

Presentación *KDD*

Ing. Vanesa Pinilla

Ing. Antonieta Kuz

Análisis del Dominio del Problema

Presentación del dominio del problema

Información estadística de alumnos de carreras de pregrado, grado y Posgrado de las universidades públicas y privadas argentinas.

Permite analizar la información de alumnos por áreas disciplinarias, títulos, rendimiento académico, etc.

La elección de este dominio se debió a la posibilidad de obtener un rico conocimiento a partir de los datos de ofertas académicas de alumnos de las distintas universidades del país.

Análisis del Dominio

ANALISIS DE LA INFORMACION CONTENIDA EN UN MOTOR DE BD

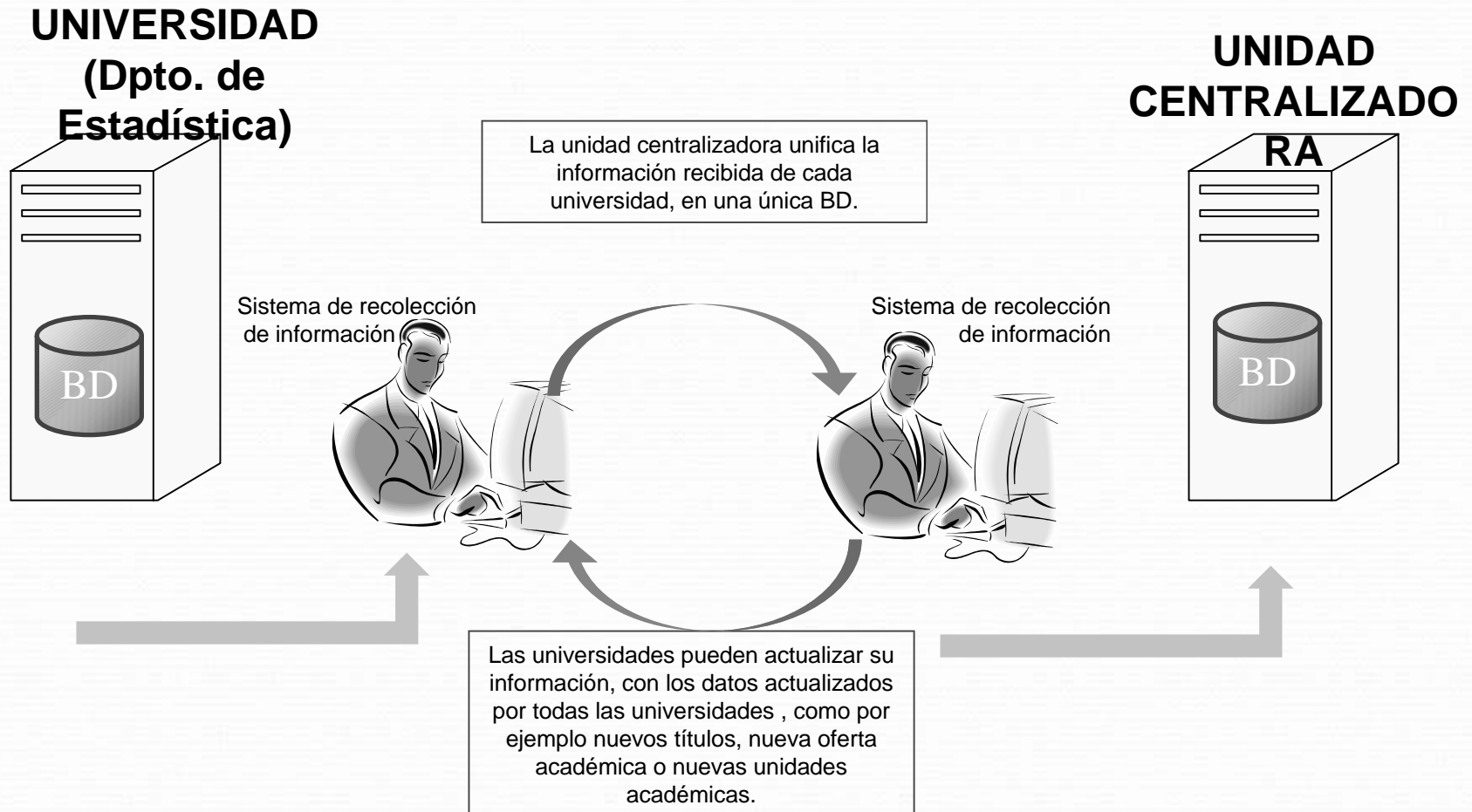
- Datos sumariados
- Modelo de BD relacional
- Datos recolectados por un sistema de información universitario
- Ingreso de datos....una vez al año
- Gran volumen de información: desde el año 1983 hasta la fecha

- Universidades informan estos datos para la asignación de su presupuesto
- La información que envían es referente a :
 - Ingreso, regularidad y egreso de los estudiantes
 - Diferentes cifras de la oferta educativa

Objetivo de recolectar esta información

El objetivo principal de recolectar toda esta información y unificarla en una unidad centralizadora, es servir de soporte para que las universidades nacionales o privadas y los institutos puedan informar sus datos estadísticos y de oferta educativa a dicha unidad, permitiendo a ambas contar con información consolidada y consistente en el momento que lo precisen. Otro objetivo fundamental es brindar una estructura de apoyo a la **gestión universitaria nacional**.

Esquema de recolección de la información



Datos estadísticos que administra

- Cantidad de alumnos (nuevos inscriptos y reinscriptos) y de egresados de las ofertas académicas de pregrado, grado y postgrado.
- Cantidad de alumnos y nuevos inscriptos, segmentados por rangos de edad.
- Cantidad de alumnos de las ofertas académicas de grado, desagregados según su situación laboral (de acuerdo con rangos de horas que trabajan).
- Cantidad de reinscriptos de las ofertas académicas de grado, desagregados por año de ingreso, por cantidad de exámenes rendidos, por cantidad de materias aprobadas y de materias aprobadas el año anterior.

Todos los datos pueden ser desagregados por oferta académica, año académico y sexo.

Otros datos que contiene la BD

- Unidades académicas de cada institución.
- Ofertas de cada unidad académica o de cada institución.
- Conceptos que permiten clasificar la información, como “ramas”, “disciplinas” o “áreas de estudio” -para la clasificación de las ofertas académicas-; o regiones, provincias y localidades para clasificar las instituciones.

Impacto de la información que se maneja

- Se trata de información utilizada por todas las Universidades Nacionales y la mayoría de las universidades privadas del país.
- Hay información cargada hasta el año 2006 de mas 87 instituciones.
- Se cuenta con información de alrededor de 1.500.000 alumnos, clasificada según los distintos criterios.

Etapas de KDD

Etapas de KDD

- Etapa de Preprocesamiento

Recolección de los datos, integración, limpieza, reducción del volumen y transformación.

- Etapa de Minería y Obtención de Patrones

Trabajar sobre los datos para extraer la esencia de ellos.

- Evaluación y uso de la salida

Etapa de Preprocesamiento

Tareas de preprocesamiento (1)

1) Analizar el modelo de datos.

Justificación:

Reconocer que parte del modelo, es mas interesante para el presente trabajo

2) Acotar el modelo de datos. Implica quitar tablas del modelo, y de ciertas tablas quitar columnas.

Justificación:

Es muy grande y tiene muchísimas tablas relacionadas.

3) Creación de una base de datos y de las tablas con las columnas equivalentes al punto anterior.

Justificación:

La necesidad de migrar los datos a un motor de base de datos y de esta manera poder filtrar solo la parte del modelo sobre el cual se realizo este trabajo, ya que los datos se obtuvieron en modo texto, como un script de comandos INSERT.

Tareas de preprocesamiento (2)

4) Migrar los datos.

Justificación:

La necesidad de migrar los datos al formato necesario para poder correr cada una de las herramientas seleccionadas para realizar el proceso de extracción del conocimiento.

5) Realizar consultas dentro del motor de Base de datos al cual se hallan migrado los datos, para luego utilizar dicho cruce de datos en la herramienta, esta consulta puede ser guardada dentro del motor, o bien migrarla a modo texto o excel, según la herramienta que se use.

Justificación:

La necesidad de cruzar información, por contar con un modelo relacional, con el fin de obtener conocimiento asociado a varios aspectos del modelo.

Tareas de preprocesamiento

Recolección de los datos

Entrada de datos:

Un dump de una base de datos Postgres, con los comandos de creación de la base y demás objetos, así como también los insert de datos a las tablas.

Áreas de preprocesamiento

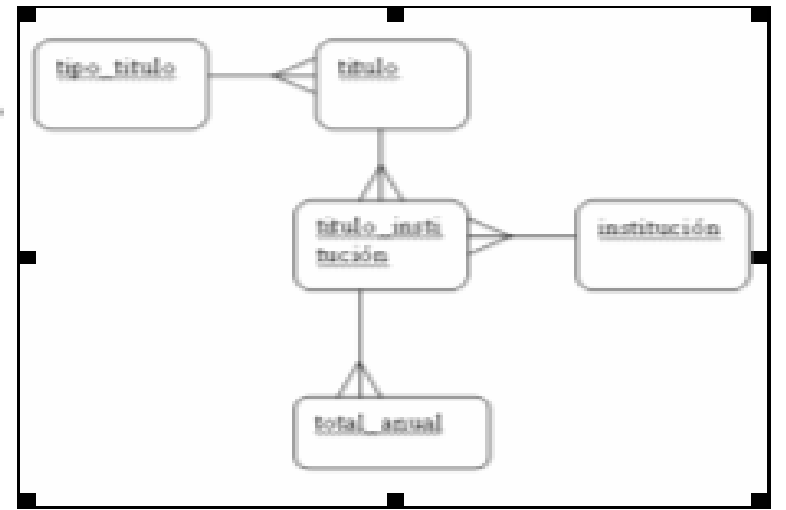
(1)

Limpieza y Selección los datos necesarios

- 1) Sobre un modelo de datos disponible, se selecciono una parte del modelo sobre la cual nos centramos, para este trabajo



- 2) Sobre cada tabla seleccionada, se filtraron columnas que no aportarían información útil, para el proceso de extracción del conocimiento.



(2)

Limpieza y Selección los datos necesarios

- 3) Migración de los datos seleccionados a SQL Server y Postgres, para usarlos respectivamente, con cada una de las herramientas seleccionadas para extraer conocimiento.
- 4) Realizar una consulta, dentro del motor de Base de datos, para cruzar la información, y llevarla al formato que necesita cada una de las herramientas utilizadas.

Etapa de Minería y Obtención de Patrones

Etapa de Minería

Trabajar sobre los datos para extraer conocimiento

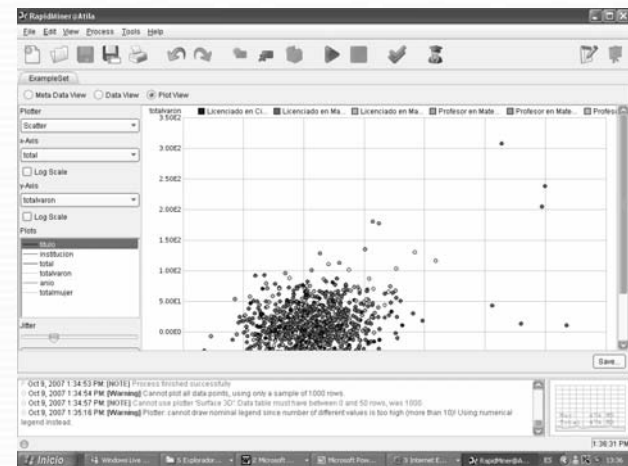
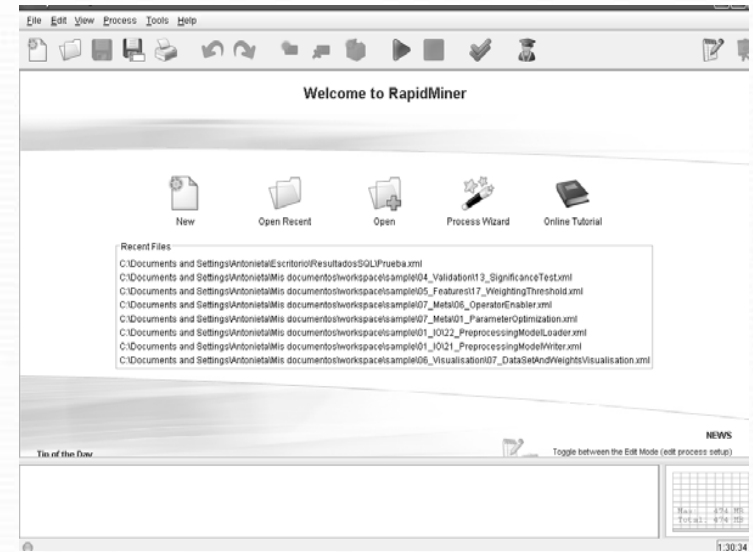
Herramientas utilizadas

- RapidMiner
- Orange Canvas
- TariyKDD
- Armado de un datawarehouse



Herramienta *RapidMiner*

- Open-Source Data Mining with the Java Software RapidMiner / YALE
- Es una herramienta de software que cubre un gran rango de tareas de datamining. Permite descubrir conocimiento a partir de un conjunto de datos.
- Ofrece los siguientes servicios: consulting, support, customization, individual system integration, and data analysis.
- Se importó el resultado de la tabla con los datos cruzados a un excel, y se usó *rapidminer*



Procesando la información para usar *RapidMiner*

- Dados los archivos, con extensión txt: area, carrera, institución disciplina, rama, tipoTítulo, título, título_institucion, total_anual.
- Se importó a SQLSERVER 7, y se uso cada archivo para armar una tabla y se le inserto la información correspondiente.
- Se consideró solo las siguientes tablas, según el dominio del problema: tipo_titulo, título, título_institucion, institucion, totalAnual. Se realizó una consulta cruzada sobre las tablas y se insertó en una tabla temporal

La consulta realizada es:

```
SELECT totalanual.total total, totalanual.totalvarones  
totalvaron, totalanual.totalmujer totalmujer,  
titulo_institucion.anioinformado anio INTO C7 FROM  
titulo_institucion  
INNER JOIN titulo ON  
titulo.idtitulo=titulo_institucion.idtitulo  
INNER JOIN institucion ON  
titulo_institucion.idinstitucion=institucion.idinstitucion  
INNER JOIN totalanual ON  
titulo.idtitulo=titulo_institucion.idtitulo AND  
titulo_institucion.idinstitucion=totalanual.idinstitucion
```

Resultados con *RapidMiner*

- Se aplicó: *BasicRuleLerner*
- *BasicRuleLerner*: forma parte del grupo de reglas supervisadas. Las capacidades de aprendizaje son: atributos polinominales, binominales, numéricos, etc. Aprende un conjunto de reglas minimizando el error entrenando sin poda.
- *Resultados, RuleModel*:

*if anio > 2006 then **Licenciado en Ciencias Matemáticas***

*if institucion = Facultad de Ciencias Exactas y Tecnología then **Licenciado en Matemática***

*if institucion = Facultad de Ciencias Exactas y Naturales then **Licenciado en Matemática***

*if institucion = Facultad de Filosofía y Letras then **Profesor en Matemática***

*if anio > 2005 then **Licenciado en Matemática***

*if institucion = Centro Universitario de Concepción then **Licenciado en Diseño de Indumentaria y Textil** else **Licenciado en Ciencias Matemáticas***

*if institucion = Subsede Académica Buenos Aires then **Profesor en Matemática***

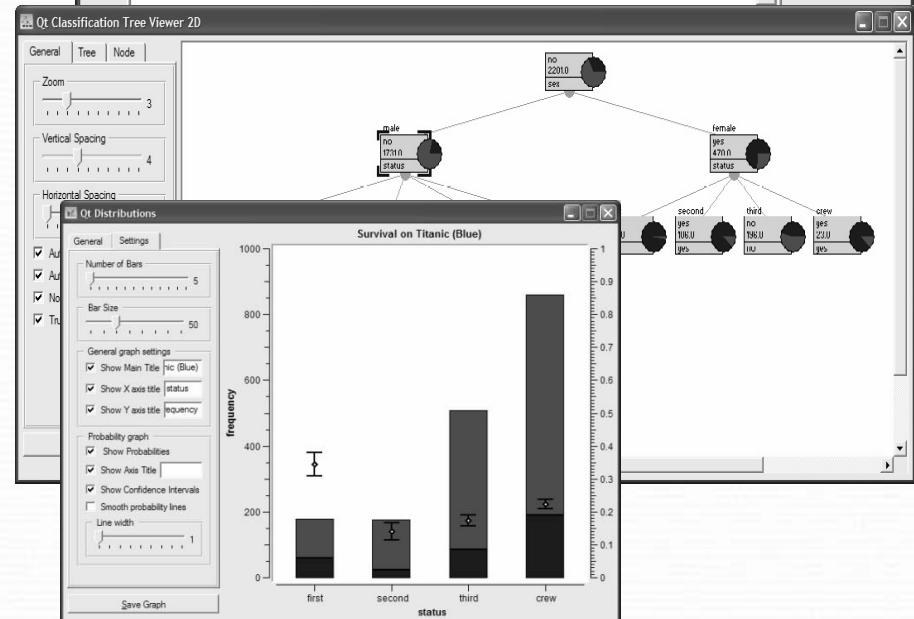
*if institucion = Facultad de Formación Docente en Ciencias then **Profesor en Matemática***

*if institucion = Unidad Académica Colegio Militar de la Nación then **Licenciado en Matemática Aplicada***

correct: 16164 out of 24278 training examples

orange

The concepts of visual programming and interactive graphical user interface components called widgets.

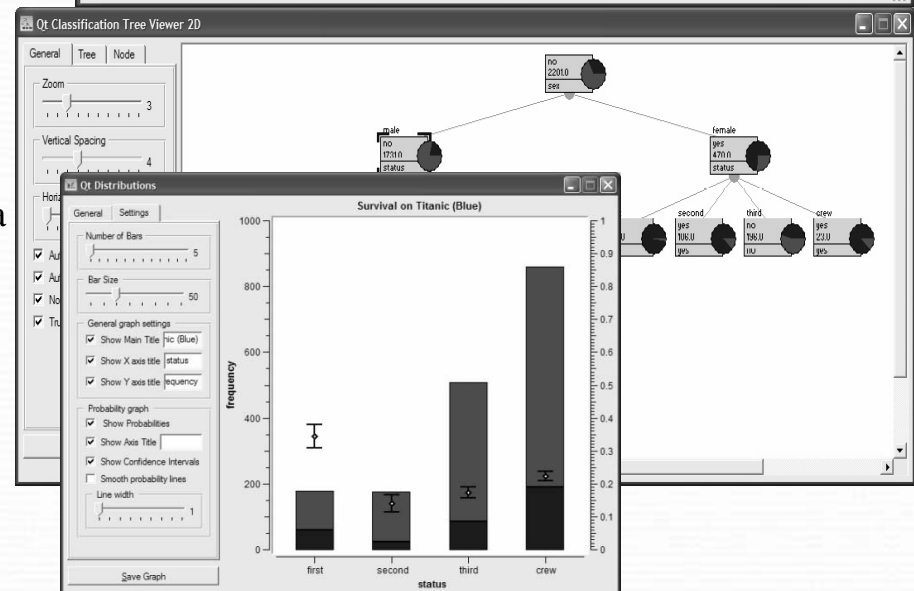
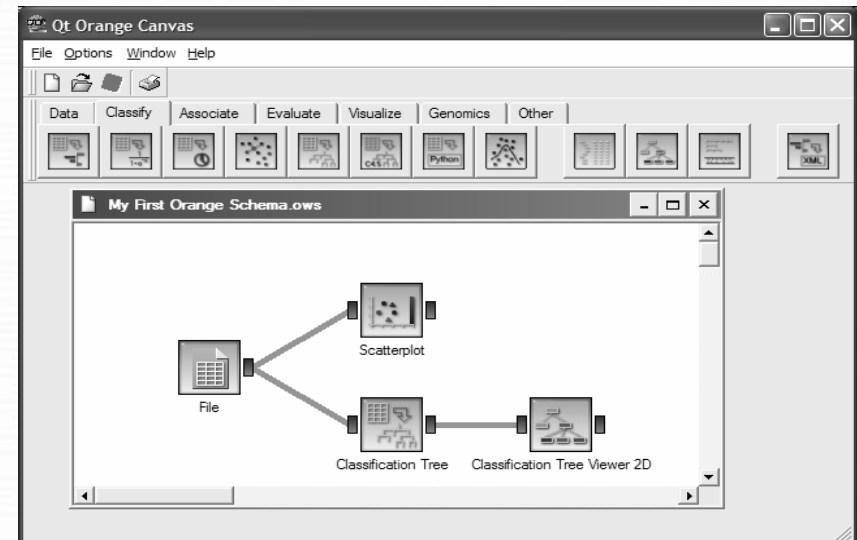


Herramienta OrangeCanvas



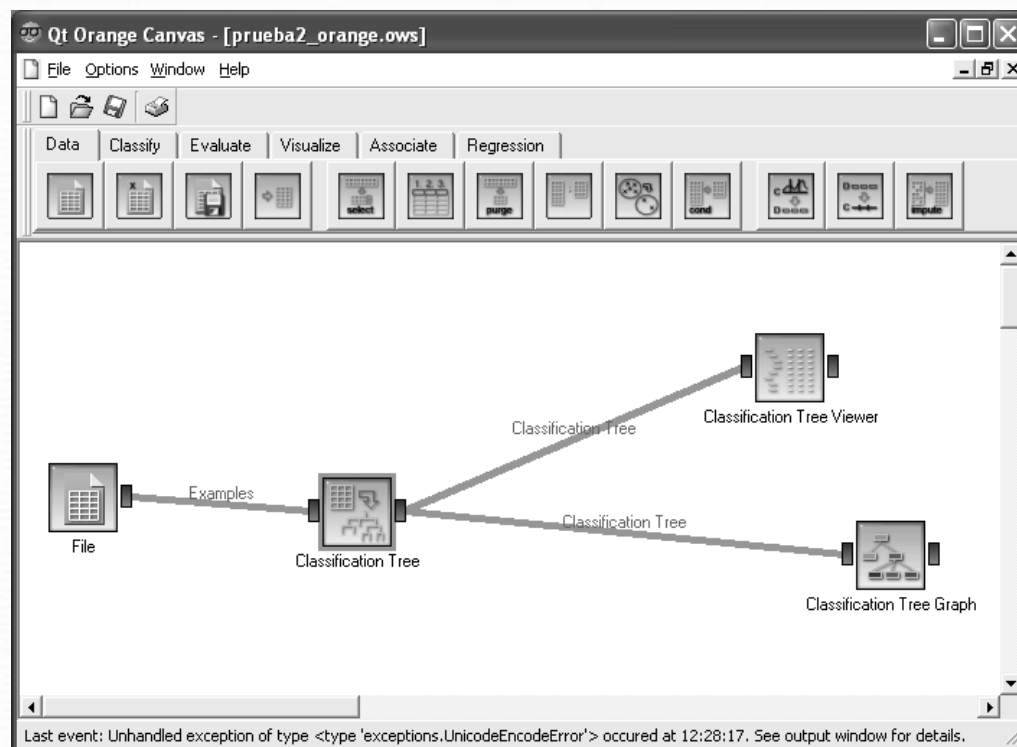
Some Features of Orange

- Orange is a component-based framework, which means you can use existing components and build your own ones.
- Preprocessing: feature subset selection, discretization, feature utility estimation for predictive tasks.
- Predictive modelling: classification trees, naive bayesian classifier, k-NN, majority classifier, support vector machines, logistic regression, rule-based classifiers (e.g., CN2).
- Data description methods: various visualizations (in widgets), self-organizing maps, hierarchical clustering, k-means clustering, multi-dimensional scaling, and other.
- Model validation techniques, that include different data sampling and validation techniques (like cross-validation, random sampling, etc.), and various statistics for model validation (classification accuracy, AUC, sensitivity, specificity, ...).



Procesando la información para usar OrangeCanvas

- Se tomó como entrada un archivo txt, exportado de una consulta armada dentro del motor de base de datos de SQLServer, al cual fueron migrados los datos anteriormente
- Se utilizó el algoritmo de árboles de clasificación.



Resultados con OrangeCanvas

orange

Classification Tree	Class
[-] <root>	"Universidad Nacional del Sur"
"-" "institucion" = "Facultad de Ciencias Físico Matemáticas e Ingeniería"	"Pontificia Universidad Católica Argentina Santa María de los Buenos Aires"
"-" "institucion" = "Facultad de Posgrado"	"Universidad Abierta Interamericana"
"-" "institucion" = "Departamento de Sistemas"	"Universidad CAECE"
"-" "institucion" = "Departamento de Escuela de Posgrado"	"Universidad CAECE"
"-" "institucion" = "Facultad de Informática"	"Universidad Champagnat"
"-" "institucion" = "Facultad de Matemáticas, Astronomía y Física"	"Universidad Nacional de Córdoba"
"-" "institucion" = "Departamento de Ingeniería e Investigaciones Tecnológicas"	"Universidad Nacional de La Matanza"
"-" "institucion" = "Instituto de Postgrado"	"Universidad Nacional de La Matanza"
"-" "institucion" = "Escuela de Postgrado"	"Universidad Nacional de La Matanza"
[-] "institucion" = "Facultad de Ciencias Exactas"	"Universidad Nacional de La Plata"
"-" "titulo" = "Doctor en Ciencias de la Computación"	"Universidad Nacional del Centro de la Provincia de Buenos Aires"
"-" "titulo" = "Doctor de la Facultad de Ciencias Exactas de la Universidad..."	"Universidad Nacional de La Plata"
"-" "titulo" = "Magister en Automatización de Oficinas"	"Universidad Nacional de La Plata"
"-" "titulo" = "Magister en Redes de Datos-Director"	"Universidad Nacional de La Plata"
"-" "titulo" = "Magister en Ingeniería de Software"	"Universidad Nacional de La Plata"
"-" "titulo" = "Magister en Ingeniería de Sistemas"	"Universidad Nacional del Centro de la Provincia de Buenos Aires"
"-" "institucion" = "Facultad de Ciencias Exactas, Físicas y Naturales"	"Universidad Nacional de San Juan"
"-" "institucion" = "Facultad de Cs. Físico-Matemáticas y Naturales"	"Universidad Nacional de San Luis"
"-" "institucion" = "Facultad de Ciencias Exactas y Tecnológicas"	"Universidad Nacional de Santiago del Estero"
"-" "institucion" = "Facultad de Ciencias Exactas y Tecnología"	"Universidad Nacional de Tucumán"
"-" "institucion" = "Facultad de Humanidades"	"Universidad Nacional del Comahue"
"-" "institucion" = "Facultad de Ingeniería y Ciencias Hídricas"	"Universidad Nacional del Litoral"
"-" "institucion" = "Facultad de Ciencias Exactas y Naturales y Agrimensura"	"Universidad Nacional del Nordeste"
"-" "institucion" = "Departamento de Ingeniería"	"Universidad Nacional del Sur"
"-" "institucion" = "Departamento de Ciencias de la Computación"	"Universidad Nacional del Sur"
"-" "institucion" = "Departamento de Ingeniería Eléctrica"	"Universidad Nacional del Sur"
"-" "institucion" = "Facultad Regional Santa Fé"	"Universidad Tecnológica Nacional"
"-" "institucion" = "Facultad Regional San Rafael"	"Universidad Tecnológica Nacional"
"-" "institucion" = "Facultad Regional Mendoza"	"Universidad Tecnológica Nacional"
"-" "institucion" = "Facultad Regional Concepción del Uruguay"	"Universidad Tecnológica Nacional"
"-" "institucion" = "Facultad Regional Buenos Aires"	"Universidad Tecnológica Nacional"
"-" "institucion" = "Facultad Regional Tucumán"	"Universidad Tecnológica Nacional"
"-" "institucion" = "Facultad de Tecnología Informática"	"Universidad de Belgrano"
"-" "institucion" = "Facultad de Arquitectura y Urbanismo"	"Universidad de Belgrano"
"-" "institucion" = "Facultad de Informática, Ciencias de la Comunicación y Té..."	"Universidad de Morón"

Herramienta *TariyKDD*



The TariyKDD Project

Herramienta genérica para el descubrimiento de conocimiento
Débilmente acoplada a un sistema gestor de base de datos.
Desarrollada en conjunto por el Grupo de Investigación GRiAS - Línea
KDD de la Universidad de Nariño (Pasto - Nariño - Colombia -
Suramérica), el Grupo de Usuarios GNU+Linux - UdeNar y ParqueSoft
Pasto, liberada bajo licencia GPL v2.0.

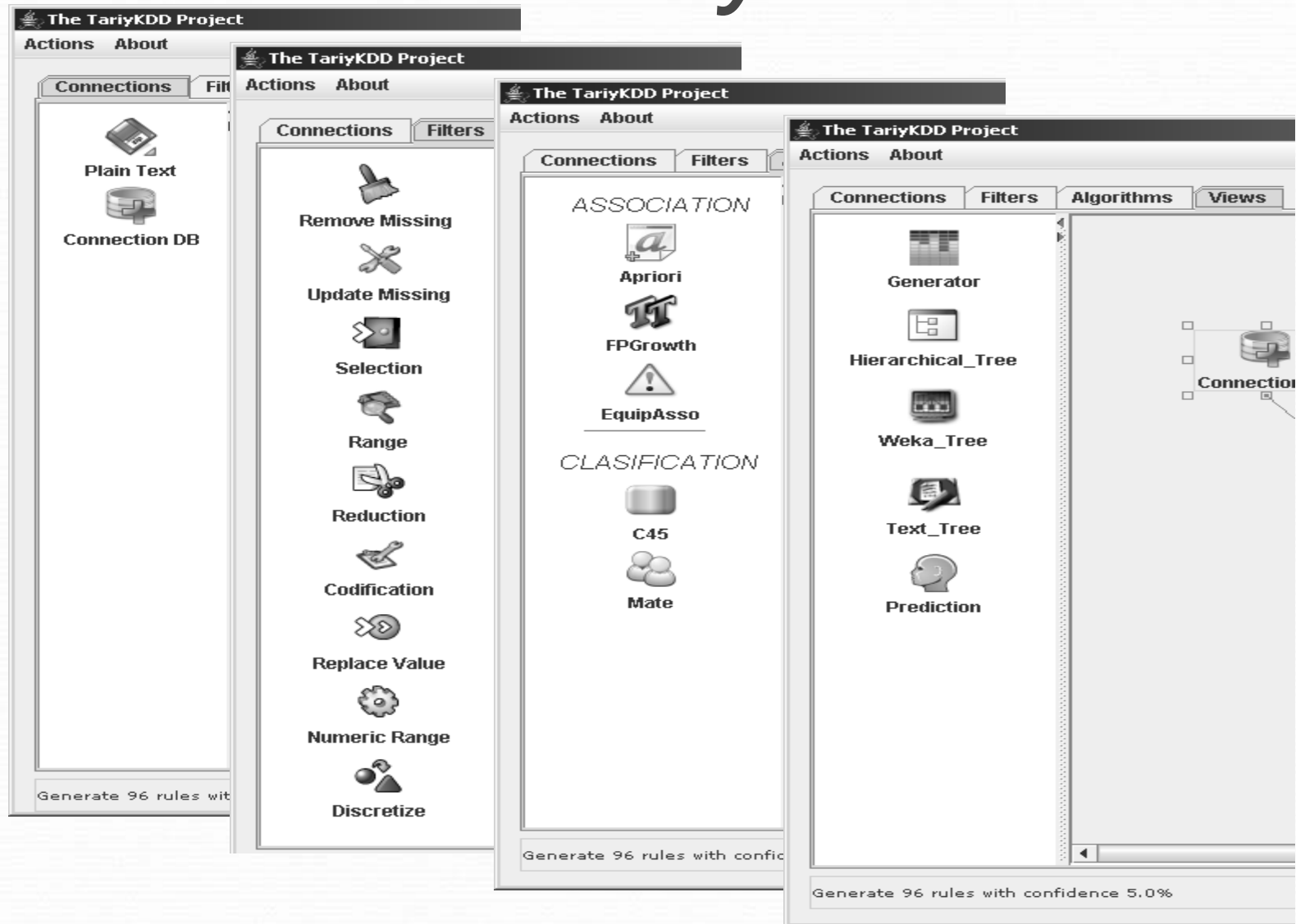
Características

Esta es una herramienta que brinda soporte a todo el proceso
KDD (Descubrimiento de conocimiento en bases de datos)
desde la conexión a bases de datos o archivos planos, pre-
procesamiento de datos, algoritmos de asociación y
clasificación y rutinas de visualización de resultados.

Herramienta *TariyKDD*



The TariyKDD Project

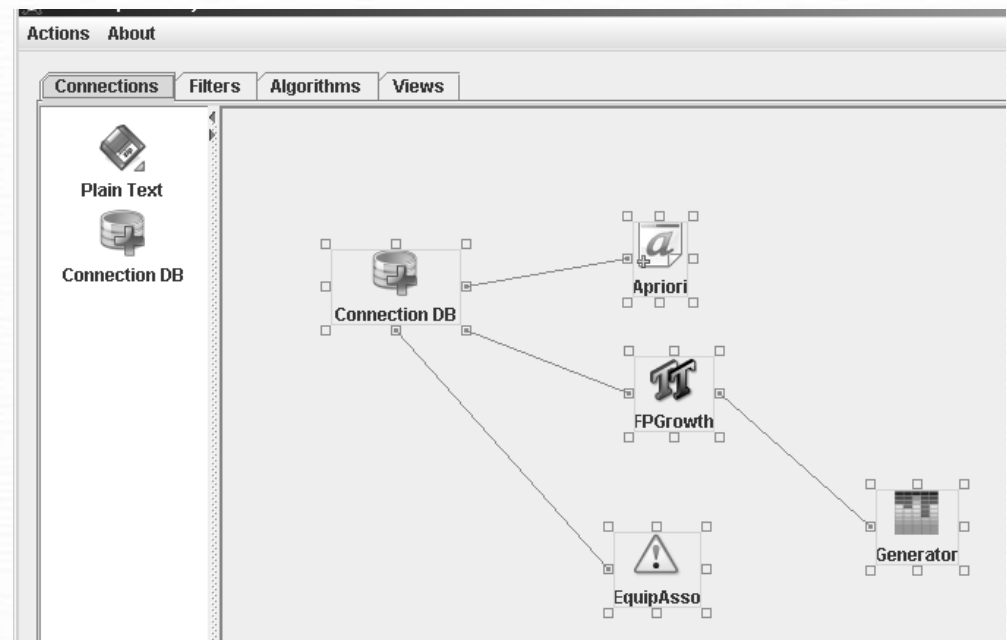


Resultados con TariyKDD

- Se realizó mediante la interfaz de la herramienta una conexión al motor de BD Postgres, donde previamente fueron migraron los datos.
- Dentro de la herramienta misma, se hizo una consulta cruzando varias tablas, donde se seleccionaron solo las columnas que nos interesan obtener.

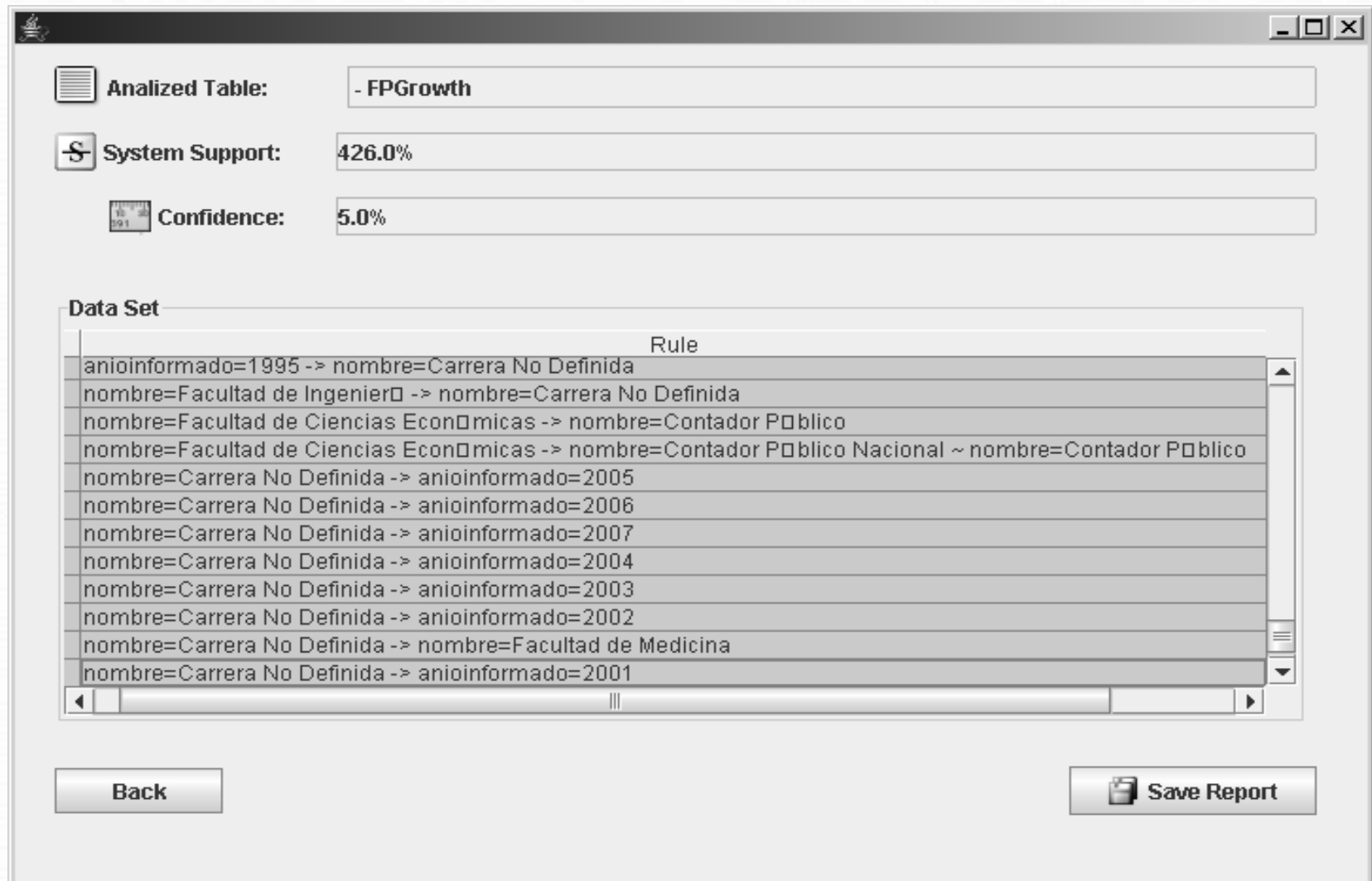
```
SELECT carrera.nombre, titulo.nombre,
institucion.nombre,
titulo_institucion.anioinformado
FROM carrera, titulo, institucion,
titulo_institucion
WHERE carrera.idcarrera =
titulo.idcarrera AND
titulo_institucion.idtitulo = titulo.idtitulo
AND titulo_institucion.idinstitucion =
institucion.idinstitucion
```

- Se utilizó para la técnica de reglas de asociación, el algoritmo A priori, FPGrowth y EquipAsso.



Resultados con *TariyKDD*

Visualización de la salida del algoritmo



The screenshot shows a software window titled "TariyKDD" with a standard Windows-style title bar. The interface is divided into several sections:

- Configuration Section:** Contains three rows of settings, each with an icon on the left and a text field on the right.
 - Row 1: Icon of a document with a list, label "Analyzed Table:", value "- FPGrowth".
 - Row 2: Icon of a plug, label "System Support:", value "426.0%".
 - Row 3: Icon of a bar chart, label "Confidence:", value "5.0%".
- Data Set Section:** A section header followed by a table of results.
- Table:** A table with two columns: "Data Set" (empty) and "Rule". It contains 12 rows of association rules. The first row is highlighted in grey.
- Buttons:** At the bottom, there are two buttons: "Back" on the left and "Save Report" on the right, which includes a floppy disk icon.

Data Set	Rule
	anioinformado=1995 -> nombre=Carrera No Definida
	nombre=Facultad de Ingenier□ -> nombre=Carrera No Definida
	nombre=Facultad de Ciencias Econ□micas -> nombre=Contador P□blico
	nombre=Facultad de Ciencias Econ□micas -> nombre=Contador P□blico Nacional ~ nombre=Contador P□blico
	nombre=Carrera No Definida -> anioinformado=2005
	nombre=Carrera No Definida -> anioinformado=2006
	nombre=Carrera No Definida -> anioinformado=2007
	nombre=Carrera No Definida -> anioinformado=2004
	nombre=Carrera No Definida -> anioinformado=2003
	nombre=Carrera No Definida -> anioinformado=2002
	nombre=Carrera No Definida -> nombre=Facultad de Medicina
	nombre=Carrera No Definida -> anioinformado=2001

Resultados con TaryKDD

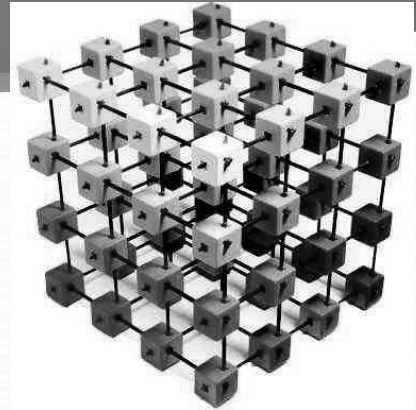


The TaryKDD Project

Reglas Obtenidas

1	nombre=Abogado -> nombre=Abogac	100.0	
2	nombre=Arquitecto -> nombre=Arquitectura y Urbanismo	100.0	
3	nombre=Licenciado en Ciencias de la Educaci -> nombre=Ciencias de la Educaci	100.0	
4	nombre=Profesor en Ciencias de la Educaci -> nombre=Ciencias de la Educaci	100.0	
5	nombre=Licenciado en Comunicaci n Social -> nombre=Comunicaci n Social	100.0	
6	nombre=Contador P blico -> nombre=Contador P blico Nacional	100.0	
7	nombre=Licenciado en Econom -> nombre=Econom	100.0	
8	nombre=Licenciado en Enfermer -> nombre=Enfermer	100.0	
9	nombre=Licenciado en Filosof -> nombre=Filosof	100.0	
10	nombre=Licenciado en Geograf -> nombre=Geograf	100.0	
.....			
28	nombre=Licenciado en Turismo -> nombre=Turismo y Hoteler	100.0	
29	nombre=Contador P blico ~ nombre=Facultad de Ciencias Econ micas -> nombre=Contador P blico Nacional	100.0	
30	nombre=Ingenier a en Construcciones -> nombre=Ingeniero en Construcciones	99.206345	
.....			
62	anioinformado=2005 -> nombre=Carrera No Definida	31.260256	
63	anioinformado=2006 -> nombre=Carrera No Definida	31.208927	
64	nombre=Gesti n y Administraci n de Empresas -> nombre=Licenciado en Administraci	31.18242	
65	anioinformado=2007 -> nombre=Carrera No Definida	28.672361	
66	anioinformado=2004 -> nombre=Carrera No Definida	28.181664	
67	nombre=Letras -> nombre=Profesor en Letras	27.627628	
68	anioinformado=2003 -> nombre=Carrera No Definida	27.317417	
69	anioinformado=1999 -> nombre=Carrera No Definida	25.950258	
70	anioinformado=2000 -> nombre=Carrera No Definida	25.943705	

Datawarehouse

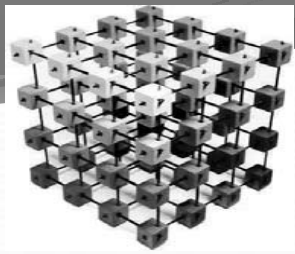


OLAP (On-Line Analytical Processing) es una estructura (tablas relacionales incorporadas) que se desnormaliza con el fin de aumentar la velocidad.

Ventajas: Permite analizar la información de la manera que se quiera, y dinámicamente armar los reportes que se necesiten, según lo que se requiera en el momento.

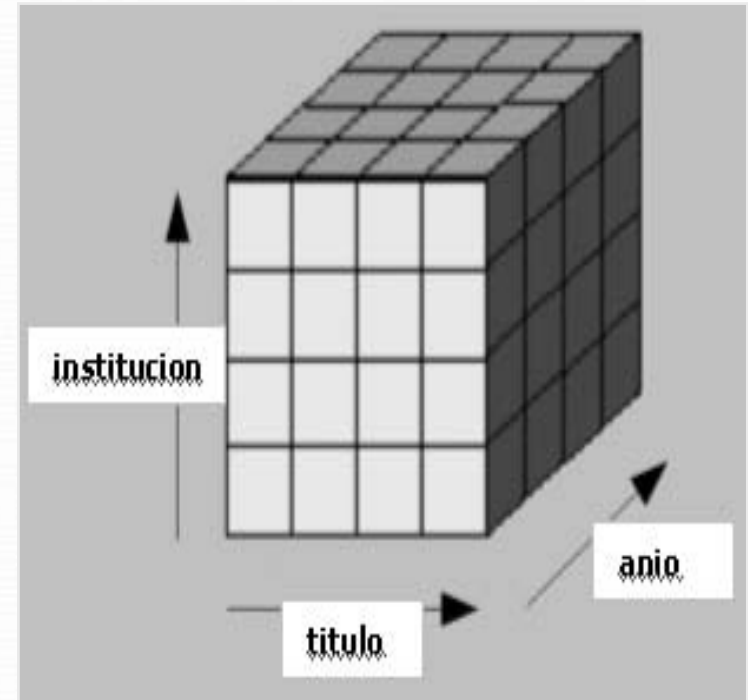
Requerimientos: Es necesario realizar un buen diseño del cubo, para poder satisfacer todas las necesidades de información que se requieran, esto es definir bien la estructura, las dimensiones y las medidas.

Procesando la información con *Data Warehousing*



Para este modelo se pueden obtener mejores resultados aplicando Data Warehousing usando SQLSERVER

- Describimos **qué** deseamos ver y a continuación **cómo** deseamos ver.
- Medidas: Especifica que lo que se desea ver, totalanual, totalmujer, totalvaron, totaldereinscriptosvarones, cantidaddeofertas, egresadosmujeres, egresadosvarones
- Dimensiones: Especifica cómo se desean ver los datos, anio_informado, ofertaacademica (nivel, titulo, unidadAcademica, institucion), lugar (region, provincia, localidad)



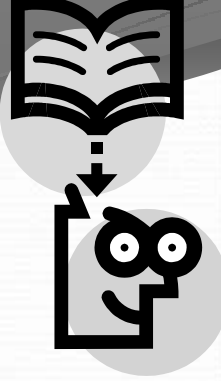
Resultados con el *Datawarehouse*

Institución	Título	Año	1981	1982	1983
		Total de Egresados	Total de Egresados	Total de Egresados	Total de Egresados
Escuela Universitaria de Teología	Profesor en Catequesis				
	Profesor en Teología				
Facultad Latinoamericana de Ciencias Sociales	Doctor en Ciencias Sociales				
	Especialista en Ciencia Política y Sociología				
	Especialista en Ciencias Sociales con mención en Salud				
	Especialista en Constructivismo y Educación				
	Especialista en Estudios Sociales Agrarios				
	Especialista en Gestión y Conducción del Sistema Educativo y sus Instituciones				
	Especialista en Política, Evaluación y Gerencia Social				
	Especialista en Políticas Educativas				
	Especialista en Relaciones Internacionales				
	Especialista en Relaciones y Negociaciones Internacionales				
	Magister en Ciencia Política y Sociología				
	Magister en Ciencias Sociales con Mención en Educación				
	Magister en Ciencias Sociales con mención en Salud				
	Magister en Diseño y Gestión de Programas Sociales				
	Magister en Estudios Sociales Agrarios				
	Magister en Procesos Cognitivos y Aprendizaje				
	Magister en Relaciones Internacionales				
Instituto de Enseñanza Superior del Ejército	Analista Administrativo Contable				
	Bachiller Universitario en Ingeniería				
	Bachiller Universitario en Relaciones Internacionales				
	CBC - Contador Público				
	CBC - Licenciado en Conducción y Gestión Operativa				
	Ciclo Básico Escuela Superior Técnica				
	Contador Público				
	Enfermera/o Universitaria/o				
	Especialista en Conducción y Gestión Estratégica				
	Especialista en Criptografía y Seguridad Teleinformática				
	Especialista en Derecho Militar				
	Especialista en Gestión Ambiental				
	Ingeniero en Construcciones				

Resultados con el *Datawarehouse*

											▲
Año	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	
Rama - Carrera	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de Alumnos	Total de
Ciencias Aplicadas	23047	122184	124799	152133	176911	199081	208255	213349	207546		
Ciencias Básicas	3633	14165	16048	19842	23640	26735	26491	22936	23665		
Ciencias de la Salud	19380	52051	51864	68148	71926	83806	97145	109886	114842		
Ciencias Humanas	11141	28858	32585	49815	61260	70187	79242	79705	79629		
Ciencias Sociales	17277	100261	111298	151562	189111	200333	206341	225570	234147		
Sin Rama	0	780	1375	1941	1705	1671	1177	1551	1483		

Análisis de los resultados obtenidos



Por el dominio del problema tratado y la naturaleza de los datos, como la evolución de las ofertas académicas por universidad a lo largo de los años, creemos que la *técnica de Minería* con la que podemos obtener mayor conocimiento e información estadística es con un DataWarehouse.

Con las otras técnicas tales como clasificación, árboles de decisión y reglas de asociación, podemos agrupar los datos para ver a que conjunto pertenecen o descubrir relaciones entre los datos, pero para este dominio en particular, el conocimiento extraído con estas técnicas resulto ya conocido.