases "Datos Alumnos", "Act00", "Act06" y "Comp1994", cada una por separado, utilizando herramientas analíticas y gráficas. Dichas bases fueron modificadas ligeramente en esta etapa de modo que únicamente se tomaran en cuenta las generaciones de 2002 a 2011 ya que por el objeto de estudio y el criterio de deserción (4 semestres no inscritos) las generaciones 2012 y 2013 no aportan información relevante pero si afectan el peso de algunos resultados.

Las cinco bases de datos contienen variables numéricas y categóricas por lo que el análisis no se puede realizar con las mismas herramientas ya que existen técnicas especiales para cada tipo, por consiguiente, en cada base se realizará el análisis en tres fases: a) Global, b) De las variables categóricas y c) De las variables numéricas.

Base "Datos Alumnos":

Análisis Global

Se comienza por examinar la base y sus atributos, conociendo los nombres de cada una de las columnas de la tabla y su dimensión que consta de 4407 registros y 27 columnas cuyos nombres aparecen en la Figura 3.6.

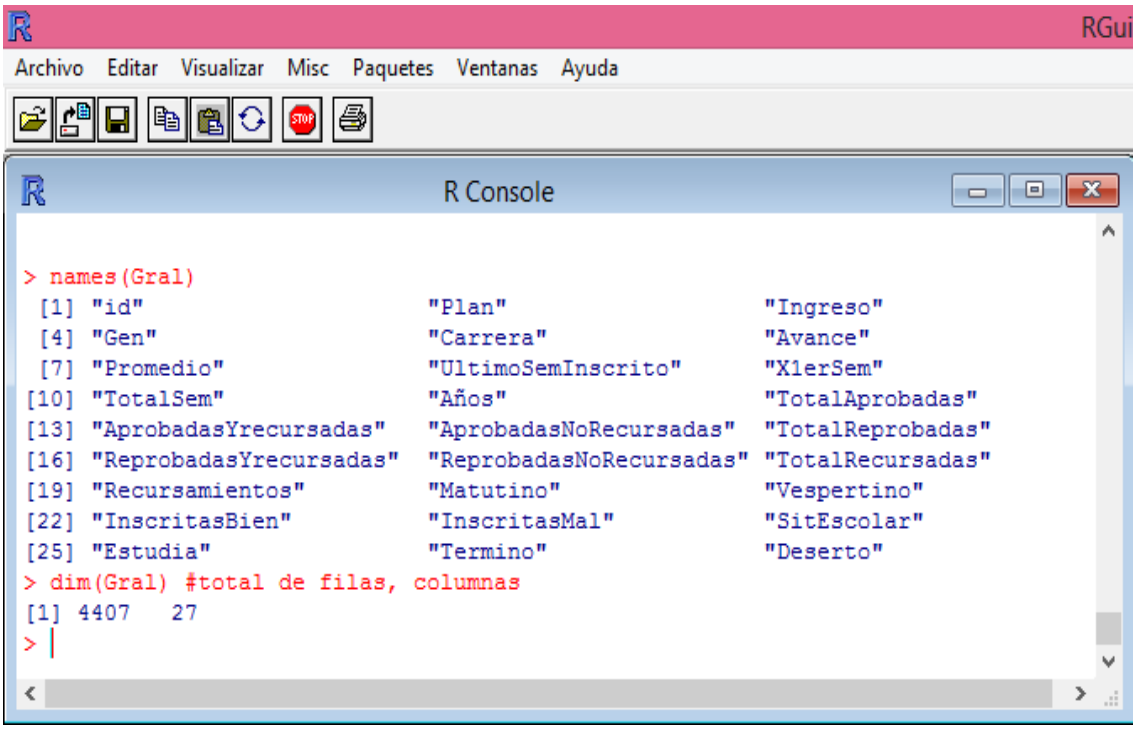
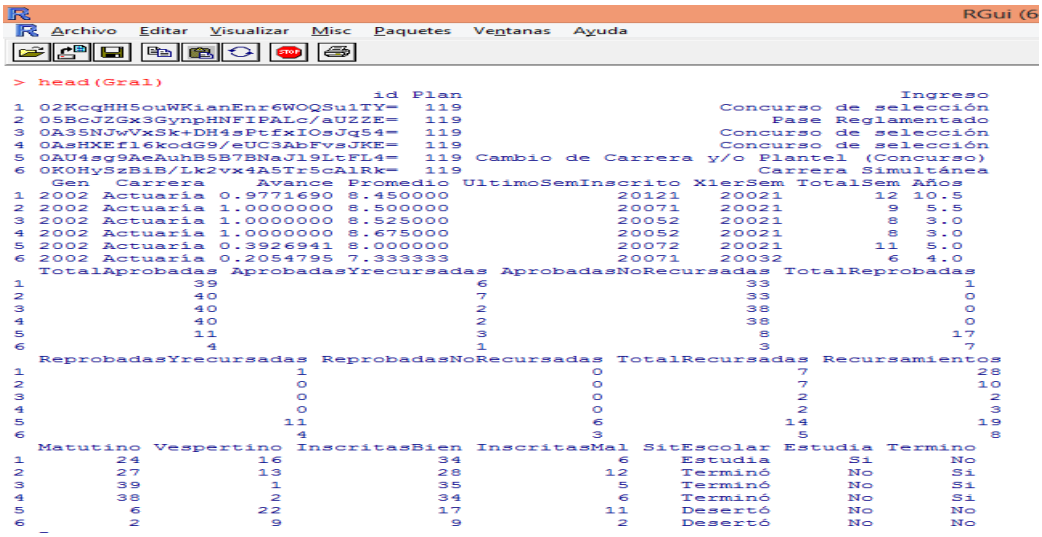


Figura 3.6 Nombres y dimensión de la tabla.

A continuación se obtiene la información que contiene la tabla en sus primeros 6 registros para conocer cómo está acomodada su información (Ver Figura 3.7), así mismo se realiza un análisis del tipo de datos que contiene cada una de las columnas, sabiendo si son variables de tipo categórico o numérico.

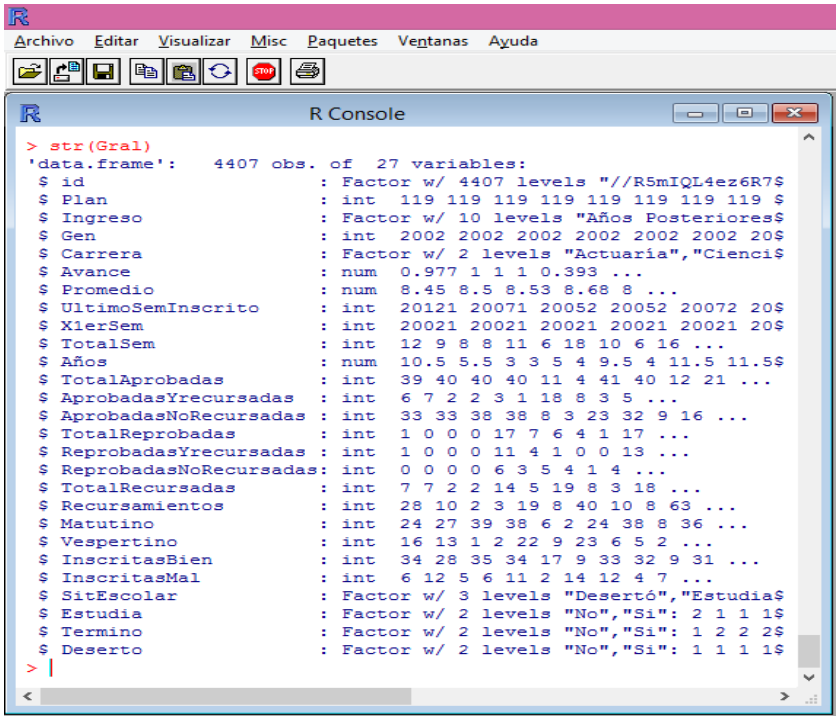
En la Figura 3.8 se puede observar que existen varios campos que R interpreta como numéricos (Int) y que en realidad son campos categóricos (Factor), por lo cual se procede a modificar los campos de Gen, Plan, UltimoSemInscrito y 1erSem para que el programa pueda identificarlos como datos categóricos y se corrobora que el programa realizó la operación correctamente, lo cual puede verse en la Figura 3.9.



```
RGui (6
Archivo  Editar  Visualizar  Misc  Paquetes  Ventanas  Ayuda

> head(Gral)
                                id Plan                                  Ingreso
1 02KcqHH5ouWKianEnr6WOQSu1TY=  119                      Concurso de selección
2 05BcJZGx3GynpHNFIPALc/aUZZE=  119                         Pase Reglamentado
3 0A35NJwVxSk+DH4sPtfxIOsJq54=  119                      Concurso de selección
4 0AsHXEf16kodG9/eUC3AbFvsJKE=  119                      Concurso de selección
5 0AU4sg9AeAuhB5B7BNaJl9LtFL4=  119 Cambio de Carrera y/o Plantel (Concurso)
6 0KOHySzBiB/Lk2vx4A5Tr5cAlRk=  119                          Carrera Simultánea
  Gen   Carrera    Avance Promedio UltimoSemInscrito X1erSem TotalSem Años
1 2002 Actuaria 0.9771690 8.450000             20121   20021       12 10.5
2 2002 Actuaria 1.0000000 8.500000             20071   20021        9  5.5
3 2002 Actuaria 1.0000000 8.525000             20052   20021        8  3.0
4 2002 Actuaria 1.0000000 8.675000             20052   20021        8  3.0
5 2002 Actuaria 0.3926941 8.000000             20072   20021       11  5.0
6 2002 Actuaria 0.2054795 7.333333             20071   20032        6  4.0
  TotalAprobadas AprobadasYrecursadas AprobadasNoRecursadas TotalReprobadas
1             39                    6                    33               1
2             40                    7                    33               0
3             40                    2                    38               0
4             40                    2                    38               0
5             11                    3                     8              17
6              4                    1                     3               7
  ReprobadasYrecursadas ReprobadasNoRecursadas TotalRecursadas Recursamientos
1                     1                     0                7             28
2                     0                     0                7             10
3                     0                     0                2              2
4                     0                     0                2              3
5                    11                     6               14             19
6                     4                     3                5              8
  Matutino Vespertino InscritasBien InscritasMal SitEscolar Estudia Termino
1       24         16            34            6    Estudia      Si      No
2       27         13            28           12    Terminó      No      Si
3       39          1            35            5    Terminó      No      Si
4       38          2            34            6    Terminó      No      Si
5        6         22            17           11    Desertó      No      No
6        2          9             9            2    Desertó      No      No
>
```

Figura 3.7 “Cabeza de la tabla Datos Alumnos”



```
Archivo  Editar  Visualizar  Misc  Paquetes  Ventanas  Ayuda

R Console

> str(Gral)
'data.frame':   4407 obs. of  27 variables:
 $ id                    : Factor w/ 4407 levels "//R5mIQL4ez6R7$
 $ Plan                  : int  119 119 119 119 119 119 119 119 $
 $ Ingreso               : Factor w/ 10 levels "Años Posteriores$
 $ Gen                   : int  2002 2002 2002 2002 2002 2002 20$
 $ Carrera               : Factor w/ 2 levels "Actuaría","Cienci$
 $ Avance                : num  0.977 1 1 1 0.393 ...
 $ Promedio              : num  8.45 8.5 8.53 8.68 8 ...
 $ UltimoSemInscrito     : int  20121 20071 20052 20052 20072 20$
 $ X1erSem               : int  20021 20021 20021 20021 20021 20$
 $ TotalSem              : int  12 9 8 8 11 6 18 10 6 16 ...
 $ Años                  : num  10.5 5.5 3 3 5 4 9.5 4 11.5 11.5$
 $ TotalAprobadas        : int  39 40 40 40 11 4 41 40 12 21 ...
 $ AprobadasYrecursadas  : int  6 7 2 2 3 1 18 8 3 5 ...
 $ AprobadasNoRecursadas : int  33 33 38 38 8 3 23 32 9 16 ...
 $ TotalReprobadas       : int  1 0 0 0 17 7 6 4 1 17 ...
 $ ReprobadasYrecursadas : int  1 0 0 0 11 4 1 0 0 13 ...
 $ ReprobadasNoRecursadas: int  0 0 0 0 6 3 5 4 1 4 ...
 $ TotalRecursadas       : int  7 7 2 2 14 5 19 8 3 18 ...
 $ Recursamientos        : int  28 10 2 3 19 8 40 10 8 63 ...
 $ Matutino              : int  24 27 39 38 6 2 24 38 8 36 ...
 $ Vespertino            : int  16 13 1 2 22 9 23 6 5 2 ...
 $ InscritasBien         : int  34 28 35 34 17 9 33 32 9 31 ...
 $ InscritasMal          : int  6 12 5 6 11 2 14 12 4 7 ...
 $ SitEscolar            : Factor w/ 3 levels "Desertó","Estudia$
 $ Estudia               : Factor w/ 2 levels "No","Si": 2 1 1 1$
 $ Termino               : Factor w/ 2 levels "No","Si": 1 2 2 2$
 $ Deserto               : Factor w/ 2 levels "No","Si": 1 1 1 1$
>
```

Figura 3.8 Tipo de información contenida en la tabla.

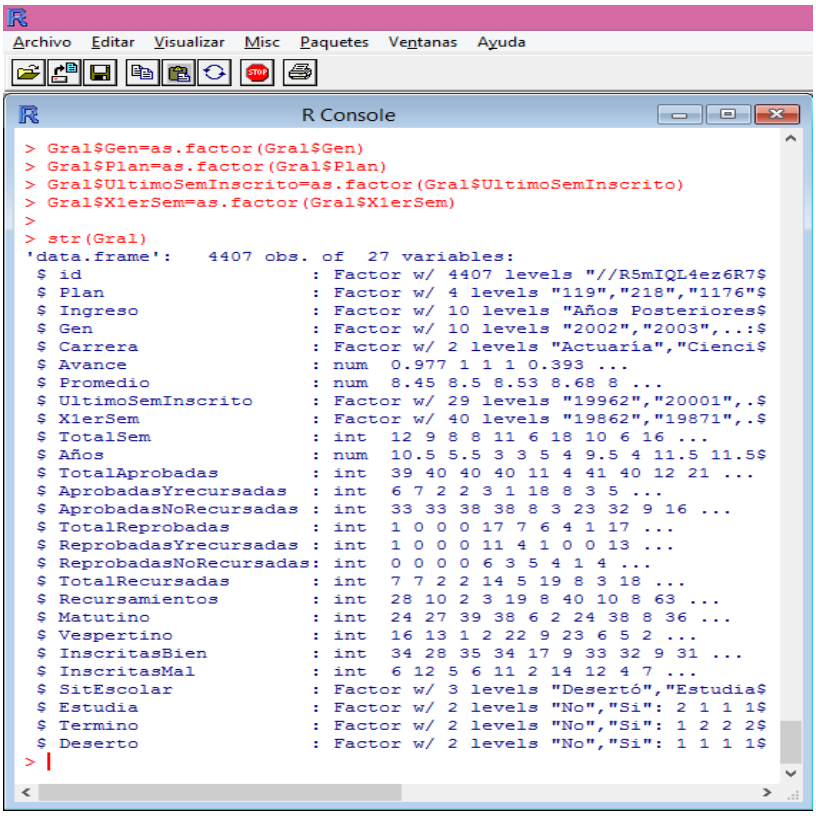


Figura 3.9 Tipo de información de la tabla después de la modificación adecuada

Una vez realizada esta modificación se puede proseguir a realizar el análisis descriptivo de la información, primeramente se obtiene un resumen estadístico con la función summary() de R.

El resumen estadístico arroja varios resultados interesantes que nos ayudan a conocer mejor el comportamiento de los alumnos de las carreras de Actuaría y Ciencias de la Computación en la Facultad de Ciencias. En las figuras 3.10a y 3.10b se puede observar un análisis global de 4407 alumnos con 26 atributos cada uno colocados en columnas.

Se observa que de acuerdo al plan se tienen 1217 alumnos de Actuaría plan 2000, 2285 alumnos de Actuaría plan 2006, 902 alumnos de Ciencias de la Computación plan 1994 y 3 alumnos de Ciencias de la computación plan 2013, lo que nos dice que en nuestra base de datos, el 79.46% de los alumnos son de la carrera de Actuaría (3502 alumnos) y el 20.54% son de Ciencias de la Computación (905 alumnos).

En el resumen del tipo de ingreso se observa que la mayoría de los alumnos que ingresan a la facultad lo hacen por la modalidad de pase reglamentario, siendo éstos un 65% de la población aproximadamente, mientras que el 29% lo hacen por

alguna modalidad de concurso de selección y el resto por las diferentes modalidades que ofrece la facultad para ingresar a ella.

Otro dato interesante que se puede observar es que en la totalidad de semestres inscritos el promedio es de 7.9 semestres por lo que se puede suponer que la mayoría de los alumnos estuvieron o están inscritos en la facultad 4 años, concluyendo su carrera o continuando con sus estudios. Pero el dato que rebota es el del máximo de semestres inscritos que es 26 por lo que se investigó ese registro y se trata de un alumno de Ciencias de la Computación que concluyó ya sus estudios y fue segunda carrera por lo que puede suponerse que las materias de tronco común de su primera carrea fueron inscritas semestres antes de incorporarse a computación y al ser revalidadas arroja un mayor número a esta variable. Mismo caso para los alumnos que rebasan los 12 semestres inscritos.

La variable años confirma que la estancia de los alumnos en la facultad es de 4 años en promedio, pero tenemos un dato con 24 años, al corroborar la información se observa que es un alumno que ingresó por examen de selección, generación 2007 y con primer semestre inscrito 19862, por lo que se intuye que existe un error en su primer semestre inscrito. Se modifica y se le colocan los 6.5 años según sus semestres inscritos que son 13.

En cuanto al promedio, notamos que la media se encuentra en 7.41, dato que arroja una población de alumnos con un aprovechamiento escolar, como el primer cuartil es de 7.42, los alumnos con un promedio bajo son pocos, misma situación con los alumnos de excelencia con calificaciones superiores al tercer cuartil de 8.6.

Ahora en el apartado de recursamientos se observa que el total de materias recursadas tiene un promedio es de 7.3, lo que representa un 20% aproximadamente de la carga académica. Así también el número de recursamientos tiene un promedio de 12.13, lo cual podría indicar que las materias que son recursadas tienden a tener más de un recursamiento.

También se puede corroborar que la mayoría de los alumnos prefiere inscribir materias en el turno matutino debido a que la media de materias inscritas en ese turno (21) es muy superior a la media del turno vespertino (7.78). Y contrario tal vez a lo que podría suponerse, la tendencia en inscribir materias en el semestre que le corresponde según el plan de estudios (par o impar), con promedio de 22 materias, es muy superior a la de inscribir materias en semestre erróneo cuyo promedio es de 7 materias.

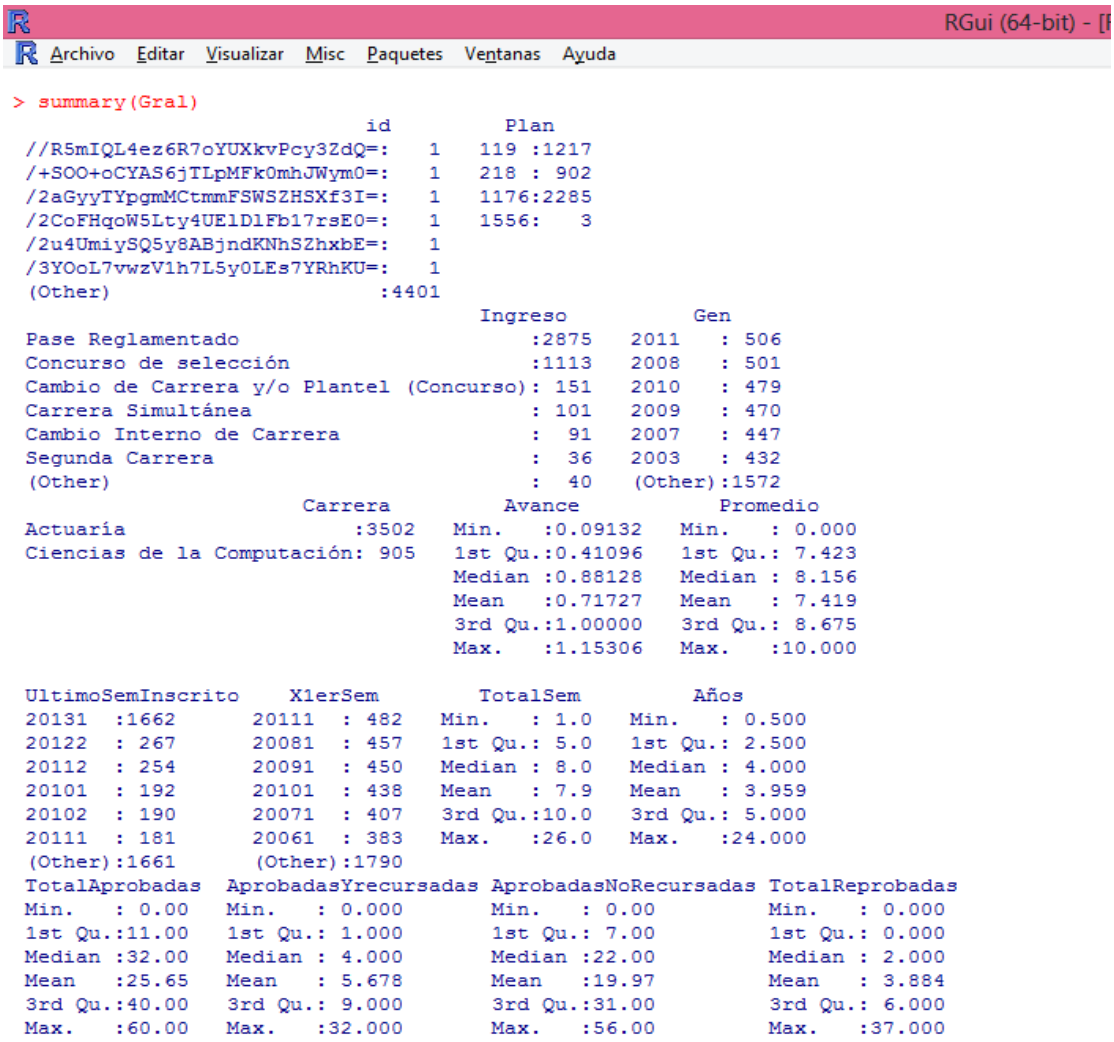
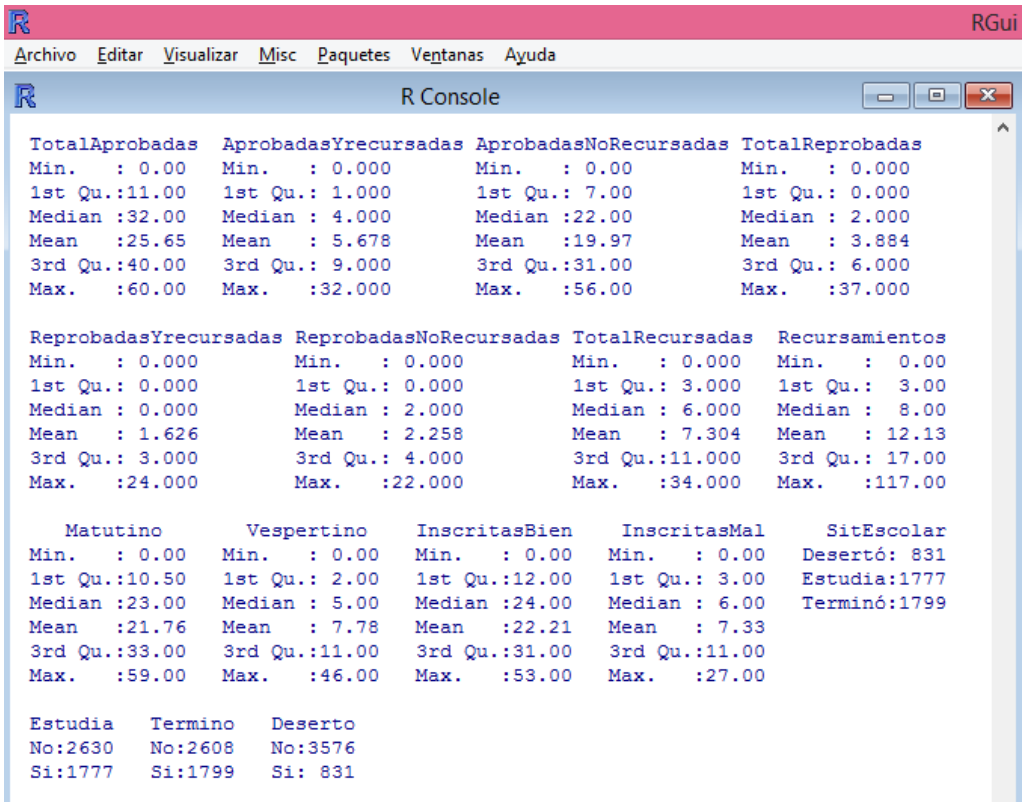


Figura 3.10a Resumen estadístico de la base “Datos Alumnos” (Parte 1)

Por supuesto no se puede olvidar el dato más importante en este trabajo, la situación escolar, en la cual tenemos que de 4407 alumnos de las generaciones 2002 a 2011, el 40.82% de los alumnos han concluido su carrera, el 40.32% sigue estudiando y el 18.85% han desertado de acuerdo al criterio mencionado al inicio de este capítulo. Más adelante se analizarán estas cifras por carrera.




```
TotalAprobadas  AprobadasYrecursadas AprobadasNoRecursadas TotalReprobadas
Min.   : 0.00   Min.   : 0.000       Min.   : 0.00         Min.   : 0.000
1st Qu.:11.00   1st Qu.: 1.000       1st Qu.: 7.00         1st Qu.: 0.000
Median :32.00   Median : 4.000       Median :22.00         Median : 2.000
Mean   :25.65   Mean   : 5.678       Mean   :19.97         Mean   : 3.884
3rd Qu.:40.00   3rd Qu.: 9.000       3rd Qu.:31.00         3rd Qu.: 6.000
Max.   :60.00   Max.   :32.000       Max.   :56.00         Max.   :37.000

ReprobadasYrecursadas ReprobadasNoRecursadas TotalRecursadas  Recursamientos
Min.   : 0.000        Min.   : 0.000         Min.   : 0.000   Min.   :  0.00
1st Qu.: 0.000        1st Qu.: 0.000         1st Qu.: 3.000   1st Qu.:  3.00
Median : 0.000        Median : 2.000         Median : 6.000   Median :  8.00
Mean   : 1.626        Mean   : 2.258         Mean   : 7.304   Mean   : 12.13
3rd Qu.: 3.000        3rd Qu.: 4.000         3rd Qu.:11.000   3rd Qu.: 17.00
Max.   :24.000        Max.   :22.000         Max.   :34.000   Max.   :117.00

   Matutino       Vespertino    InscritasBien    InscritasMal     SitEscolar
Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Desertó: 831
1st Qu.:10.50   1st Qu.: 2.00   1st Qu.:12.00   1st Qu.: 3.00   Estudia:1777
Median :23.00   Median : 5.00   Median :24.00   Median : 6.00   Terminó:1799
Mean   :21.76   Mean   : 7.78   Mean   :22.21   Mean   : 7.33
3rd Qu.:33.00   3rd Qu.:11.00   3rd Qu.:31.00   3rd Qu.:11.00
Max.   :59.00   Max.   :46.00   Max.   :53.00   Max.   :27.00

Estudia   Termino   Deserto
No:2630   No:2608   No:3576
Si:1777   Si:1799   Si: 831
```

Figura 3.10b Resumen estadístico de la base “Datos Alumnos” (Parte 2)

En la figura 3.11 se encuentran algunas gráficas de barras que nos ayudan a ver de una manera más clara los atributos de carrera, situación escolar, plan de estudios y generación en donde las barras representan cada una de las generaciones y su población. Y en la gráfica 3.12 se muestra dos gráficos con el tipo de ingreso, donde la mayor población entra a la facultad por pase reglamentado.

Figura 3.11 Gráficos de barra de la base “Datos Alumnos”

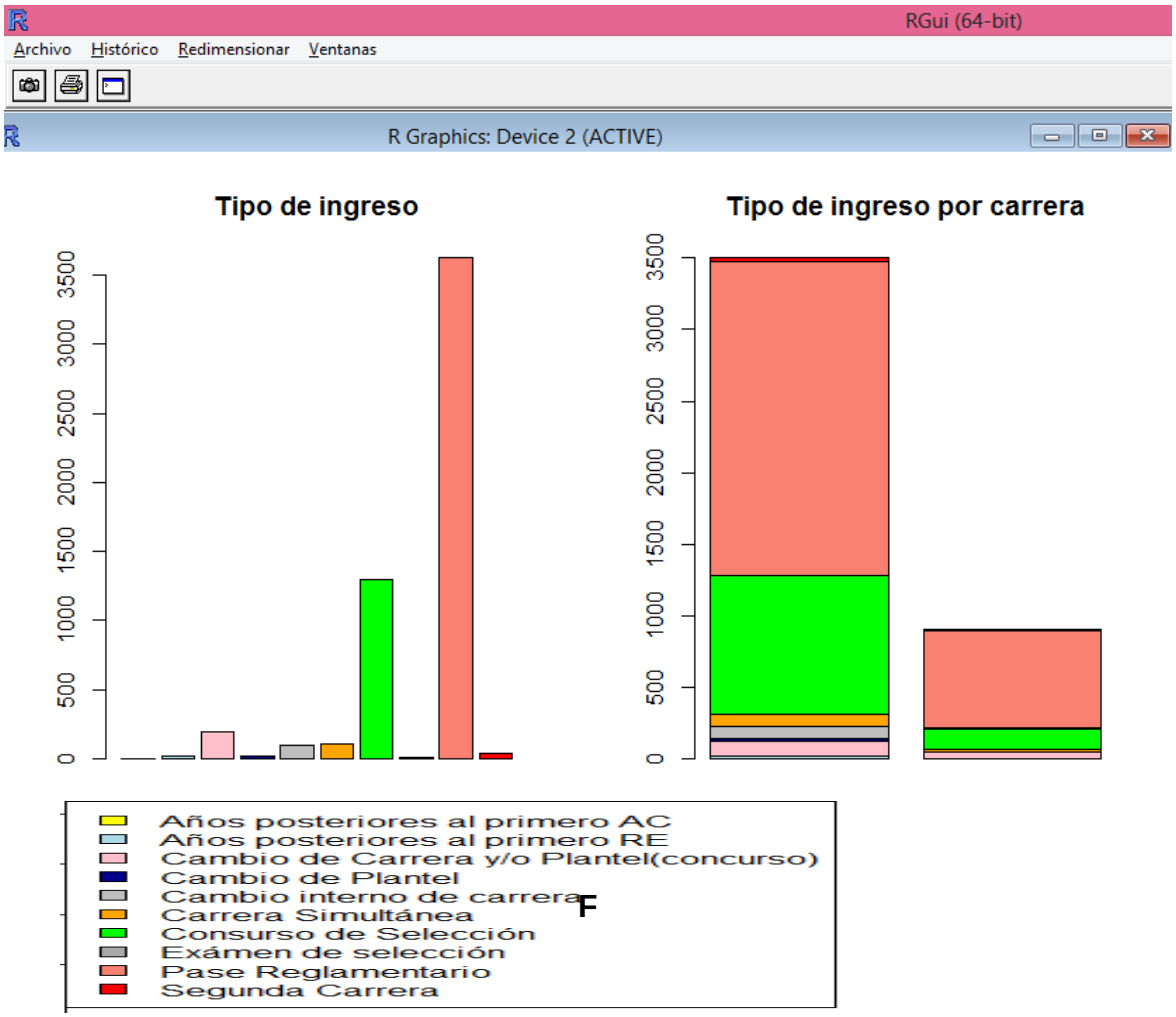


Figura 3.12 Grafico de barras del Tipo de ingreso a la Facultad de Ciencias

Para conocer las cantidades exactas de los gráficos de la figura 3.12 se presentan las tablas de frecuencia de cada una en la siguiente figura. Aquí se puede resumir que en general, el 65.24% de los alumnos ingresan por pase reglamentado, el 25.41% ingresan por modalidad de selección, el 5.88% por cambio de carrera o plantel y el resto por otras modalidades. De modo más específico, en Actuaría el 62.56% ingresan por pase reglamentado y el 27.7% por algún tipo de selección, el resto en otras modalidades; en Ciencias de la computación75.6% ingresan por pase reglamentado y el 16.6% por algún tipo de selección y el resto por otras modalidades.

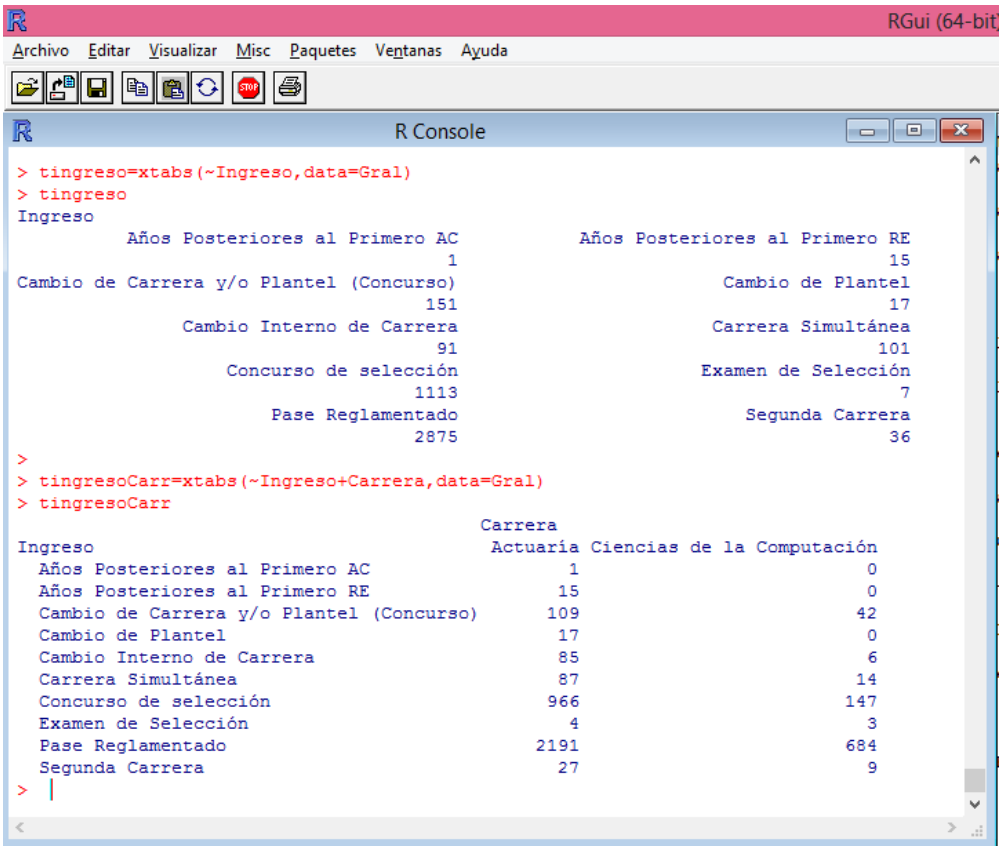


Figura 3.13 Tipo de ingreso a la Facultad de Ciencias

Análisis de las Variables Categóricas

Para este tipo de datos se realiza un análisis con tablas de contingencia para conocer la frecuencia de los datos en cada cruce de variables, una vez realizada ésta operación es importante conocer si las variables que se cruzan mantienen algún tipo de asociación, para ello se realiza una prueba de Ji-Cuadrada, en caso de tener salir positiva la asociación, se obtiene el coeficiente de Cramer. (Recordar que en la prueba Ji-Cuadrada la hipótesis nula es que no hay asociación entre las variables y se rechaza cuando el P-value <0.5 y el coeficiente de Cramer indica una fuerte asociación cuando es mayor a 0.3).

En la Figura 3.14a se puede observar de manera detallada la relación Plan-Ingreso, Plan-Gen y Plan-Carrera, ya que muestra la frecuencia de alumnos en cada categoría, pero no aportan nueva información únicamente se vuelve más específica. Se observa que las generaciones con mayor ingreso de alumnos fueron 2008 y 2011 para Actuaría, mientras que para Ciencias de la Computación fueron las generaciones 2003 y 2011.

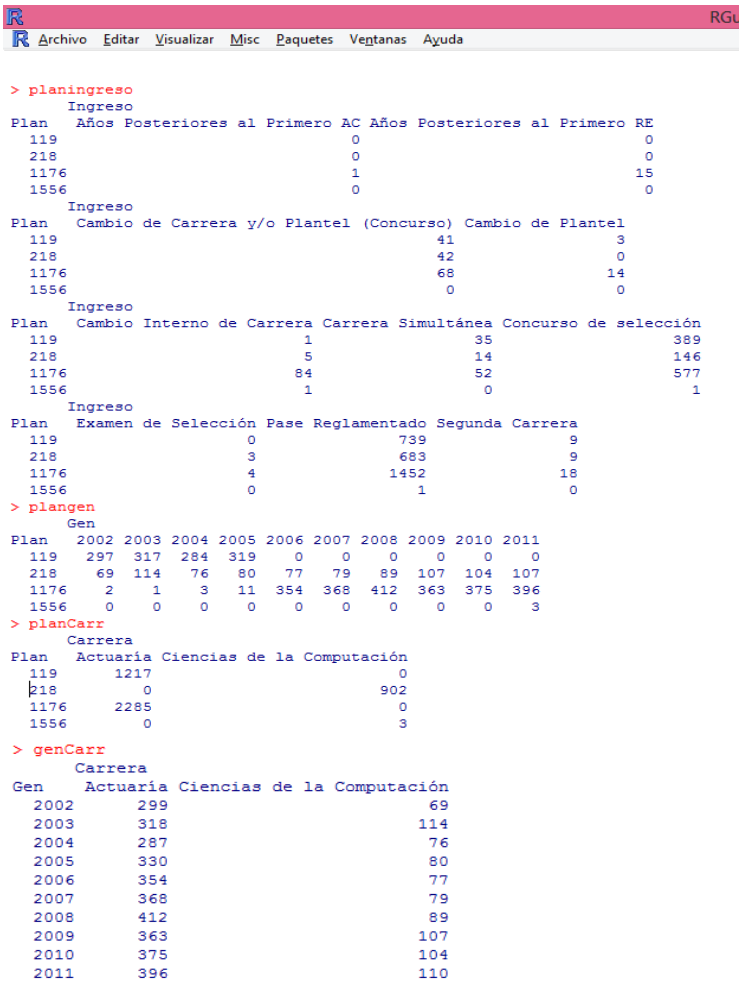


Figura 3.14a Tablas de contingencia de la base “Datos Alumnos” (Parte 1)

De acuerdo al plan de estudios y la situación escolar, de los alumnos de Actuaría han desertado 294 del plan 2000 y 262 del plan 2006, mientras que en Ciencias de la Computación han desertado 275 del plan 1994 y 0 del plan 2013. Por lo anterior se resume que el 15.87% de los actuarios han desertado y el 30.38% de los alumnos de ciencias de la computación han desertado (Ver Figura 3.14b).

Las generaciones con mayor deserción son 2002 y 2003 para Actuaría con 87 y 82 deserciones respectivamente y para Ciencias de la Computación es la generación 2003 con 50 deserciones.

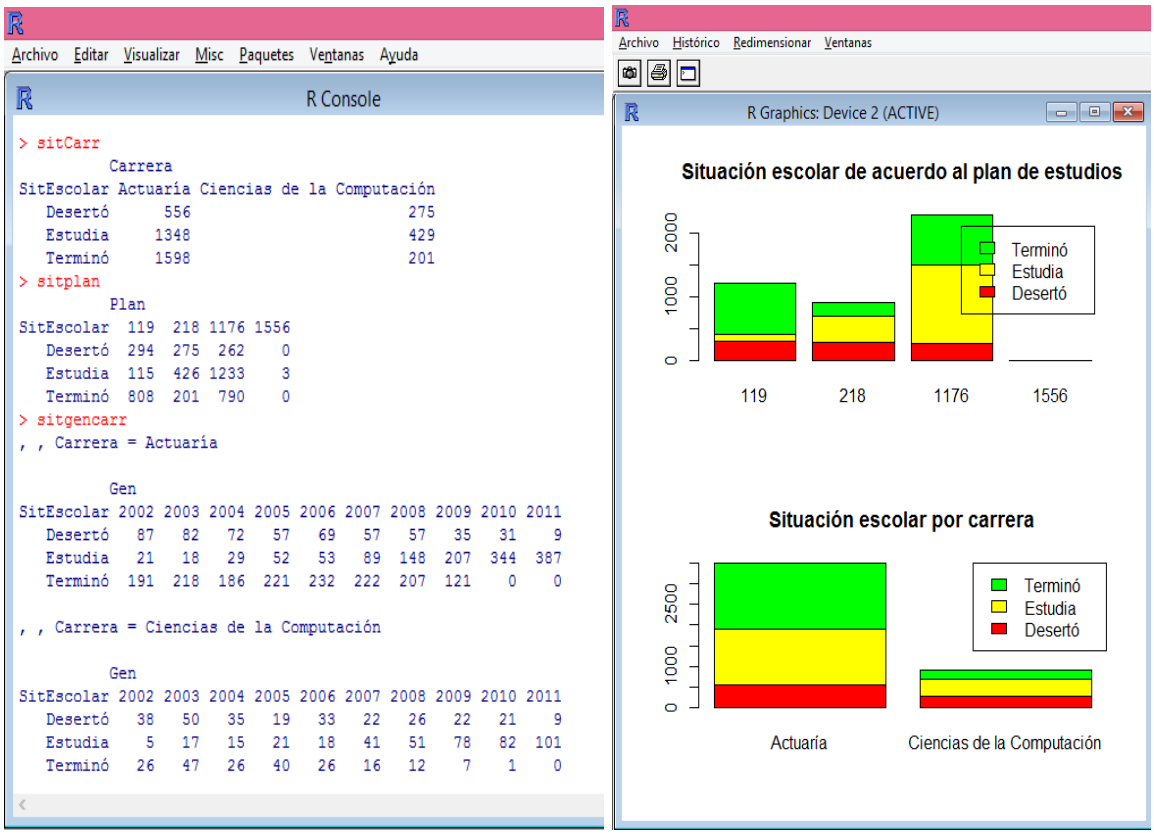
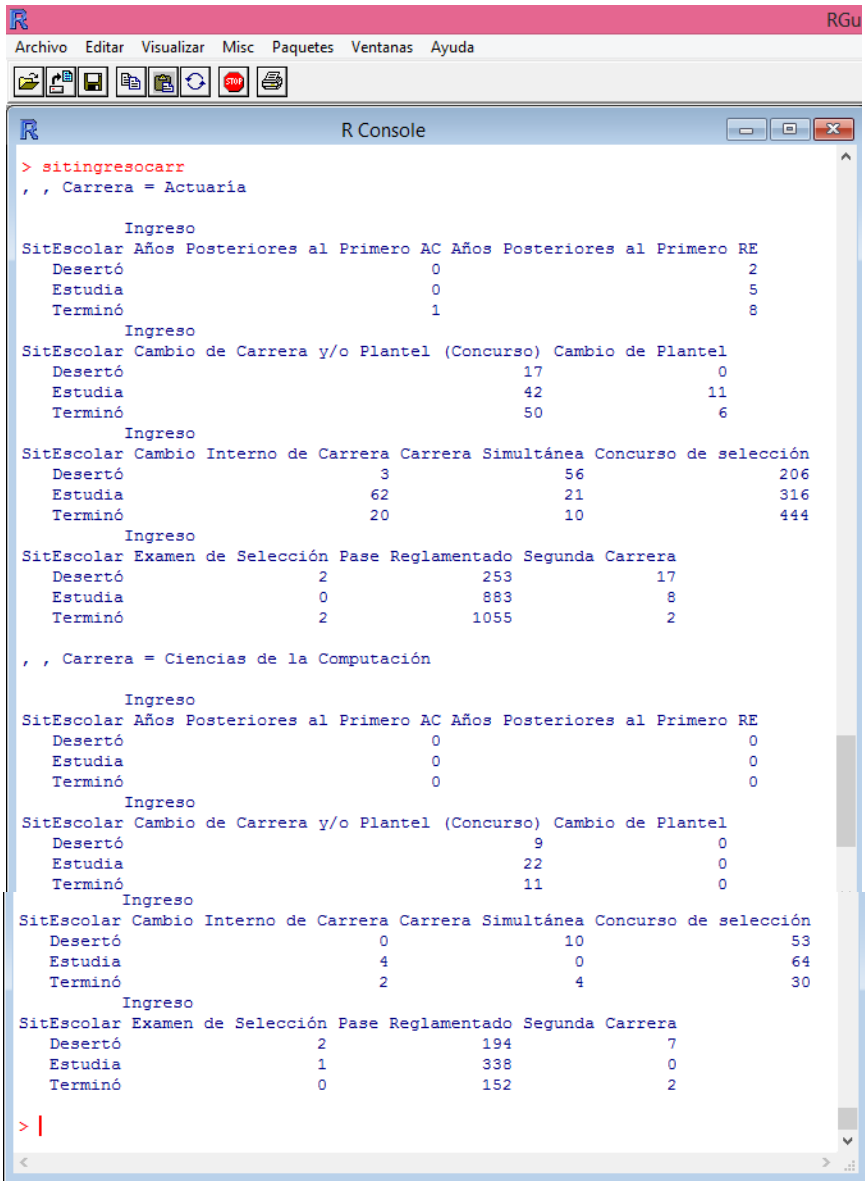


Figura 3.14b Tablas de contingencia de la base "Datos Alumnos" (Parte 2)

Unos datos de mucho interés son probablemente los índices de deserción por carrera por forma de ingreso, dicho análisis se realiza en la Figura 3.14c y se obtienen las siguientes estadísticas:

En la carrera de Actuaría, el mayor índice de deserción se encuentra en el ingreso por carrera simultánea con 64.36%, seguida del ingreso por segunda carrera con 62.96%, concurso de selección y pase reglamentado tienen índices de 20.68% y 11.5% respectivamente. La modalidad de cambio interno de carrera es la de menor deserción con un índice del 3.5%.

En Ciencias de la computación, los mayores índices están en las modalidades de segunda carrera (77.77%) y carrera simultánea (71.42%), concurso de selección y pase reglamentado tienen índices de 36.05% y 28.36% respectivamente. La modalidad de cambio interno de carrera es la de menor deserción (0%).




```
> sitingresocarr
, , Carrera = Actuaría

          Ingreso
SitEscolar Años Posteriores al Primero AC Años Posteriores al Primero RE
   Desertó                             0                               2
   Estudia                             0                               5
   Terminó                             1                               8
          Ingreso
SitEscolar Cambio de Carrera y/o Plantel (Concurso) Cambio de Plantel
   Desertó                                       17                 0
   Estudia                                       42                11
   Terminó                                       50                 6
          Ingreso
SitEscolar Cambio Interno de Carrera Carrera Simultánea Concurso de selección
   Desertó                         3                 56                   206
   Estudia                        62                 21                   316
   Terminó                        20                 10                   444
          Ingreso
SitEscolar Examen de Selección Pase Reglamentado Segunda Carrera
   Desertó                  2               253              17
   Estudia                  0               883               8
   Terminó                  2              1055               2

, , Carrera = Ciencias de la Computación

          Ingreso
SitEscolar Años Posteriores al Primero AC Años Posteriores al Primero RE
   Desertó                             0                               0
   Estudia                             0                               0
   Terminó                             0                               0
          Ingreso
SitEscolar Cambio de Carrera y/o Plantel (Concurso) Cambio de Plantel
   Desertó                                        9                 0
   Estudia                                       22                 0
   Terminó                                       11                 0
          Ingreso
SitEscolar Cambio Interno de Carrera Carrera Simultánea Concurso de selección
   Desertó                         0                 10                    53
   Estudia                         4                  0                    64
   Terminó                         2                  4                    30
          Ingreso
SitEscolar Examen de Selección Pase Reglamentado Segunda Carrera
   Desertó                  2               194               7
   Estudia                  1               338               0
   Terminó                  0               152               2

>
```

Figura 3.14c Tablas de contingencia de la base “Datos Alumnos” (Parte 3)

Se realizan a continuación las pruebas Ji-cuadrada en cada uno de los cruces posibles, dando como resultado en todas un P-value menor a 0.5 lo cual nos indica que existe algún tipo de asociación entre las variables, se procede entonces a obtener el coeficiente de Cramer en donde los cruces con un índice mayor a 0.3 fueron Plan-Carrera (1), Plan-Gen (0.51), SitEscolar-Gen (0.47) y SitEscolar-Plan (0.311), esto nos indica que estas variables están fuertemente relacionadas y resta ver si esta relación implica que las variables están dando el mismo tipo de información o es precisamente esa relación la que buscaremos analizar. La única relación que podría dar la misma información es Plan-Carrera por lo que la variable Carrera no se incluirá para explicar el modelo.

En las figuras 3.15a, 3.15b y 3.15c se pueden ver los resultados obtenidos en los P-values y coeficientes de Cramer para los 10 cruces de variables realizados.

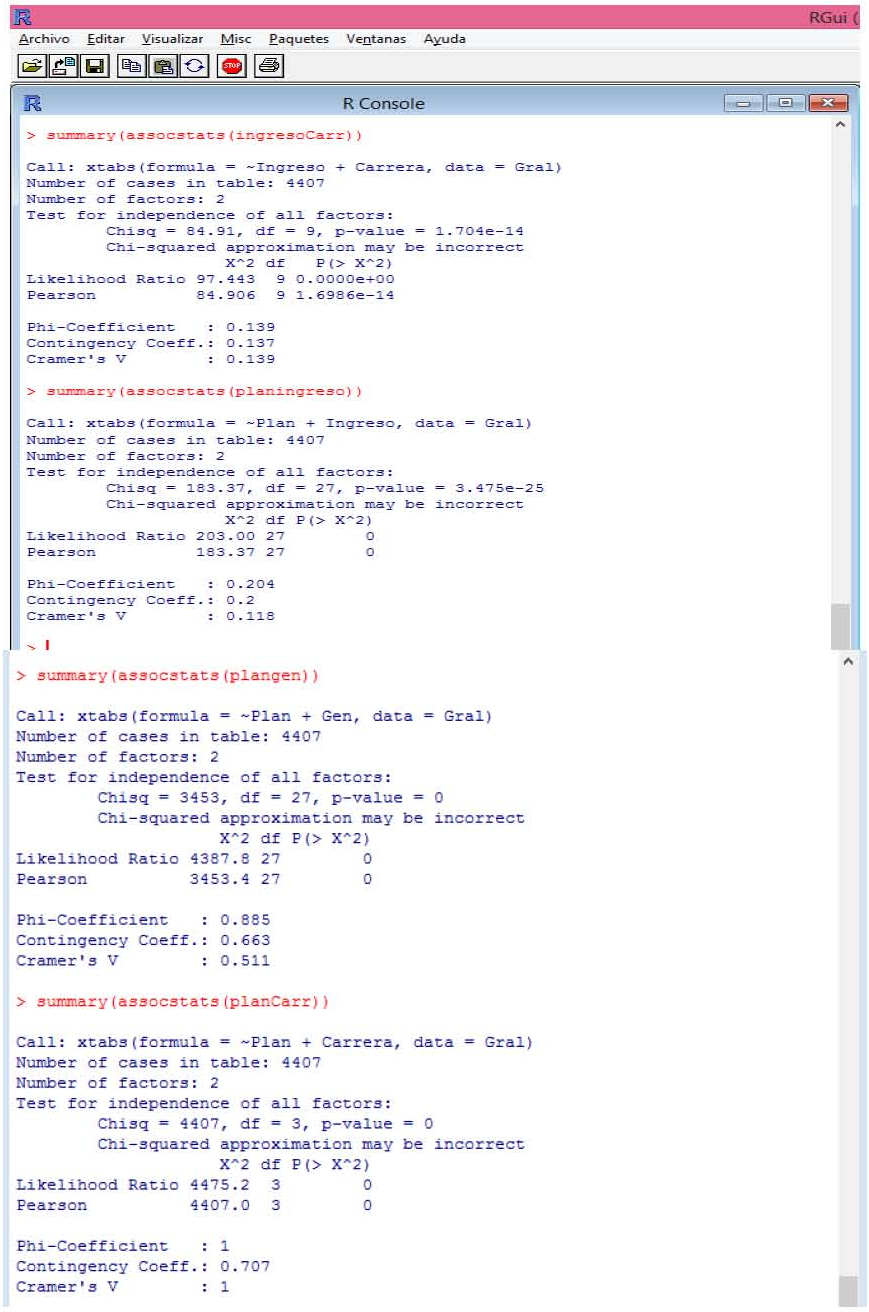


Figura 3.15a Análisis de Ji-cuadrada y coeficiente de Cramer (Parte1)

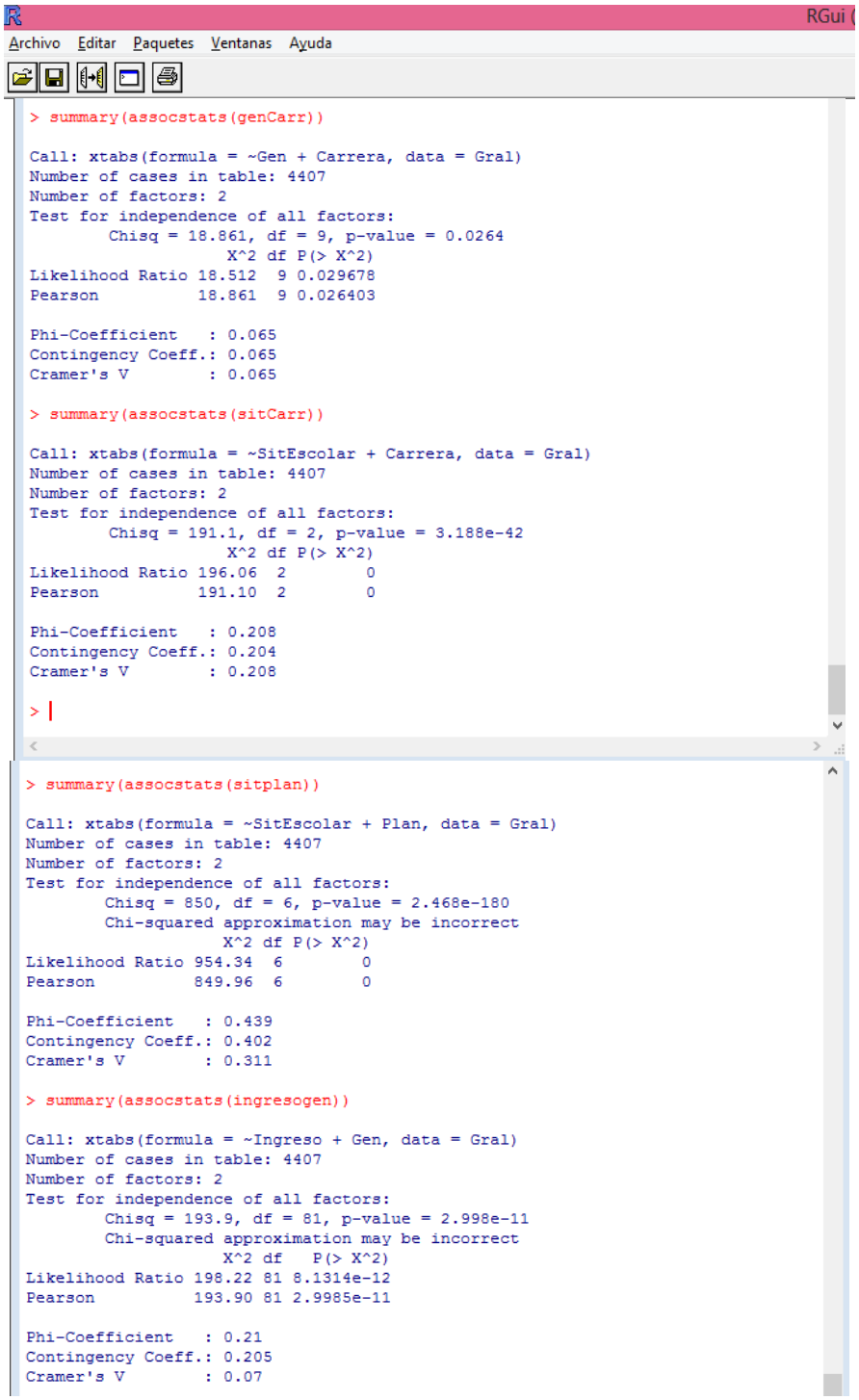


Figura 3.15b Análisis de Ji-cuadrada y coeficiente de Cramer (Parte 2)

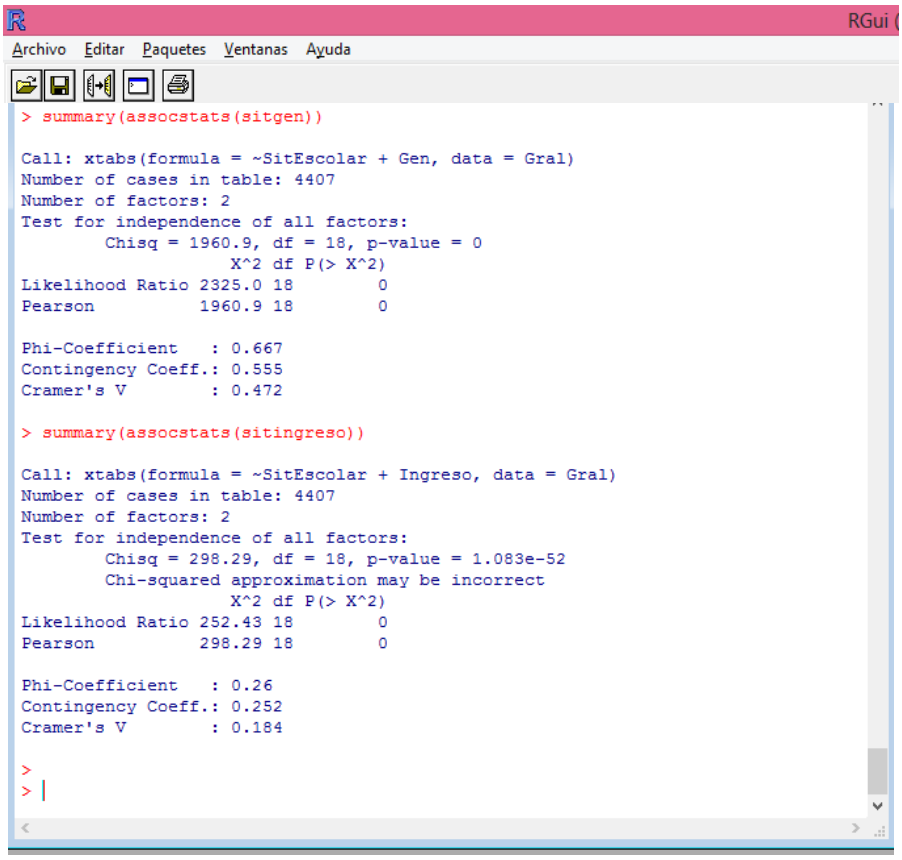


Figura 3.15c Análisis de Ji-cuadrada y coeficiente de Cramer (Parte3)

Análisis de las Variables Numéricas

Para las variables categóricas, es importante conocer la distribución de los datos y que tan dispersos se encuentran los datos y en caso de tener sesgo (con distribución cargada a la derecha o izuierda) o estar demasiado dispersos aplicarles algún tipo de transformación antes de introducirlos en el modelo, así mismo hacer un análisis de correlación para sacar variables que contengan el mismo tipo de información y finalmente un análisis de valores extremos. Las variables numéricas a analizar en esta sección son 16: Avance, PromedioGral, TotalSem, Años, TotalAprobadas, AprobadasYrecursadas, AprobadasNoRecursadas, TotalReprobadas, ReprobadasYrecursadas, ReprobadasNoRecursadas, TotalRecursadas, Recursamientos, Matutino, Vespertino, InscritasBien e InscritasMal.

Para conocer las medias del comportamiento de los alumnos a la hora de desertar, se realiza la tabla mostrada en la Figura 3.16a, aquí se puede ver que la media de deserción se encuentra en el 25.82% de avance académico, con promedios sumamente bajos (4.13) en un tiempo de 3.9 semestres con un total de materias aprobadas de 5.2, la mayoría inscritas en el turno matutino e inscritas

mayormente en los semestres correspondientes. Por otro lado los alumnos que terminan la carrera llevan un promedio bueno (8.5), terminando la carrera en 4.5 años aproximadamente, con la mayor parte de las materias inscritas en el turno matutino e inscritas en el semestre que corresponde a cada materia.

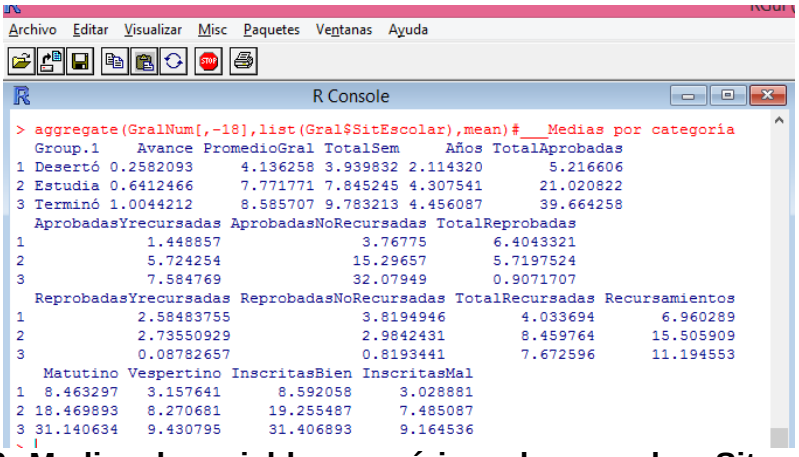


Figura 3.16a Medias de variables numéricas de acuerdo a Situación Escolar.

Si se analiza la misma información pero por carrera, se observa que en Actuaría la población deserta con media de 22.82% de avance académico con 6 materias aprobadas aproximadamente y los alumnos que concluyen su carrera lo hacen en 9.5 semestres. En Ciencias de la computación los alumnos están desertando con media de 31.87% de avance académico con 4 materias aprobadas y los alumnos concluyen su carrera en 11 semestres (Ver Figura 3.16b).

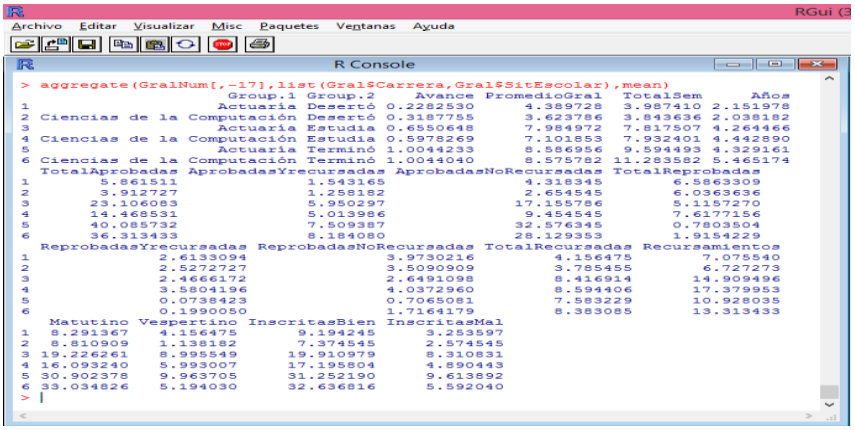


Figura 3.16b Medias de variables numéricas de acuerdo a Situación Escolar y Carrera.

A continuación se realiza un análisis para ver la dispersión de los datos, sesgos y densidad de los mismos (ver Figura 3.17), recordar que la desviación estándar mostrará que tan dispersos están los datos, el coeficiente de Kurtosis indicara que tan puntiaguda es la curva de distribución (media=0, alta>0, baja<0) y el coeficiente de Skewnes o de asimetría, indicará el sesgo (izquierda<0, derecha>0).

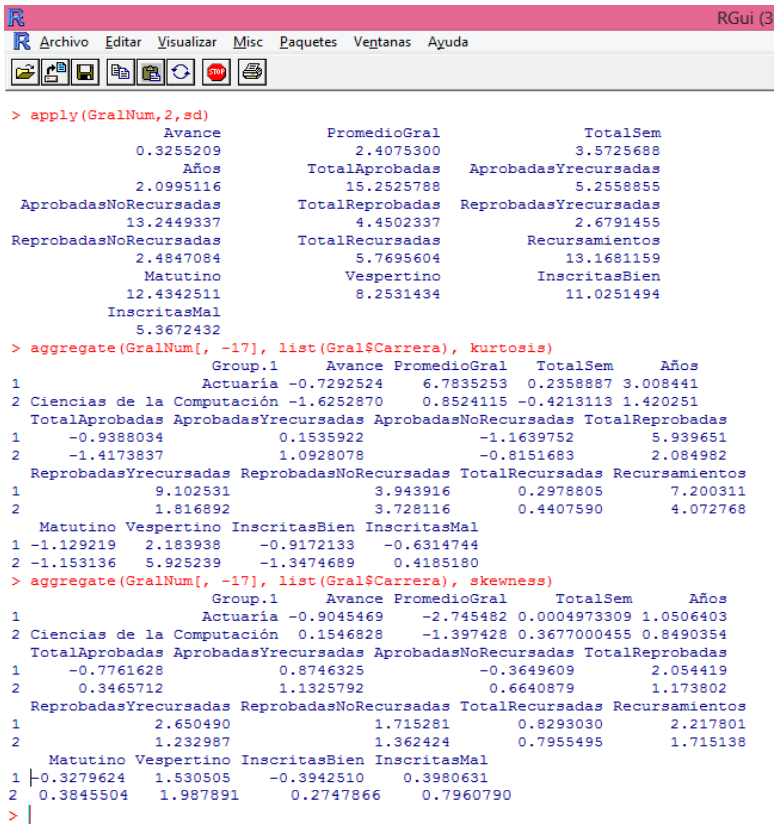


Figura 3.17 Desviación estándar, Coeficiente de Kurtosis y de Asimetría.

Se puede notar que los datos que se encuentran más dispersos son los de las variables Total Aprobadas, Recursamientos, AprobadasNoRecursadas, Matutino, Vespertino e Inscritas Bien, por lo que son las variables propensas a la aplicación de una transformación (ver Figura 3.18).

Las curvas con más pico son para Actuaría ReprobadasYrecursadas, Recursamientos y PromedioGral, mientras que en Ciencias de la Computación la curva con más picos es Vespertino. En general las curvas más picudas son ReprobadasYrecursadas, AprobadasNoRecursadas, y RepprobadasNo Recursadas, por lo que son las variables propensas a la aplicación de una transformación por tener picos muy elevados en sus densidades (Ver Figura 3.19a y 3.19b)

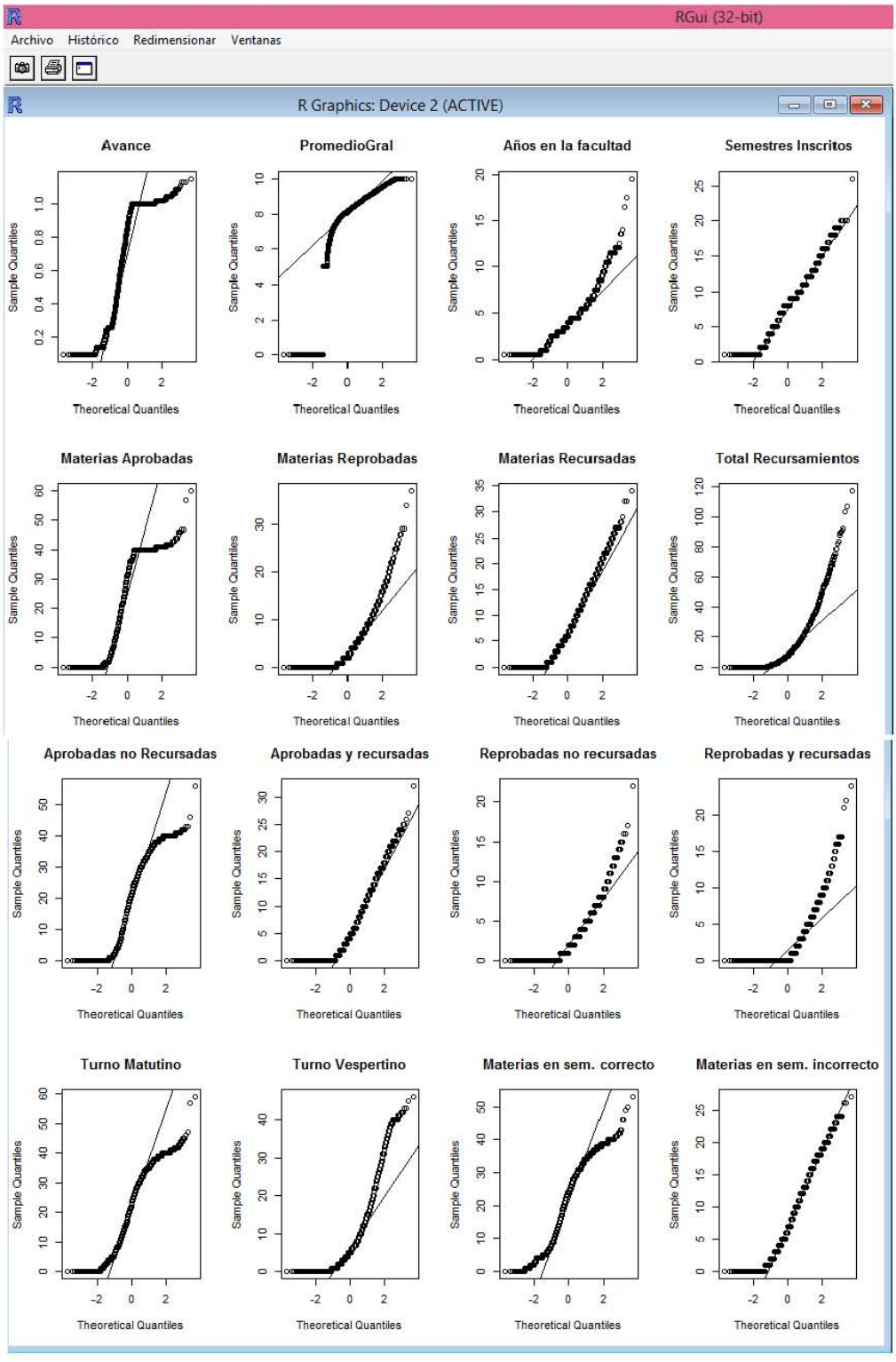


Figura 3.18 Comparación de los datos con la distribución normal.

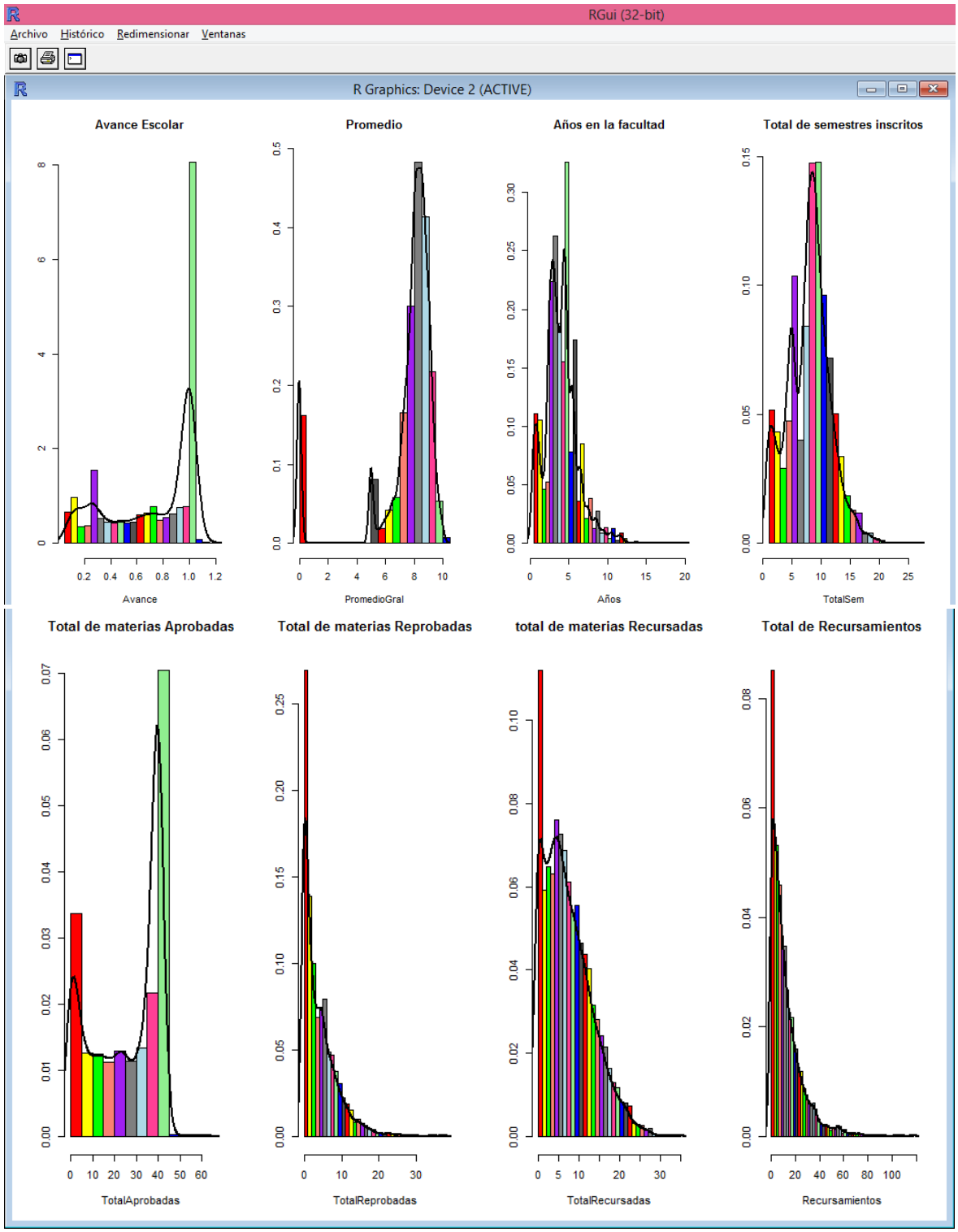


Figura 3.19a Ajuste de distribuciones de los datos (Parte 1)

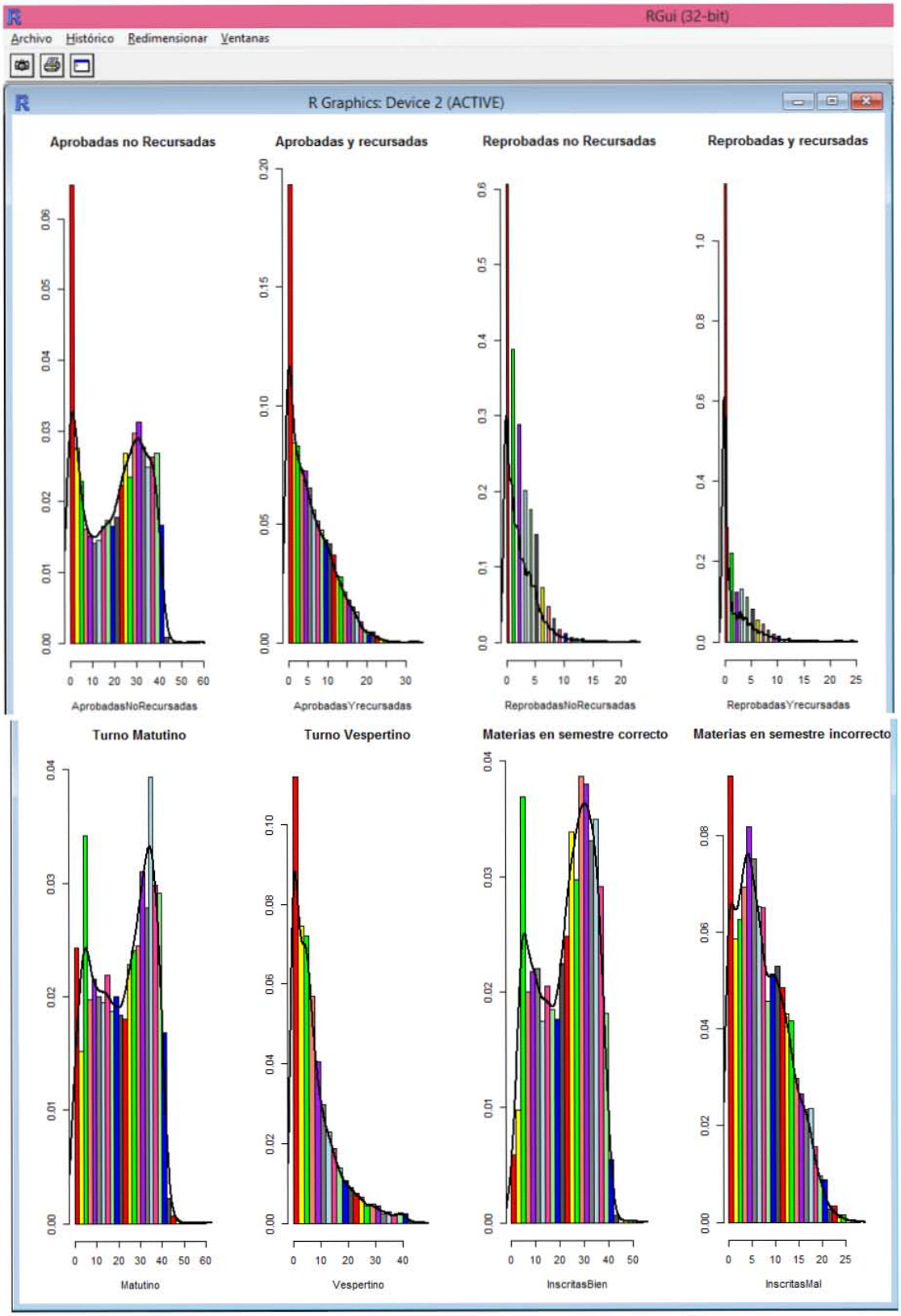


Figura 3.19b Ajuste de distribuciones de los datos (Parte 2)

Ahora con ayuda de los coeficienes de Skewness obtenidos en la Figura 3.17 y reflejados en los histogramas siguientes (Figuras 3.20a, 3.20b y 3.20c), resulta que las variables Años, Recursamientos, TotalRecursadas, TotalReprobadas, ReprobadasYrecursadas, AprobadasYrecursadas, AprobadasNoRecursadas, ReprobadasNoRecursadas y Vespertino tienen sesgo hacia la izquierda por lo que son candidatas a aplicarles la transformación y=log(x). La variable Promedio tiene sesgo hacia la derecha por lo que es candidata a aplicarle la transformación $y=x^2$.

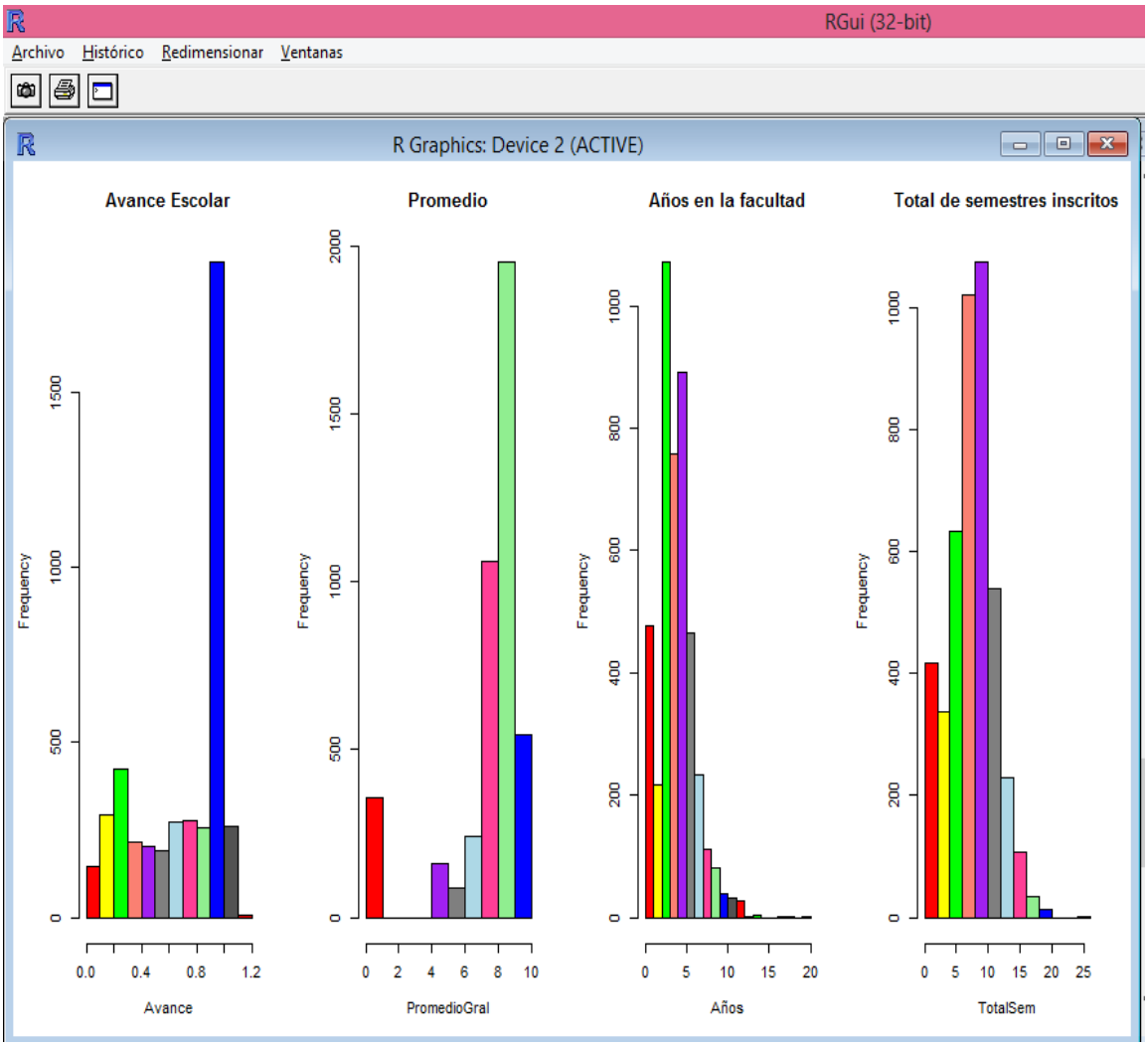


Figura 3.20a Histogramas de variables numéricas (Parte 1)

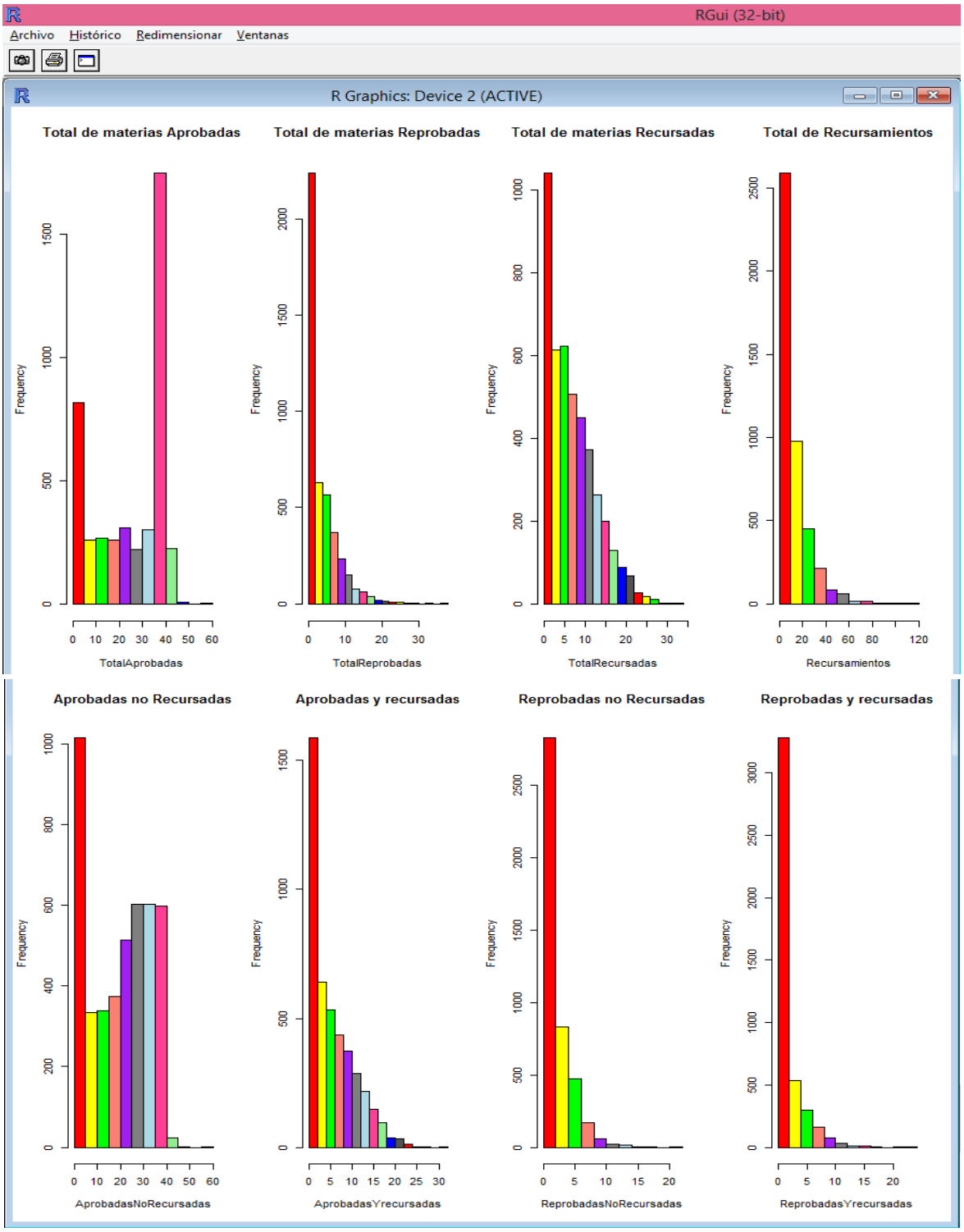


Figura 3.20b Histogramas de variables numéricas (Parte 2)

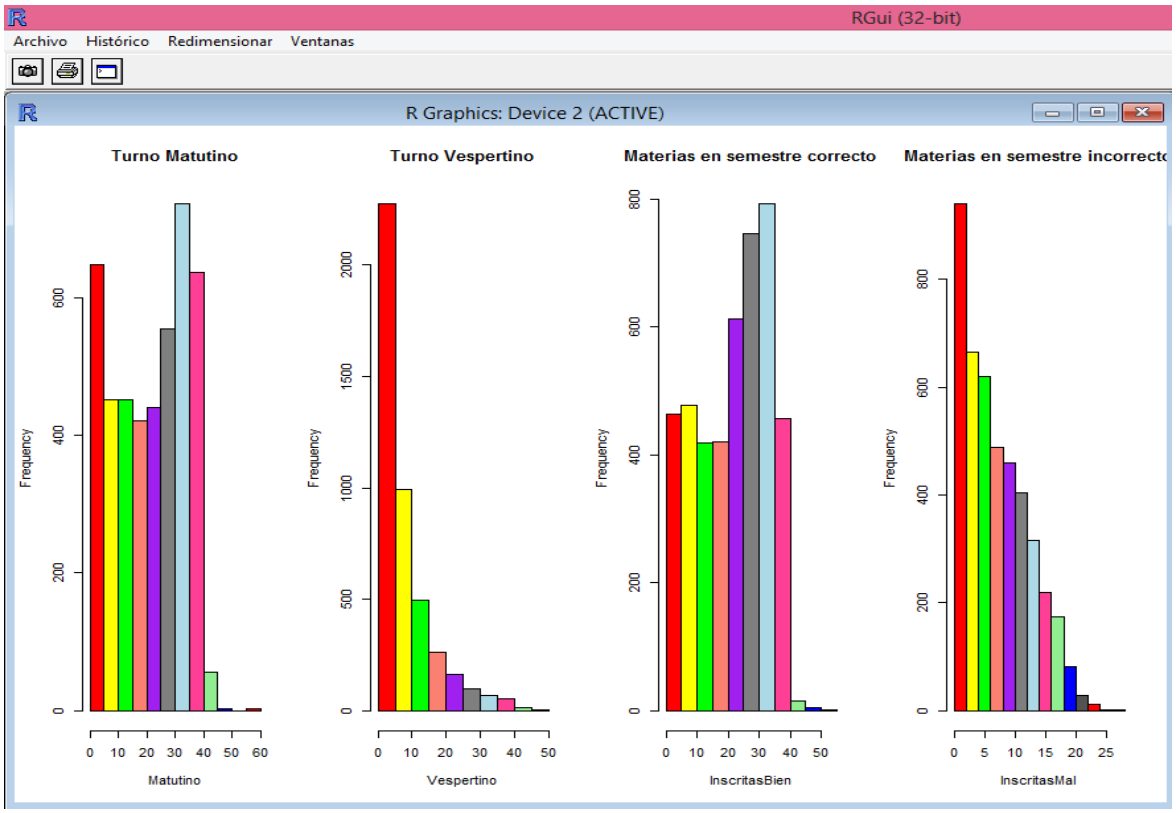
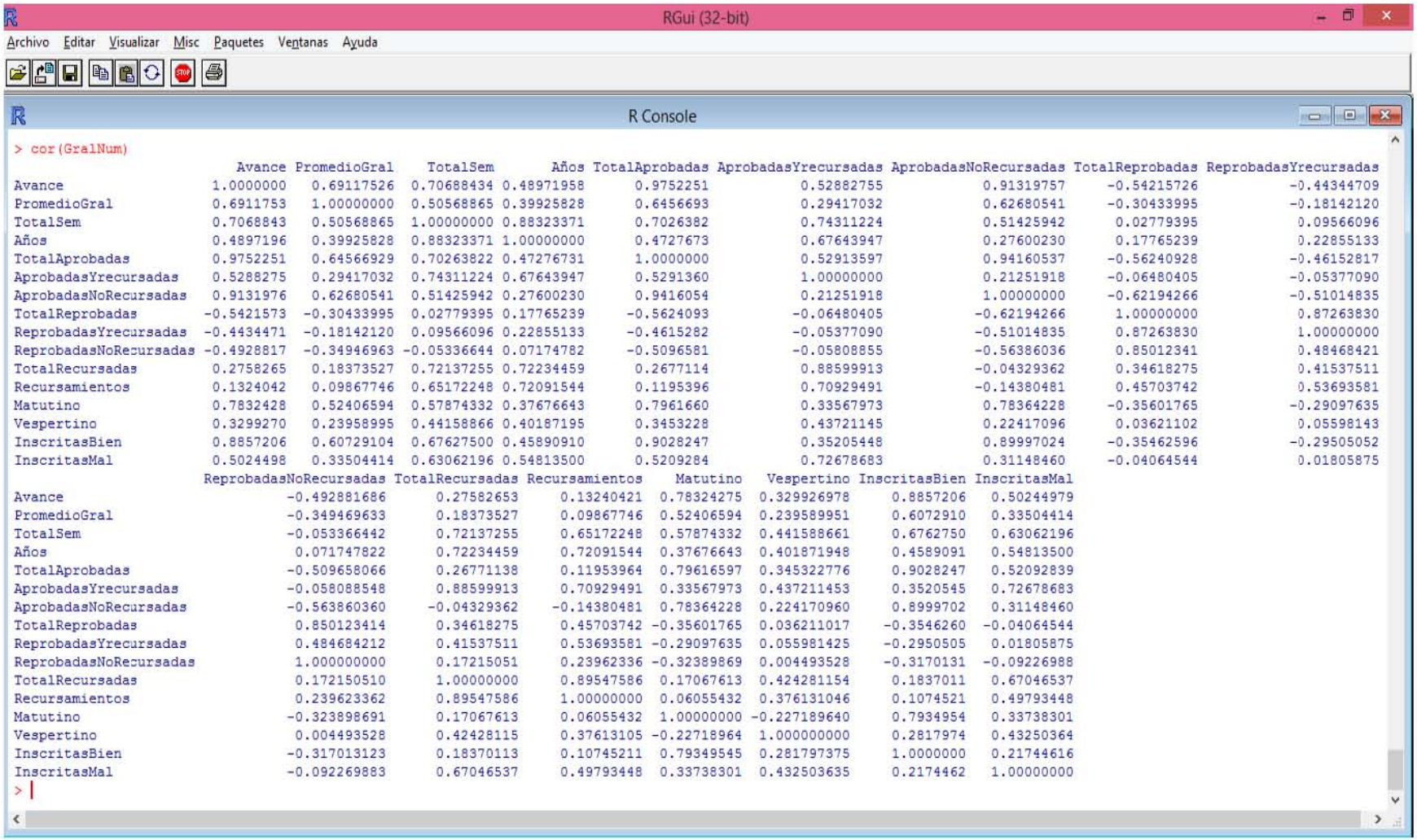


Figura 3.20c Histogramas de variables numéricas (Parte 3)

En el análisis de correlación se considerarán variables altamente relacionadas aquellas cuyos coeficientes sean menores a -0.7 ó mayores a 0.7 (recordar que los coeficientes de correlación se encuentran entre menos uno y uno).

Existen dos maneras de interpretar los resultados, la primera es que en una correlación alta, las variables presentan de cierta forma la misma información, con lo cual una de las variables no se introduce al modelo, en este caso tenemos las variables de la Tabla 3.1, la otra forma de interpretación es que probablemente una variable sea dependiente de la otra o guardan una relación explicativa, en dicho caso se conservan ambas variables, en este caso tenemos las variables de la Tabla 3.2. Los valores presentados en ambas tablas son obtenidos del análisis presentado en la Ver Figura 3.21.




```
> cor(GralNum)
                    Avance PromedioGral    TotalSem       Años TotalAprobadas AprobadasYrecursadas AprobadasNoRecursadas TotalReprobadas ReprobadasYrecursadas
Avance           1.0000000   0.69117526 0.70688434 0.48971958      0.9752251          0.52882755           0.91319757     -0.54215726          -0.44344709
PromedioGral     0.6911753   1.00000000 0.50568865 0.39925828      0.6456693          0.29417032           0.62680541     -0.30433995          -0.18142120
TotalSem         0.7068843   0.50568865 1.00000000 0.88323371      0.7026382          0.74311224           0.51425942      0.02779395           0.09566096
Años             0.4897196   0.39925828 0.88323371 1.00000000      0.4727673          0.67643947           0.27600230      0.17765239           0.22855133
TotalAprobadas   0.9752251   0.64566929 0.70263822 0.47276731      1.0000000          0.52913597           0.94160537     -0.56240928          -0.46152817
AprobadasYrecursadas 0.5288275 0.29417032 0.74311224 0.67643947    0.5291360          1.00000000           0.21251918     -0.06480405          -0.05377090
AprobadasNoRecursadas 0.9131976 0.62680541 0.51425942 0.27600230   0.9416054          0.21251918           1.00000000     -0.62194266          -0.51014835
TotalReprobadas  -0.5421573  -0.30433995 0.02779395 0.17765239     -0.5624093         -0.06480405          -0.62194266     1.00000000           0.87263830
ReprobadasYrecursadas -0.4434471 -0.18142120 0.09566096 0.22855133 -0.4615282         -0.05377090          -0.51014835     0.87263830           1.00000000
ReprobadasNoRecursadas -0.4928817 -0.34946963 -0.05336644 0.07174782 -0.5096581       -0.05808855          -0.56386036     0.85012341           0.48468421
TotalRecursadas  0.2758265   0.18373527 0.72137255 0.72234459      0.2677114          0.88599913           -0.04329362     0.34618275           0.41537511
Recursamientos   0.1324042   0.09867746 0.65172248 0.72091544      0.1195396          0.70929491           -0.14380481     0.45703742           0.53693581
Matutino         0.7832428   0.52406594 0.57874332 0.37676643      0.7961660          0.33567973           0.78364228      -0.35601765          -0.29097635
Vespertino       0.3299270   0.23958995 0.44158866 0.40187195      0.3453228          0.43721145           0.22417096      0.03621102           0.05598143
InscritasBien    0.8857206   0.60729104 0.67627500 0.45890910      0.9028247          0.35205448           0.89997024      -0.35462596          -0.29505052
InscritasMal     0.5024498   0.33504414 0.63062196 0.54813500      0.5209284          0.72678683           0.31148460      -0.04064544          0.01805875
                 ReprobadasNoRecursadas TotalRecursadas Recursamientos    Matutino  Vespertino InscritasBien InscritasMal
Avance                      -0.492881686     0.27582653     0.13240421  0.78324275  0.329926978    0.8857206   0.50244979
PromedioGral                -0.349469633     0.18373527     0.09867746  0.52406594  0.239589951    0.6072910   0.33504414
TotalSem                    -0.053366442     0.72137255     0.65172248  0.57874332  0.441588661    0.6762750   0.63062196
Años                         0.071747822     0.72234459     0.72091544  0.37676643  0.401871948    0.4589091   0.54813500
TotalAprobadas              -0.509658066     0.26771138     0.11953964  0.79616597  0.345322776    0.9028247   0.52092839
AprobadasYrecursadas        -0.058088548     0.88599913     0.70929491  0.33567973  0.437211453    0.3520545   0.72678683
AprobadasNoRecursadas       -0.563860360    -0.04329362    -0.14380481  0.78364228  0.224170960    0.8999702   0.31148460
TotalReprobadas              0.850123414     0.34618275     0.45703742 -0.35601765  0.036211017   -0.3546260  -0.04064544
ReprobadasYrecursadas        0.484684212     0.41537511     0.53693581 -0.29097635  0.055981425   -0.2950505   0.01805875
ReprobadasNoRecursadas       1.000000000     0.17215051     0.23962336 -0.32389869  0.004493528   -0.3170131  -0.09226988
TotalRecursadas              0.172150510     1.00000000     0.89547586  0.17067613  0.424281154    0.1837011   0.67046537
Recursamientos               0.239623362     0.89547586     1.00000000  0.06055432  0.376131046    0.1074521   0.49793448
Matutino                    -0.323898691     0.17067613     0.06055432  1.00000000 -0.227189640    0.7934954   0.33738301
Vespertino                   0.004493528     0.42428115     0.37613105 -0.22718964  1.000000000    0.2817974   0.43250364
InscritasBien               -0.317013123     0.18370113     0.10745211  0.79349545  0.281797375    1.0000000   0.21744616
InscritasMal                -0.092269883     0.67046537     0.49793448  0.33738301  0.432503635    0.2174462   1.00000000
>
```

Figura 3.21 Matriz de correlación de variables numéricas.

VARIABLE 1	VARIABLE 2 (variable que no entrará al modelo)	Coeficiente de correlación
Avance	TotalAprobadas	0.97
TotalAprobadas	AprobadasNoRecursadas	0.94
TotalReprobadas	ReprobadasNoRecursadas	0.85

Tabla 3.1 Coeficientes de correlación altos de variables con el mismo tipo de información

VARIABLE 1	VARIABLE 2	Coeficiente de correlación
Avance	AprobadasNoRecursadas	0.91
TotalAprobadas	InscritasBien	0.90
AprobadasYrecursadas	TotalRecursadas	0.88
Avance	InscritasBien	0.88
TotalSem	Años	0.88
TotalReprobadas	ReprobadasYrecursadas	0.87
TotalAprobadas	Matutino	0.79
Avance	Matutino	0.78
TotalSem	AprobadasYrecursadas	0.74
TotalSem	TotalRecursadas	0.72
Años	TotalRecursadas	0.72
Años	Recursamientos	0.72
AprobadasYrecursadas	InscritasMal	0.72

Tabla 3.2 Índices de correlación altos de variables con diferente información.

Con la finalidad de descubrir más sobre cómo es que se están concentrando los datos, se realizan los diagramas de caja y bigotes de cada variable numérica, los cuales se interpretan de manera que el bigote más largo indica que existe mayor dispersión de los datos del lado de ese bigote que en el otro lado representado por el otro bigote, la caja representa los datos que se encuentran entre el cuartil 25 y el cuartil 75, aquí es donde se concentra la mayor parte de la población, la línea gruesa que divide la caja representa la media o promedio de la variable de análisis; la parte más ancha de la caja indica que en ese rango existe mayor dispersión que en la otra mitad de la caja.

Por último, los puntos que quedan fuera de los bigotes son los datos atípicos presentes en la muestra.

De acuerdo con los diagramas de la Figura 3.22 se puede realizar un análisis para cada variable pero a continuación sólo se realizará el análisis detallado de los diagramas que aportan información nueva.

Las variables Años, TotalRecursadas, Recursamientos, AprobadasYrecursadas, ReprobadasYrecursadas, contienen varios “outliers” o datos atípicos, algunos se minimizarán con las transformaciones propuestas en párrafos anteriores, estos

valores pudieran tener un tratamiento especial dependiendo de los modelos de minería a aplicar, en este caso no es necesario realizar más cambios ya que los modelos que se realizarán no se ven afectados por estos valores extremos.

Los datos atípicos de la variable Vespertino pueden representar en sí los alumnos que llevan y llevaron su carrera total o mayormente en dicho turno y no datos que indiquen comportamientos extremos.

Por otro lado, se observa que la cantidad de materias inscritas en los semestres correctos es mayor a las materias inscritas en semestres incorrectos.

Para finalizar el análisis de esta primera base de datos, se presentan en la Figura 3.23 tres diagramas de dispersión que ilustran a la población general y su comportamiento de acuerdo a su situación escolar.

En el primer diagrama se puede observar que la mayor parte de los alumnos desertan antes del 30% del avance académico como se concluyó anteriormente pero también resaltan algunos puntos que estando casi al 100% de la carrera deciden abandonarla.

En el segundo diagrama rescatamos como dato novedoso que existen alumnos que concluyen su carrera en la mitad de tiempo marcado por la facultad.

En el tercer diagrama se corrobora que gran parte de la población se encuentra en el turno matutino pero también existen muchos alumnos que prefieren tener un horario mixto.

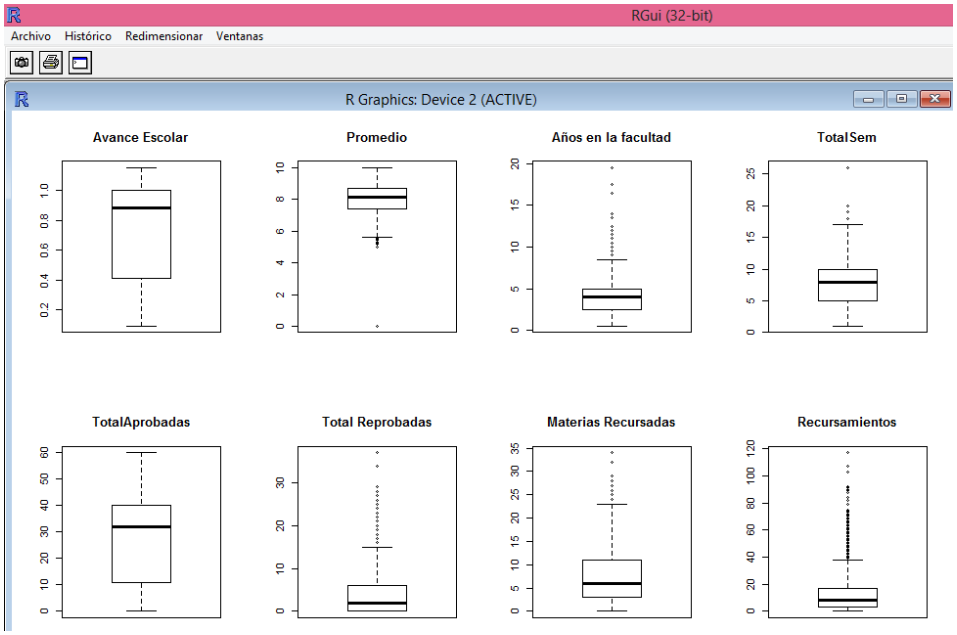


Figura 3.22a Diagramas de caja y bigotes de las variables numéricas

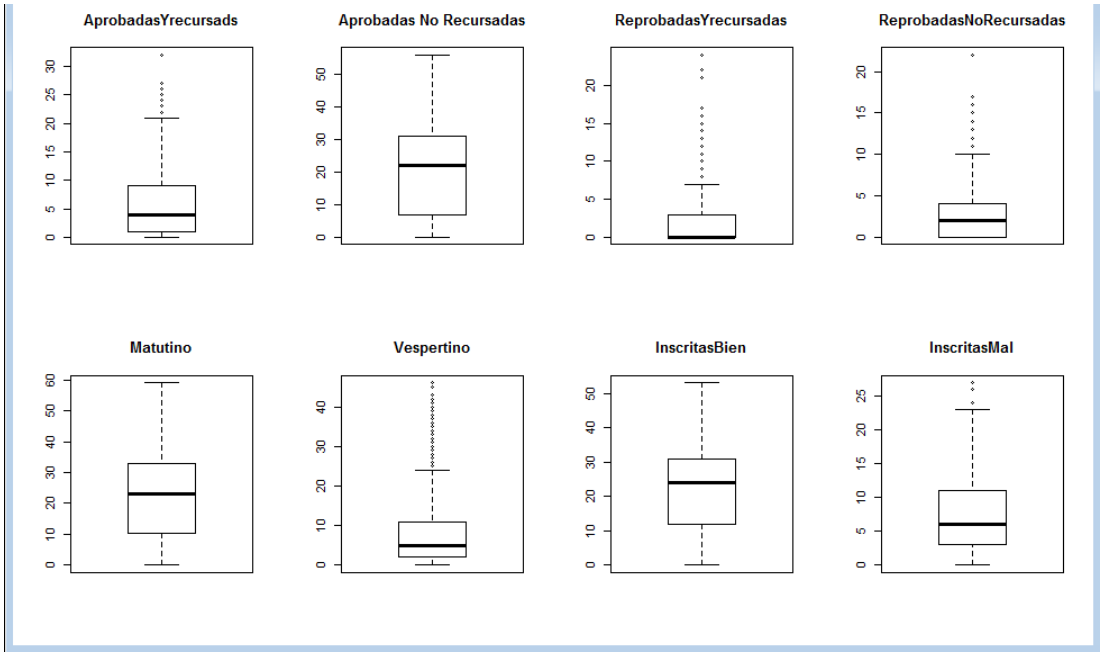


Figura 3.22b Diagramas de caja y bigotes de las variables numéricas

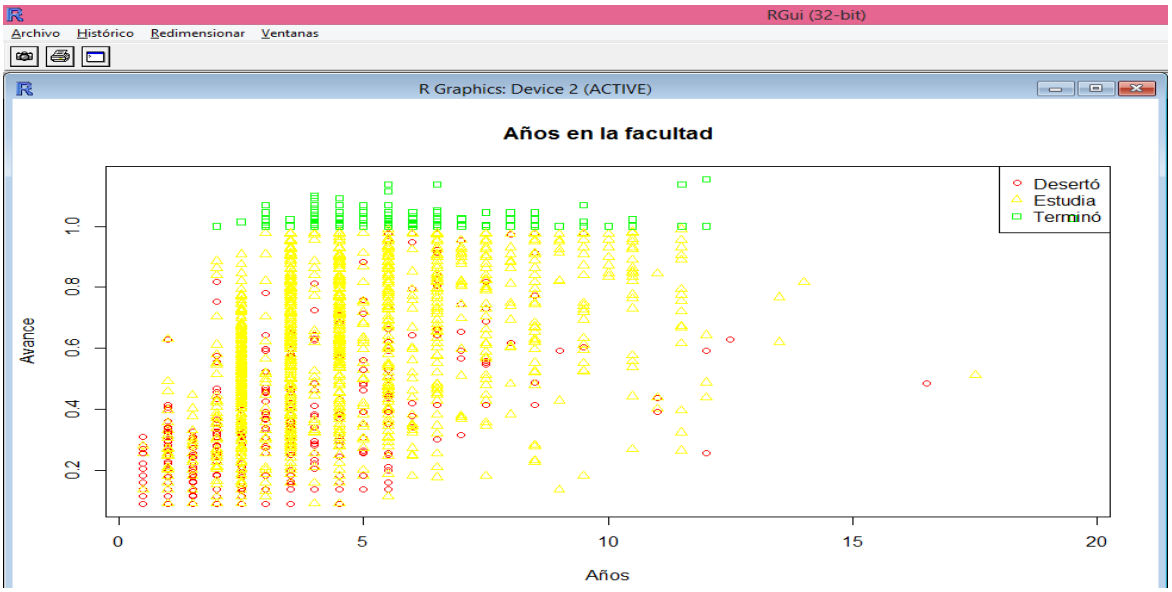


Figura 3.23a Gráficas de dispersión.

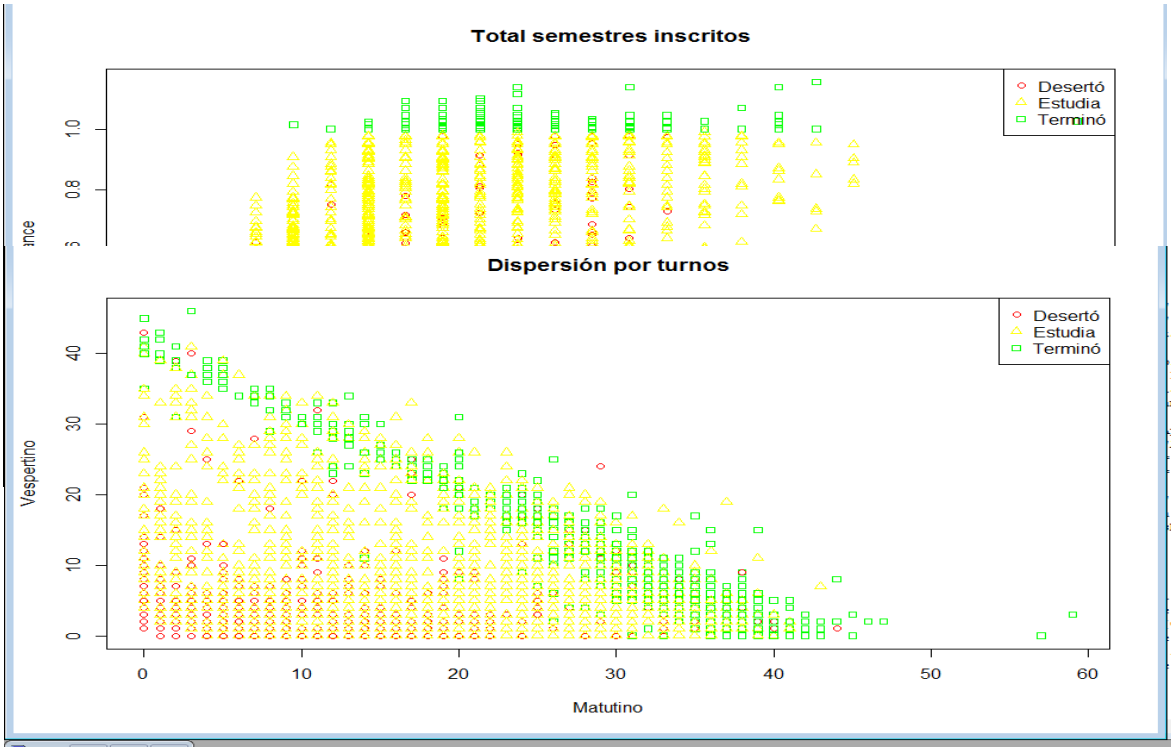


Figura 3.23b Gráficas de dispersión.

Bases "Act00, Act06 y Comp1994":

El objetivo de análisis de estas tres bases de datos es indagar sobre otros factores que inciden en la culminación de la carrera, factores que tengan que ver directamente con las materias, el tipo de inscripción y las veces que se cursó.
La base Comp2013 no se analiza ya que al tomar en cuenta hasta la generación 2011 para el estudio, estos datos quedan excluidos.
Para las tres bases de datos se realiza un estudio más sencillo que para la base "Datos Alumnos" debido a que el tipo de información que contienen es menor y columnas con la misma información se pueden ver a simple vista.

Act00:

Esta base tiene un tamaño de 1173 filas y 101 columnas cuyos nombres son Id, SitEscolar, Estudia, Termino, Deserto, los nombres de las 32 materias obligatorias correspondientes al plan 2000, 32 columnas "Grupo" referentes al grupo inscrito en cada materia y 32 columnas "Tipo" referentes al tipo de inscripción de cada materia. Todas las columnas contienen datos de tipo categórico.

Los nombres de las materias y los números que les corresponden son:

Número	Materia
1	Álgebra Superior I
2	Cálculo Diferencial e Integral I
3	Geometría Analítica I
4	Matemáticas Financieras
5	Problemas Socio-Económicos de México
6	Álgebra Superior II
7	Cálculo Diferencial e Integral II
8	Geometría Analítica II
9	Programación I
10	Álgebra Lineal I
11	Cálculo Diferencial e Integral III
12	Probabilidad I
13	Programación II
14	Teoría del Seguro
15	Cálculo Diferencial e Integral IV
16	Ecuaciones Diferenciales I
17	Finanzas I
18	Matemáticas Actuariales del Seguro de Personas I
19	Probabilidad II
20	Análisis Matemático I
21	Estadística I
22	Finanzas II
23	Investigación de Operaciones
24	Matemáticas Actuariales del Seguro de Personas II
25	Economía I
26	Estadística II
27	Matemáticas Actuariales del Seguro de Daños
28	Análisis Numérico
29	Demografía I
30	Seguridad Social
31	Pensiones Privadas
32	Teoría del Riesgo

Al efectuar la función summary() se obtienen frecuencias de cada variable, las más importantes se encuentran en la Figura 3.24 y con las cuales se obtiene esta información:

Existe una tendencia de inscripción en un grupo específico en materias como Cálculo Diferencial e Integral 4, donde el grupo 4118 presenta un total de 114 alumnos inscritos y los demás grupos cuentan con un máximo de 54 alumnos; en Análisis Matemático, el grupo 4118 presenta 153 inscritos y los demás grupos no pasan de 75 alumnos; en Finanzas II el grupo 6072 tiene 133 y los demás tienen un máximo de 90 alumnos; en Seguridad Social el grupo 6114 tiene 144 inscritos

y los demás no pasan de 88 alumnos. Esta información refiere a que el grupo en el que se inscriben los alumnos podría llegar a influir en su situación escolar.

Con ayuda de las columnas “Tipo”, se obtienen los índices de reprobación y aprobación en donde las materias con mayor índice de reprobación son Probabilidad I (23.14%), Cálculo Diferencial e Integral II (22.05%), Cálculo Diferencial e Integral I (20.88%), Álgebra Lineal (20.36%) e Investigación de Operaciones (18.55%). Mientras que las de mayor índice de aprobación son Teoría del Seguro (97.12%), Problemas Socio-Económicos de México (95.82%), Matemáticas Actuariales del Seguro de Personas II (94.45%) y Estadística II (90.08%).

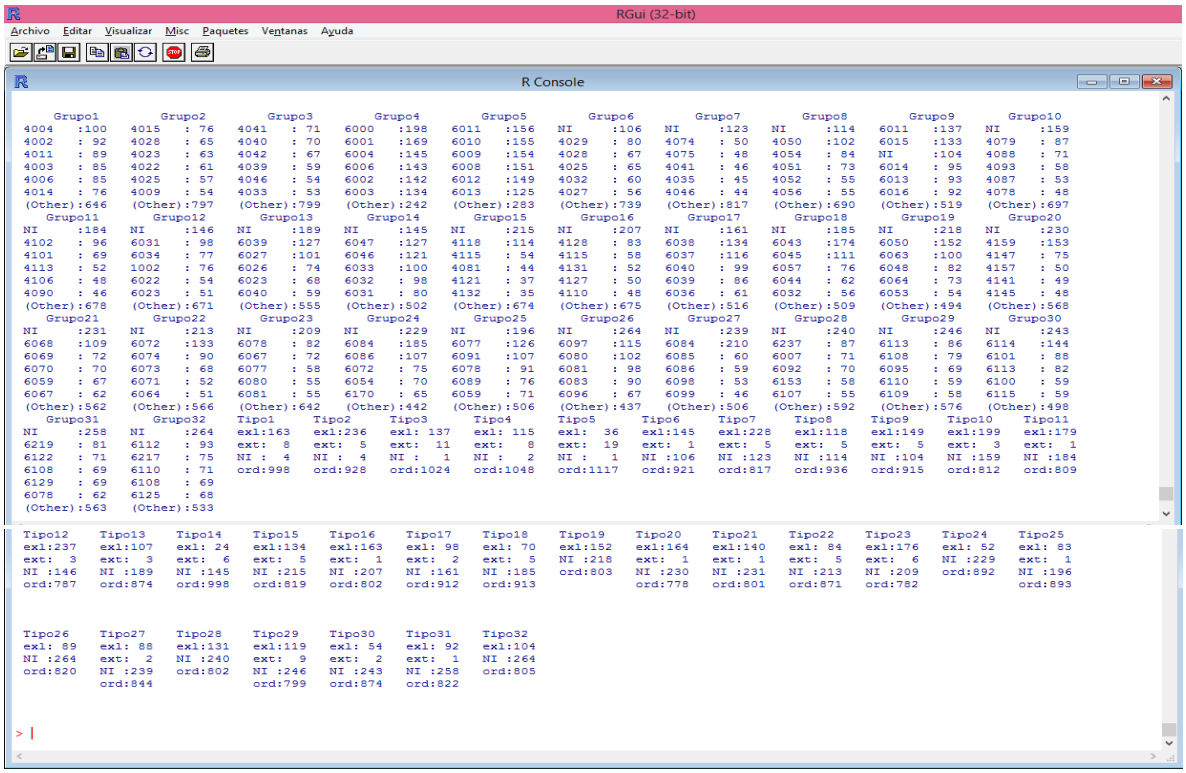


Figura 3.24 Frecuencia de variables “Grupo” y “Tipo” para Actuaría plan 2000

Act06:

La base tiene un tamaño de 2215 filas y 113 columnas cuyos nombres son Id, SitEscolar, Estudia, Termino, Deserto, los nombres de las 36 materias obligatorias correspondientes al plan 2000, 36 columnas “Grupo” referentes al grupo inscrito en cada materia y 36 columnas “Tipo” referentes al tipo de inscripción de cada materia. Todas las columnas contienen datos de tipo categórico.

Los nombres de las materias y los números que les corresponden son:

Número	Materia
1	Álgebra Superior I
2	Cálculo Diferencial e Integral I
3	Contabilidad
4	Geometría Analítica I
5	Problemas Socio-Económicos de México
6	Álgebra Superior II
7	Cálculo Diferencial e Integral II
8	Geometría Analítica II
9	Matemáticas Financieras
10	Programación I
11	Álgebra Lineal I
12	Cálculo Diferencial e Integral III
13	Probabilidad I
14	Programación II
15	Teoría del Seguro
16	Cálculo Diferencial e Integral IV
17	Ecuaciones Diferenciales I
18	Finanzas I
19	Matemáticas Actuariales del Seguro de Personas I
20	Probabilidad II
21	Análisis Matemático I
22	Estadística I
23	Finanzas II
24	Investigación de Operaciones
25	Matemáticas Actuariales del Seguro de Personas II
26	Economía I
27	Estadística II
28	Matemáticas Actuariales del Seguro de Daños
29	Procesos Estocásticos
30	Análisis Numérico
31	Demografía I
32	Estadística III
33	Seguridad Social
34	Administración Actuarial
35	Pensiones Privadas
36	Teoría del Riesgo

En el resumen estadístico se obtiene la siguiente información, la cual se puede corroborar con los datos de la Figura 3.25:

Igual que en el plan 2000, en este plan también existe una tendencia de inscripción en un grupo específico. Cálculo Diferencial e Integral III presenta una inscripción de 146 alumnos en el grupo 4061, mientras que otros grupos tienen menos de 96 inscripciones; Matemáticas Financieras es la que más llama la atención ya que las inscripciones en ciertos grupos son superiores a 200 alumnos; Análisis Matemático y Probabilidad I presentan cierta tendencia extrema de inscripción.

Las materias con mayor índice de reprobación son Probabilidad I (17.51%), Cálculo Diferencial e Integral II (17.25%) y Cálculo Diferencial e Integral III (14.95%). Las materias de mayor índice de aprobación son Seguridad Social (98.77%), Geometría Analítica I (98.35%), Teoría del Seguro (98.35%) y Matemáticas Actuariales del Seguro de Personas II (98.15%).

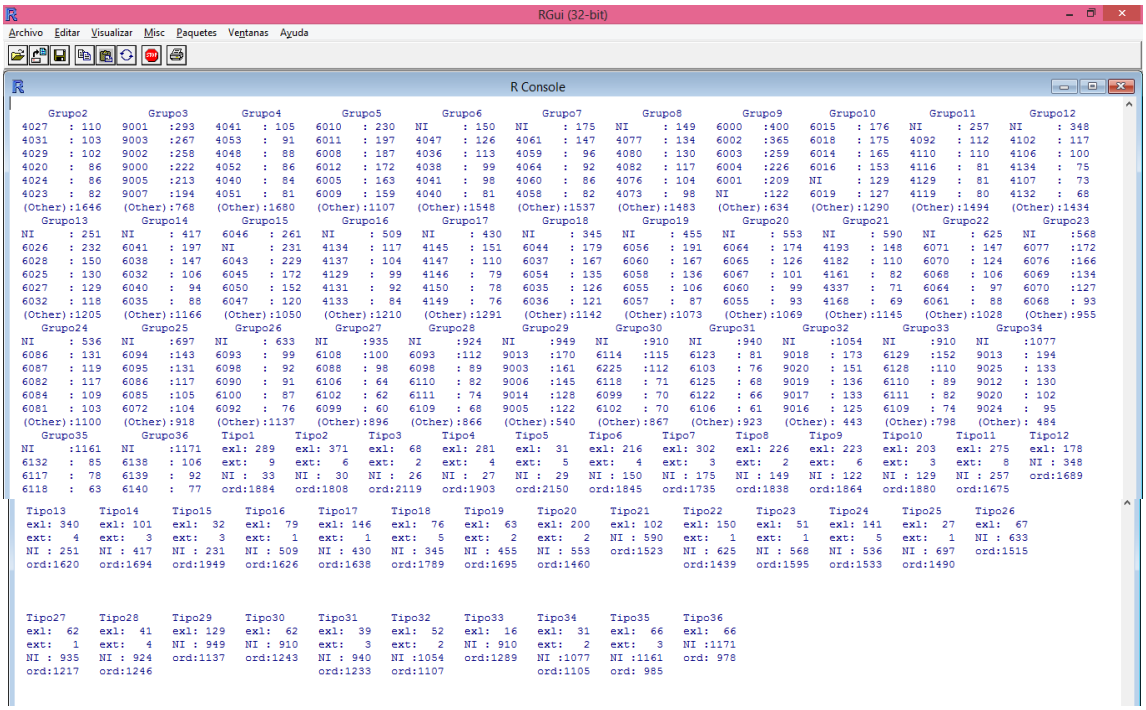


Figura 3.25 Frecuencia de variables "Grupo" y "Tipo" para Actuaría plan 2000

Comp1994:

La base tiene un tamaño de 879 filas y 103 columnas cuyos nombres son Id, Termino, Deserto, los nombres de las 25 materias obligatorias correspondientes al plan 1994 con 0 si la materia no se aprobó y 1 si fue aprobada; 25 columnas "Grupo" con la clave del grupo inscrito en cada materia; 25 columnas "Tipo" referentes al tipo de inscripción de cada materia (ord, ext, extl).

Los nombres de las materias y los números que les corresponden son:

Número	Materia
1	Álgebra Superior I
2	Cálculo Diferencial e Integral I
3	Geometría Analítica I
4	Matemáticas Discretas
5	Álgebra Superior II
6	Cálculo Diferencial e Integral II
7	Geometría Analítica II
8	Introducción a Ciencias de la Computación I
9	Álgebra Lineal I
10	Cálculo Diferencial e Integral III
11	Introducción a Ciencias de la Computación II
12	Probabilidad y Estadística
13	Álgebra Lineal II
14	Análisis Lógico
15	Cálculo Diferencial e Integral IV
16	Diseño de Sistemas Digitales
17	Análisis de Algoritmos I
18	Arquitectura de Computadoras
19	Teoría de la Computación
20	Inteligencia Artificial
21	Lenguajes de Programación y sus Paradigmas
22	Sistemas de Bases de Datos
23	Sistemas Operativos
24	Análisis Numérico I
25	Ingeniería de Software

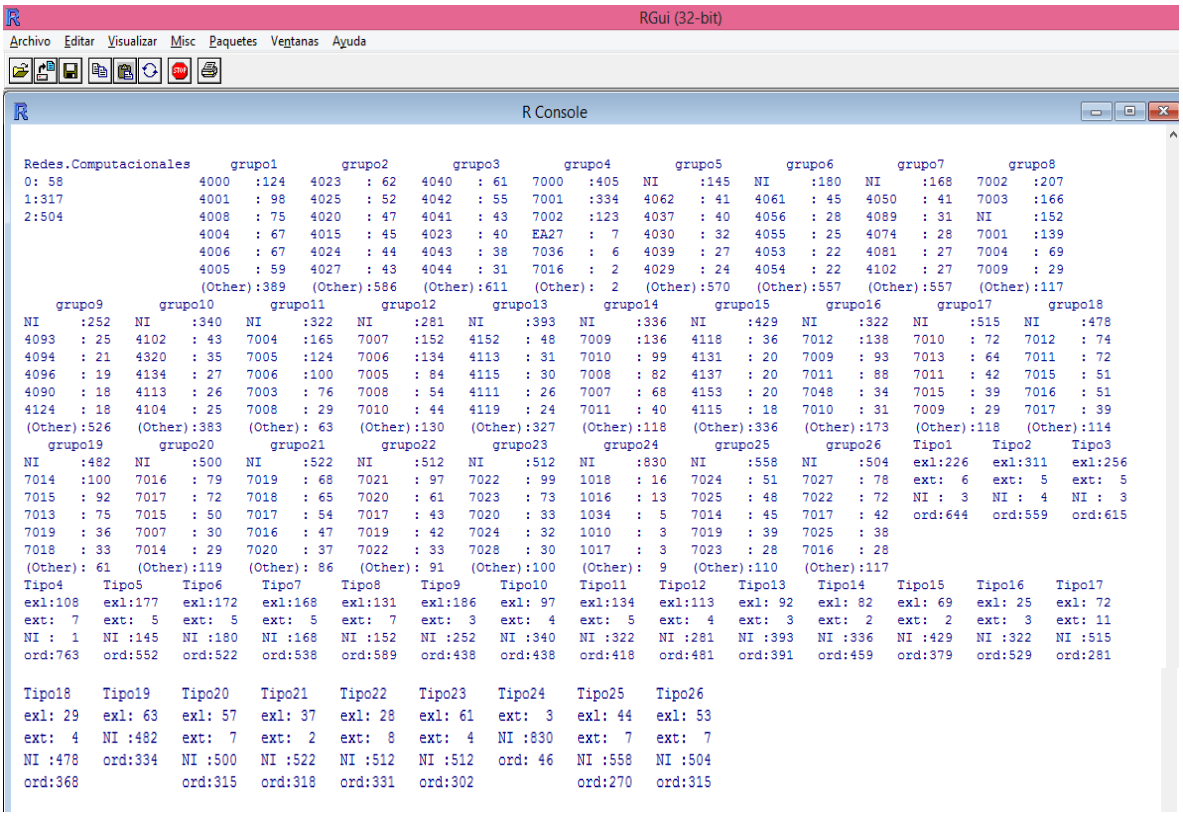
De la misma forma que las bases de historiales de actuaría se realiza primeramente la transformación de las variables al categóricas o numéricas según corresponda, posteriormente se ejecuta un resumen estadístico en el que se obtiene la siguiente información (ver Figura 3.26):

Las tendencias en la inscripción de grupos se presentan en las materias de Diseño de Sistemas Digitales en donde el grupo 7012 tiene 138 alumnos inscritos y los demás por debajo de 93; llama la atención Introducción a Ciencias de la Computación II con tres grupos que superan los 100 inscritos mientras que otros grupos tienen menos de 76 alumnos; en a materia de Álgebra Lineal I, también existe una diferencia considerable en tres grupos con más de 130 alumnos mientras los otros tienen menos de 70 inscritos.

Las materias con mayor índice de reprobación son Cálculo Diferencial e Integral I (36.11%) y Álgebra Lineal (30.14%), Geometría Analítica I (28.65%) y Álgebra Superior I (26.48%). Las materias de mayor índice de aprobación son Diseño de

Sistemas Digitales (94.8%), Análisis Numérico (93.87%), Arquitectura de Computadoras (91.77%) y Sistemas de Bases de Datos (90.2%).

En este análisis las variables “i” reflejan la misma información sobre los índices de aprobación y reprobación que las variables “tipo”, sin embargo contienen información mucho más valiosa para los modelos. Por ejemplo materias como Álgebra superior II, Cálculo Diferencial en Integral II, Análisis de algoritmos, Álgebra Lineal I y II tienen alumnos con un total de inscripciones a la materia superiores a 10, lo cual nos hace suponer que en estas materias podría existir algún tipo de filtro, trataremos de corroborar esta información con el modelo de minería del capítulo 4.




```
Redes.Computacionales     grupo1        grupo2        grupo3        grupo4        grupo5        grupo6        grupo7        grupo8
0: 58                  4000   :124   4023   : 62   4040   : 61   7000   :405   NI     :145   NI     :180   NI     :168   7002   :207
1:317                  4001   : 98   4025   : 52   4042   : 55   7001   :334   4062   : 41   4061   : 45   4050   : 41   7003   :166
2:504                  4008   : 75   4020   : 47   4041   : 43   7002   :123   4037   : 40   4056   : 28   4089   : 31   NI     :152
                       4004   : 67   4015   : 45   4023   : 40   EA27   :  7   4030   : 32   4055   : 25   4074   : 28   7001   :139
                       4006   : 67   4024   : 44   4043   : 38   7036   :  6   4039   : 27   4053   : 22   4081   : 27   7004   : 69
                       4005   : 59   4027   : 43   4044   : 31   7016   :  2   4029   : 24   4054   : 22   4102   : 27   7009   : 29
                       (Other):389   (Other):586   (Other):611   (Other):  2   (Other):570   (Other):557   (Other):557   (Other):117
    grupo9       grupo10       grupo11       grupo12       grupo13       grupo14       grupo15       grupo16       grupo17       grupo18
NI     :252   NI     :340   NI     :322   NI     :281   NI     :393   NI     :336   NI     :429   NI     :322   NI     :515   NI     :478
4093   : 25   4102   : 43   7004   :165   7007   :152   4152   : 48   7009   :136   4118   : 36   7012   :138   7010   : 72   7012   : 74
4094   : 21   4320   : 35   7005   :124   7006   :134   4113   : 31   7010   : 99   4131   : 20   7009   : 93   7013   : 64   7011   : 72
4096   : 19   4134   : 27   7006   :100   7005   : 84   4115   : 30   7008   : 82   4137   : 20   7011   : 88   7011   : 42   7015   : 51
4090   : 18   4113   : 26   7003   : 76   7008   : 54   4111   : 26   7007   : 68   4153   : 20   7048   : 34   7015   : 39   7016   : 51
4124   : 18   4104   : 25   7008   : 29   7010   : 44   4119   : 24   7011   : 40   4115   : 18   7010   : 31   7009   : 29   7017   : 39
(Other):526   (Other):383   (Other): 63   (Other):130   (Other):327   (Other):118   (Other):336   (Other):173   (Other):118   (Other):114
   grupo19       grupo20       grupo21       grupo22       grupo23       grupo24       grupo25       grupo26       Tipo1         Tipo2         Tipo3
NI     :482   NI     :500   NI     :522   NI     :512   NI     :512   NI     :830   NI     :558   NI     :504   exl:226   exl:311   exl:256
7014   :100   7016   : 79   7019   : 68   7021   : 97   7022   : 99   1018   : 16   7024   : 51   7027   : 78   ext:  6   ext:  5   ext:  5
7015   : 92   7017   : 72   7018   : 65   7020   : 61   7023   : 73   1016   : 13   7025   : 48   7022   : 72   NI :  3   NI :  4   NI :  3
7013   : 75   7015   : 50   7017   : 54   7017   : 43   7020   : 33   1034   :  5   7014   : 45   7017   : 42   ord:644   ord:559   ord:615
7019   : 36   7007   : 30   7016   : 47   7019   : 42   7024   : 32   1010   :  3   7019   : 39   7025   : 38
7018   : 33   7014   : 29   7020   : 37   7022   : 33   7028   : 30   1017   :  3   7023   : 28   7016   : 28
(Other): 61   (Other):119   (Other): 86   (Other): 91   (Other):100   (Other):  9   (Other):110   (Other):117
Tipo4     Tipo5     Tipo6     Tipo7     Tipo8     Tipo9     Tipo10    Tipo11    Tipo12    Tipo13    Tipo14    Tipo15    Tipo16    Tipo17
exl:108   exl:177   exl:172   exl:168   exl:131   exl:186   exl: 97   exl:134   exl:113   exl: 92   exl: 82   exl: 69   exl: 25   exl: 72
ext:  7   ext:  5   ext:  5   ext:  5   ext:  7   ext:  3   ext:  4   ext:  5   ext:  4   ext:  3   ext:  2   ext:  2   ext:  3   ext: 11
NI :  1   NI :145   NI :180   NI :168   NI :152   NI :252   NI :340   NI :322   NI :281   NI :393   NI :336   NI :429   NI :322   NI :515
ord:763   ord:552   ord:522   ord:538   ord:589   ord:438   ord:438   ord:418   ord:481   ord:391   ord:459   ord:379   ord:529   ord:281

Tipo18    Tipo19    Tipo20    Tipo21    Tipo22    Tipo23    Tipo24    Tipo25    Tipo26
exl: 29   exl: 63   exl: 57   exl: 37   exl: 28   exl: 61   ext:  3   exl: 44   exl: 53
ext:  4   NI :482   ext:  7   ext:  2   ext:  8   ext:  4   NI :830   ext:  7   ext:  7
NI :478   ord:334   NI :500   NI :522   NI :512   NI :512   ord: 46   NI :558   NI :504
ord:368             ord:315   ord:318   ord:331   ord:302             ord:270   ord:315
```

Figura 3.26a. Frecuencia de variables “Grupo” y “Tipo” para Ciencias de la Computación plan 1994

```
Tipo24    Tipo25          i1              i2              i3
ext:  2   exl: 28   Min.   :0.000   Min.   : 0.000   Min.   : 0.000
NI :575   ext:  6   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 1.000
ord: 38   NI :388   Median :2.000   Median : 2.000   Median : 2.000
          ord:193   Mean   :2.062   Mean   : 2.402   Mean   : 2.189
                    3rd Qu.:3.000   3rd Qu.: 3.000   3rd Qu.: 3.000
                    Max.   :9.000   Max.   :10.000   Max.   :11.000

       i4              i5              i6              i7
Min.   : 0.000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000
1st Qu.: 1.000   1st Qu.: 1.00   1st Qu.: 1.00   1st Qu.: 1.000
Median : 1.000   Median : 1.00   Median : 1.00   Median : 1.000
Mean   : 1.527   Mean   : 1.73   Mean   : 1.68   Mean   : 1.538
3rd Qu.: 2.000   3rd Qu.: 2.00   3rd Qu.: 2.00   3rd Qu.: 2.000
Max.   :10.000   Max.   :13.00   Max.   :15.00   Max.   :10.000

       i8              i9              i10             i11
Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 1.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
Median : 1.000   Median : 1.000   Median : 1.000   Median : 1.000
Mean   : 1.509   Mean   : 1.558   Mean   : 1.127   Mean   : 1.275
3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 2.000
Max.   :14.000   Max.   :18.000   Max.   :10.000   Max.   :10.000

      i12             i13             i14             i15
Min.   :0.000   Min.   : 0.0000   Min.   :0.000   Min.   :0.0000
1st Qu.:0.000   1st Qu.: 0.0000   1st Qu.:0.000   1st Qu.:0.0000
Median :1.000   Median : 1.0000   Median :1.000   Median :1.0000
Mean   :1.208   Mean   : 0.9967   Mean   :1.018   Mean   :0.8634
3rd Qu.:2.000   3rd Qu.: 1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
```

Figura 3.26b. Frecuencia de variables "Grupo" y "Tipo" para Ciencias de la Computación plan 1994.

Gráficas Generales:

A continuación se presentan tres gráficas de la situación de los alumnos por materia en cada uno de los planes antes analizados.

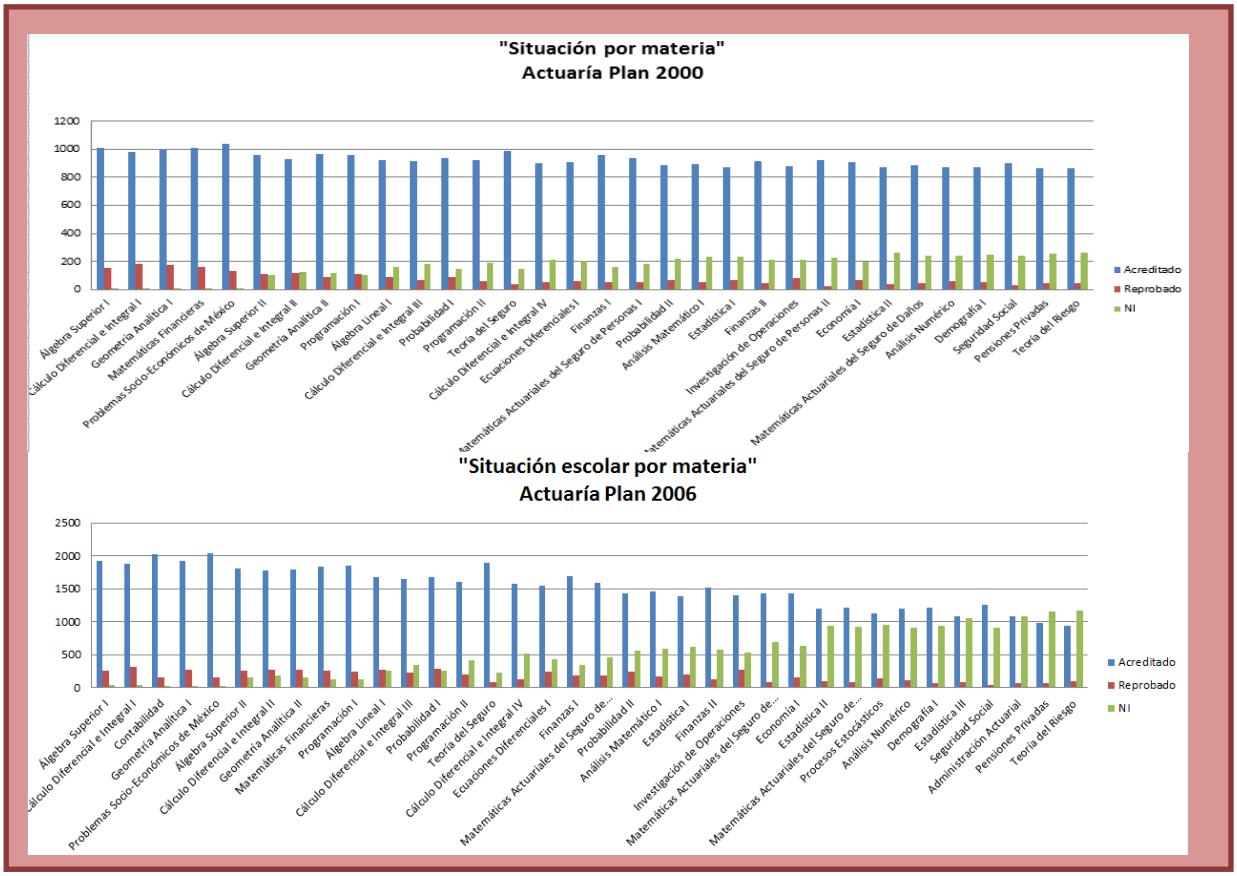


Figura 3.27a Graficas de situación por materia de los tres planes analizados

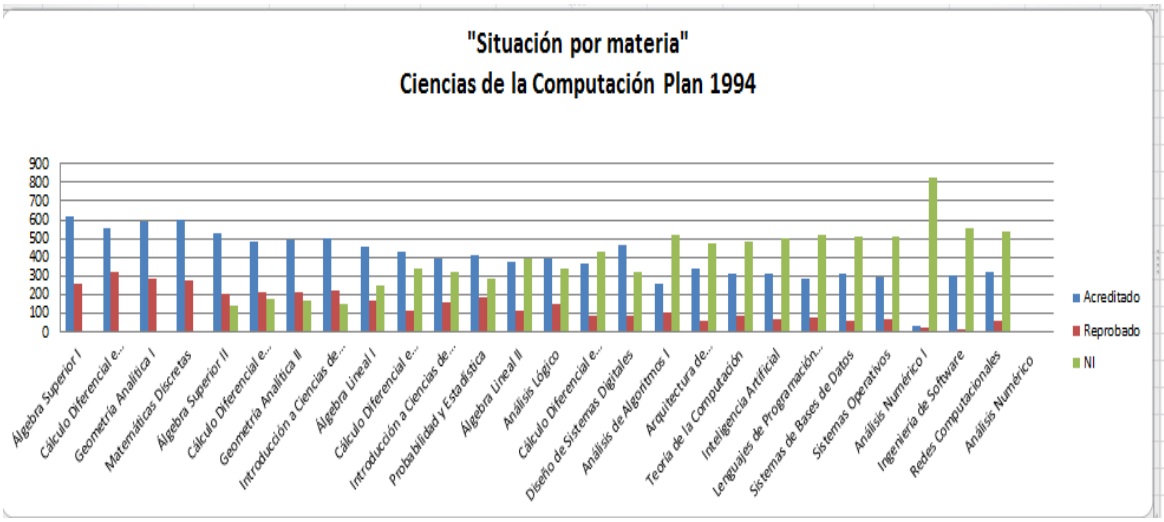


Figura 3.27b Graficas de situación por materia de los tres planes analizados

3.5 Decisiones tomadas

Concluido el análisis exploratorio se toman las siguientes decisiones:

De la base “Datos Alumnos":

- Las variables Carrera, TotalAprobadas, AprobadasNoRecursadas y ReprobadasNoRecursadas no serán ingresadas a los modelos.
- Se modifica el valor de la variable Años del alumno con 24 años y se coloca 6.5 que corresponde al tiempo según sus semestres inscritos.
- Se aplicará la transformación logaritmo para ingresarlas a los modelos a las variables Años, Vespertino, TotalReprobadas, TotalRecursadas, ArobadasYrecursadas, ReproadasYrecursadas y Recursamientos.
- Se aplicará la transformación $y=x^2$ a la variable Promedio.
- Se quitarán de la base los alumnos cuya forma de ingreso es Segunda Carrera o Carrera simultánea debido a que son alumnos que afectan y mueven mucho los datos y que debido a que existe otra carrera en sus vidas, su situación escolar no se comporta igual que los alumnos que cursan una única carrera (objetos de este estudio estudio).

De las bases “Act00, Act06, Comp1994, se toman las siguientes decisiones:

- Se decide sustituir las variables tipo por variables “i” que contendrán la información sobre la cantidad de veces que el alumno inscribió la materia, equivalente a “tipo” porque con más de dos inscripciones pasa de ser ordinario a extraordinario largo. Las variables “i” aportan mayor información que puede ser muy útil en los modelos.

CAPITULO 4. “APLICACIÓN DE MINERÍA DE DATOS. Modelación, Evaluación e Implementación.”

4.1 Introducción

En este capítulo se procederá a la elaboración de los modelos, eligiendo la técnica de Árboles de Decisión utilizando primeramente el algoritmo CART para su elaboración y posteriormente usando Bosques Aleatorios. Todo esto se realizará en R[32].

Se eligieron los árboles de decisión para este trabajo por a su gran capacidad de clasificación, además de que la interpretación de sus reglas es más sencilla gracias a la forma gráfica como pueden presentarse, y con la finalidad de encontrar patrones que ayuden a cubrir el objetivo de la tesis.

La forma de modelar los datos será de la siguiente manera, se realizarán modelos de árboles de decisión para la base “Datos Alumnos”, realizando un árbol para cada una de dos variables objetivo, Terminó y Desertó; haciendo el análisis de cada carrera por separado.

Se utiliza el algoritmo CART ya que es el algoritmo que sigue R para la elaboración de árboles de decisión, recordar que éste algoritmo realiza modelos para variables objetivo binarias, y precisamente la variable objetivo de esta investigación cubre esa característica.

Al terminar con estos modelos se utilizará la técnica de Bosque Aleatorio (Random Forest) para encontrar reglas de asociación para las bases “Act00”, “Act06” y “Comp1994”, se realizará un modelo supervisado con dos diferentes variables objetivo, Terminó y Desertó.

Los bosques aleatorios son elegidos debido a que son una herramienta clasificatoria muy potente que además sirve para bases de datos que tienen muchas variables, como es el caso de los historiales académicos de las bases antes mencionadas.

Una vez realizados los modelos se procederá a la evaluación de los mismos con algunas técnicas descritas en el Capítulo 2.

4.2 Árboles de decisión por método CART

Para realizar la modelación de los datos se utilizará tanto la consola estándar de R como la interfaz del minero de datos que utiliza la librería “rattle()”. (Figura 4.0).

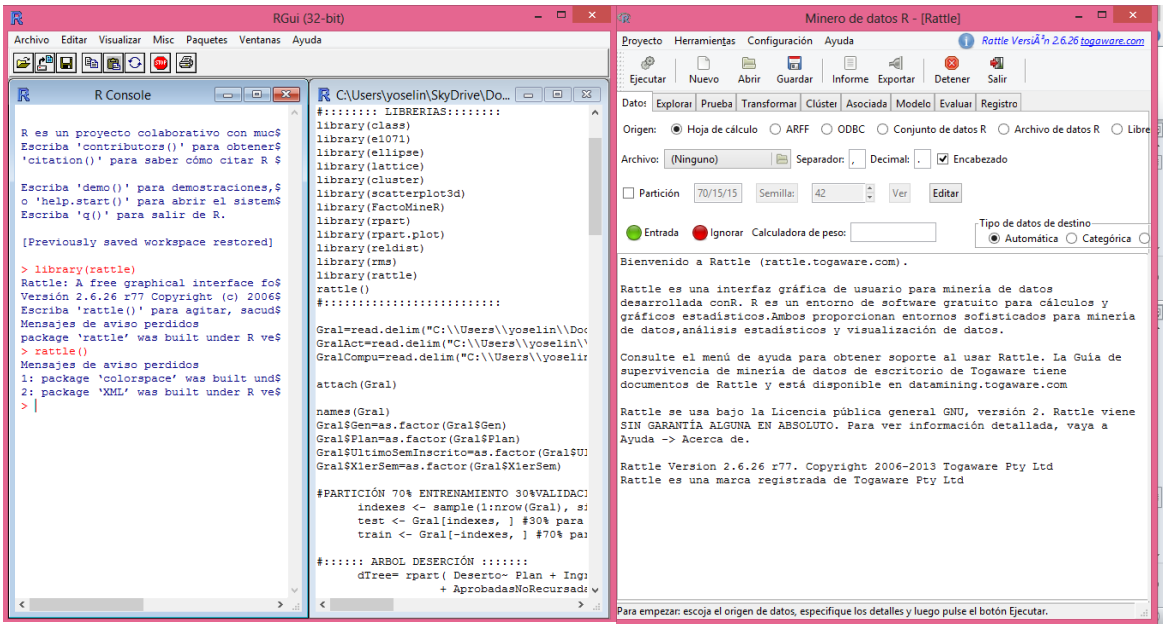


Figura 4.0 Ventanas de consola programador y minero de R.

R realiza los árboles de decisión con el algoritmo CART mediante la función “rpart()” y el bosque aleatorio mediante la función “randomForest”(), las propiedades y características de estas funciones son descritas en el Anexo 1.

R arroja realiza de manera automática un primer árbol, con parámetros en la función predeterminados, a este árbol lo llamaremos árbol maximal y es el resultado de tener una holgura muy grande en el número de capas por lo que está sobreajustado por ello, se requiere llevar a cabo la poda del árbol.

Los resultados de la función rpart son arrojados en forma de diagrama analítico (en código) y a partir de éste realiza un diagrama gráfico del mismo.

También arroja como parte de los resultados una tabla que contiene las variables CP, nsplit, rel error, xerror, xstd; “CP” es un punto de corte, es un índice que indica la complejidad del modelo o límite de tolerancia “nsplit” indica el número de capas que tiene el árbol, “rel error” es el error relativo del modelo estimado con los datos de entrenamiento, “xerror” y “xstd” son estimaciones resultantes de una validación cruzada que realiza R internamente. (Figura 4.1)

Dos de estas variables se utilizan para realizar la poda del árbol, se elige el xerror mínimo para tratar de obtener el árbol óptimo (el mejor de todos y que mejor se

ajusta a los datos con el menor error y mejores resultados) y a partir de ese error se selecciona el CP correspondiente a ese renglón para ponerlo como parámetro en la sección complejidad.

En la parte analítica el árbol se presentan las variables que entran y sus ramificaciones en los dos renglones siguientes, es un diagrama en forma de cascada. La forma de interpretación de la parte analítica consiste en identificar los nodos hoja, marcados por el programa con un asterisco (*) y seguir las reglas hacia arriba.

Por ejemplo en la Figura 4.1 se muestra un ejemplo de árbol con los datos de los alumnos y variable objetivo Desertó y la tabla donde aparece el Xerror y el CP. La interpretación es la siguiente:

La primera variable que entra es Avance, este nodo se ramifica de modo que si el Avance es mayor que 0.37 se genera un nodo hoja (final) que indica que el 97% de los datos cumplen esa condición y no desertan, ahora si el avance es menor a 0.37 entonces entra la variable Recursamientos, si éstos son mayores o iguales a 0.87 entonces entra la variable Plan que tiene dos nodos hoja que divide a los datos en dos grupos el primero 2006 y 2013 y el segundo con otros planes; ahora si los recursamientos son menores a 0.87 entonces se genera otro nodo hoja que indica que el 84% de los datos cumplen esa condición y terminan desertando.

El mínimo de los xerror (0.42043) se encentra en el nsplit número 3, por lo que al realizar la poda se tomaría el CP de 0.002600 y se colocaría en el argumento de complejidad.

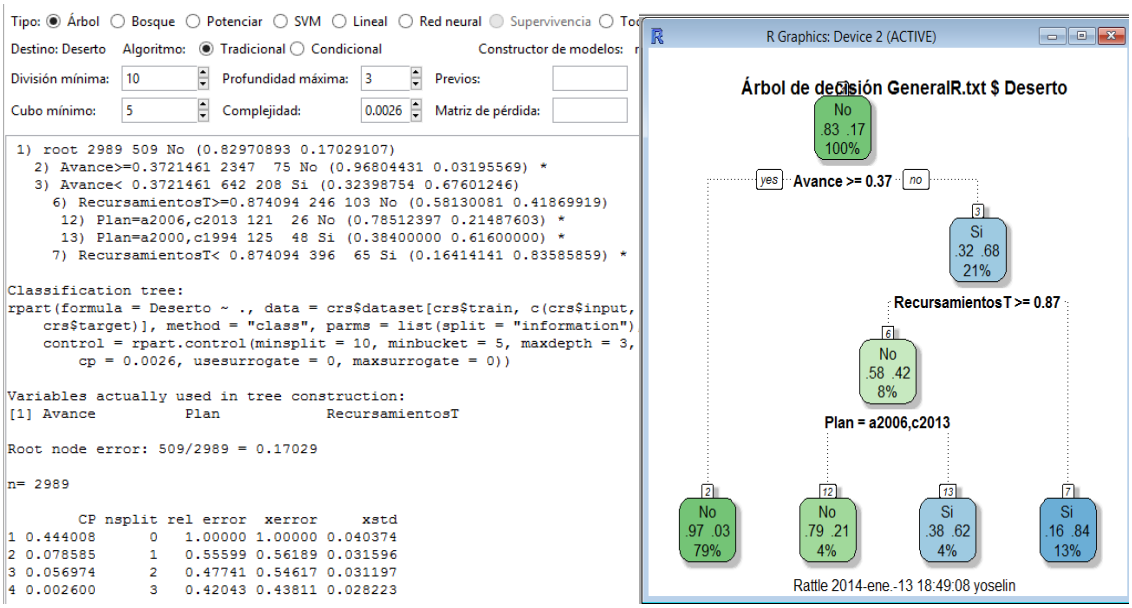


Figura 4.1 Ejemplo de resultados arrojados por R de un árbol de decisión.

Una limitación de la herramienta radica en que a la hora de elaborar el gráfico del árbol, si éste es demasiado grande, el dibujo se ajusta al tamaño de la ventana de la computadora por lo que puede ser que la imagen resulte muy pequeña para y no se alcance a apreciar el contenido, y no se puede realizar ningún tipo de zoom.

Para fines de este trabajo, se incluirán ambos resultados, ya que si no se alcanza a apreciar en el dibujo del árbol, el lector puede referenciarse al diagrama codificado para el entendimiento.

Después de explicado esto, se procede a modelar los árboles. Para comenzar se realiza una partición de los datos 70% para entrenamiento y 30% para validación, esta división es realizada de manera aleatoria en R.

Las variables objetivo serán Terminó y Desertó, cada una en modelos separados y las variables explicativas serán Plan, Ingreso, Gen, Avance, TotalAprobadas, Matutino, InscritasBien, InscritasMal, TotalSem, VespertinoT, YearsT, TotRepT, TotRecT, RecursamientosT, ApRecT, PromedioT. Las variables con una "T" al final del nombre son las resultantes de variables anteriores aplicándoles la transformación correspondiente. (Figura 4.2)

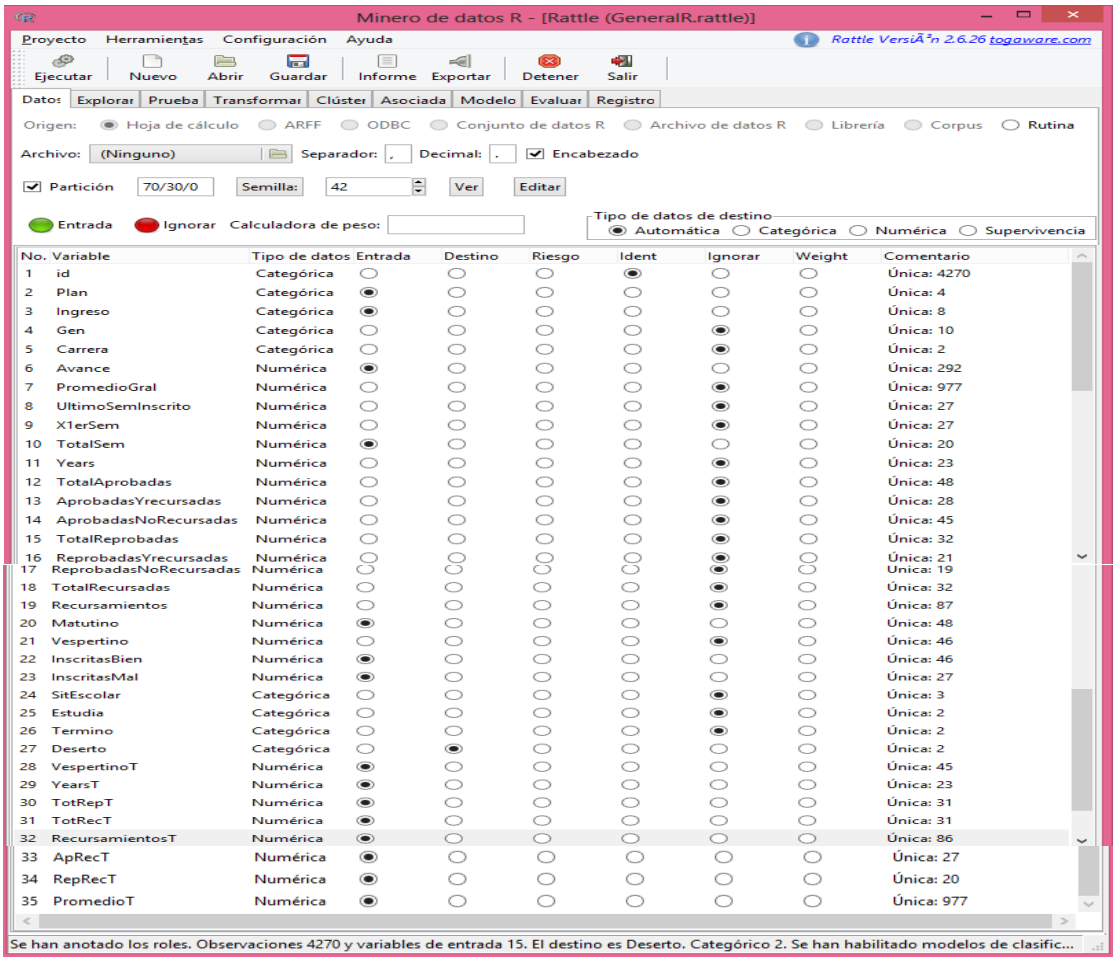


Figura 4.2 Tarea asignada a cada variable de la base "Datos Alumnos"

No se debe olvidar que las variables que tienen una “T” al final de su nombre, son aquellas que transformamos con alguna función por lo que para conocer sus valores reales se debe regresa a conocer sus valores originales.

En la carrera de Ciencias de la Computación resultan árboles más pequeños por lo que el análisis que se hace es más detallado regla por regla, en cambio en Actuaría se hace el análisis de las reglas más importantes y se resume la información en general.

Árbol 1: Actuaría, variable objetivo Desertó

Se realiza un primer modelo de árbol maximal el está sobreajustado, para eliminar esto, se realiza una poda del árbol tomando en cuenta el “xerror” mínimo arrojado por el programa de 0.32024 correspondiente a un cp de 0.0018, dato que se pondrá en los parámetros para crear el árbol óptimo.

Las variables que explican mejor el fenómeno de deserción de los alumnos de actuaría y que por ende terminan como parte de las reglas en el árbol óptimo son Avance, Plan, TotalSem, AñosT, Ingreso, PromedioT, TotRecT, InscritasBien, RecursamientosT, TotRepT, Matutino, RepRecT, VespertinoT.

Resultan 35 reglas de las cuales 17 son positivas a la deserción de los alumnos, por ello se enfoca el análisis a estas 17.

El nodo raíz tiene como variable Avance con punto de corte 38.12%, de aquí los alumnos con un avance superior a ese porcentaje generan tres reglas en las que encontramos los siguientes patrones:

- Un alumno del plan 2006 que deserta lo hace con menos de 6 años en la facultad, con un avance menor al 1% y con menos de 5 materias reprobadas.
- Un alumno del plan 200 deserta en menos de 5 años en la facultad, con más de 5 materias reprobadas, menos de 14 semestres inscritos y un avance menor al 1%, recursando menos de 2 materias.

La otra parte del corte generado por el nodo raíz, deserta de la carrera con un porcentaje menor al 38% generan 13 reglas de las cuales se concluye que:

- Un alumno del plan 2006 deserta con menos de siete materias recursadas y un avance menor al 23%, inscribiendo más de diez materias en el turno matutino y con más de 8 inscritas correctamente según lo marca el plan de estudios, tienen menos de 20 recursamientos en total y un promedio mayor a 5.3.

- Alumnos del plan 2000 con un avance mayor al 29% desertan después de estar 5 años en la facultad, con más de 5 materias recursadas y un promedio mayor a 5. Pero también desertan antes de llegar a esos 5 años inscritos.
- Los alumnos con un tipo de ingreso por Cambio de Carrera y/o Plantel y Concurso de selección del plan 2006 desertan con un avance menor al 30%, con más de 2 años de haber ingresado a la carrera, menos de 21 recursamientos en total, más de 4 materias recursadas y con un promedio mayor a 5.
- Alumnos que ingresaron a la Facultad por Concurso de selección y Pase Reglamentado del plan 2006 desertan con menos de 3 materias reprobadas y recursadas, un promedio superior 7.9, después de 3 semestres inscritos.
- Cuando el promedio es menor a 5 y mayor a 3, después de estar al menos un año en la Facultad, con avance menor al 11%, con menos de 12 materias inscritas en el turno matutino y más de 9.5 materias inscritas en el semestre correcto el alumno tiene altas probabilidades de abandonar su carrera.

En general los alumnos de la carrera de Actuaría desertan con menos del 38% de avance académico, lo que dice que un alumno que pasa de este porcentaje de créditos muy probablemente ya no decida abandonar la carrera. Además los alumnos desertores son personas que no aguantan la presión de estar cursando muchas veces las materias para lograr acreditarlas.

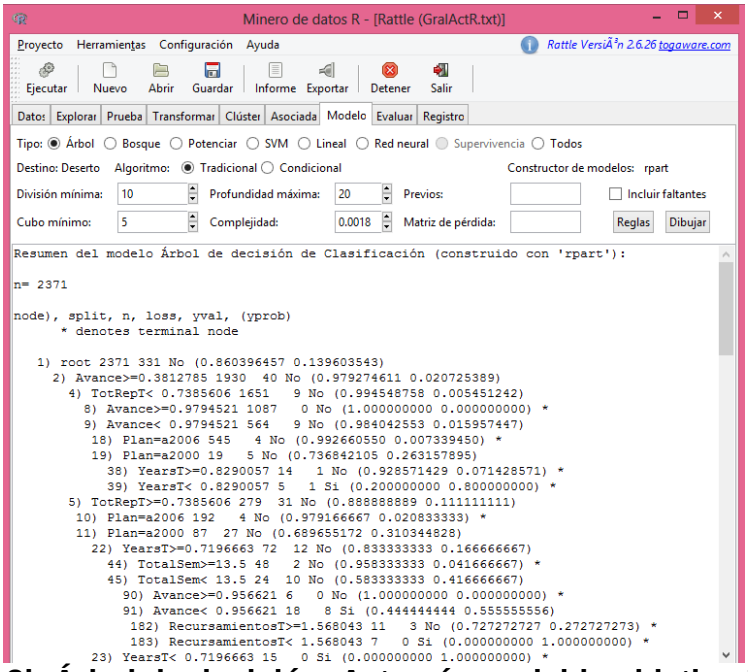
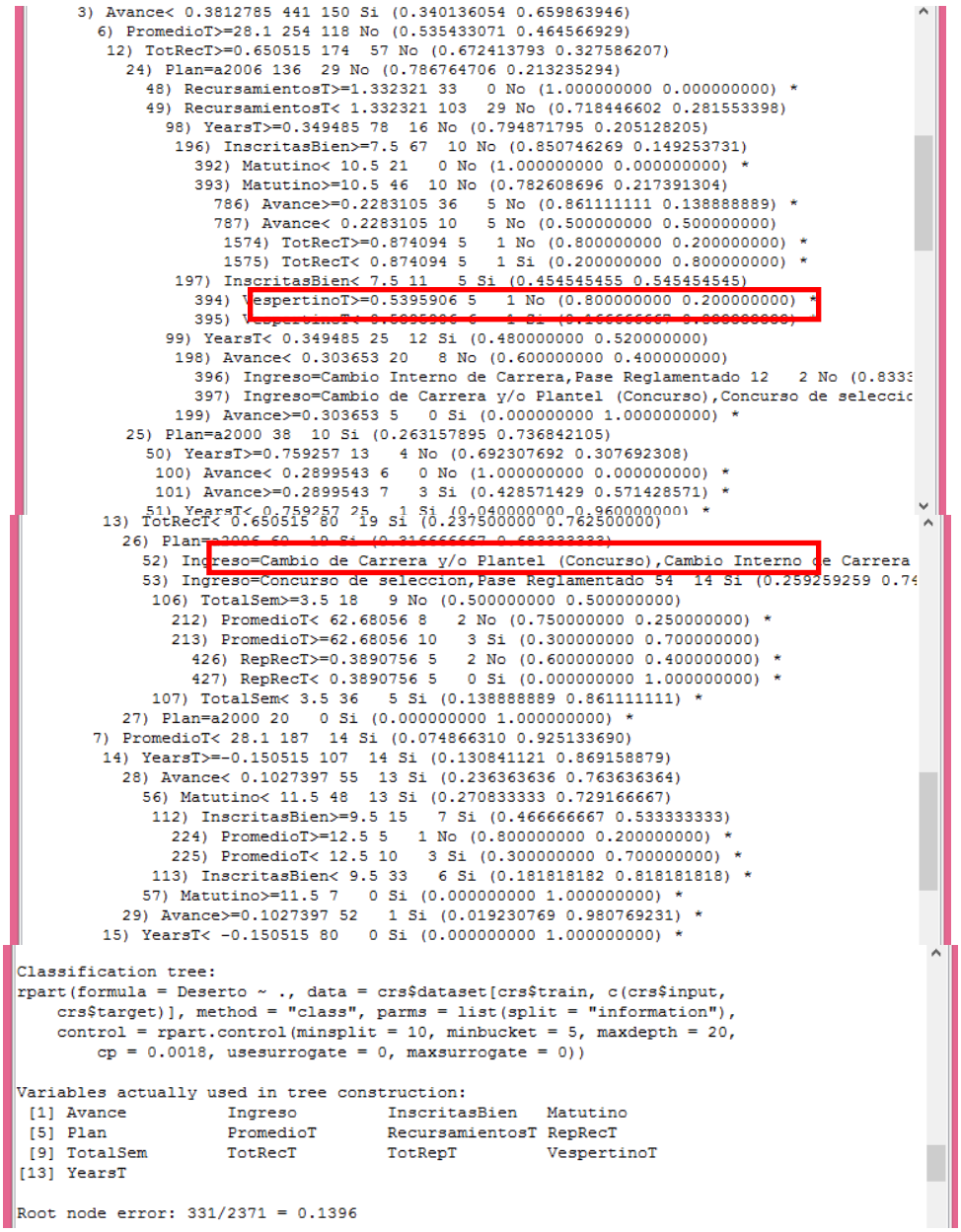


Figura 4.3b Árbol de decisión, Actuaría, variable objetivo Desertó.



```
      3) Avance< 0.3812785 441 150 Si (0.340136054 0.659863946)
        6) PromedioT>=28.1 254 118 No (0.535433071 0.464566929)
         12) TotRecT>=0.650515 174  57 No (0.672413793 0.327586207)
           24) Plan=a2006 136  29 No (0.786764706 0.213235294)
             48) RecursamientosT>=1.332321 33   0 No (1.000000000 0.000000000) *
             49) RecursamientosT< 1.332321 103  29 No (0.718446602 0.281553398)
               98) YearsT>=0.349485 78  16 No (0.794871795 0.205128205)
                196) InscritasBien>=7.5 67  10 No (0.850746269 0.149253731)
                  392) Matutino< 10.5 21   0 No (1.000000000 0.000000000) *
                  393) Matutino>=10.5 46  10 No (0.782608696 0.217391304)
                    786) Avance>=0.2283105 36   5 No (0.861111111 0.138888889) *
                    787) Avance< 0.2283105 10   5 No (0.500000000 0.500000000)
                     1574) TotRecT>=0.874094 5   1 No (0.800000000 0.200000000) *
                     1575) TotRecT< 0.874094 5   1 Si (0.200000000 0.800000000) *
                197) InscritasBien< 7.5 11   5 Si (0.454545455 0.545454545)
                  394) VespertinoT>=0.5395906 5   1 No (0.800000000 0.200000000) *
                  395) VespertinoT< 0.5395906 6   1 Si (0.166666667 0.833333333)
               99) YearsT< 0.349485 25  12 Si (0.480000000 0.520000000)
                198) Avance< 0.303653 20   8 No (0.600000000 0.400000000)
                  396) Ingreso=Cambio Interno de Carrera,Pase Reglamentado 12   2 No (0.8333
                  397) Ingreso=Cambio de Carrera y/o Plantel (Concurso),Concurso de seleccic
                199) Avance>=0.303653 5   0 Si (0.000000000 1.000000000) *
           25) Plan=a2000 38  10 Si (0.263157895 0.736842105)
             50) YearsT>=0.759257 13   4 No (0.692307692 0.307692308)
              100) Avance< 0.2899543 6   0 No (1.000000000 0.000000000) *
              101) Avance>=0.2899543 7   3 Si (0.428571429 0.571428571) *
             51) YearsT< 0.759257 25   1 Si (0.040000000 0.960000000) *
         13) TotRecT< 0.650515 80  19 Si (0.237500000 0.762500000)
           26) Plan=a2006 69  19 Si (0.275362319 0.724637681)
             52) Ingreso=Cambio de Carrera y/o Plantel (Concurso),Cambio Interno de Carrera
             53) Ingreso=Concurso de seleccion,Pase Reglamentado 54  14 Si (0.259259259 0.74
               106) TotalSem>=3.5 18   9 No (0.500000000 0.500000000)
                 212) PromedioT< 62.68056 8   2 No (0.750000000 0.250000000) *
                 213) PromedioT>=62.68056 10   3 Si (0.300000000 0.700000000)
                   426) RepRecT>=0.3890756 5   2 No (0.600000000 0.400000000) *
                   427) RepRecT< 0.3890756 5   0 Si (0.000000000 1.000000000) *
               107) TotalSem< 3.5 36   5 Si (0.138888889 0.861111111) *
           27) Plan=a2000 20   0 Si (0.000000000 1.000000000) *
        7) PromedioT< 28.1 187  14 Si (0.074866310 0.925133690)
         14) YearsT>=-0.150515 107  14 Si (0.130841121 0.869158879)
           28) Avance< 0.1027397 55  13 Si (0.236363636 0.763636364)
             56) Matutino< 11.5 48  13 Si (0.270833333 0.729166667)
              112) InscritasBien>=9.5 15   7 Si (0.466666667 0.533333333)
                224) PromedioT>=12.5 5   1 No (0.800000000 0.200000000) *
                225) PromedioT< 12.5 10   3 Si (0.300000000 0.700000000) *
              113) InscritasBien< 9.5 33   6 Si (0.181818182 0.818181818) *
             57) Matutino>=11.5 7   0 Si (0.000000000 1.000000000) *
           29) Avance>=0.1027397 52   1 Si (0.019230769 0.980769231) *
         15) YearsT< -0.150515 80   0 Si (0.000000000 1.000000000) *

Classification tree:
rpart(formula = Deserto ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(minsplit = 10, minbucket = 5, maxdepth = 20,
        cp = 0.0018, usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
 [1] Avance          Ingreso         InscritasBien   Matutino
 [5] Plan            PromedioT       RecursamientosT RepRecT
 [9] TotalSem        TotRecT         TotRepT         VespertinoT
[13] YearsT

Root node error: 331/2371 = 0.1396
```

Figura 4.3b Árbol de decisión, Actuaría, variable objetivo Desertó.

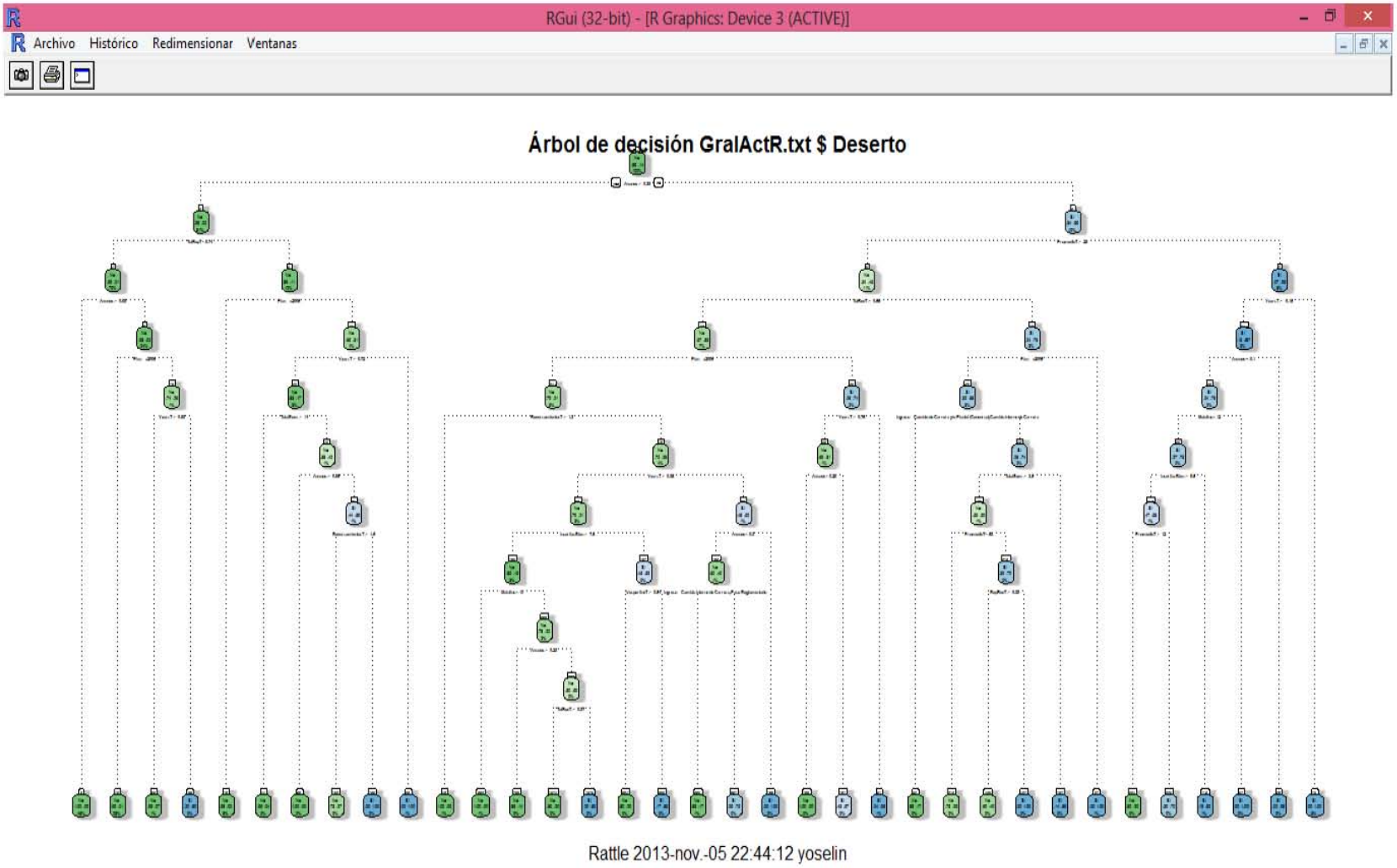


Figura 4.4 Diagrama de árbol de decisión, Actuaría, variable objetivo Desertó.

Árbol 2: Actuaría, variable objetivo Terminó

Para la creación de éste modelo se realiza una división de 70% de los datos de entrenamiento y 30% para validación.

El árbol maximal resulta con un total de 13 capas, al realizar la poda del árbol, se encuentra un “xerror” mínimo de 0.0013 correspondiente a un cp de 0.007813 el cual se le asignará como valor en la fórmula para encontrar el árbol óptimo mostrado en las figuras 4.5 y 4.6.

Resultan 24 reglas, nos enfocaremos a analizar las 12 con respuesta positiva a que los alumnos terminen.

Las variables con mayor peso y reflejadas en el árbol son Ingreso, InscriasBien, InscritasMal, Matutino, Plan, RepRecT, TotalSem, TotRecT, TotRepT, YearsT.

La variable que inicia la división del árbol es la de mayor importancia que es Total de semestres y el punto de corte es 7.5, que nos indica a los alumnos que terminaron en un tiempo menor al establecido en el plan de estudios. Para estos alumnos se obtienen dos reglas de las que se concluye lo siguiente:

Son alumnos con más de 24 materias inscritas en el turno matutino, del plan 2006 y por tipo de ingreso de Cambio de Carrera (todas las modalidades) y Pase Reglamentado, con materias inscritas en un semestre que no le corresponde mayores a 6 y con menos de 4 materias reprobadas en total.

Entonces se espera que un alumno que cumpla con estas características sea propenso a terminar su carrera en menos del tiempo establecido, con excelente historial académico pero no se sabe la calidad (promedio) con que salgan. Nótese que esta es una tendencia actual ya que son alumnos del plan 2006.

De los alumnos que tienen un total de semestres superiores o iguales a los establecidos por el programa (8 semestres), se generan 10 reglas de clasificación que se pueden ver claramente en el diagrama de la figura 4.6 y de las cuales se obtienen los siguientes patrones:

- La culminación de la carrera está relacionada directamente con el número de materias reprobadas y recursadas es decir los alumnos que no dejan de intentar pasar las materias a pesar de reprobarlas logran terminar.
- Las materias inscritas bien juegan un papel importante, pero no determinante para que un alumno termine ya que de ella se desprenden otras características que combinadas hacen que se logre el objetivo.
- Los alumnos del plan 2000 que terminan, tienen pocas materias reprobadas, menos de 24 materias inscritas bien y más de 16 materias inscritas en semestres incorrectos.

- Los alumnos que inscriben más de 24 materias en semestres correctos con más de 2 materias reprobadas y del plan 2000 terminan en menos de 10 años mientras que los del plan 2006 terminan en menos de 6 años.
- El turno juega un papel importante ya que la mayoría de los alumnos tienen inscritas sus materias en el turno matutino, lo que ayuda en la conclusión de la carrera.

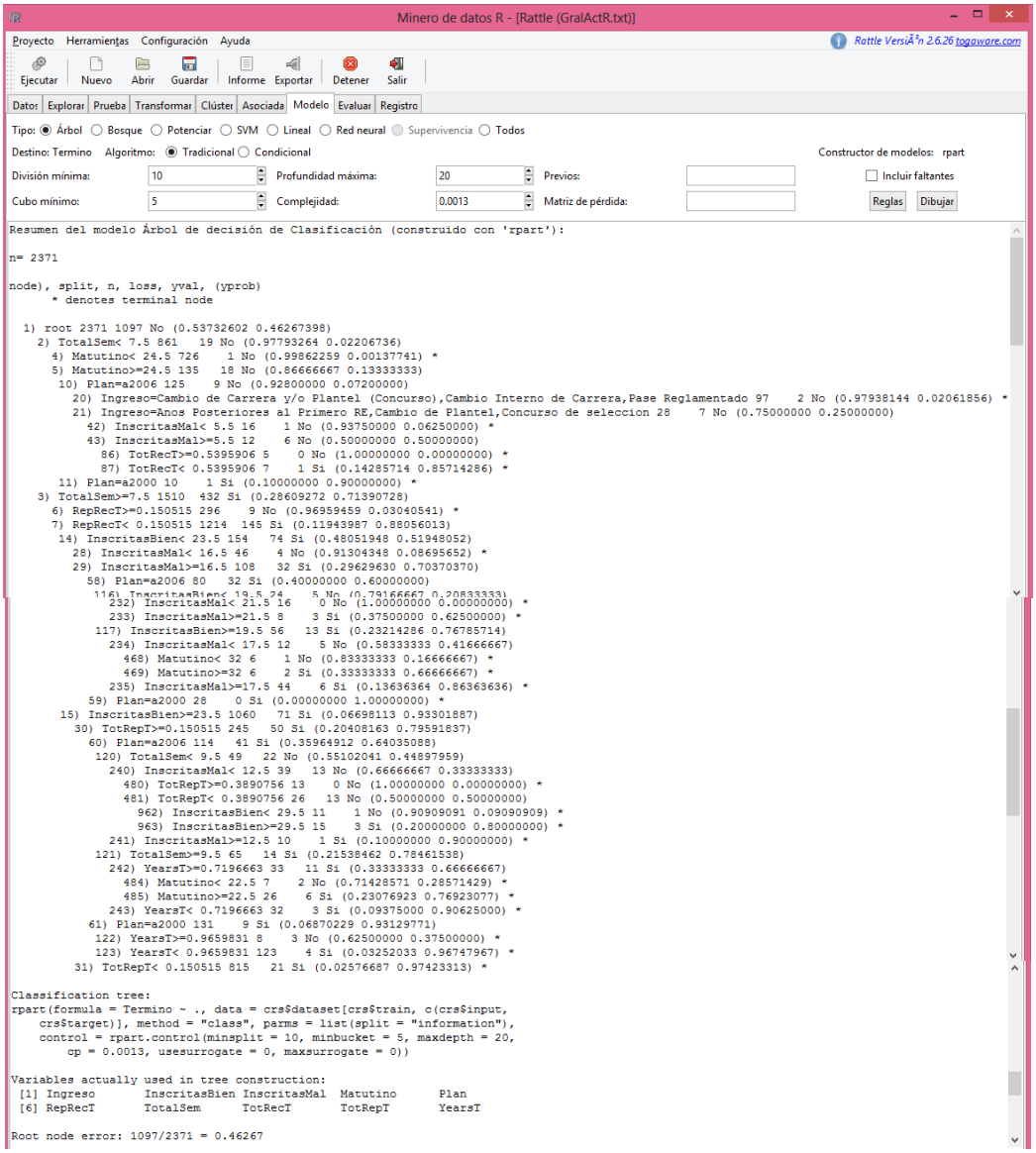


Figura 4.5 Árbol de decisión, Actuaría, variable objetivo Terminó.

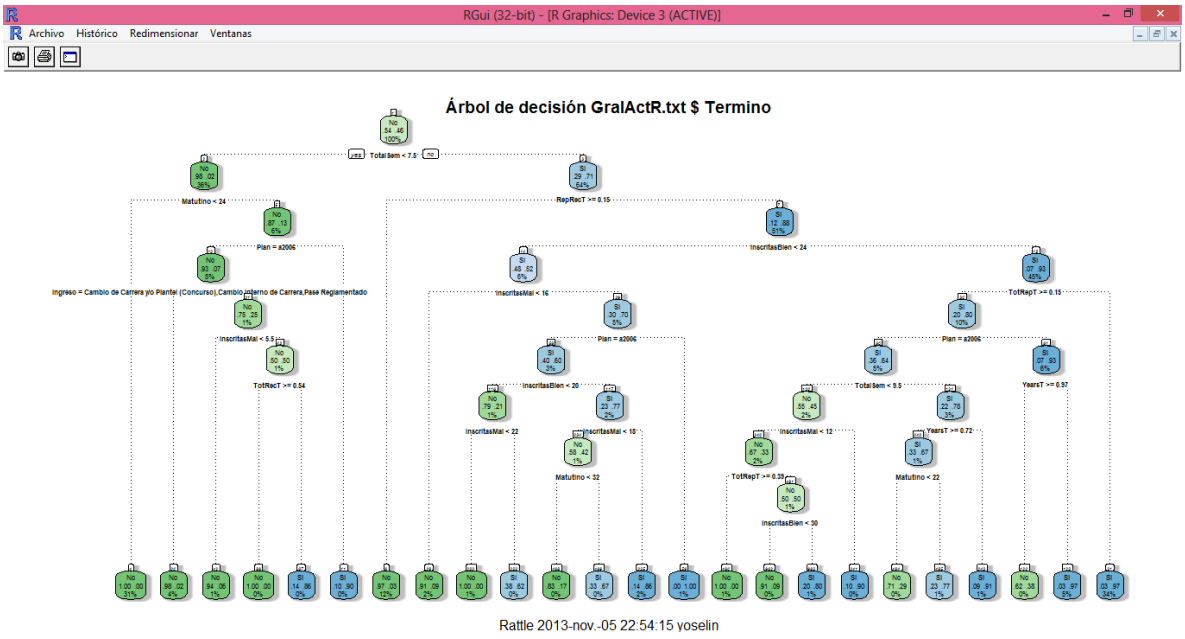


Figura 4.6 Diagrama de árbol de decisión, Actuaría, variable objetivo Terminó.

Árbol 3: Ciencias de la Computación, variable objetivo Desertó

Para la creación de éste modelo y debido a que la base de datos es pequeña y con el objetivo de reducir el error, se realiza una división de 80% de los datos de entrenamiento y 20% para validación.

El árbol maximal resulta con un total de 10 capas, al realizar la poda del árbol, se encuentra un "xerror" mínimo de 0.38421 correspondiente a un cp de 0.0078 el cual se le asignará como valor en la fórmula del árbol óptimo dando como resultado las figuras 4.7 y 4.8 de las que se obtiene la siguiente información.

Las variables más importantes y que son tomadas en cuenta por el programa para la construcción del modelo son Avance, InscritasMal, PromedioT, RecursamientosT, VespertinoT y AñosT.

Se obtienen un total de 12 reglas, de las cuales 6 son el objeto de estudio ya que contiene una respuesta positiva a la deserción de los alumnos, dichas reglas muestran que un alumno deserta cuando:

1. El avance académico es menor al 32%, las materias inscritas en el turno vespertino son menores a 6, un promedio mayor a 3.53, materias inscritas mal mayores a 2, después de 2 años, tiene un total de recursamientos menores o iguales a 9 pero menores a 15.

2. El avance académico es menor al 32%, las materias inscritas en el turno vespertino son menores a 6, tiene un promedio mayor a 3.53 y más de 3 recursamientos, cuando tiene materias inscritas en semestres incorrectos superiores a 2 y menos de 3 años en la facultad.
3. El avance académico es menor al 32%, con materias en vespertino menores a 6 y un promedio mayor a 4 y con menos de 3 recursamientos en total.
4. El avance académico es menor al 32%, las materias inscritas en el turno vespertino son menores a 6 y un promedio menor a 3.56.

De este modelo se puede concluir en general que los alumnos desertan en los primeros años de la carrera, con un avance menor al 32%, con promedio extremadamente bajo y en realidad después de reprobar sus primeras materias y no poder pasarlas al tercer recursamiento.

Claramente el turno impacta en la deserción de los alumnos, pero no se puede decir que es un factor determinante ya que al ser poca la población en la carrera de Ciencias de la Computación y la mayoría estar inscritos en el turno matutino, se refleja en el modelo como si no hubiera deserción en la tarde, pero es porque el número de alumnos en ese turno no tiene el peso suficiente para encontrar un patrón.

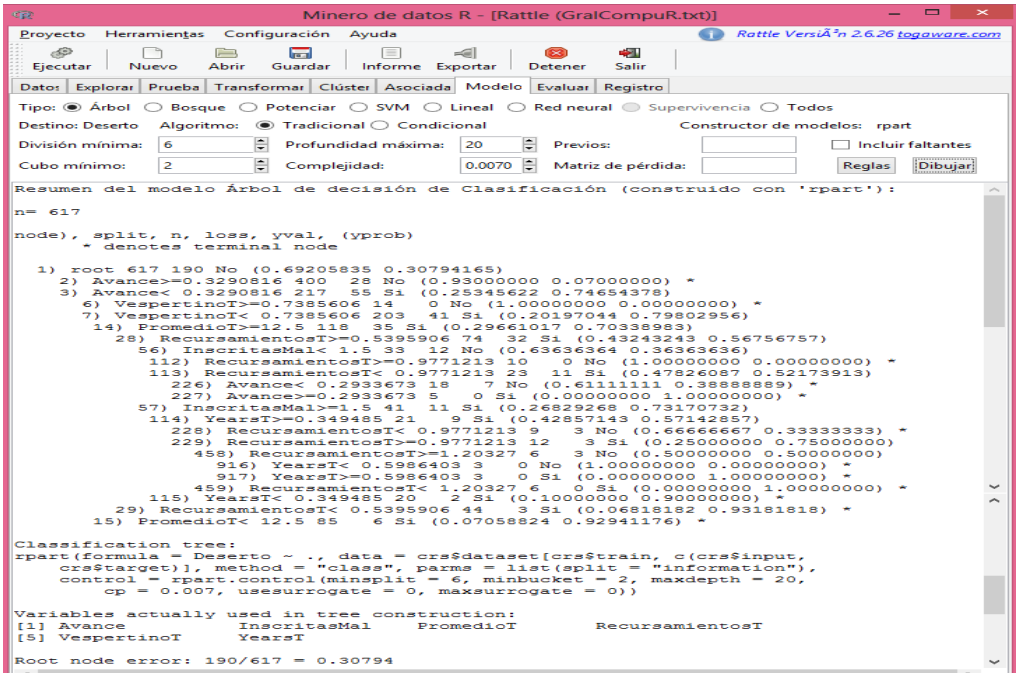


Figura 4.7 Árbol de decisión, Ciencias de la Computación, variable objetivo Desertó.

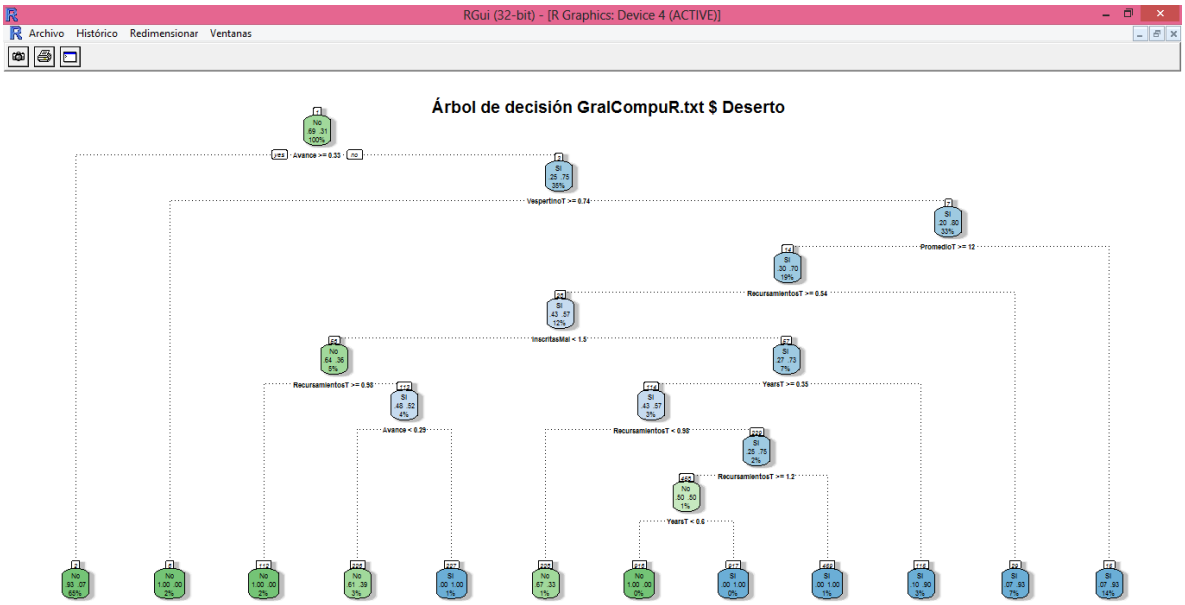


Fig 4.8 Diagrama de árbol de decisión, Ciencias de la Computación, variable objetivo Desertó.

Árbol 4: Ciencias de la Computación, variable objetivo Terminó

Para la creación de éste modelo y debido a que la base de datos es pequeña y con el objetivo de reducir el error, se realiza una división de 80% de los datos de entrenamiento y 20% para validación.

El árbol maximal resulta con un total de 8 capas, al realizar la poda del árbol, se encuentra un “xerror” mínimo de 0.2437 correspondiente a un cp de 0.003 el cual se le asignará como valor en la fórmula del árbol óptimo.

Como puede observarse en el diagrama de la figura 4.10 el modelo arroja 15 reglas o patrones de comportamiento de los alumnos para clasificarlos descritos en código en la figura 4.9, de las cuales 7 son respuestas positivas a la variable objetivo Terminó y son justo estas las que importan ya que el que no haya terminado implica o que desertó o que sigue estudiando.

Las variables que entraron al modelo y tienen impacto en la conclusión de la carrera de Ciencias de la Computación son InscritasBien, Matutino, RepRecT, TotalSem, TotRecT, TotRepT y VespertinoT.

Figura 4.9 Árbol de decisión, Ciencias de la Computación, variable objetivo Terminó.

Se describen a continuación las reglas relevantes del modelo donde un alumno concluye su carrera si:

1. El número de materias inscritas en el turno matutino es menor a 26, el número de materias inscritas en el turno vespertino es mayor a 13 y el total de materias reprobadas es menor a 3. Es decir un alumno que no reprueba y con horario mixto o vespertino mayormente.

2. El número de materias inscritas en el turno matutino es mayor o igual a 36 y, las materias reprobadas y recursadas son más de 3 y con un total de

semestres mayor a 12. De aquí concluimos que los alumnos del turno matutino terminan su carrera en seis años.

3. El número de materias inscritas en el turno matutino es mayor o igual a 36, tiene menos de dos materias reprobadas y recursadas, el total de semestres es mayor a 7.5, el total de materias reprobadas es mayor a una, las materias inscritas bien son menos de 34, tiene más de 12 recursamientos. Es decir son alumnos del turno matutino con gran tenacidad para cursar varias veces sus materias y que no están inscribiendo las materias de acuerdo al plan de estudios.

4. Los alumnos estudian mayormente en el turno matutino.

Como se puede observar las reglas no revelan mucha información nueva, solo el hecho de que los alumnos del turno matutino son más propensos a terminar su carrera y que el inscribir materias en semestres correctos no es una regla que impacte directamente en el resultado.

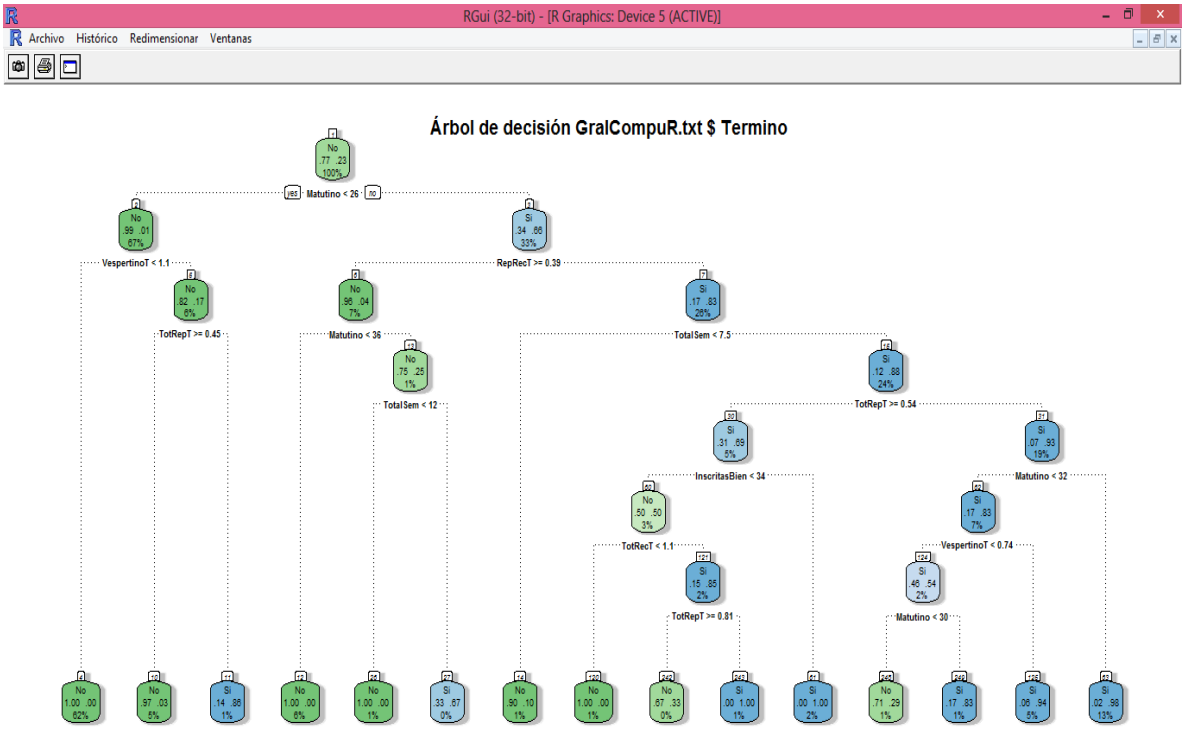


Figura 4.10 Diagrama de árbol de decisión, Ciencias de la Computación, variable objetivo Terminó.

4.3 Árboles de decisión usando Bosque Aleatorio

Como se vio en el capítulo 2, ésta técnica sirve para detectar interacciones entre las variables, la visualización de las reglas es sencilla y su aprendizaje es rápido. Por ello se eligió como herramienta de análisis de las bases “Act00”, “Act06” y “Comp1994”.

Se realizaron varias pruebas con las variables de la totalidad de las materias, grupos inscritos y cantidad de inscripción y los resultados arrojados demostraban que las variables “grupo” tenían gran peso para el agrupamiento de los datos pero no arrojaban resultados relevantes, solo metían ruido. Además algunas materias presentaban una gran cantidad de categorías en su grupo por lo que la técnica no se podía aplicar. Se tomó entonces la decisión de no incluir dichas variables en los modelos.

Debe notarse que para que un alumno concluya debe haber aprobado la totalidad de las materias, por lo que las variables que se meten en los modelos cuya variable objetivo es “Terminó”, son únicamente las tipo “i” para conocer qué materias son las que cursan más o menos y que están siendo factor en la culminación.

Para cada base de datos se realizaron varias pruebas, resultando los mejores Bosques Aleatorios los siguientes:

4.3.1 Base Comp1994

Variable Objetivo “Terminó”, Bosque1:

Los parámetros utilizados fueron los siguientes: número de árboles a construir igual a 50 y número de variables para tomar en cuenta en cada división igual a 8.

De esos árboles, se analiza el modelo 45 que resulta con más reglas (51 con 24 de respuesta positiva).

Para concluir la carrera de Ciencias de la Computación las materias clave son: Geometría Analítica I, Cálculo Diferencial e Integral I, Introducción a ciencias de la Computación II, Álgebra Superior I y Cálculo Diferencial e Integral IV ya que un alumno que concluye la carrera las mantiene con un promedio menor de 1.5 inscripciones, es decir están inscritas en forma ordinaria.

Por otro lado las materias en las que los alumnos que terminan deben tener mayor tenacidad y probablemente cursarlas varias veces en extraordinario largo son: Teoría de la Computación y Redes Computacionales.

Los alumnos que están concluyendo su carrera reflejan en varias reglas que en materias como Cálculo Diferencial e Integral I y IV, Probabilidad y Estadística y

Análisis Lógico, el número de veces que el alumno inscribió para poder acreditar la materia es mayor a 2, es decir que son extraordinarios largos.

Alumnos que culminan y cursaron más de una vez Geometría Analítica I, también lo hicieron en las materias de Cálculo Diferencial e Integral IV y Álgebra Superior II; otro patrón encontrado es que los algunos que solo cursan más de una vez Análisis Lógico también lo hacen en Cálculo Diferencial e Integral I y Probabilidad y Estadística, ambas reglas podrían indicar que las materias están ligadas altamente en conocimiento.

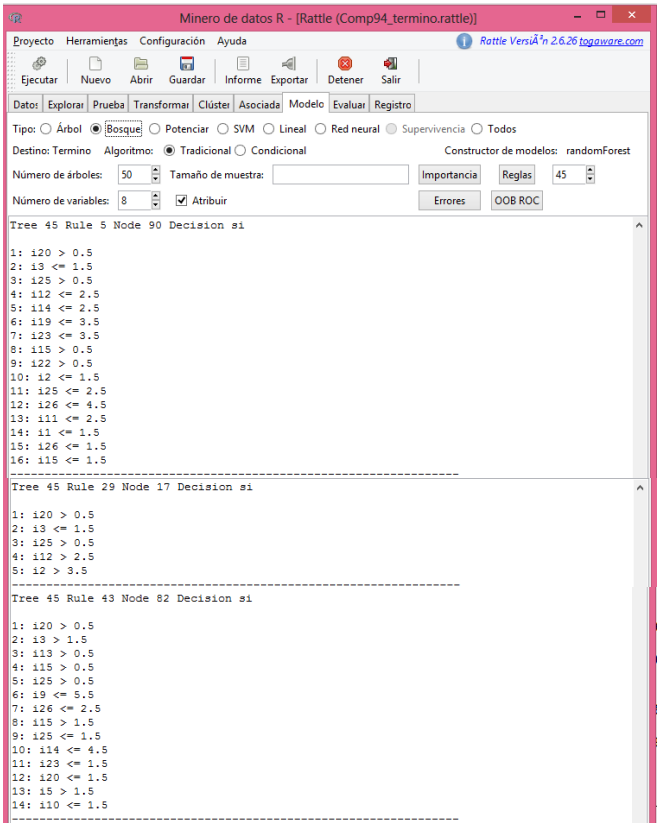


Figura 4.11 Reglas más sobresalientes para variable objetivo Terminó, Ciencias de la Computación

Variable Objetivo "Desertó", Bosque 2:

Los resultados que arroja el modelo pueden ser interpretados como las materias que están siendo filtros y clave para que el alumno deserte y no concluya su carrera.

Este es el caso de las materias de Álgebra Lineal I, Álgebra Superior II, Cálculo Diferencial e Integral III, Probabilidad y Estadística, Diseño de Sistemas Digitales,

Análisis Lógico e Introducción a Ciencias de la Computación I. Dichas materias aparecen en el 80% de las reglas que terminan en una deserción. (Figura 4.12)

Por otra parte la regla 27 indica que los alumnos que tienen más de 2.5 promedio de inscripciones en las materias de Cálculo Diferencial e Integral II (i6), Álgebra Superior I (i1) e Introducción a Ciencias de la Computación II (i11) tienden a desertar más, esto quiere decir que un alumno que no logra acreditar esas materias en más de 3 intentos tiende a desertar de la carrera. (Figura 4.12a y 4.12b)

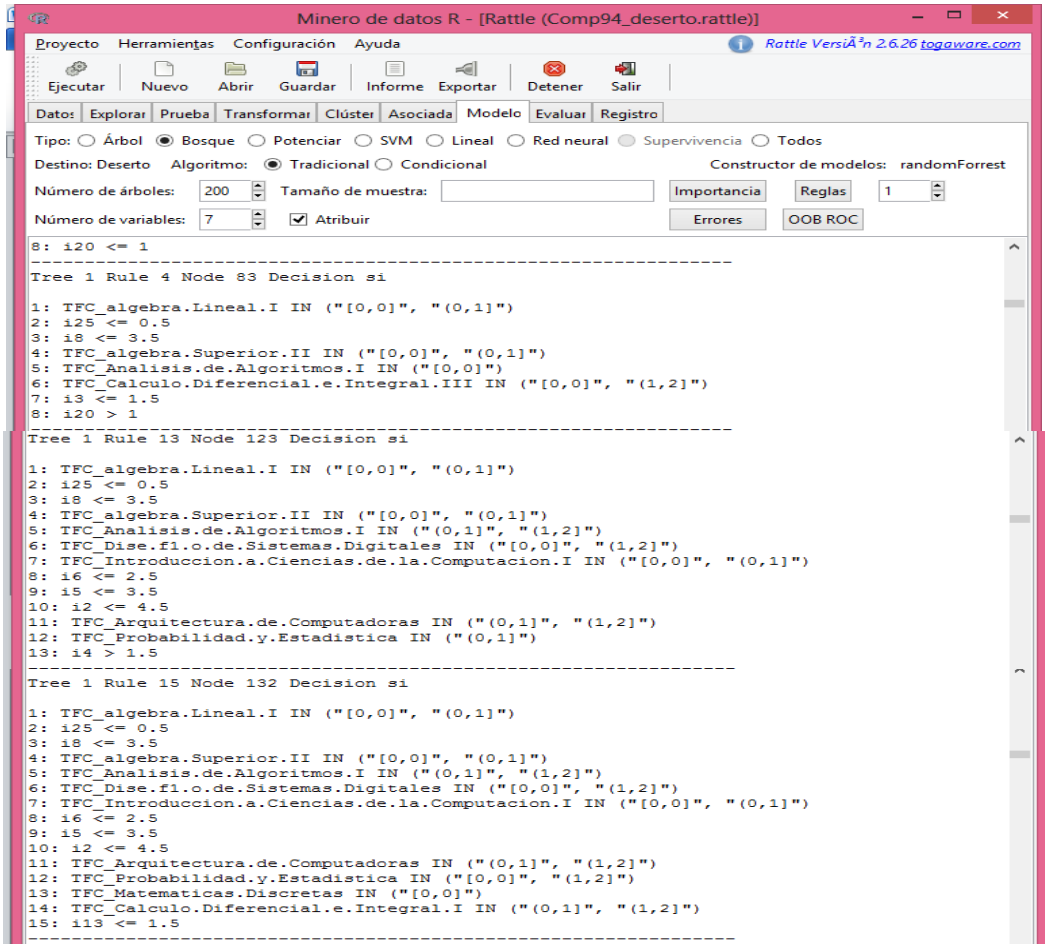


Figura 4.12a Reglas más sobresalientes para variable objetivo Desertó, Ciencias de la Computación (Parte 1)

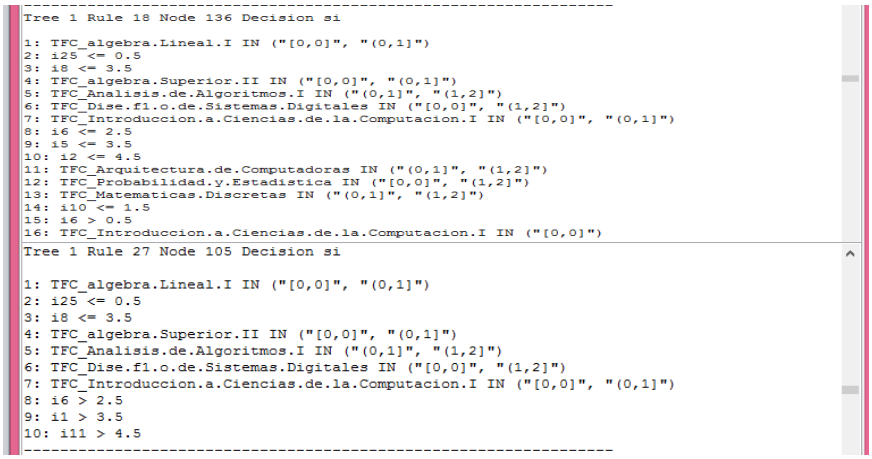


Figura 4.12b Reglas más sobresalientes para variable objetivo Desertó, Ciencias de la Computación (Parte 2)

4.3.2 Base Act00

Variable Objetivo "Terminó", Bosque3:

Se crean 80 árboles, con 5 variables para tomar en cuenta en cada división.

Dada la cantidad de árboles generados, se analizan las reglas de cada uno y resultan reglas muy parecidas, por lo que se explicará en este trabajo el modelo número 1 que arroja el mayor número de reglas.

En total se obtienen 48 reglas de las cuales resultan 21 con respuesta positiva de las cuales se rescatan las siguientes características.

Las variables más influyentes para obtener respuesta favorable son *i32, i26, i28, i23, i15, i21, i10, i13, i22, i9, i11, i29 e i19,* correspondientes a las materias de Teoría del riesgo, Estadística II, Análisis numérico, Investigación de Operaciones, Álgebra Lineal I, Programación II, Finanzas II, Programación I, Cálculo Diferencial e Integral III, Demografía I y Probabilidad II.

Todo alumno que termina debe haber cursado al menos una vez Teoría del riesgo, lo cual es muy lógico ya que es una materia obligatoria, pero lo rescatable es que es la materia con más peso y se toma como punto de partida.

La cantidad de inscripciones en cada materia varía pero las materias que necesitan ser inscritas un número menor igual a 3 inscripciones para que un alumno tenga más probabilidad de terminar son Estadística II, Investigación de Operaciones, Cálculo Diferencial e Integral III, Programación I, Programación II, Cálculo Diferencial e Integral I, Seguridad Social, Cálculo Diferencial e Integral II, Probabilidad I y Álgebra Lineal I.

Mientras que en algunas reglas arrojadas las materias de Estadística III, Investigación de Operaciones, Cálculo Diferencial e Integral IV, Álgebra Lineal I y Programación II tienen una cantidad de inscripciones elevadas y un alumno que termina debe tener mucha tenacidad para recursarlas un número de veces mayor o igual a 4.

Las materias que más llaman la atención son Cálculo Diferencial e Integral IV que influye en la mayoría de las reglas y cuyo punto de corte es de 6.5 inscripciones y Álgebra Lineal I con un punto de corte de 11.5 inscripciones, de lo que podemos concluir que un alumno que termina tiene problemas para acreditar estas dos materias con una probabilidad muy elevada, siendo también los filtros más importantes, cosa que se comprobará al analizar la variable Desertó.

En la Figura 4.13 se muestran algunas de las reglas más ilustrativas de estos resultados.

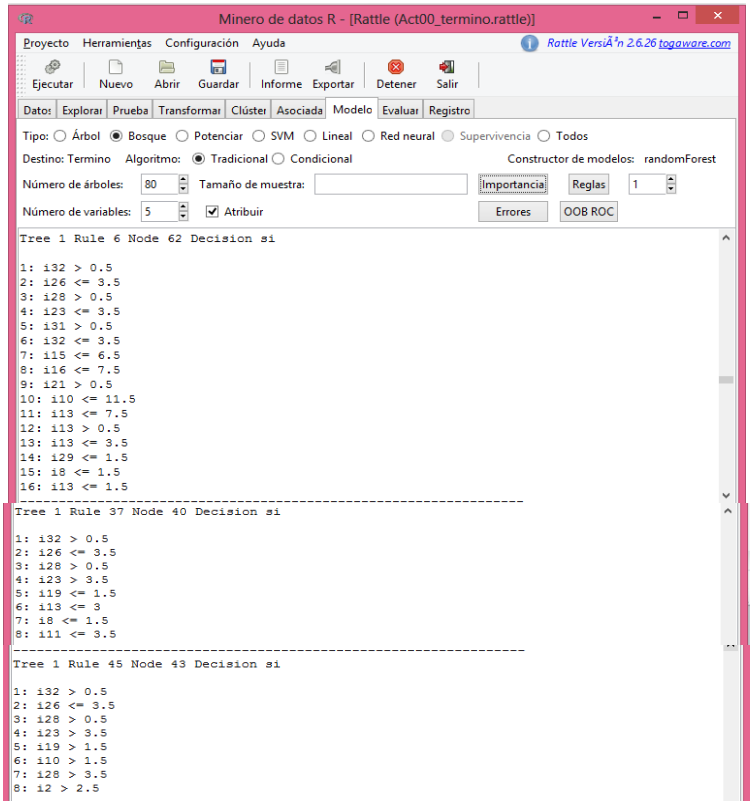


Figura 4.13 Reglas más sobresalientes para variable objetivo Terminó, Actuaría Plan 2000

Variable Objetivo "Desertó", Bosque4:

Se crean 200 árboles con un número de variables mínimas para cada división igual a 7, se analizan las reglas de cada uno y el modelo número 1 arroja el mayor número de reglas (40 con 20 de respuesta positiva).

La materia con mayor peso para la toma de decisiones es Pensiones Privadas, como es una de las materias de los últimos semestres, la división es de quienes llegaron a inscribir la materia y quienes ya no llegaron a inscribirla.

De los alumnos que no llegaron a inscribir la materia resultan los filtros de Probabilidad II y de materias que después de cursarlas más de 2 veces el alumno se da por vencido, estas materias son Álgebra Superior I, Cálculo Diferencial e Integral II, Geometría Analítica II, Álgebra Lineal I y Cálculo Diferencial e integral III. Esta parte de la población deserta antes de llegar a la mitad de la carrera.

Ahora de los alumnos que logran inscribir Pensiones Privadas ya sea que llegaron a su último semestre o que por alguna razón adelantaron la inscripción de la materia se generan reglas en donde saltan a primera vista otros filtros importantes en las materias de Cálculo Diferencial e Integral III y IV, Análisis Numérico y Matemáticas Actuariales del seguro de personas II. Así mismo resalta que los alumnos que desertan tienen inscripciones superiores a 3 en las materias de Álgebra Superior II, Análisis Matemático I, y Geometría Analítica I. Esta parte de la población deserta con más del 40% de la carrera acreditada.

En las siguientes dos figuras se muestran algunas reglas representativas.

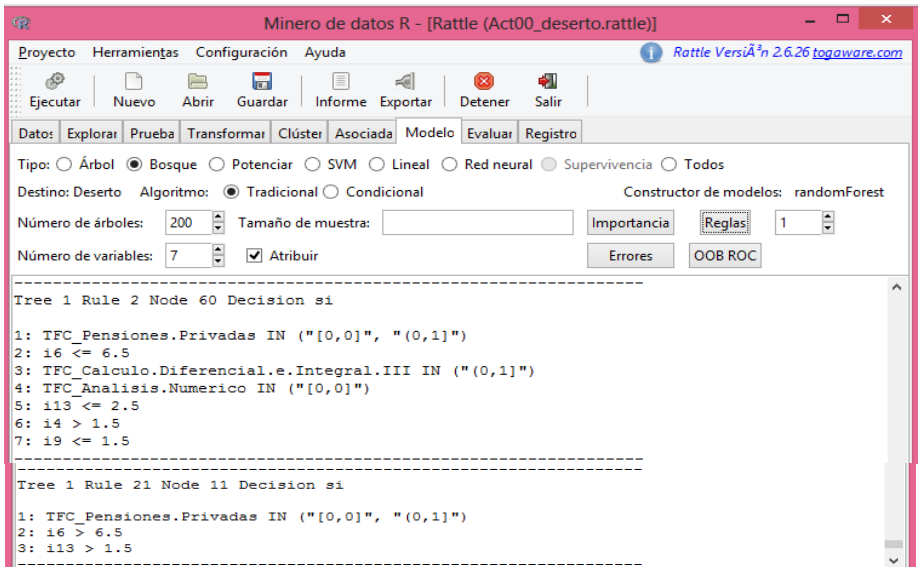


Figura 4.14a Reglas más sobresalientes para variable objetivo Desertó, Actuaría Plan 2000 (Parte 1)

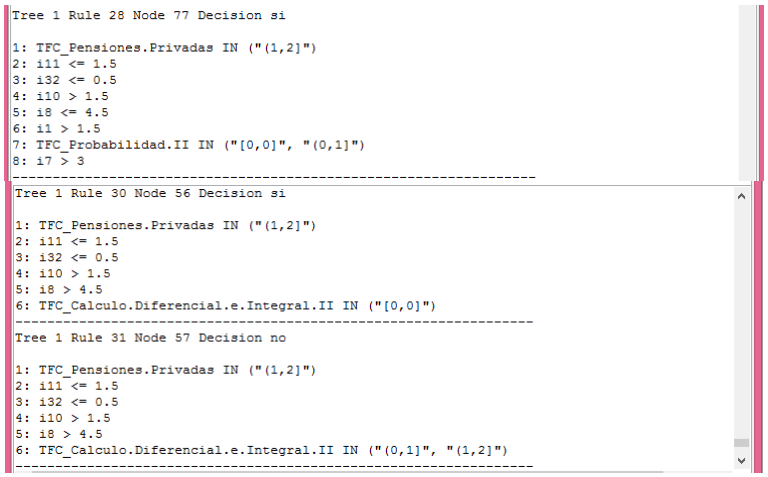


Figura 4.14b Reglas más sobresalientes para variable objetivo Desertó, Actuaría Plan 2000 (Parte 2)

4.3.3 Base Act06

Variable Objetivo "Terminó", Bosque5:

Se crean 100 árboles, con 6 variables para tomar en cuenta en cada división.

De los árboles generados, se explicará en este trabajo el modelo número 50 que arroja el mayor número de reglas.

En total se obtienen 112 reglas de las cuales 54 son de respuesta positiva y se rescata las siguientes características.

La variable de mayor peso es la *i27* seguida de la *i10, i36* e *i32,* correspondientes a las materias de Estadística II, Programación I, Teoría del Riesgo y Estadística III.

Para que un alumno tenga más probabilidad de terminar, las materias que necesitan ser inscritas un número menor igual a 2, es decir ser cursadas como ordinario son Álgebra Superior I, Contabilidad, Geometría Analítica, Matemáticas Financieras, Programación I y II, Cálculo Diferencial e Integral IV, Probabilidad I, Análisis Matemático I, Ecuaciones Diferenciales I, Teoría del Riesgo y Pensiones Privadas.

Por otro lado las materias que un alumno que termina normalmente pasa en extraordinario largo es decir inscripciones mayores o igual a 3 son Geometría Analítica II, Finanzas I, Finanzas II y Procesos Estocásticos, pero estos recursamientos deben ser menores a 5 inscripciones.

En la figura 4.15 se presentan algunas de las reglas más sobresalientes de este modelo.

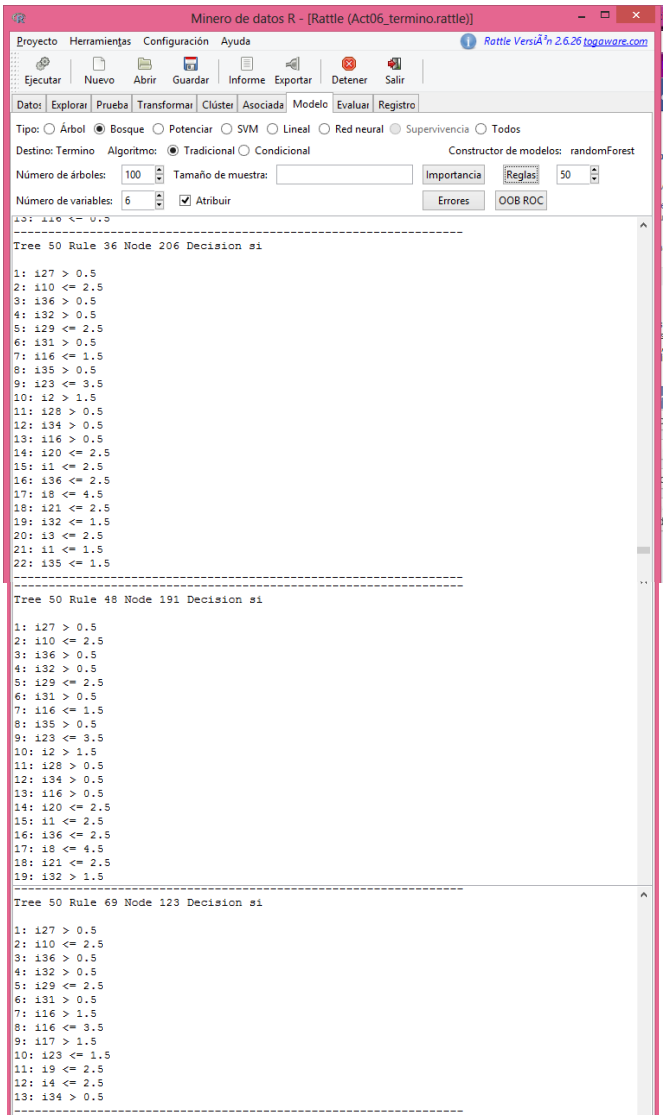


Figura 4.15 Reglas más sobresalientes para variable objetivo Terminó, Actuaría Plan 2006

Variable Objetivo "Desertó", Bosque6:

Se crean 100 árboles con un número de variables mínimas para cada división igual a 7, se analizan las reglas de cada uno, el modelo número 4 arroja el mayor número de reglas (96 con 41 de respuesta positiva).

La materia con mayor peso para la toma de decisiones es Álgebra Superior II, desde la que se hace una partición entre los que la inscribieron y los que no la inscribieron.

Los alumnos que no llegaron a inscribir Álgebra Superior II desertan cuando recursan la materia de Geometría Analítica II más de una vez, Problemas Socio-Económicos de México más de dos veces y Programación I más de una y menos de 4 veces.

De los alumnos que si inscribieron Álgebra Superior II se obtienen reglas que arrojan como filtros las materias de Probabilidad I, Economía I, Estadística II, Cálculo Diferencial e Integral IV, Geometría Analítica II, Ecuaciones Diferenciales I y Álgebra Lineal I.

Las materias cuyo número de inscripción son superiores a 2 y causan deserción son las siguientes: Problemas Socio-Económicos de México, Matemáticas Actuariales del seguro de Personas I, Álgebra Lineal I y Programación I, donde se puede ver que no son de las materias con carga matemática muy grane pero aun así están afectando los alumnos al no poder acreditarlas se rinden y terminan desertando.

Una de las reglas arroja que alumnos desertores que inscribieron la materia de Ecuaciones Diferenciales I y cursaron más de 5 veces Álgebra Lineal I, también son alumnos con problemas para Acreditar Estadística II y con más de dos inscripciones en Teoría del Seguro y Estadística III.

Una combinación interesante es que los alumnos desertan cuando reprueban más de 2 veces Calculo Diferencial e Integral IV y Matemáticas Financieras y más de 4 veces Álgebra Superior II.

En la Figura 4.16 igual que en los modelos anteriores se presentan algunas de reglas sobresalientes.

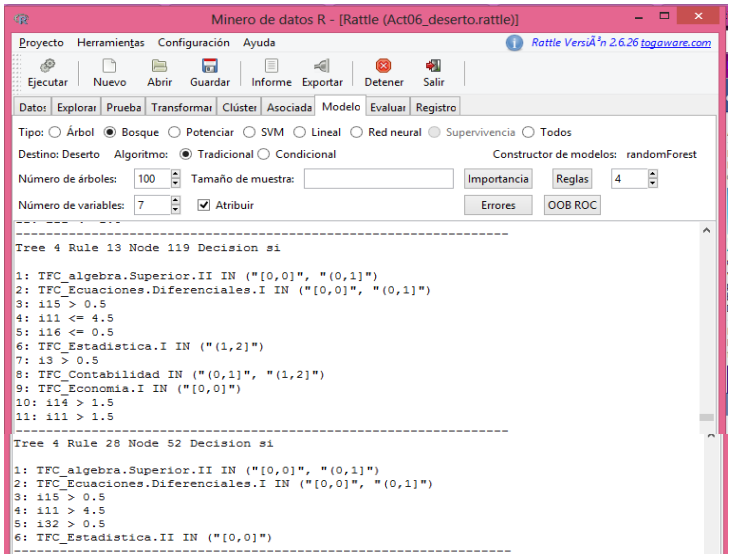


Figura 4.16a Reglas más sobresalientes para variable objetivo Desertó, Actuaría Plan 2006

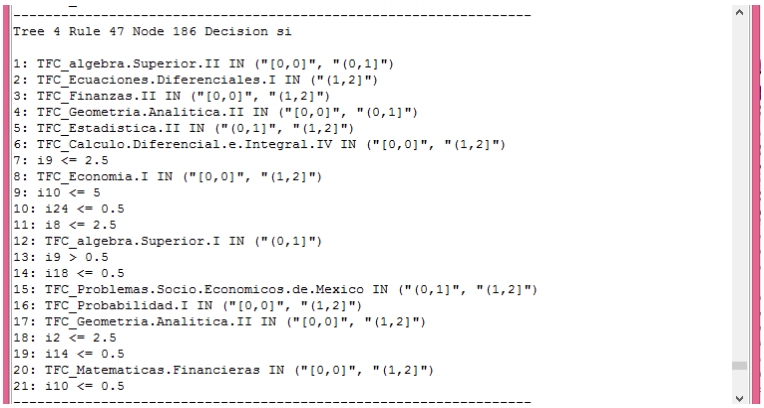


Figura 4.16b Reglas más sobresalientes para variable objetivo Desertó, Actuaría Plan 2006

4.4 Evaluación de los modelos.

La última parte de la modelación es precisamente la evaluación de los árboles de decisión creados en este trabajo, para ello se utilizaron algunas de las técnicas descritas en el Capítulo 2, se eligieron gracias a su facilidad de entendimiento y alta capacidad de evaluación. Estas técnicas son la matriz de error para los modelos, error genera e índice ROC.

Recordar que para la evaluación de los modelos, el índice ROC los clasifica como malos cuando el índice está en el intervalo [0.5,0.6), un test regular cuando está en el intervalo [0.6,0.75), un test bueno cuando está en el intervalo [0.75, 0.9), un test muy bueno cuando está en el intervalo [0.9,0.97) y un test excelente cuando está en el intervalo [0.97,1).

Además de estas técnicas y previo a plasmar el modelo final se crearon muchos más modelos que se estuvieron comparando entre sí y comparando los errores arrojados para cada uno se eligió el de menores índices de error y mayor asertividad. A continuación se presentan los resultados de las evaluaciones de los modelos plasmados las páginas anteriores.

Viendo las figuras que se presentan en la evaluación de cada modelo es muy fácil sacar la conclusión de la evaluación, aun así se explicarán brevemente. En las figuras de evaluación por matriz de error se expresa la información resultante al evaluar los datos apartados para validación, la primera matriz expresa las cantidades de datos en cada categoría y la segunda matriz de la figura expresa las cantidades en forma de porcentajes.

Árbol 1: Actuaría variable objetivo Desertó

Al aplicar la matriz de error se observa que el modelo es bastante bueno ya que al ejecutar la parte de validación arroja 53 predicciones incorrectas de 1016, Dando así un error de 5.2% que es bastante aceptable. (Figura 4.17)

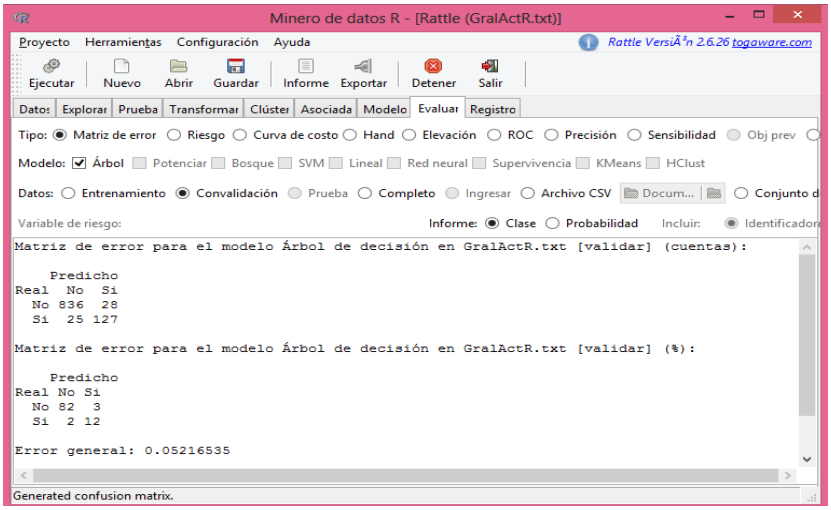


Figura 4.17 Matriz de error para el modelo del Árbol 1

El índice ROC resulta de 0.9694 lo cual quiere decir que es un muy buen modelo, así mismo se puede ver en la gráfica cómo la curva es bastante elevada.

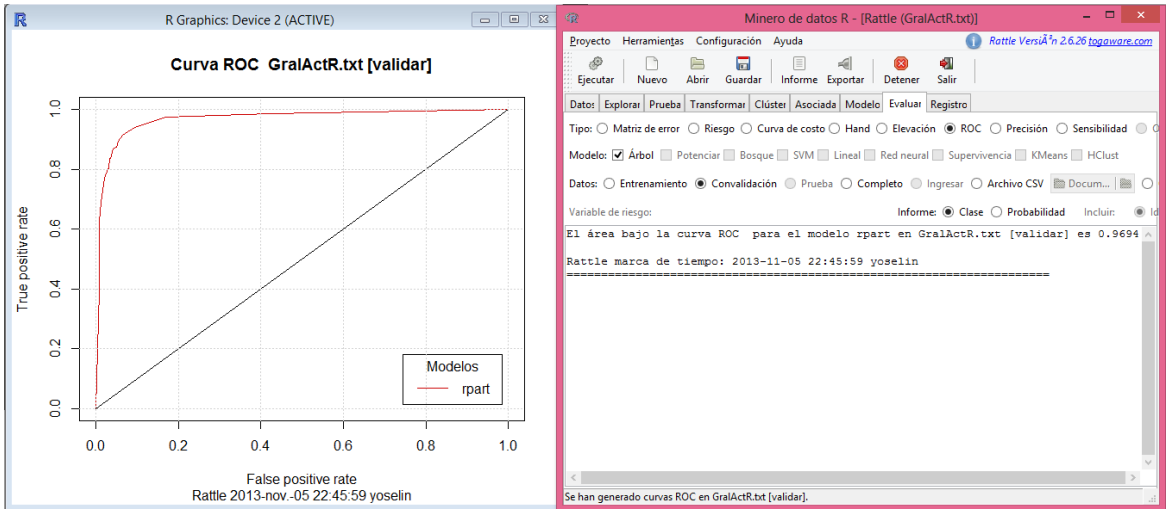


Figura 4.18 Índice y curva ROC para el modelo del Árbol 1

Arbol 2: Actuaría, variable objetivo Terminó

En la matriz de error resulta un modelo bueno, donde se equivoca 50 de los 1016 datos utilizados en validación. Dando así un error de 4.9% que es excelente. (Figura 4.19)

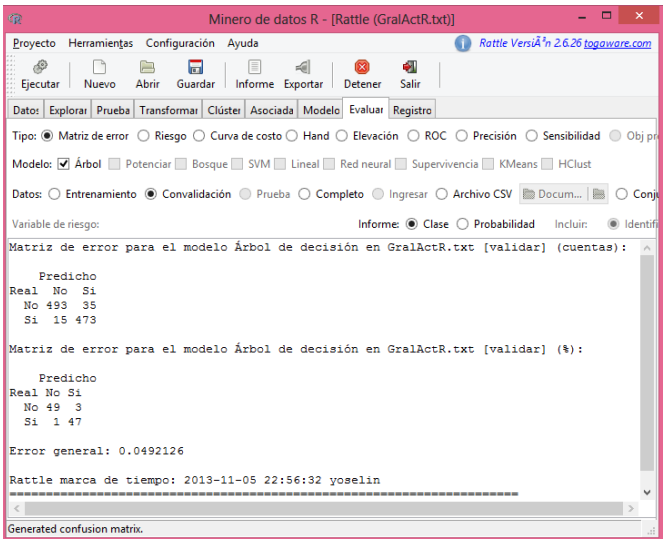


Figura 4.19 Matriz de error para el modelo del Árbol 2

El índice ROC es de 0.9727 lo que indica que es un modelo excelente, esto lo podemos corroborar en la Figura 4.4.4 en donde se observa la curva y el indicador.

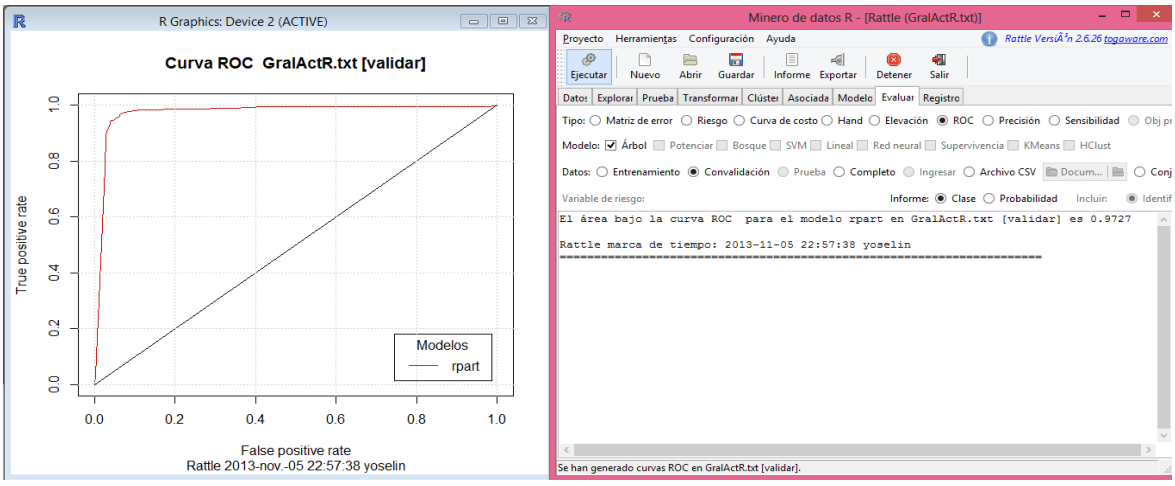


Figura 4.20 Índice y curva ROC para el modelo del Árbol 2

Árbol 3: Ciencias de la computación, variable objetivo Desertó

Al elaborar la matriz de error, el modelo acierta 119 de 132 ocasiones en los datos de validación. Al ser pocos los datos el error resulta más grande (9.8%) pero aun así el modelo es bastante aceptable. (Figura 4.21)

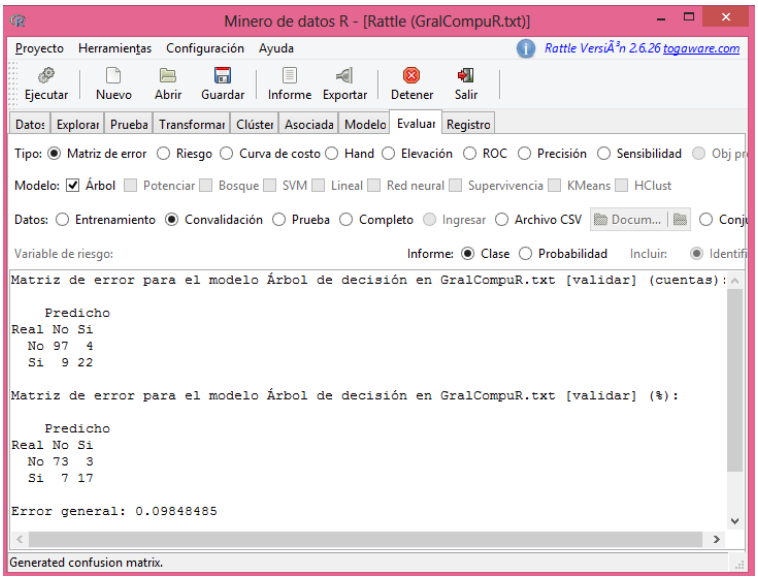


Figura 4.21 Matriz de error para el modelo del Árbol 3

El índice ROC es de 0.83 que lo califica como un modelo excelente.

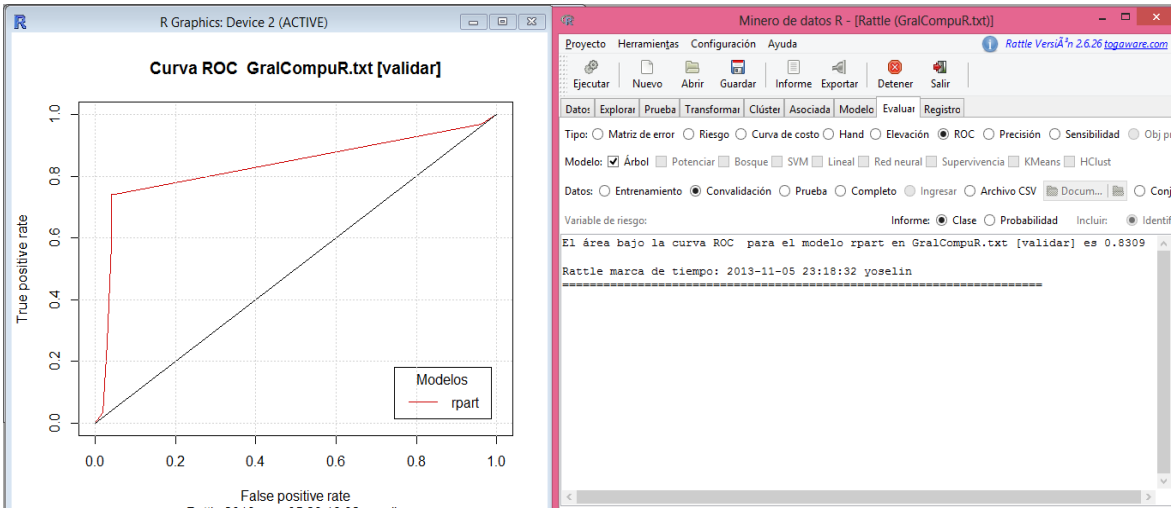


Figura 4.22 Índice y curva ROC para el modelo del Árbol 3

Árbol 4: Ciencias de la computación, variable objetivo Terminó.

La matriz de error se muestra que en la base para test únicamente se equivoca en la predicción de 8 de 176 datos de validación. Dando un error de 4.54% por lo que de acuerdo con ésta evaluación el modelo es excelente.

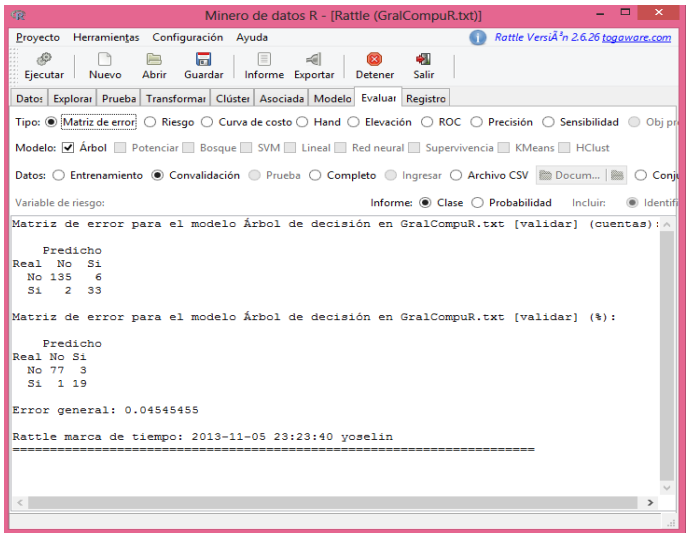


Figura 4.23 Matriz de error para el modelo del Árbol 4

El índice ROC resulta de 0.9472 y como consecuencia una curva muy cercana a uno, entonces esta evaluación también indica que es un modelo excelente.

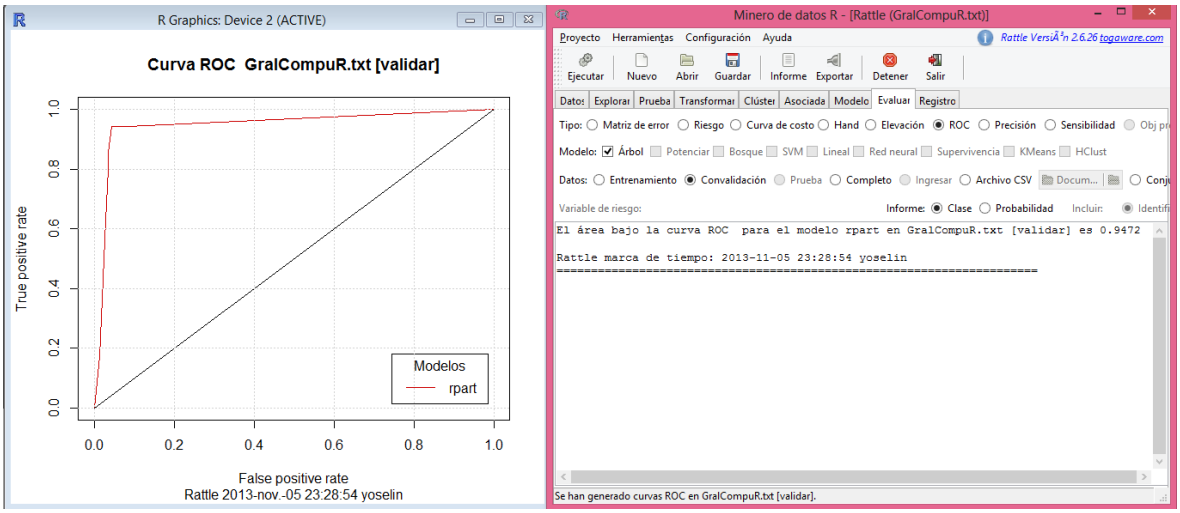


Figura 4.24 Índice y curva de ROC para el modelo del Árbol 4

Bosque 1, Ciencias de la Computación, variable objetivo Terminó:

Al aplicar la evaluación de matriz de error resulta en los datos de validación se equivoca al predecir 15 de 263 datos, y al evaluar con Índice ROC éste resulta de 0.96 por lo que se concluye que este es un muy buen modelo.

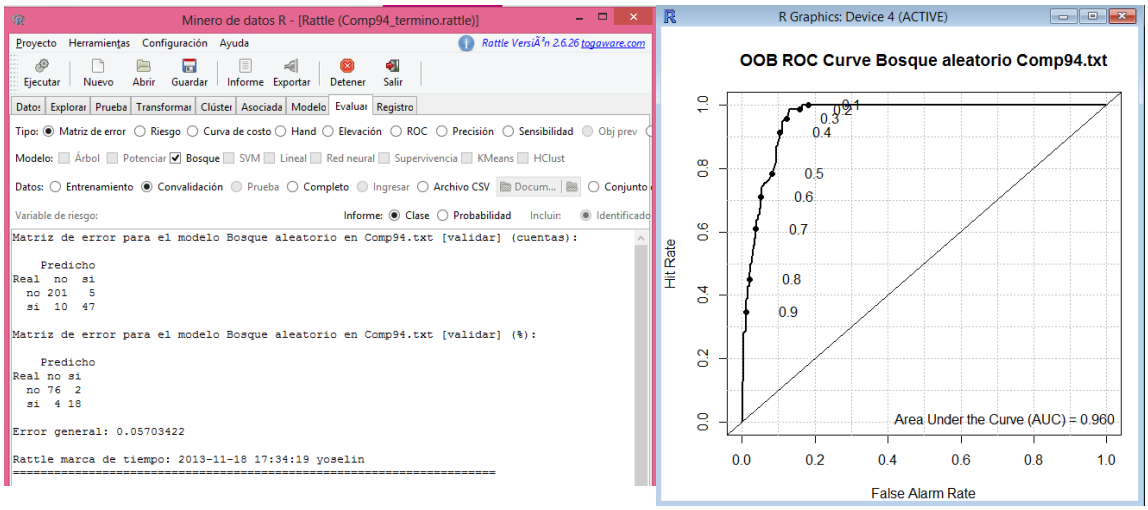


Figura 4.25 Matriz de error, índice y curva ROC para el modelo de Bosque 1

Bosque 2, Ciencias de la Computación, variable objetivo Desertó:

Al aplicar la evaluación por matriz de error resulta que para los datos de validación acierta en 216 de 263 predicciones, dando un error de 17.8% pudiendo ser a causa de que son pocos datos.

El índice ROC resulta de 0.905 por lo que consideraremos que es un modelo muy bueno y combinándolo con los resultados de la matriz se concluye que es un modelo bueno.

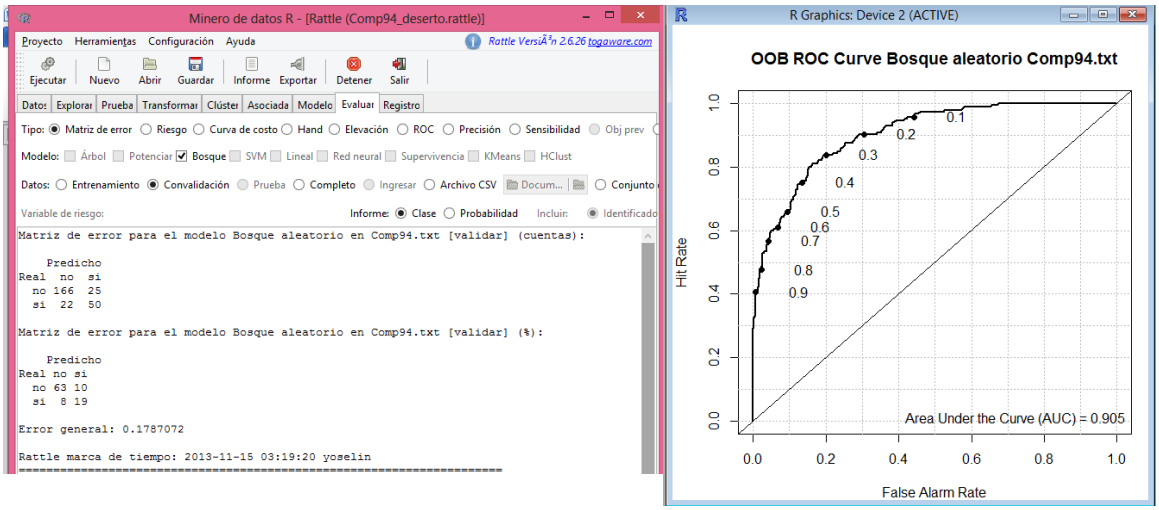


Figura 4.25 Matriz de error, índice y curva ROC para el modelo de Bosque 2

Bosque 3, Actuaría 2000, variable objetivo Terminó:

La matriz de error muestra que para los datos de validación acierta en 332 de 351 predicciones, dando un error de 5.4% que es bastante aceptable y por lo tanto el modelo es muy bueno.

El índice ROC resulta de 0.98 por lo que consideraremos que es un modelo excelente y combinándolo con los resultados de la matriz se concluye que es un modelo muy bueno y acertado.

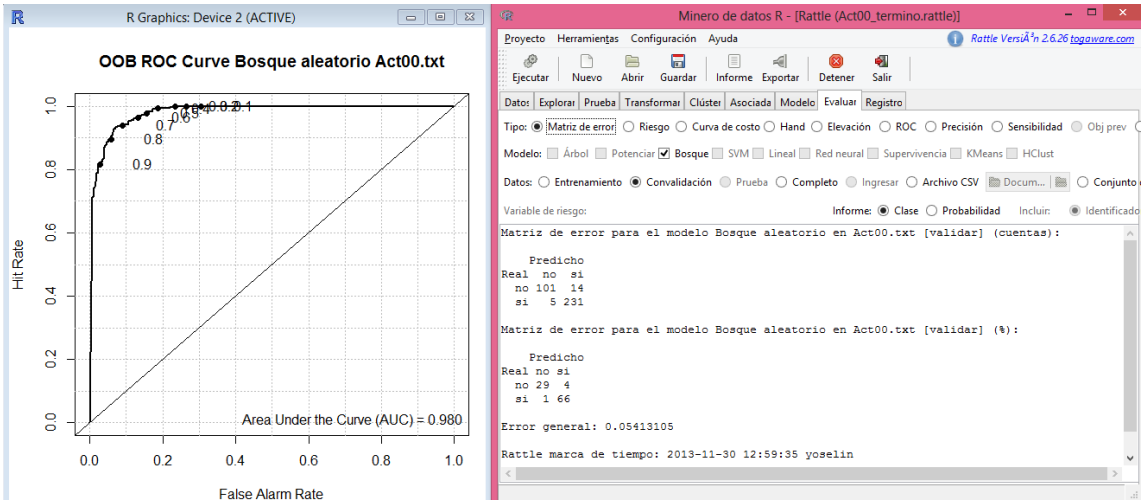


Figura 4.26 Matriz de error, índice y curva ROC para el modelo de Bosque 3

Bosque 4, Actuaría 2000, variable objetivo Desertó:

La matriz de error muestra que para los datos de validación acierta en el 96% de las predicciones y por lo tanto el modelo es muy bueno.

El índice ROC resulta de 0.986 por lo que consideraremos que es un modelo excelente y combinándolo con los resultados de la matriz se concluye que es un modelo bueno.

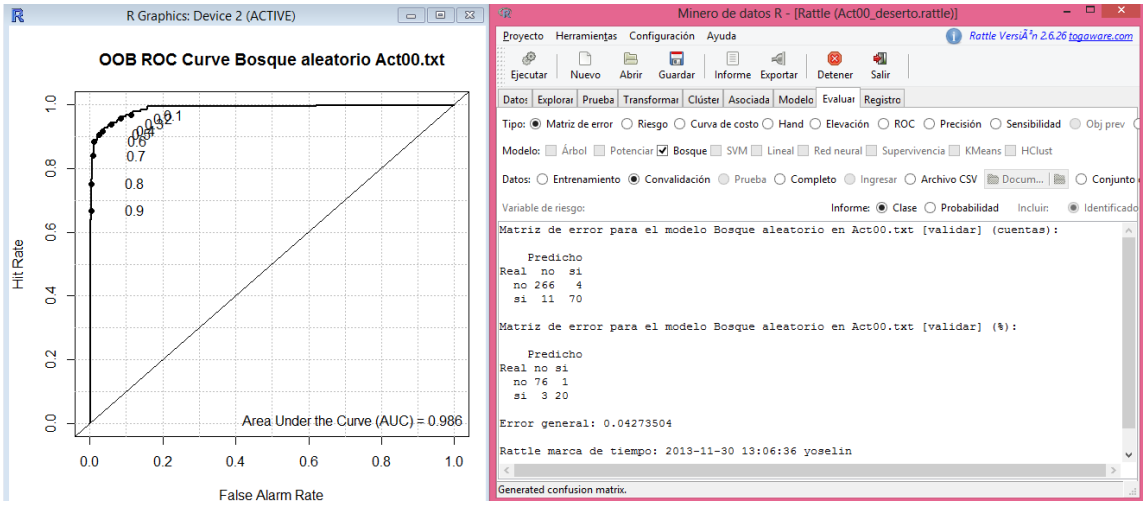


Figura 4.27 Matriz de error, índice y curva ROC para el modelo de Bosque 4

Bosque 5, Actuaría 2006, variable objetivo Terminó:

El modelo tiene un porcentaje de aciertos del 97% en los datos ocupados para validación, lo que arroja un error del 2.2% que refleja un modelo excelente.

El índice ROC da de 0.999 que permite evaluarlo también como un modelo excelente. Conclusión que se toma para este modelo.

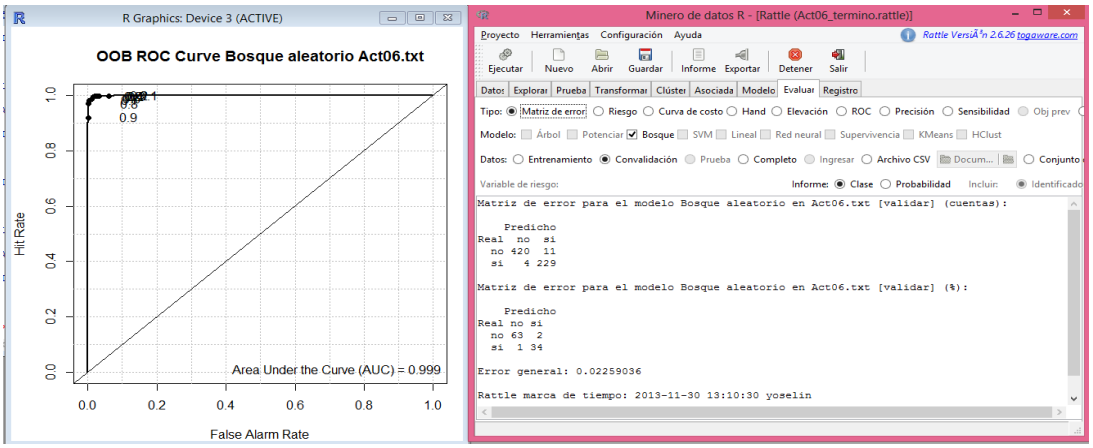


Figura 4.28 Matriz de error, índice y curva ROC para el modelo de Bosque 5

Bosque 6,Actuaría 2006, variable objetivo Desertó:

El modelo tiene un porcentaje de aciertos del 94% en los datos ocupados para validación, lo que arroja un error del 6.02% que es aceptable y evalúa al modelo como bueno.

El índice ROC da de 0.953 que permite evaluarlo también como un modelo muy bueno y en conjunto con la evaluación de matríz de error, se concluye que es bueno.

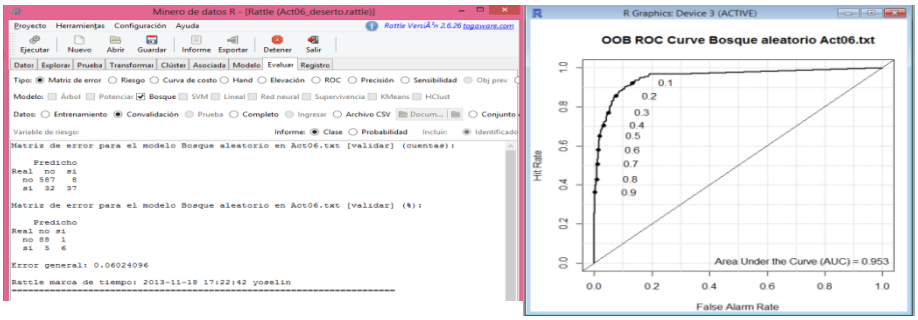


Figura 4.29 Matriz de error, índice y curva ROC para el modelo de Bosque 6

CONCLUSIONES

Durante todo el proceso de elaboración de la tesis no sólo se generó conocimiento nuevo sobre qué es la Minería de Datos, las principales técnicas y su base teórica y matemática que la sustenta sino que se logró cubrir el objetivo principal al generar conocimiento del comportamiento de los alumnos de la Lic. en Actuaría y la Lic. en Ciencias de la Computación a lo largo de su trayectoria escolar en la Facultad de Ciencias.

Como parte del conocimiento adquirido sobre minería de datos se rescatan los siguientes puntos:

- La minería de datos es una técnica de análisis de la información que sirve para generar patrones y pronósticos a partir de una base de datos dada y con el objetivo de diagnosticar actividades y procesos de una organización, así como ayudar en la toma de decisiones para el crecimiento e incremento de utilidades de la misma. [pp.11]

- La aplicación de inteligencia en los negocios es un proceso integral ya que requiere de personas especialistas en varias áreas de conocimiento y que éstas puedan interactuar entre sí, como por ejemplo actuarios, administradores de bases de datos, economistas, mercadólogos, estadisticos, matemáticos, inversionistas, especialistas en tecnologías de la información, personal de apoyo para la aplicación de estrategias, gerentes, etc.

- La parte más básica para poder iniciar un proceso de minería radica en conocer y entender el negocio y cuáles son los objetivos a seguir, además de contar con una base de datos que genere información suficiente, de lo contrario es necesario que la organización adquiera un plan de acción para crear su propia base de datos.

- La diferencia entre la metodología SEMMA y CRISP-DM radica en que la primera se centra más en las características técnicas del desarrollo del proceso, mientras la segunda mantiene una perspectiva basada en los objetivos organizacionales del proyecto. [pp23]

- La estadística y la minería de datos en conjunto forman una base sólida, eficiente que genera grandes resultados que probablemente por sí solas no podrían llegar a encontrar. Sin embargo la minería de datos depende en mayor proporción de la estadística ya que para el análisis de la

información y para realizar cierto tipo de proyecciones utiliza técnicas estadísticas que por sí solas generan resultados importantes.[pp23]

- Cada una de las técnicas de minería tiene un objetivo específico que cumplir y para saber cuál aplicar en cada estudio es necesario primero determinar qué tarea se va a desempeñar, diferenciando también si se requiere detectar patrones o para predicciones, posteriormente saber si es un estudio supervisado (con variable objetivo) o no supervisado, la dimensión de la base de datos, la complejidad requerida para el estudio y el tipo de presentación de los resultados.

De la parte aplicada, en base a la experiencia al realizar el estudio se determina lo siguiente:

- La tarea para obtener los resultados mostrados en el capítulo cuatro fue ardua ya que a pesar de que la parte de modelación parece sencilla, se requirió un trabajo previo de recolección y limpieza de los datos para generar la vista minable (base de datos introducida al modelo) que en varias ocasiones requirió más de una modificación.[pp.67]

- La preparación de los datos juega una parte fundamental para el éxito de los modelos ya que una vista minable aunque esté mal hecha probablemente genere resultados pero éstos serán completamente equivocados, sin sentido o sin grandes resultados.[pp. 64]

- Otro factor determinante en la eficiencia del proceso es la herramienta de software utilizado, ya que a pesar de que existen varios programas especializados cada uno tiene sus limitantes tanto en la holgura de los atributos para cada técnica como en la forma como se presentan los resultados (reportes y gráficas). En este caso las limitantes más importante de R son en primer lugar los gráficos ya que en particular el dibujo de un árbol grande no se alcanza a distinguir y no existe ninguna herramienta para aplicar un zoom y en segundo lugar el software no es tan interactivo como otros cuya licencia de uso tiene altos costos.[pp.117]

- Para poder realizar un análisis más amplio fue necesaria la creación de muchas nuevas variables que no estaban contempladas en la base de datos original, por lo que probablemente para estudios futuros se requiera realizarlas nuevamente o implementar nuevas, lo que se puede proponer es la solicitud de una mayor cantidad de información para el análisis en caso de existir y en caso de no existir variables como las

aquí creadas de “Años”, “SemInscritos”, “Recursamientos”, etc., crearlas a fin de ayudar a estudios posteriores. [pp.68]

En base a los resultados obtenidos al aplicar los modelos:

- En general los alumnos de la Facultad de Ciencias egresan con promedios superiores a 8.[pp.88]

- Los alumnos que desertan en ambas carreras lo hacen normalmente antes de tener 30% de los créditos pero también existe una parte de la población de Actuaría que deserta después de tener el 40% de avance académico debido a materias filtro de los últimos semestres.
  [pp. 114, pp.120 y pp.131]

- De la parte estadística podemos ver que los alumnos de la Facultad de ciencias egresan en un rango de ocho a diez semestres, los alumnos que superan esta cantidad de semestres tienen una alta probabilidad de deserción.[pp.88]

- La cantidad de recursamientos que un alumno tiene al concluir su carrera indica que son personas con gran tenacidad ya que a pesar de reprobar muchas veces, vuelven a intentarlo hasta lograrlo y personas que no tienen ésta característica son más propensos a desertar y hacerlo en los primeros semestres de la carrera.[pp.119]

- En la Licenciatura en Actuaría existe una diferencia significativa entre los dos planes de estudio, una de ellas es la reciente tendencia del plan 2006 de adelantar materias para concluir en un menor tiempo la carrera, fenómeno que no sucedía en el plan 2000. [pp.118]

- La seriación de las materias no juega un papel determinante para aprobar o reprobar una materia y que esto cause una deserción ya que en ambas carreras las materias que llevan a un alumno a desertar no tienen grandes temas en común, pero si tienen un gran peso de teoría matemática.[pp. 133 y pp.126]

- En ambas carreras se pudieron detectar filtros importantes de algunas materias por lo que una posible solución sería revisar los temarios de esas materias para posibles modificaciones o en los planes de estudio incluir otras materias previas para que el alumno llegue con un conocimiento más especializado.[pp126,pp.129 y pp.133]

Ahora bien la idea de este trabajo radica en que los resultados obtenidos en el modelaje puedan ayudar a la Facultad de Ciencias para elaborar planes de acción para disminuir los índices de deserción.

Como parte de la fase de implementación del modelo, se recomiendan algunos cursos de acción que podrían servir a la Facultad de Ciencias para ayudar a los alumnos a continuar con sus estudios y reducir así el índice de deserción en las carreras de Actuaría y Ciencias de la Computación. Cabe señalar que queda a criterio de la Facultad de Ciencias tomar como punto de partida estas recomendaciones o crear nuevas ideas basadas en estos y otros resultados para implementar diversas soluciones. Se propone:

- Crear un Taller para cada una de las carreras, parecido al Taller de Matemáticas de modo que se den asesorías y una especie de cursos de regularización a los alumnos en las materias que resultaron ser filtros y que tienen mayor índice de reprobación. Teniendo personal especializado sólo en estas materias.

  Poniendo a cargo a algún personal de tiempo completo y que coordine a alumnos que quieran realizar su servicio social y que sean expertos en las materias filtro y de mayor reprobación, aplicando una entrevista para que solo aquellos capacitados en el tema puedan dar asesorías y cursos con temarios apegados a los de la Facultad de Ciencias.

- Elaborar algún programa que ayude a llevar un control y seguimiento de los alumnos, filtrando a aquellos que cumplan con las características de deserción arrojadas por este estudio y enviar un citatorio con dichos alumnos en el área de Psicología de la Facultad para conocer los motivos que lo están orillando a tener esos resultados. Posteriormente canalizarlos ya sea a atención psicológica o a los talleres de manera obligatoria y monitoreando el cumplimiento de esas acciones.

  Por ejemplo, resultó que : “Un alumno del plan 200 deserta en menos de 5 años en la facultad, con más de 5 materias reprobadas, menos de 14 semestres inscritos y un avance menor al 1%, recursando menos de 2 materias”[pp.114]. Entonces al identificar a un alumno del plan 2006 con más de 5 materias reprobadas con avance menor al 1% es canalizado al área correspondiente.

- Sería conveniente tener una base de datos especial únicamente con variables útiles para el análisis de deserción de los alumnos, independiente a la base general, para que cada vez que se requiera actualizar los resultados, evaluar los modelos ya hechos anteriormente o realizar nuevos, el encargado de realizar el análisis no tenga que volver a crear la base de datos desde un inicio.

Como se puede imaginar se pueden realizar muchos más cursos de acción y demás estudios de minería para detectar más patrones en bases de datos parecidas de otras carreras o en otras áreas de oportunidad, pero resultaría ambicioso realizarlo en un solo trabajo, por lo que se deja abierta la puerta para que otros integrantes del equipo de análisis de la Facultad de Ciencias realicen estudios en esas otras áreas o profundicen el análisis que aquí se presentó.

ANEXO 1 “Microsoft Excel 2010”

“Excel es un programa de hojas de cálculo de Microsoft Office system. Permite crear y aplicar formato a libros (un conjunto de hojas de cálculo) para analizar datos y tomar decisiones fundadas sobre aspectos de su negocio. Concretamente, se puede usar para hacer un seguimiento de datos, crear modelos para analizar datos, escribir fórmulas para realizar cálculos con dichos datos, dinamizar los datos de diversas maneras y presentarlos en una variedad de gráficos con aspecto profesional.” [31]

Para fines de este trabajo se realizó un trabajo conjunto entre las funciones básicas de Excel y su programador Microsoft Visual Basic.

Las funciones son ecuaciones para realizar ciertos cálculos, éstos pueden ser operaciones numéricas o tareas realizadas a un texto; dichas ecuaciones devuelven la información solicitada y pueden estar anidadas unas con otras. Una fórmula siempre comienza con el signo igual (=).

Las funciones utilizadas para este trabajo y sus características se describen en la siguiente tabla:

FUNCIÓN	DESCRIPCIÓN
CONSULTAV(valor_buscado, matriz_buscar_en, indicador_columnas,ordenado)	Busca un valor en la columna de la izquierda de una tabla y, a continuación, devuelve un valor en la misma fila desde una columna que el usuario especifica en la tabla.
SI (prueba_lógica, valor_si_verdadero, valor_si_falso)	La función **SI** devuelve un valor si una condición especificada se evalúa como VERDADERO y otro valor si se evalúa como FALSO.
ESERROR(valor)	Devuelve el valor lógico VERDADERO si **valor** hace referencia a cualquier valor de error, como #N/A, #¡VALOR!, #¡REF!, #¡DIV/0!, #NUM!, #¿NOMBRE? o #¡NULO!; de lo contrario, devuelve FALSO.
FILTROS	Sirven para buscar un subconjunto de datos de un rango y trabajar con el mismo. Un rango filtrado muestra sólo las filas que cumplen el criterio o criterios que e especifiquen para una coumna.

Tabla A1 Fórmulas de Microsoft Excel 2010 utilizadas.

Para lograr cambiar el formato de las bases otorgadas y crear algunas variables se utilizó Microsoft Visual Basic de Excel y el código utilizado se muestra en la sección 3.3 a lo largo de la explicación de la preparación de los datos.

ANEXO 2 “R Project for Statistical Computing”

“R es un lenguaje y un entorno para computación y gráficos estadísticos. Es un proyecto GNU, que es similar al lenguaje S y al que se ha desarrollado en los Laboratorios Bell (antes AT & T, ahora Lucent Technologies) por John Chambers y colegas.” [32]

“R está disponible como software libre bajo los términos de la Licencia Pública General de GNU de la Free Software Foundation en forma de código fuente. Compila y ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS”[32]

El programa provee una gran variedad de funciones, gráficas y técnicas para la manipulación, modificación a través de diversos cálculos, y almacenamiento de diversas bases de datos.

R puede extenderse en sus tareas a través de paquetes y librerías contenidas en ellos. Ocho de estos paquetes están disponibles dentro del mismo programa y muchos otros a través de la conexión con la página de “r-cran Project”[32].

En este trabajo se utilizaron los paquetes básicos contenidos en el programa y se tuvieron que instalar los paquetes especializados para realizar el análisis exploratorio y modelos de minería de datos, éstos paquetes son: arules, e1071, party, cluster, rpart, png, mgcv, nnet, class,ellipse, vcd, lattice, scatterplot3d, FactoMineR y rattle.

El código utilizado para realizar el análisis exploratorio de la información se presenta a continuación:

```
####  LIBRERIAS  ####
library(class)          library(e1071)          library(ellipse)
library(vcd)            library(MASS)           library(lattice)
library(cluster)        library(scatterplot3d)  library(FactoMineR)
library(rattle)         rattle()
########################
Gral=read.delim("C:\\ Documents\\BasesTesis\\General.txt")
GralAct=read.delim("C:\\Documents\\BasesTesis\\Gral_Act.txt")
GralCompu=read.delim("C:\\Documents\\BasesTesis\\Gral_Compu.txt")
attach(Gral)
names(Gral)
dim(Gral) #total de filas, columnas
head(Gral)
str(Gral) # Examina el tipo de datos que tenemos
#_____Se transforman algunas variables para que las
interprete como categoricas
Gral$Gen=as.factor(Gral$Gen)
Gral$Plan=as.factor(Gral$Plan)
Gral$UltimoSemInscrito=as.factor(Gral$UltimoSemInscrito)
Gral$X1erSem=as.factor(Gral$X1erSem)
str(Gral
#_____Resumen Estadístico
summary(Gral)
par(mfrow=c(2,2))
color=c("red","yellow","green","salmon","purple","gray50","lightblue","violetred1","lightgreen","blue","gray32"
```

```
barplot(carreraG,main="Carrera",col=color)
barplot(sitG,main="Situación Escolar",col=color)
barplot(planG,main="Plan de Estudios",xlab="Tipo de
plan",ylab="Frecuencia",col=color,args.legend=list(x="topright")), legend("topright",legend=c("2000
Actuaría", "1994 Ciencias de la Computación","2006 Actuaría", "2013 Ciencias de la
Computación"),col=color,pch = c(16,16,16,16))
barplot(genG,main="Generación",col=color)

par(mfrow=c(1,1))
barplot(ingresoG,main="Tipo de  ingreso", col=c("Yellow","lightBlue", "Pink", "darkblue", "Grey",
"orange", "Green", "DarkGrey", "salmon","Red"), legend.text=c("Años posteriores al primero AC","Años
posteriores al primero RE","Cambio de Carrera y/o Plantel(concurso)","Cambio de Plantel",
        "Cambio interno de carrera","Carrera Simultánea","Consurso de Selección","Exámen de
selección","Pase Reglamentario","Segunda Carrera"), args.legend = list(x = "topleft"))

barplot(ingresoCarr,main="Tipo de ingreso por  carrera", col=c("Yellow", "lightBlue", "pink", "darkblue",
"Grey", "orange","Green","DarkGrey","salmon","Red") ,legend.text=c("Años posteriores al primero
AC","Años posteriores al primero RE","Cambio de Carrera y/o Plantel(concurso)","Cambio de Plantel",
        "Cambio interno de carrera","Carrera Simultánea","Consurso de Selección","Exámen de
selección","Pase Reglamentario","Segunda Carrera"),  args.legend = list(x = "topright"))

        #### DATOS CATEGORICOS ####

#_____________________________________Tablas de contingencia y pruebas de dependencia

        carreraG=xtabs(~ Carrera, data=Gral)
        sitG =xtabs(~ SitEscolar, data=Gral)
        planG=xtabs(~ Plan, data=Gral)
        genG=xtabs(~ Gen, data=Gral)
        ingresoG=xtabs(~Ingreso,data=Gral)
        planingreso=xtabs(~Plan+Ingreso,data=Gral)
        plangen=xtabs(~Plan+Gen,data=Gral)
        ingresoCarr=xtabs(~Ingreso+Carrera,data=Gral)
        planCarr=xtabs(~Plan+Carrera,data=Gral)
        genCarr=xtabs(~Gen+Carrera, data=Gral)
        sitCarr=xtabs(~ SitEscolar + Carrera , data=Gral)  #grafica
        sitplan= xtabs(~ SitEscolar + Plan,data=Gral)
        ingresogen=xtabs(~Ingreso+Gen,data=Gral)
        sitgen=xtabs(~ SitEscolar + Gen, data=Gral)
        sitingreso=xtabs(~ SitEscolar + Ingreso, data=Gral)
        ingresogencarr=xtabs(~Ingreso+Gen+Carrera,data=Gral)
        sitgencarr=xtabs(~ SitEscolar + Gen + Carrera, data=Gral)
        sitingresocarr=xtabs(~ SitEscolar + Ingreso+Carrera, data=Gral)
ingresoG
ingresoCarr
planingreso
plangen
planCarr
genCarr
sitCarr
sitplan
        par(mfrow=c(2,1))
        barplot(sitplan, main="Situación escolar de acuerdo al plan de estudios",
legend.text=c("Desertó", "Estudia","Terminó"),col=color)
        barplot(sitCarr, main="Situación escolar por
carrera",legend.text=c("Desertó","Estudia","Terminó"),args.legend = list(x = "topright"),col=color)
        ingresogencarr
        sitgencarr
        sitingresocarr
```

```
summary(assocstats(ingresoCarr))
summary(assocstats(planingreso))
summary(assocstats(plangen))
summary(assocstats(planCarr))
summary(assocstats(genCarr))
summary(assocstats(sitCarr))
summary(assocstats(sitplan))
summary(assocstats(ingresogen))
summary(assocstats(sitgen))
summary(assocstats(sitingreso))

#### DATOS NUMÉRICOS ####

GralNum=cbind(Avance,PromedioGral,TotalSem,Años,TotalAprobadas,AprobadasYrecursadas,AprobadasNoRecursadas,TotalReprobadas,ReprobadasYrecursadas,
ReprobadasNoRecursadas,TotalRecursadas,Recursamientos,Matutino,Vespertino,InscritasBien,InscritasMal)
dim(GralNum)
aggregate(GralNum[,-17],list(Gral$SitEscolar),mean)
aggregate(GralNum [,-17], list(Gral$Carrera,Gral$SitEscolar), mean)
apply(GralNum,2,sd)
aggregate(GralNum[, -17], list(Gral$Carrera), kurtosis)
aggregate(GralNum[, -17], list(Gral$Carrera), skewness)
par(mfrow=c(2,4))
qqnorm(Avance, main="Avance")
qqline(Avance)
qqnorm(PromedioGral, main="PromedioGral")
qqline(PromedioGral)
qqnorm(Años, main="Años en la facultad")
qqline(Años)
qqnorm(TotalSem, main="Semestres Inscritos")
qqline(TotalSem)
qqnorm(TotalAprobadas, main="Materias Aprobadas")
qqline(TotalAprobadas)
qqnorm(TotalReprobadas,main="Materias Reprobadas")
qqline(TotalReprobadas)
qqnorm(TotalRecursadas,main="Materias Recursadas")
qqline(TotalRecursadas)
qqnorm(Recursamientos, main="Total Recursamientos")
qqline(Recursamientos)
qqnorm(AprobadasNoRecursadas,main="Aprobadas no Recursadas")
qqline(AprobadasNoRecursadas)
qqnorm(AprobadasYrecursadas,main="Aprobadas y recursadas")
qqline(AprobadasYrecursadas)
qqnorm(ReprobadasNoRecursadas,main="Reprobadas no recursadas")
qqline(ReprobadasNoRecursadas)
qqnorm(ReprobadasYrecursadas, main="Reprobadas y recursadas")
qqline(ReprobadasYrecursadas)
qqnorm(Matutino, main="Turno Matutino")
qqline(Matutino)
qqnorm(Vespertino, main="Turno Vespertino")
qqline(Vespertino)
qqnorm(InscritasBien,main="Materias en sem. correcto")
qqline(InscritasBien)
qqnorm(InscritasMal, main="Materias en sem. incorrecto")
qqline(InscritasMal)
```

```
                                                        #_________Graficos
par(mfrow=c(1,4))
                                                        #________Ajuste de densidad
truehist(Avance,main="Avance Escolar",col=color)
        lines(density(Avance),lwd = 2)
truehist(PromedioGral, main="Promedio",col=color)
        lines(density(PromedioGral),lwd = 2)
truehist(Años, main="Años en la facultad",col=color)
        lines(density(Años),lwd = 2)
truehist(TotalSem, main="Total de semestres inscritos",col=color)
        lines(density(TotalSem),lwd = 2)

truehist(TotalAprobadas,main="Total de materias Aprobadas",col=color)
        lines(density(TotalAprobadas),lwd = 2)
truehist(TotalReprobadas, main="Total de materias Reprobadas", col=color)
        lines(density(TotalReprobadas),lwd = 2)
truehist(TotalRecursadas, main="total de materias Recursadas",col=color)
        lines(density(TotalRecursadas),lwd = 2)
truehist(Recursamientos, main="Total de Recursamientos",col=color )
        lines(density(Recursamientos),lwd = 2)

truehist(AprobadasNoRecursadas,main="Aprobadas no Recursadas", col=color)
        lines(density(AprobadasNoRecursadas),lwd = 2)
truehist(AprobadasYrecursadas, main="Aprobadas y recursadas",col=color)
        lines(density(AprobadasYrecursadas),lwd = 2)
truehist(ReprobadasNoRecursadas,main="Reprobadas no Recursadas", col=color)
        lines(density(ReprobadasNoRecursadas),lwd = 2)
truehist(ReprobadasYrecursadas, main="Reprobadas y recursadas",col=color)
        lines(density(ReprobadasYrecursadas),lwd = 2)

truehist(Matutino,main="Turno Matutino",col=color)
        lines(density(Matutino),lwd = 2)
truehist(Vespertino,main="Turno Vespertino",col=color)
        lines(density(Vespertino),lwd = 2)
truehist(InscritasBien, main="Materias en semestre correcto",col=color)
        lines(density(InscritasBien),lwd = 2)
truehist(InscritasMal,main="Materias en semestre incorrecto",col=color)
        lines(density(InscritasMal),lwd = 2)

par(mfrow=c(1,4))

hist(Avance,main="Avance Escolar",col=color)
hist(PromedioGral, main="Promedio",col=color)
hist(Años, main="Años en la facultad",col=color)
hist(TotalSem, main="Total de semestres inscritos",col=color)

hist(TotalAprobadas,main="Total de materias Aprobadas",col=color)
hist(TotalReprobadas, main="Total de materias Reprobadas", col=color)
hist(TotalRecursadas, main="Total de materias Recursadas",col=color)
hist(Recursamientos, main="Total de Recursamientos",col=color )

hist(AprobadasNoRecursadas,main="Aprobadas no Recursadas", col=color)
hist(AprobadasYrecursadas, main="Aprobadas y recursadas",col=color)
hist(ReprobadasNoRecursadas,main="Reprobadas no Recursadas", col=color)
hist(ReprobadasYrecursadas, main="Reprobadas y recursadas",col=color)

hist(Matutino,main="Turno Matutino",col=color)
hist(Vespertino,main="Turno Vespertino",col=color)
hist(InscritasBien, main="Materias en semestre correcto",col=color)
hist(InscritasMal,main="Materias en semestre incorrecto",col=color)
```

```
cor(GralNum)                                    #_____________Análisis de Coorelación
                                                #_____________________________boxplot

par(mfrow=c(2,4))
boxplot(Avance,main="Avance Escolar")
boxplot(PromedioGral,main="Promedio")
boxplot(Años,main="Años en la facultad")
boxplot(TotalSem,main="TotalSem")
boxplot(TotalAprobadas,main="TotalAprobadas")
boxplot(TotalReprobadas,main="Total Reprobadas")
boxplot(TotalRecursadas,main="Materias Recursadas")
boxplot(Recursamientos,main="Recursamientos")
boxplot(AprobadasYrecursadas,main="AprobadasYrecursads")
boxplot(AprobadasNoRecursadas,main="Aprobadas No Recursadas")
boxplot(ReprobadasYrecursadas, main="ReprobadasYrecursadas")
boxplot(ReprobadasNoRecursadas,main="ReprobadasNoRecursadas")
boxplot(Matutino,main="Matutino")
boxplot(Vespertino,main="Vespertino")
boxplot(InscritasBien,main="InscritasBien")
boxplot(InscritasMal,main="InscritasMal")
                                                #_____________________________Dispersión
par(mfrow=c(1,1))
plot(Años,Avance, pch=c(1,2,22)[Gral$SitEscolar],col=color[Gral$SitEscolar],main="Años en la
facultad")
        legend("topright",legend=c("Desertó", "Estudia","Terminó"),col=color,pch = c(1,2,22))
plot(TotalSem,Avance, pch=c(1,2,22)[Gral$SitEscolar],col=color[Gral$SitEscolar],main="Total
semestres inscritos")
        legend("topright",legend=c("Desertó", "Estudia","Terminó"),col=color,pch = c(1,2,22))
plot(Matutino,Vespertino, pch=c(1,2,22)[Gral$SitEscolar],col=color[Gral$SitEscolar],main="Dispersión
por turnos")
        legend("topright",legend=c("Desertó", "Estudia","Terminó"),col=color,pch = c(1,2,22))
```

Para la parte de modelado se utilizó la librería rattle() por lo que no se creó ningún código. A pesar de ello se describirán las funciones que R utiliza para crear los árboles de decisión mediante el algoritmo CART y bosque aleatorio.

MÉTODO	FUNCIÓN	CARACTERÍSTICAS
CART	**rpart (**formula, data, weights, subset, na.action = na.rpart, method, model = FALSE, x = FALSE, y = TRUE, parms, CP,...**)**[32]	Utilizada para crear árboles de decisión, el tipo de árbol (clasificación o predicción) se establece en el argumento "model" el número de capas se establece en "nsplit" y la complejidad en "CP"
BOSQUE ALEATORIO	**randomForest(** x, y=NULL, ntree=500, replace=TRUE, classwt=NULL, cutoff, strata, sampsize = if (replace) nrow(x) else ceiling(.632*nrow(x)), nodesize = if (!is.null(y) && is.factor(y)) 5 else 1, maxnodes = NULL…**)**[32]	Implementa el algoritmo de bosque aleatorio para la clasificación y regresión. También se puede utilizar en el método no supervisado para evaluar proximidades entre los puntos de datos.[32]

FUENTES DE INFORMACIÓN

[1] Aldehuela Lucena, María. “Métodos estadísticos y de minería de datos”. Proyecto de Fin de Carrera, Universidad Pontificia Comillas. Madrid España. Junio 2005.
http://www.iit.upcomillas.es/pfc/resumenes/42b98a68079d0.pdf

[2] Análisis de Regresión Lineal: El procedimiento Regresión Lineal.
http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/18reglin_SPSS.pdf

[3] Casella, G. and Berger, R.L. Statistical Inference, 2nd Edition. Duxbury Press, 2002.

[4] CRISP-DM, Una metodología para proyectos de Minería de Datos.
http://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

[5] Cruz Arrrela, Liliana. “Minería de Datos con aplicaciones”. Tesis para obtener el título de Lic. En Matemáticas Aplicadas y computación. FES Acatlán, UNAM. Marzo 2010

[6] Curso de Minería de Datos.
http://es.wikiversity.org/wiki/Curso_de_Minería_de_datos

[7] Dr. Saed Sayad. An Introduction to Data Mining.
http://www.saedsayad.com/

[8] García Reyes, Roberto. “Minería de datos para la toma de decisiones e inteligencia de negocios: Aplicaciones en la mercadotecnia”. Tesis para obtener el título de Actuario. Facultad de Ciencias, UNAM. 2012

[9] García Cambronero, Cristina; Gómez Moreno Irene. “Algoritmos de aprendizaje: KNN y KMEANS”. http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf

[10] Gómez, Flor, “Sistemas Operativos”, Evolución histórica. noviembre 2012. http://florysel.blogspot.mx/2012/11/13-evolucion-historica.html

[11] Hernández Orallo, José, Et. Al. “Introducción a la minería de datos”, Pearson Prentice Hall, México 2008.

[12] Ibarra García, Ernesto Pathros, Et. Al. “Estudio de minería de datos en la facultad de ingeniería”. Tesis para obtener el título de Ing. en Computación. Junio 2009.

[13] IBM, Página web. http://pic.dhe.ibm.com/infocenter/idm/docv3/index.jsp?topic=%2Fcom.ibm.datatools.dimensional.ui.doc%2Ftopics%2Fc_dm_star_schemas.html

[14] Ing. Juan Miguel Moine, Et. Al. Estudio comparativo de metodologías para minería de datos. Grupo de investigación en Minería de Datos, UTN Rosario, Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires, Facultad de Informática. http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1

[15] Inteligencia Artificial, Redes Neuronales. http://inteligenciaartificialudb.blogspot.mx/2008/01/redes-neuronales.html

[16] Larose, Daniel T. "Discovering Knowledge in Data. An introduction to Data Mining". John Wiley & Sons, Inc., Hoboken, New Jersey, 2005

[17] Lezcano Ramón, David E. Trabajo de Investigación Bibliográfica: Minería de Datos. Universidad Nacional del Nordeste, Departamento de Informática. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosLezcano.pdf

[18] Lindgren, B.W. Statistical Theory. Champan & Hall. 4th Edition.1993.

[19] Marín Diazaraque, Juan Miguel, Profesor Titular de la Universidad Carlos III de Madrid. Material Docente, Minería de Datos. http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/DM2008.html

[20] Pérez López, César, Santín González Daniel. "Minería de datos: técnicas y herramientas". Thomson. Madrid, España. 2008.

[21] R and Data Mining. http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

[22] Refaat, Momdouh. "Data Preparation for Data Mining Using SAS". Morgan Kaufmann Publishers is an imprint of Elsevier. United States of America, 2007

[23] Rodríguez Montequín, Ma. Teresa, Et al. Metodologías para la realización de proyectos de data mining, Universidad de Oviedo. http://aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

[24] Rodríguez González, Ansel Yoan, Et al. Minería de Reglas de Asociación sobre Datos Mezclados. http://ccc.inaoep.mx/portalfiles/file/CCC-09-001.pdf

[25] Rojas, María Isabel. “Monografía de Adscripción: Data Warehouse”. Universidad Nacional del Nordeste. Facultad de Ciencias Exactas y Naturales y Agrimensura, Argentina 2009
http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonoAdsDiseno.pdf

[26] Sinnexus. Bussines Intelligence. Informática estratégica. http://www.sinnexus.com/business_intelligence/datawarehouse.aspx

[27] Vallejos, Sofía J. “Minería de Datos”. Trabajo de Adscripción, Universidad Nacional del Nordeste, Facultad de Ciencias Exactas, Naturales y Agrimensra. 2006.

[28] Virseda Benito, Fernando; Román Carrillo, Javier. “Minería de datos y aplicaciones”. http://www.it.uc3m.es/~jvillena/irc/practicas/06-07/22.pdf

[29] Xindong Wu, Vipin Kumar. “The Top Ten Algorithms in Data Mining”. Champman & Hall / CRC. Data Mining Knowledge Discovery Series.United States 2009.

[30] Yanchang, Zhao. “R and Data Mining: Examples and Case Studies”. April 2013.
http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

[31] Página oficial de Microsoft Office en México.http://office.microsoft.com, http://office.microsoft.com/es-mx/excel-help/introduccion-a-excel-2010-HA010370218.aspx

[32] R-Cran Project. http://www.r-project.org/
http://cran.r-project.org/web/packages/rpart/rpart.pdf
http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[33] Base de Datos otorgada por la Facultad de Ciencias a través de la Secretaría General de los Historiales Académicos de las carreras de Actuaría y Ciencias de la Computación.