Escola d' Estiu de Doctorat fac. Pedagogia

Taller Minería de datos aplicados a la educación

2ª parte

Presentación del software

PASW Modeler

27 de junio de 2011

Mercedes Torrado

Departamento Métodos de Investigación y Diagnóstico en Educación (MIDE)

Para citar la obra:
Torrado, M. (2011) *Minería de datos aplicados a la educación: software PASW MODELER.* Barcelona: Universidad de Barcelona. Deposito Digital http://hdl.handle.net/2445/19862

SPSS Inc (2010) ***Manual de minería interna de base de datos de PASW Modeler 13***

Escola
d' Estiu
de Doctorat
Fac. Pedagogia

Presentación del software PASW Modeler

1.- ¿Qué nos permite hacer este software?

2.- ¿Con qué terminología trabaja?

3.- ¿Qué preguntas nos debemos hacer?

4.- ¿Qué proceso debemos seguir?

1¿Qué nos permite hacer este software?

- **Definición del PASW Modeler**

Es un conjunto de herramientas de minería de datos que permite desarrollar rápidamente ***modelos predictivos*** *y utilizarlos para* ***mejorar la toma de decisiones***

Permite extraer ***información*** *de grandes bases de datos*

- **Definición de MODELO**

*Es un **conjunto de reglas**, fórmulas o ecuaciones que puede utilizarse para pronosticar un resultado basándose en un conjunto de campos o variables de entrada*

1¿Qué nos permite hacer este software?

- **Objetivo del programa**

El objetivo principal del *PASW Modeler* es la capacidad de pronosticar un resultado a partir del análisis predictivo y la comprensión del proceso

1¿Qué nos permite hacer este software?

- **Por ejemplo**

Identificar alumnos universitarios con mayor probabilidad de no persistir en la universidad en base a una serie de características personales y psicoeducativas

2 ¿Con qué terminología trabaja?

- **Terminología común**
  - **ESCALA DE MEDIDA**

Spss-win	PASW Modeler
Escala	Rango
N/a	Discreto
Nominal	Conjunto
Ordinal	Conjunto ordenado
N/a	Sin tipo
N/a	Por defecto

- **Terminología común**
  - **TÉRMINO DE PAPELES**

Variable independiente o predictora

“entrada”

Variable dependiente

“salida” “objetivo”

Eliminación o anulación

“Ninguno”

2 ¿Con qué terminología trabaja?

- **Terminología común**
  - Término de atributo, campo o variable
  - Término de registro, ejemplo o caso
  - “ruido” datos con errores

3.- ¿Qué preguntas nos debemos hacer?

- ¿Cuál es el problema?
- ¿Dónde tenemos los datos?
- ¿Qué datos son los más importantes?
- ¿Qué ruidos existen?
- ¿Qué técnica de modelado es la más adecuada?
- ¿El modelo es adecuado?

4.- ¿Qué proceso debemos seguir?

TIPOLOGÍA DE MODELADOS

- ***Modelos de clasificación***

A partir del valor de uno o más campos de entrada (VI) predicen el valor de uno o más variables de salida (VD)

- ***Modelos de asociación***

Permiten pronosticar varios resultados. Encuentran patrones de asociación entre datos

- ***Modelos de segmentación***

Dividen los datos en segmentos o conglomerados con características similares y se desconoce el resultado

- ***Modelos de cribados***

Permiten identificar anomalías / casos extraños o valores atípicos

4.- ¿Qué proceso debemos seguir?

- **Creación de un modelo**
  - Nodo de origen
  - Nodo de tipo
  - Nodo de modelado
  - Nodo de tabla y análisis
- **Exploración de un modelo**
  - Nugget del modelo (examinar) visor
- **Evaluación de un modelo**
  - Análisis

Glosario básico
Ejemplo práctico

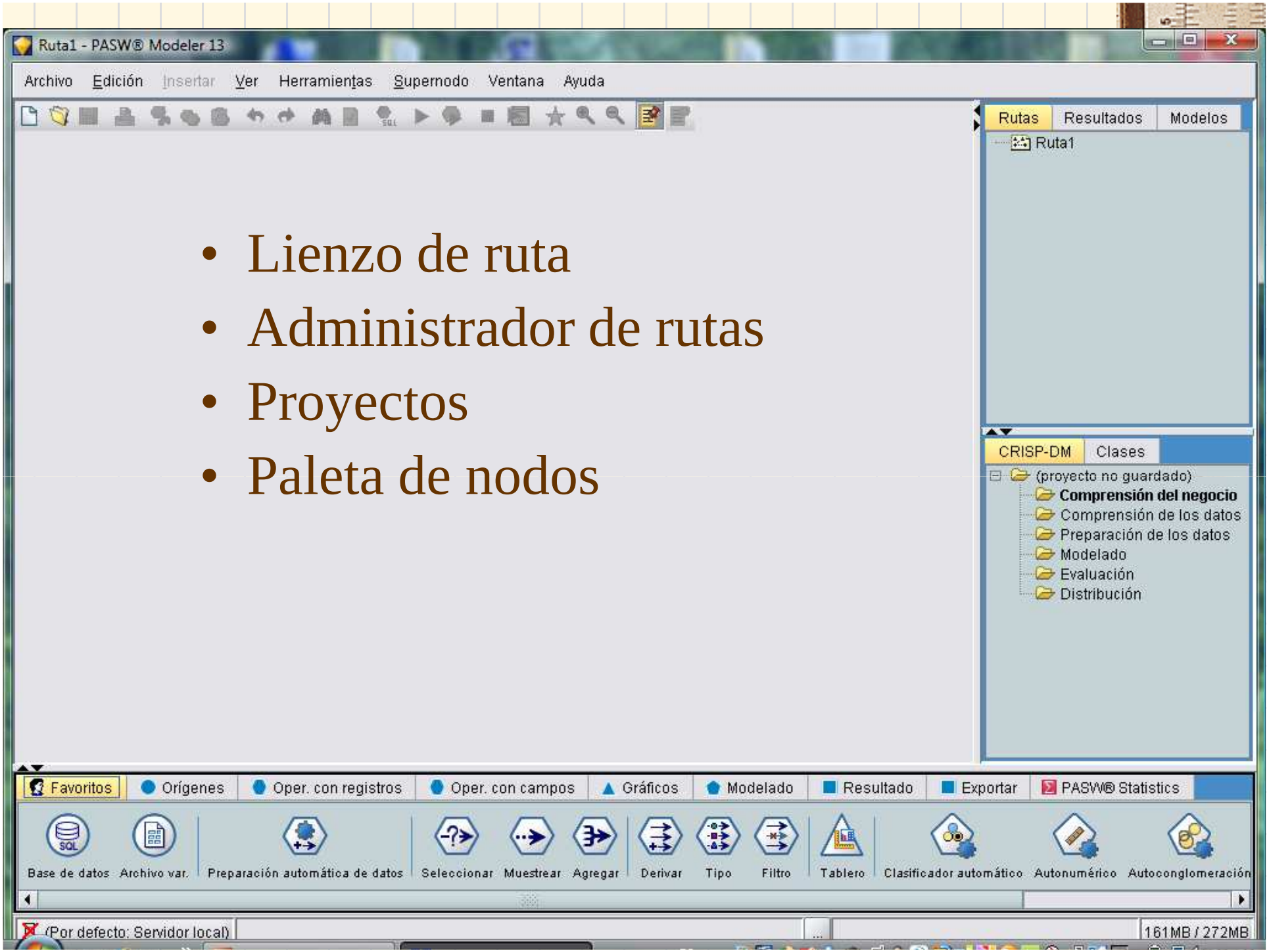
Glosario básico

- Lienzo de ruta
- Administrador de rutas
- Proyectos
- Paleta de nodos
- Icono o nodo
- Rutas
- Flujo de datos
- Nuggets

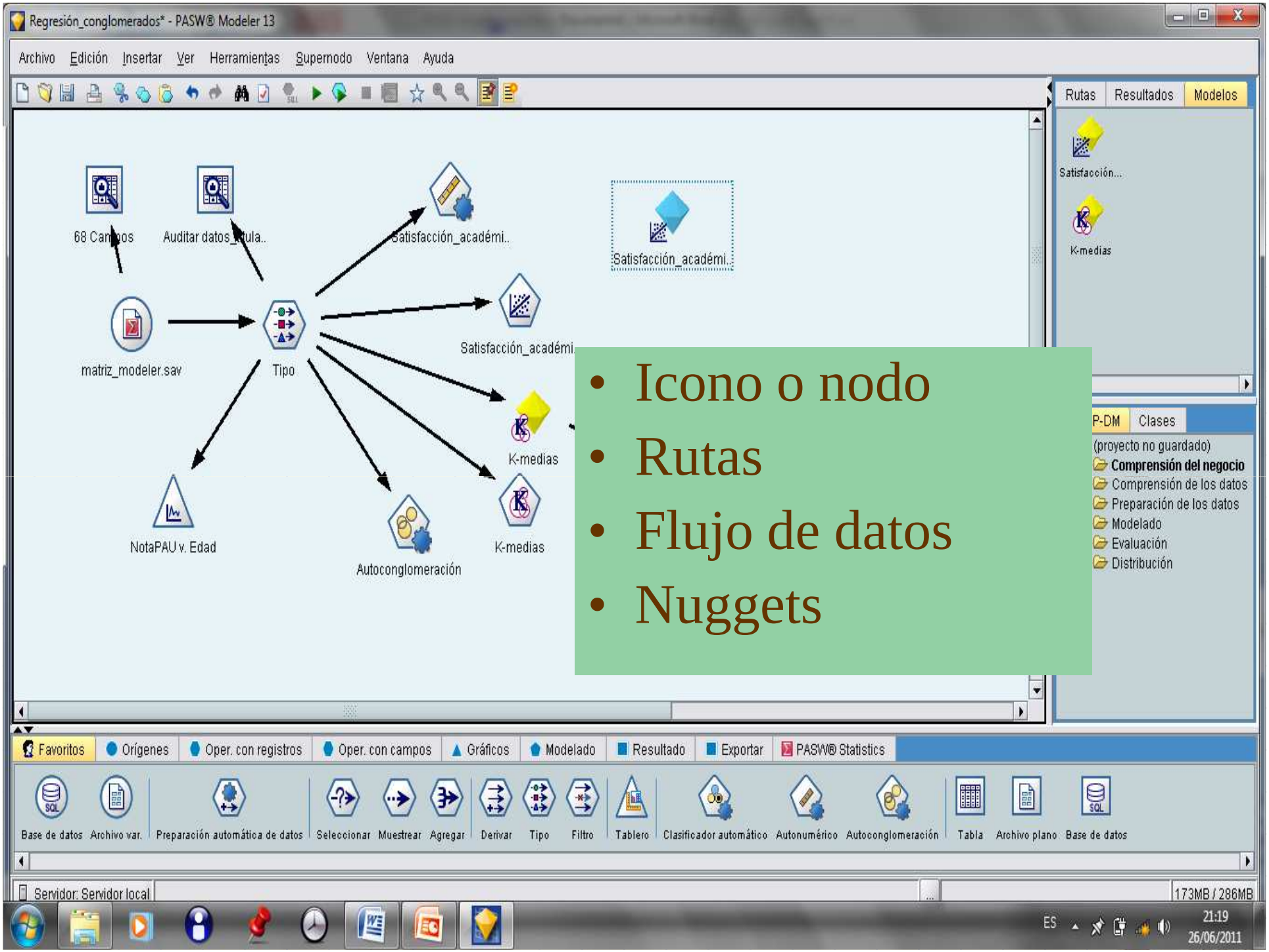
Creación de un modelo

Para crear un modelo se necesitan mínimo tres elementos

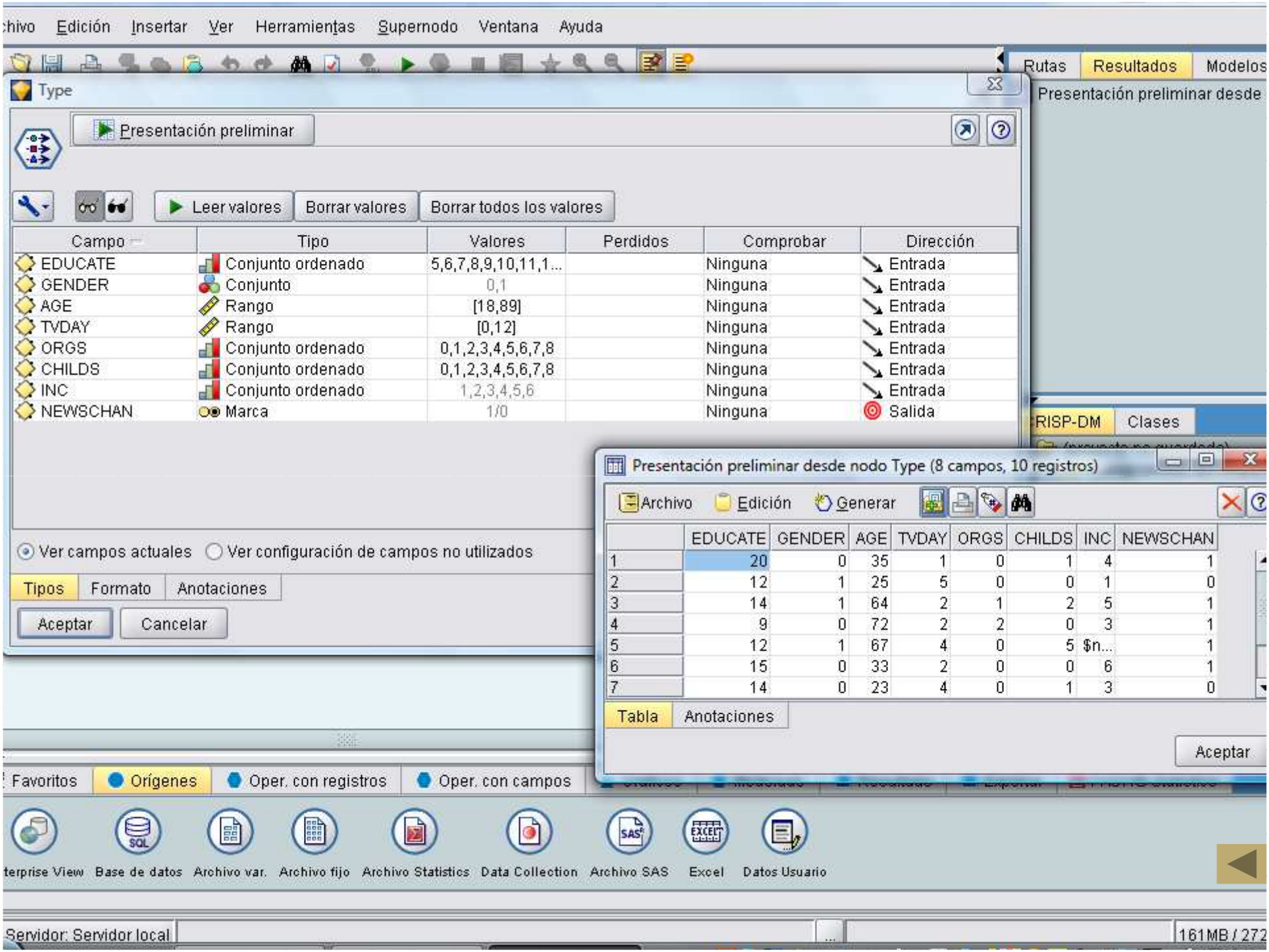
- UN NODO DE ORIGEN que lea los datos
- UN NODO DE ORIGEN O NODO TIPO que especifique propiedades del campo, tipo de datos, el papel de cada campo (origen o predictor en el modelado)
- UN NODO DE MODELADO que genere un nugget de modelado cuando se ejecute la ruta
- UN NODO DE TABLA /ANÁLISIS para ver los resultados de puntuación después de crear el nugget de modelo y añadirlo a la ruta

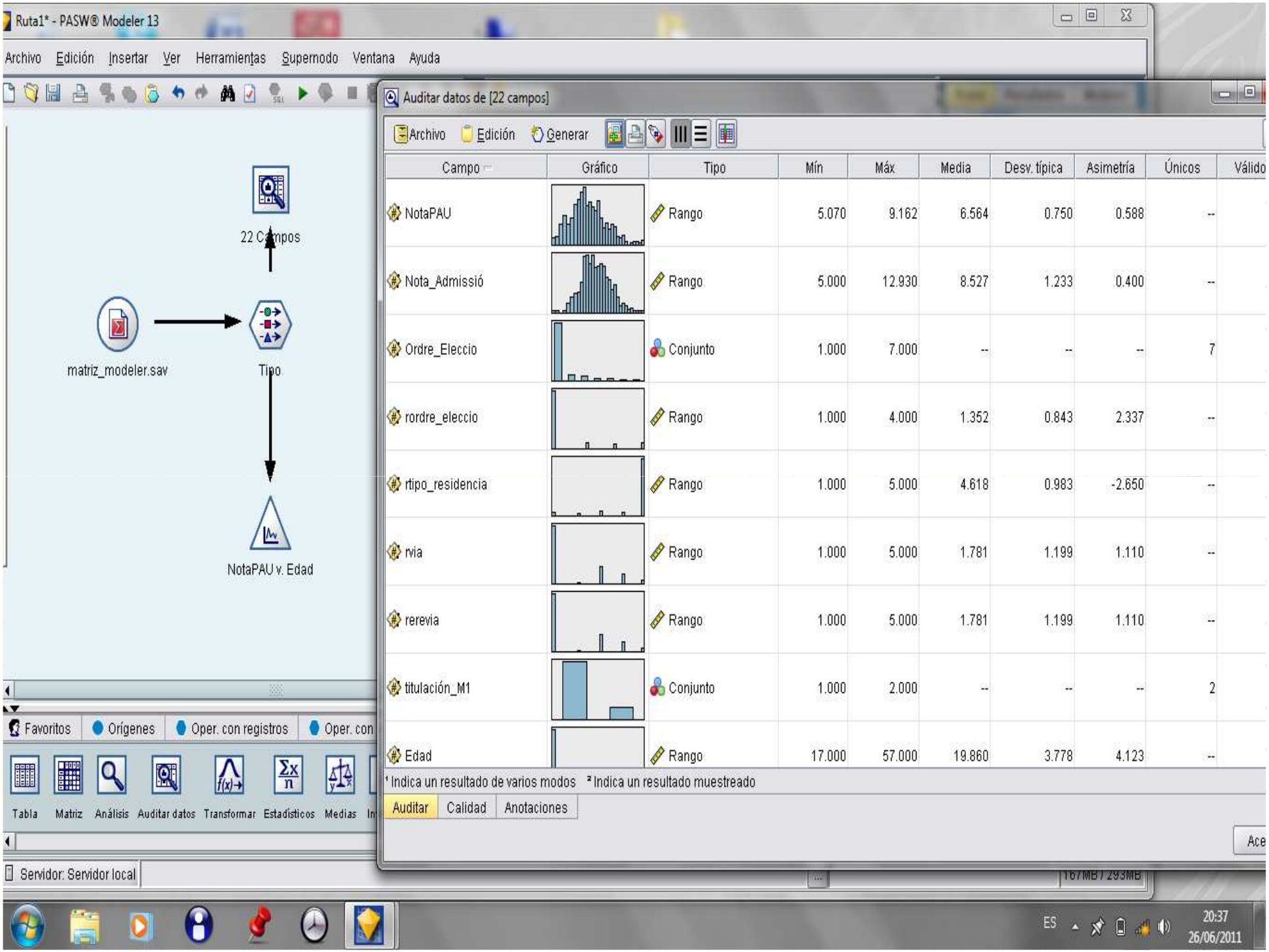


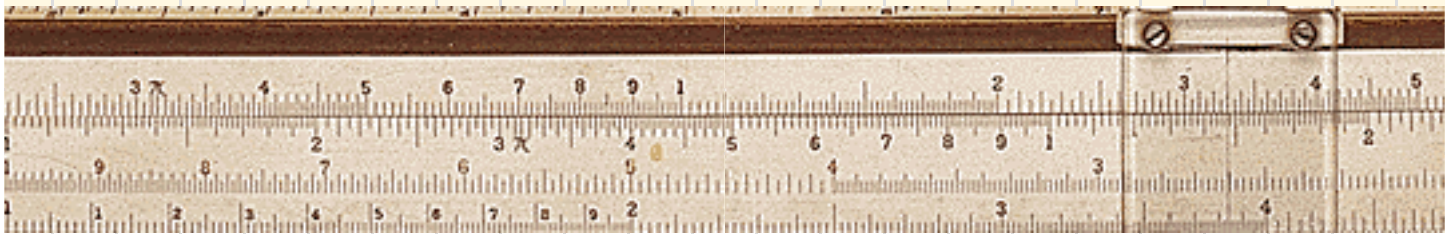
- Lienzo de ruta
- Administrador de rutas
- Proyectos
- Paleta de nodos



- Icono o nodo
- Rutas
- Flujo de datos
- Nuggets

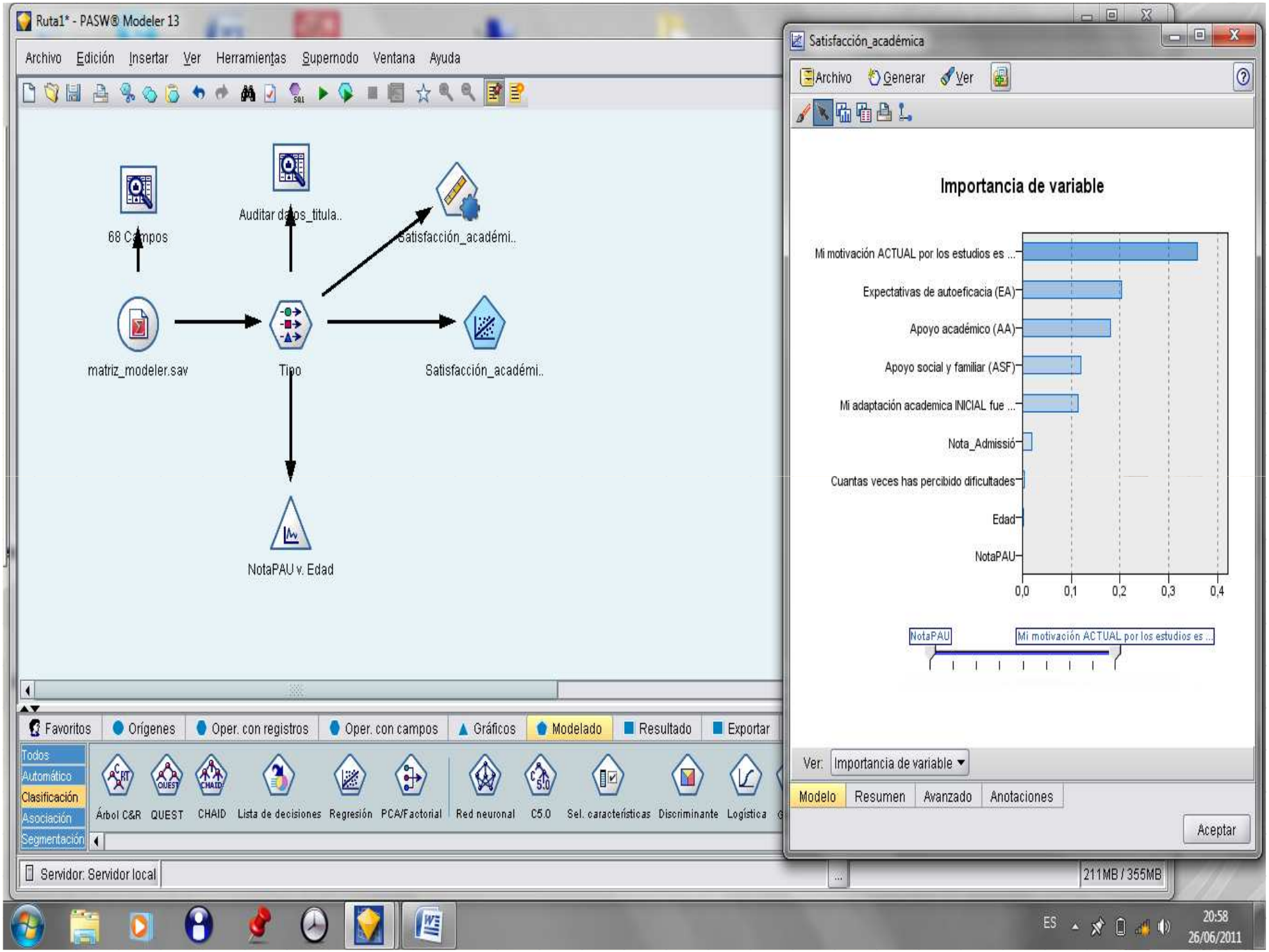

hivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
Rutas Resultados Modelos
Presentación preliminar desde
Type
Presentación preliminar
Leer valores
Borrar valores
Borrar todos los valores
Campo Tipo Valores Perdidos Comprobar Dirección
EDUCATE Conjunto ordenado 5,6,7,8,9,10,11,1... Ninguna Entrada
GENDER Conjunto 0,1 Ninguna Entrada
AGE Rango [18,89] Ninguna Entrada
TVDAY Rango [0,12] Ninguna Entrada
ORGS Conjunto ordenado 0,1,2,3,4,5,6,7,8 Ninguna Entrada
CHILDS Conjunto ordenado 0,1,2,3,4,5,6,7,8 Ninguna Entrada
INC Conjunto ordenado 1,2,3,4,5,6 Ninguna Entrada
NEWSCHAN Marca 1/0 Ninguna Salida
RISP-DM Clases
Ver campos actuales
Ver configuración de campos no utilizados
Tipos Formato Anotaciones
Aceptar Cancelar
Presentación preliminar desde nodo Type (8 campos, 10 registros)
Archivo Edición Generar
EDUCATE GENDER AGE TVDAY ORGS CHILDS INC NEWSCHAN
1 20 0 35 1 0 1 4 1
2 12 1 25 5 0 0 1 0
3 14 1 64 2 1 2 5 1
4 9 0 72 2 2 0 3 1
5 12 1 67 4 0 5 $n... 1
6 15 0 33 2 0 0 6 1
7 14 0 23 4 0 1 3 0
Tabla Anotaciones
Aceptar
Favoritos Orígenes Oper. con registros Oper. con campos
terprise View Base de datos Archivo var. Archivo fijo Archivo Statistics Data Collection Archivo SAS Excel Datos Usuario
Servidor: Servidor local
161MB / 272


Ruta1* - PASW® Modeler 13
Archivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
22 Campos
matriz_modeler.sav
Tipo
NotaPAU v. Edad
Auditar datos de [22 campos]
Archivo Edición Generar
Campo Gráfico Tipo Mín Máx Media Desv. típica Asimetría Únicos Válido
NotaPAU Rango 5.070 9.162 6.564 0.750 0.588 --
Nota_Admissió Rango 5.000 12.930 8.527 1.233 0.400 --
Ordre_Eleccio Conjunto 1.000 7.000 -- -- -- 7
rordre_eleccio Rango 1.000 4.000 1.352 0.843 2.337 --
rtipo_residencia Rango 1.000 5.000 4.618 0.983 -2.650 --
rvia Rango 1.000 5.000 1.781 1.199 1.110 --
rerevia Rango 1.000 5.000 1.781 1.199 1.110 --
titulación_M1 Conjunto 1.000 2.000 -- -- -- 2
Edad Rango 17.000 57.000 19.860 3.778 4.123 --
Indica un resultado de varios modos Indica un resultado muestreado
Auditar Calidad Anotaciones
Favoritos Orígenes Oper. con registros
Tabla Matriz Análisis Auditar datos Transformar Estadísticos Medias
Servidor: Servidor local
ES 20:37 26/06/2011



Explorar el modelo

Cuando se finaliza la ejecución del modelo el nugget se añade a la paleta de modelos. Para ver los detalles se pulsa el botón derecho del ratón y se selecciona “examinar”. La ficha muestra los detalles del modelo (variables, valores, etc.) La ficha visor muestra también un gráfico de la importancia de la variable en la estimación del modelo.


Ruta1* - PASW® Modeler 13
Archivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
68 Campos
Auditar datos_titula..
Satisfacción_académi..
matriz_modeler.sav
Tipo
Satisfacción_académi..
NotaPAU v. Edad
Favoritos Orígenes Oper. con registros Oper. con campos Gráficos Modelado Resultado Exportar
Todos
Automático
Clasificación
Asociación
Segmentación
Árbol C&R QUEST CHAID Lista de decisiones Regresión PCA/Factorial Red neuronal C5.0 Sel. características Discriminante Logística
Servidor: Servidor local
211MB / 355MB
Satisfacción_académica
Archivo Generar Ver
Importancia de variable
Mi motivación ACTUAL por los estudios es ...
Expectativas de autoeficacia (EA)
Apoyo académico (AA)
Apoyo social y familiar (ASF)
Mi adaptación academica INICIAL fue ...
Nota_Admissió
Cuantas veces has percibido dificultades
Edad
NotaPAU
0,0 0,1 0,2 0,3 0,4
NotaPAU
Mi motivación ACTUAL por los estudios es ...
Ver: Importancia de variable
Modelo Resumen Avanzado Anotaciones
Aceptar
ES
20:58
26/06/2011

K-medias

Archivo Generar Ver

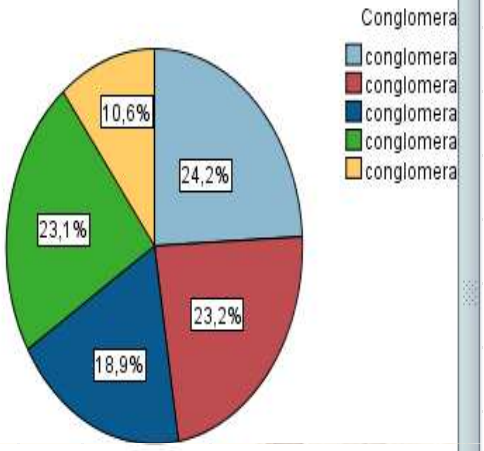
Conglomerado	conglomerado-1	conglomerado-2	conglomerado-4	conglomerado-3	conglomerado-5
Etiqueta					
Descripción					
Tamaño	24,2% (200)	23,2% (192)	23,1% (191)	18,9% (156)	10,6% (88)
Características	Apoyo académico (AA) -0,07	Apoyo académico (AA) 0,33	Apoyo académico (AA) 0,14	Apoyo académico (AA) -0,73	Apoyo académico (AA) 0,44
	Apoyo social y familiar (ASF) 0,25	Apoyo social y familiar (ASF) 0,26	Apoyo social y familiar (ASF) 0,00	Apoyo social y familiar (ASF) -0,83	Apoyo social y familiar (ASF) 0,36
	Confianza: combinar estudios con otras actividades 3,41	Confianza: combinar estudios con otras actividades 4,22	Confianza: combinar estudios con otras actividades 3,99	Confianza: combinar estudios con otras actividades 3,24	Confianza: combinar estudios con otras actividades 3,91
	Confianza: completar todos los requisitos académicos 3,63	Confianza: completar todos los requisitos académicos 4,32	Confianza: completar todos los requisitos académicos 3,88	Confianza: completar todos los requisitos académicos 3,25	Confianza: completar todos los requisitos académicos 3,98
	Confianza: comprender los contenidos de las asig...	Confianza: comprender los contenidos de las asig...	Confianza: comprender los contenidos de las asig...	Confianza: comprender los contenidos de las asig...	Confianza: comprender los contenidos de las asig...

Ver: Conglomerados Casillas: Centros de conglomerados Representación... Restablecer

Clasificar características por: Importancia global Clasificar conglomerados por: Tamaño

Modelo Resumen Anotaciones

Tamaños de conglomerados



Tamaño de conglomerado más pequeño	88 (10,6%)
Tamaño de conglomerado más grande	200 (24,2%)
Cociente de tamaños:	2,27

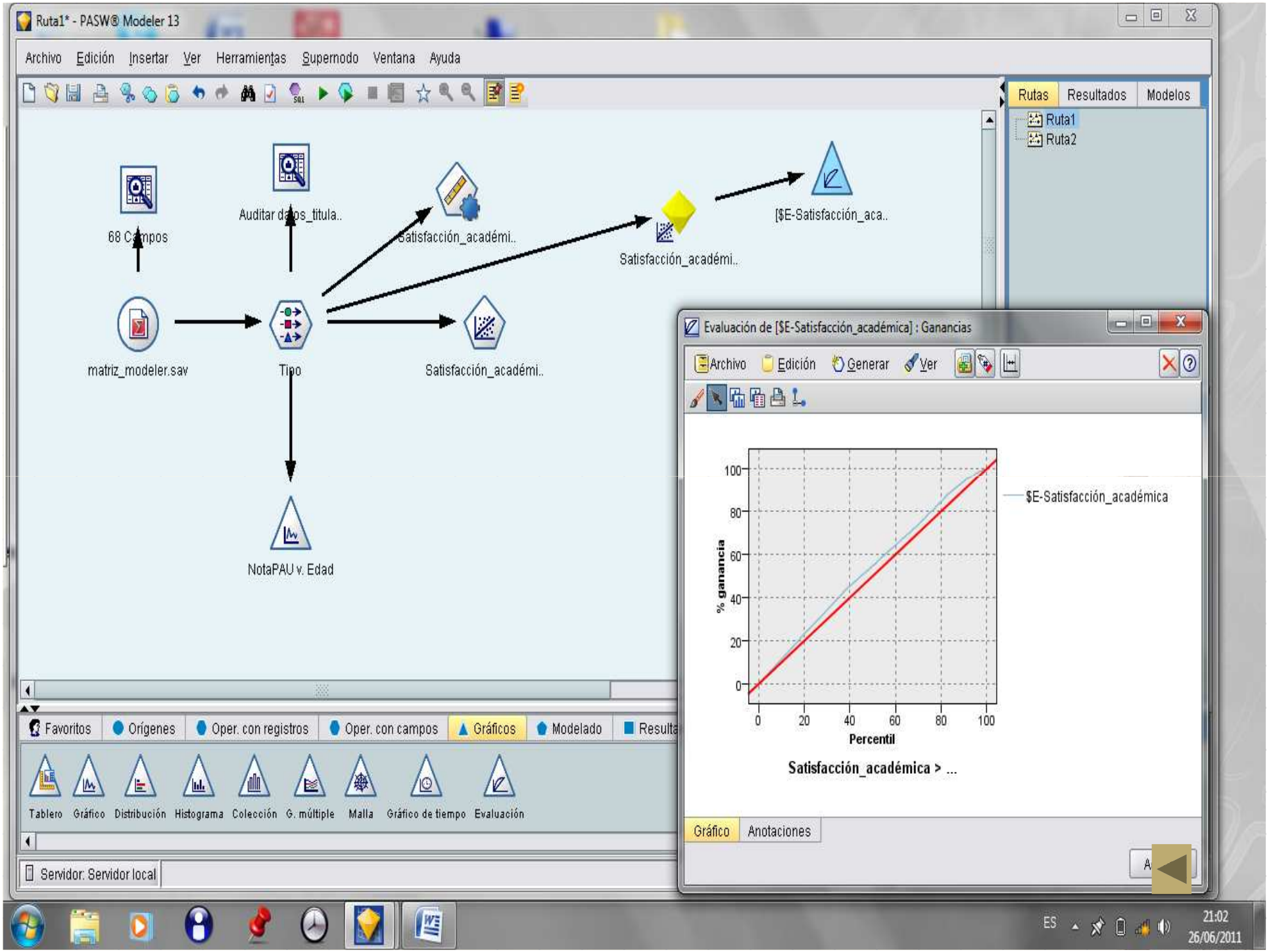
Ver: Tamaños de conglomerados

eptar

ES 21:09 26/06/2011

Evaluación del modelo

Para evaluar con que precisión trabaja en modelo, debe puntuar varios registros y comparar las respuestas pronosticadas por el modelo con los resultados reales. Se debe pulsar “añadir ruta”. Se adjunta el modelo al nodo Tipo. Para ver las puntuaciones o pronósticos se adjunta el nodo Tabla y se ejecuta. Se añadirá al final de la tabla el valor pronosticado. Para descubrir cuántos pronósticos son correctos puede leer la tabla o puede adjuntar el nodo análisis que lo hace automáticamente.


Ruta1* - PASW® Modeler 13
Archivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
Rutas Resultados Modelos
Ruta1
Ruta2
68 Campos
Auditar datos_titula..
Satisfacción_académi..
Satisfacción_académi..
[$E-Satisfacción_aca..
matriz_modeler.sav
Tipo
Satisfacción_académi..
NotaPAU v. Edad
Evaluación de [$E-Satisfacción_académica] : Ganancias
Archivo Edición Generar Ver
100
80
60
40
20
0
% ganancia
0 20 40 60 80 100
Percentil
$E-Satisfacción_académica
Satisfacción_académica > ...
Gráfico Anotaciones
Favoritos Orígenes Oper. con registros Oper. con campos Gráficos Modelado
Tablero Gráfico Distribución Histograma Colección G. múltiple Malla Gráfico de tiempo Evaluación
Servidor: Servidor local
ES
21:02
26/06/2011

Ejemplo práctico

Investigación sobre ***"La persistencia y el abandono en el primer año de universidad en ciencias sociales: bases para la mejora de la retención"*** (Ref: EDU2009-10351) que se lleva a cabo por el grupo de investigación TRALS de la Facultad de Pedagogía de la UB

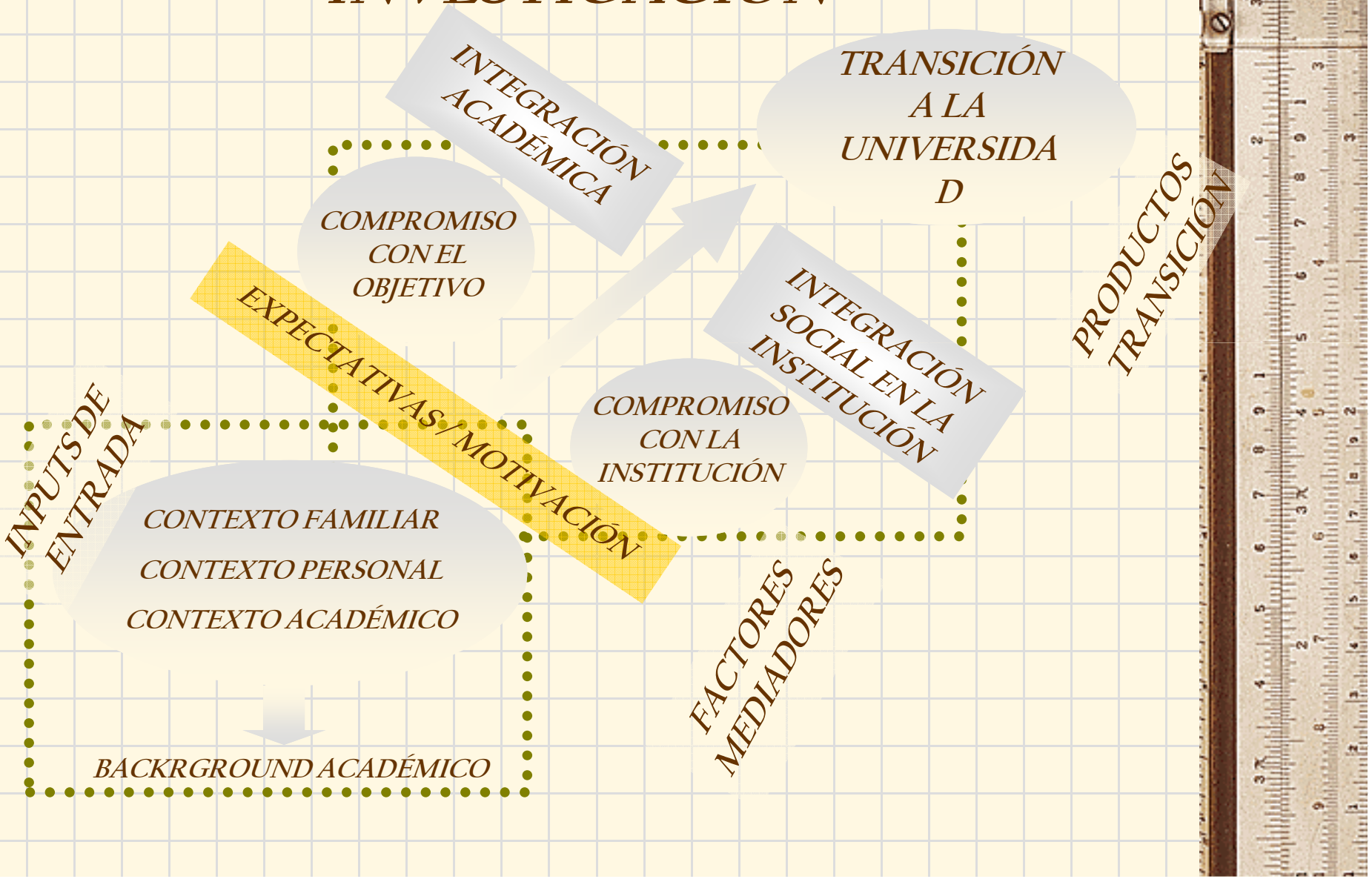
Objetivo de la investigación

- Analizar los factores personales e institucionales de persistencia académica al finalizar el primer año de universidad.

- Plantea un proceso de investigación longitudinal de carácter descriptivo y comprensivo, basado en la metodología multiestratégica y sistémica en dos niveles de análisis: Macro (conjunto de la promoción) y Micro (análisis de tipologías).

- La muestra está constituida por el total de una cohorte de estudiantes de la promoción 2010 de las titulaciones de Pedagogía y Administración de Empresas.

MODELO DE TRANSICIÓN DE LA INVESTIGACIÓN



1ª fase. Objetivo

- Analizar el proceso de integración inicial y la percepción de los estudiantes en relación a la primera etapa de transición académica (primer semestre de curso).

- Se parte del modelo de integración de Robert W. Lent, que analiza los factores sociocognitivos relacionados con la adaptación inicial a los estudios.

Dorio, I.; Figuera, P. y Torrado, M. (2011) **La integración académica y social en la universidad: la importancia de las redes sociales. Comunicación presentada en AIDIPE**

Qué se quiere demostrar

1.- Una persona se siente satisfecha en el ámbito académico de la universidad cuando se siente competente en lo que hace (expectativas de autoeficacia), sabe que va a obtener beneficios haciéndolo (expectativas de resultados), puede implicarse y participar en el desarrollo de la tarea (actividad hacia la meta) y cuenta con recursos y apoyos necesarios que le ayudaran a alcanzar sus objetivos (soporte).

Qué se quiere demostrar

2.- Los resultados sobre el proceso de integración inicial es diferente según el contexto académico.

UNA APROXIMACIÓN AL PROGRAMA DE MINERÍA DE DATOS

PASW Modeler

Modelo de clasificación/relación

- **Árboles de decisión QUEST**

Método de clasificación **binario** para generar árboles de decisión.

VI – continua VD categórica

- **Árbol de decisión CHAID**

Método de clasificación utilizando estadísticos del Chi-cuadrado para identificar las divisiones óptimas.

VI – continua/categórica VD continua/categórica

Modelo de clasificación/relación

- **Regresión Lineal**

Permite pronosticar valores en las VD a partir de la VI

Ajusta en una superficie o línea recta las discrepancias existentes entre los valores de salida reales y los pronosticados

VI – continua VD continua

- **Regresión logística**

Es similar a la regresión lineal pero parte de valores categóricos en la VD

Modelo de clasificación/relación

- **Análisis discriminante**

Permite identificar el peso de discriminación de las variables independientes

VI – continua VD categórica

- **Análisis factorial**

Permite reducir datos para reducir la complejidad de los mismos. Busca combinaciones lineales

Modelo de cribado

- **Detección de anomalías**

Identifica valores extraños o atípicos que no se ajustan a los patrones de los demás datos

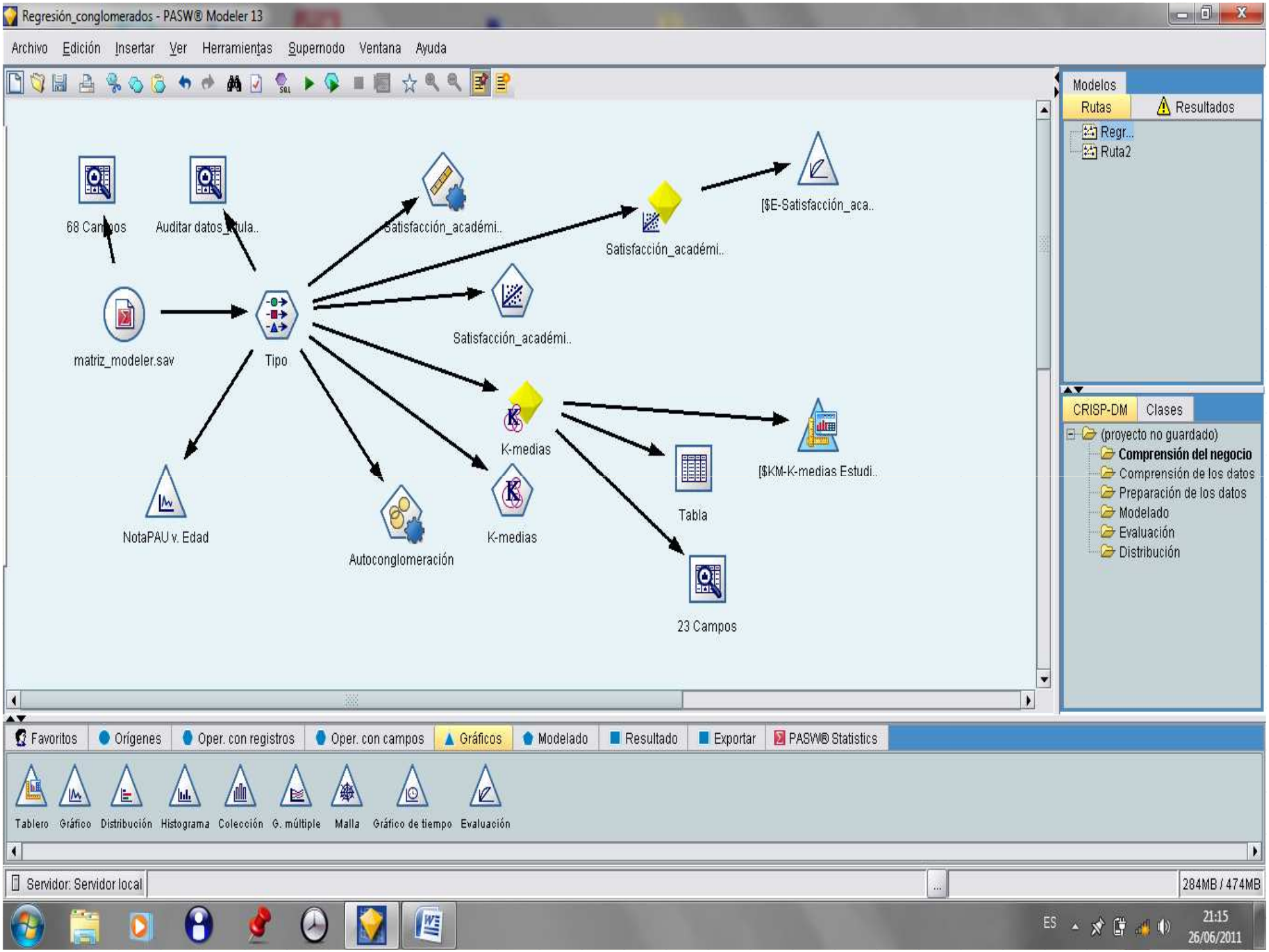
Modelo de segmentación

- **K-medias**

Agrupa conjunto de datos en conglomerados. Define un número fijo de grupos

- **Bietápico**

Es un método de dos pasos. Estima automáticamente el número óptimo de conglomerados para los datos


Regresión_conglomerados - PASW® Modeler 13
Archivo Edición Insertar Ver Herramientas Supernodo Ventana Ayuda
Modelos
Rutas
Resultados
Regr...
Ruta2
68 Campos
Auditar datos_Ruta..
Satisfacción_académi..
Satisfacción_académi..
[$E-Satisfacción_aca..
matriz_modeler.sav
Tipo
Satisfacción_académi..
K-medias
[$KM-K-medias Estudi..
Tabla
NotaPAU v. Edad
Autoconglomeración
K-medias
23 Campos
CRISP-DM
Clases
(proyecto no guardado)
Comprensión del negocio
Comprensión de los datos
Preparación de los datos
Modelado
Evaluación
Distribución
Favoritos
Orígenes
Oper. con registros
Oper. con campos
Gráficos
Modelado
Resultado
Exportar
PASW® Statistics
Tablero
Gráfico
Distribución
Histograma
Colección
G. múltiple
Malla
Gráfico de tiempo
Evaluación
Servidor: Servidor local
284MB / 474MB
ES
21:15
26/06/2011