TESIS DE GRADO
EN INGENIERÍA INDUSTRIAL

Detección de Patrones de Producción Educativa Basada en Minería de Datos

Autor: Santiago L. Molteni

Directores de Tesis:

Prof. M. Ing. Paola Britos

Prof. Dr. Enrique Sierra

2007

RESUMEN

El siguiente trabajo presenta cómo implementar Minería de Datos, por medio de la metodología CRISP-DM, aplicada al sector educativo argentino. Este proceso detecta comportamientos y patrones referentes a los datos analizados. Para esto, se selecciona información del sector educativo argentino proveniente de la Dirección Nacional de Información y Evaluación de la Calidad Educativa (DINIECE). El proyecto demuestra cómo la Minería de Datos es una técnica sólida para la detección de patrones. Aunque la metodología CRISP-DM se desarrolla como herramienta para compañías y grandes empresas, es ideal para el análisis de la información, sin importar el tamaño de la organización y de su base de conocimiento.

ABSTRACT

This project demonstrates how to implement Data Mining, by CRISP-DM methodology, on a real case based on the Argentinean education. This process detects behaviors and patterns concerning the analyzed data. For this, information from the Argentinean educational sector, spread by the National Direction of Educational Information and Quality Evaluation (DINIECE), was gathered. The project shows as well, how Data Mining is a solid technique for the detection of patterns. Although CRISP-DM methodology is developed as a tool for companies and big enterprises, it is really suitable for the analysis of information, without caring about the size of the organization and its data warehouse.

Índice

1. INTRODUCCIÓN

La investigación educativa constituye una fuente insustituible de conocimientos para la toma de decisiones en materia de política educativa [Tiramonti, 2003]. A pesar de esta evidencia, las diferentes lógicas de producción y uso del conocimiento en el ámbito académico y en el ámbito político han tendido a generar una tensión en el campo de la investigación educativa entre los criterios y procedimientos propios de la producción científica y las demandas de la política pública. Estas tensiones se han traducido en la ausencia de vinculaciones sistemáticas entre ambos espacios sociales. Adicionalmente, la ausencia de mecanismos lo suficientemente fluidos de acceso a la producción académica en su conjunto y la alta concentración geográfica de lo desarrollado en la zona metropolitana, debilita la posibilidad de socializar las producciones existentes.

Por su función específica y capacidad operativa, la Dirección Nacional de Información y Evaluación de la Calidad Educativa (DINIECE) del Ministerio de Educación, Ciencia y Tecnología de la Nación puede constituirse en un órgano capaz de promover la superación de dichas limitaciones, apoyando la conformación de instancias de fomento, intercambio y socialización de la investigación.

Para avanzar en la dirección deseada, la DINIECE ha iniciado nuevas líneas de acción. Se ha propuesto en primer lugar, el tránsito de la consolidación de los sistemas de información educativa a los que se abocó el Ministerio en la última década, hacia la generación de mecanismos de acceso público a las bases de datos en que dicha información se sistematiza: el Relevamiento Anual de Matrícula y Establecimientos; el Censo Nacional Docente; y el Operativo Nacional de Evaluación de la Calidad.

El Relevamiento Anual de Matrícula y Establecimientos ofrece información comparable a lo largo del tiempo sobre matrícula y cargos docentes de todos los tipos y niveles de educación de ambos sectores de gestión. Para esta base usuario se presenta la información correspondiente a educación común y artística.

Contemplando la problemática educativa que se menciona en la siguiente sección, se propone realizar una investigación valiéndose de los datos difundidos por la DINIECE. En dicho proyecto se plantea como objetivo identificar patrones en la producción educativa de los últimos años de todos los niveles educativos previos a los estudios universitarios. Dichos patrones estarán constituidos por variables educativas (matriculados, egresados, repetidos, modalidades), socio-económicas (sector, edades) y geográficas (ámbito).

Por otro lado, gran parte de las investigaciones difundidas en medios públicos, se desarrollan a partir de la estadística clásica. Esta técnica de análisis de datos plantea la resolución de un problema a partir del rechazo o la aprobación de una hipótesis. Sin embargo, existe la metodología de Minería de Datos como técnica complementariamente para el análisis de datos. Esta plantea requerimientos u objetivos a partir de información para obtener resultados por medio de algoritmos permitidos por dicha técnica.

Es por ello que como Proyecto Final se realiza una investigación educativa basada en los datos reales provistos por la DINIECE en sus relevamientos a nivel nacional. En dicho trabajo se aplica Minería de Datos para descubrir patrones de la producción educativa.

2. DESCRIPCIÓN DEL PROBLEMA

La Educación, como proyecto para la mejora de un país, es una de las inversiones a nivel macroeconómico con mayor tiempo de retorno [Filmus, 1999]. Según estadísticas de países primer mundistas, un proyecto educativo de largo plazo lleva, hasta obtener resultados significativos en una sociedad, 50 años [Tedesco, 2003]. En la Argentina, es objetivo decir que cada 10 años o menos hay cambios educativos que modifican planes, objetivos y proyectos planificados. Con lo cual, las medidas políticas en el marco de la educación no tienen una meta definida y ajena a visiones políticas, con objetivos de mejora continua de la formación de los estudiantes generación tras generación, independientemente de quien gobierne. Una de las principales falencias y problemáticas es la falta de investigación educativa continua, a fin de detectar variables contraproducentes y lograr consistentes planes de mejora.

A su vez, la escasez de investigación educativa utilizando herramientas y metodologías ajenas a la estadística clásica, incrementa la problemática planteada. En definitiva, no es un error aplicar la estadística clásica sobre bases de conocimientos, todo lo contrario. Sin embargo, la implementación en paralelo de otro tipo de técnicas complementarias brinda resultados de mayor objetivada y certeza para la educación.

Aprovechando la producción de información por parte de la DINIECE, sobre diferentes aspectos del sistema educativo nacional (con excepción del nivel universitario) [DINIECE, 2006], y considerando que el relevamiento de mayor amplitud es el Relevamiento Anual de Matrícula y Establecimientos (RA), se plantean algunos requisitos a analizar. Estos serán prioritarios en la determinación de los requerimientos específicos a resolver, los cuales se desarrollan en la sección correspondiente a los objetivos del proceso de minería. Entre estos se destacan:

- Detectar y fundamentar las diferencias y similitudes entre los establecimientos de los diferentes sectores de gestión (estatal y privado) de todos los niveles educativos.

- Determinar la producción educativa de los establecimientos de los distintos niveles, haciendo referencia a los sectores de gestión.

- Analizar las edades promedio de comienzo y finalización de la educación básica y terciaria.

- Estudiar la producción educativa de los establecimientos rurales, determinado los principales inconvenientes que atraviesa dicho ámbito y planteando sus respectivas causas.

- Analizar la producción educativa de las escuelas urbanas de los diferentes sectores de gestión. Identificar también su evolución en los últimos años, considerando la implementación de cambios en los planes educativos de los diferentes niveles.

- Analizar las características presupuestarias de las escuelas de los diversos ámbitos y sectores educativos.

- Determinar tendencias en cuanto a modalidades y carreras terciarias de los alumnos correspondientes a los niveles Medio/Polimodal y Superior No Universitario. Paralelamente, identificar la oferta de los establecimientos encargados de dichos niveles educativos.

3. SOLUCIÓN

Para lograr el objetivo planteado, se aplicará minería de datos a las bases de datos provistas por el MECyT, siguiendo la metodología CRISP DM aprendida en la materia electiva Sistemas de Administración de la Información [Britos *et al*, 2005]. Esta metodología posee seis (6) fases bien definidas que ayudan a la organización del proyecto, manteniendo un orden y brindando una comprensión de las tareas a realizar. Las fases son las siguientes:

Fase I: Comprensión de Negocio.
Fase II: Comprensión de los Datos.
Fase III: Preparación de los Datos.
Fase IV: Modelado.
Fase V: Evaluación.
Fase VI: Implementación.

Cabe aclarar que las fases enumeradas anteriormente son desarrolladas para la aplicación de minería de datos en un ámbito empresarial. Sin embargo, las actividades que constituyen cada una de las fases se ajustan sin inconvenientes a una investigación de estas características. Únicamente la implementación (Fase VI), no se llevará a cabo debido a que la formulación de acciones dirigidas al sector educativo sobrepasa los objetivos planteados.

3.1. Fase I: Comprensión del negocio

Esta fase requiere la valoración de varios factores, entre ellos se encuentra la comprensión de lo que es el negocio, su marco conceptual, sus objetivos, prioridades, dificultades, etc. A su vez, se debe exponer cual es la evaluación del equipo que implementa minería de datos basándose en la lectura del negocio. Esta evaluación debe informar, los recursos que forman parte del equipo, los recursos externos, las aplicaciones y bases utilizadas, los requerimientos y expectativas para alcanzar el objetivo del negocio y los riesgos que en su desarrollo pueden aparecer con su respectivo plan de contingencias.

Se deben enumerar, también, los objetivos del proceso de minería de datos. Estos objetivos son denominados requerimientos y se deben discutir con el cliente a fin de que los mismos respondan a una necesidad real del negocio. En este caso, los requerimientos serán discutidos con el tutor a cargo. Los factores claves de éxito también se deben exponer dentro de esta sección, señalando los recursos responsables para dicho logro. Finalmente se debe detallar un plan de proyecto, determinando la metodología de entregas respetando las fases de CRISP DM, con sus fechas u horas de trabajo asociadas.

Se recuerda que esta metodología fue diseñada principalmente para áreas de negocio donde la información es generada por la misma empresa o corporación. Sin embargo, debido a la naturaleza del proyecto, muchas secciones importantes que propone la metodología CRISP DM, quedan fuera de este análisis. La totalidad de las secciones de CRISP DM se pueden hallar en el anexo al final del documento.

Objetivos del negocio

El principal objetivo de esta sección es comprender completamente la perspectiva del negocio, es decir, lo que el cliente realmente quiere lograr. A menudo un cliente tiene muchos objetivos que compiten y deben ser equilibrados apropiadamente. Nuestra meta es encontrar factores importantes que pueden influir en el resultado del proyecto. Una posible consecuencia de descuidar esta fase es malgastar el tiempo y trabajo a responder preguntas que no se corresponden con el objetivo del negocio.

Escenario actual

La investigación educativa constituye una fuente insustituible de conocimientos para la toma de decisiones en materia de política educativa. A pesar de esta evidencia, las diferentes lógicas de producción y uso del conocimiento en el ámbito académico y en el

ámbito político han tendido a generar una tensión en el campo de la investigación educativa entre los criterios y procedimientos propios de la producción científica y las demandas de la política pública. Estas tensiones se han traducido en la ausencia de vinculaciones sistemáticas entre ambos espacios sociales. Adicionalmente, la ausencia de mecanismos lo suficientemente fluidos de acceso a la producción académica en su conjunto y la alta concentración geográfica de lo desarrollado en la zona metropolitana, debilita la posibilidad de socializar las producciones existentes.

Ambas situaciones son una fuente de limitaciones si se considera que el acceso de los decisores políticos a los resultados de la producción desarrollada en el circuito académico contribuye de manera decisiva a la toma de decisiones más informada; y que la socialización de las producciones entre diferentes equipos del propio campo académico, puede fortalecer la capacidad de los mismos, al tiempo que contribuir a delimitar fortalezas y áreas de vacancia en la producción investigativa del país.

Por su función específica y capacidad operativa, la Dirección Nacional de Información y Evaluación de la Calidad Educativa (DINIECE) del Ministerio de Educación, Ciencia y Tecnología de la Nación puede constituirse en un órgano capaz de promover la superación de dichas limitaciones, apoyando la conformación de instancias de fomento, intercambio y socialización de la investigación.

La DINIECE, es la unidad del Ministerio de Educación, Ciencia y Tecnología responsable del diseño y desarrollo de investigaciones vinculadas con la formulación de las políticas educativas, de las acciones de evaluación del sistema educativo nacional y del desarrollo y sustentabilidad del sistema federal de información educativa.

Su misión es brindar información oportuna y de calidad para la planificación, gestión y evaluación de la política educativa. Para ello produce, analiza y difunde información sobre diferentes aspectos del sistema educativo nacional, con excepción del nivel universitario, y desarrolla investigaciones orientadas a mejorar su calidad y equidad. El propósito es contribuir a una política educativa que promueva la igualdad en el acceso, permanencia y egreso, así como un alto rendimiento académico en sus distintos niveles.

Para avanzar en la dirección deseada, la DINIECE ha iniciado nuevas líneas de acción. Se ha propuesto en primer lugar, el tránsito de la consolidación de los sistemas de información educativa a los que se abocó el Ministerio en la última década, hacia la generación de mecanismos de acceso público a las bases de datos en que dicha información se sistematiza: el Relevamiento Anual de Matrícula y Establecimientos; el Censo Nacional Docente; y el Operativo Nacional de Evaluación de la Calidad. Esta medida constituye una vía para dar

cumplimiento al Decreto 1172/03, al garantizar el acceso a la información pública en el área educativa. En segundo lugar, espera promover el intercambio y la producción conjunta entre equipos de investigación académica y equipos técnicos de Ministerios y Secretarías provinciales y/o municipales de Educación del país.

Objetivos de negocio

La DINIECE, con la producción, análisis y exposición de los datos sobre los diferentes aspectos del sistema educativo nacional, tiene como principales objetivos:

- Contribuir al desarrollo y fortalecimiento del intercambio y cooperación entre las instituciones académicas y de gestión educativa.

- Fomentar el uso de los datos producidos regularmente por la gestión educativa y fortalecer la capacidad de análisis de esta información.

- Desarrollar conocimiento significativo en áreas relevantes del quehacer educativo actual.

Evaluación de la situación

En esta sección se realiza un estudio mas detallado sobre los recursos, supuestos y otros factores que deben ser considerados para determinar los objetivos del análisis de datos y el plan de proyecto.

Inventario de Recursos

La magnitud del proyecto lleva a tener un inventario de recursos reducido. El inventario se divide según tres tipos de recursos que son, los recursos humanos, la aplicación o software de modelización y las bases de datos.

Los recursos humanos que participan de este proyecto son los que se muestran en la tabla 1.

RR. HH.		
Nombre del puesto	Funciones principales	Disponibilidad de tiempo
Tutor ***Prof. M. Ing. Paola Britos***	• Brindar soporte sobre la metodología de CRISP DM y de otros temas relacionados con el proyecto. • Evaluar los resultados obtenidos.	Part - time
Tutor – Experto en educación ***Prof. Dr. Enrique Sierra***	• Evaluar los resultados obtenidos basándose en el marco educativo actual.	Part - time
Autor ***Santiago Luís Molteni***	• Implementar la metodología CRISP DM. • Presentar resultados a los tutores. • Preparar la documentación del total del proyecto.	Full time

Tabla 1. Recursos humanos

La aplicación encargada del modelado y de la ejecución de los algoritmos seleccionados proporciona interfaz visual al mundo de los datos, los estadísticos y los algoritmos complejos. Cada proceso se representa con un icono, o nodo, que se conecta para formar una ruta que representa el flujo de datos a través de una serie de procesos.

Como aplicación de minería de datos ofrece un método estratégico para encontrar relaciones útiles entre grandes conjuntos de datos. Al contrario que los métodos estadísticos más tradicionales, no es necesario saber lo que se está buscando al comenzar. Puede explorar los datos, mediante el ajuste de diferentes modelos y la investigación de diferentes relaciones, hasta que encuentre la información que resulte útil. De todas formas, en este proyecto se define a priori cuales son los requerimientos que justifican su utilización.

En cuanto a las bases de datos, estas son las que se encuentran difundidas públicamente por la DINIECE. Las mismas se encuentran comprimidas en la página Web de la DINIECE y son de uso público. Los archivos dentro de esta carpeta tiene una extensión .db (.database), con lo cual se la transforma en archivos .xls.

Requisitos

Para poder alcanzar los objetivos establecidos se deben cumplir con determinados requisitos:

- La información proveniente de la DINIECE debe ser confiable e inalterable. Solo puede ingresar este tipo de información los empleados que allí trabaja, de acuerdo a las funciones y roles a los que fueron designados en cada RA anual. Para ello se debe contar con personal idóneo. Este requerimiento es obligatorio.

- El conocimiento del experto en educación es un factor importante que aporta a este proyecto un marco de objetividad en la evaluación de los resultados obtenidos de la aplicación de la metodología CRISP DM. Este requerimiento es obligatorio.

- Se deberá contar con un software con gran capacidad de procesamiento y flexibilidad para lograr un análisis consistente y eficaz de los distintos niveles que se observan en las bases de datos. Este requerimiento es deseable.

Expectativas

En cuanto a las expectativas que genera este proyecto son:

- Brindar información objetiva al Instituto Tecnológico de Buenos Aires (ITBA) sobre ámbito educativo nacional, basada en datos recolectados por la DINIECE.

- Lograr presentar dicha investigación fuera del marco del proyecto final de ingeniería del ITBA, aportando mi conocimiento y análisis a otros sectores de investigación de materia educativa.

- Ampliar mi conocimiento sobre la educación en nuestro país. Comprender como esta compuesta la producción educativa hoy en día, cuales son los niveles que requieren mayor atención y que clase de acciones son requerida para su mejora.

- Tener el know-how del manejo de la aplicación, que amplía el conocimiento sobre las herramientas que sustentan a la metodología CRISP DM. Poder aplicar dicho conocimiento en otros rubros e instancias de mi carrera profesional.

Riesgos y Contingencias

A continuación se listan los riesgos que pueden traer inconvenientes o que a causa de ellos el proyecto pueda fallar. También se listan, junto a cada riesgo, el objetivo que se desea cumplir y su correspondiente plan de contingencia que será llevado a cabo en caso de que ocurra.

Objetivo	Riesgo	Plan de Contingencia
Obtener resultados consistentes con los requerimientos planteados.	El formato de las tablas de las bases sea de solo lectura (Read only), sin poder modificar los campos o agregar nuevos.	Exportar las tablas a nuevas planillas de Excel. En último caso, trabajar con la información existente sesgando los requerimientos.
Obtener conclusiones señalando un posible accionar en el ámbito educativo.	Ausencia del análisis de los resultados obtenidos por parte del experto en educación.	Las conclusiones se obtendrán con la ayuda del el tutor no experto en el ámbito educativo, basándonos en comportamientos o patrones que resulten significativos y representativos.
Utilización de todos los datos provistos por las bases.	Datos errados, ausentes, inconsistentes.	Estos datos serán eliminados y se reportará el error cometido por la DINIECE en carga de datos, dentro del requerimiento que impacte.
Utilizar el software seleccionado para la modelación y ejecución	Complicaciones de licencia u otra clase de problemas relacionados con la aplicación utilizada para el DM.	Utilización de las macros provistas por la cátedra de Sistemas de Administración de la Información para realizar el modelado y ejecución del mismo.

Tabla 2. Objetivos, riesgos y plan de contingencias.

Objetivos del proceso de minería de datos

En esta tarea se definen las metas a alcanzar con el desarrollo del proyecto de Exploración de Información. Considerando la amplia información que poseen las bases difundidas por la DINIECE, y que las mismas se agrupan según los niveles educativos, se plantean requerimientos para cada uno de estos grupos. A su vez, se enumeran algunos factores críticos que influyen altamente en la concreción de los objetivos.

Objetivos de minería de datos

Los rendimientos intencionales del proyecto de minería de datos que habilitan el logro de los objetivos del negocio, se obtienen de la resolución de los posteriores requerimientos.

Dado que los objetos del negocio no poseen otro fin que el análisis de las tablas difundidas para desarrollar conocimiento significativo sobre las áreas de estudio, los objetivos planteados para este proyecto se obtienen a partir de la información que se visualizó en el primer contacto con la base y de los requisitos expuestos en la descripción del problema. Con lo cual, a continuación se enuncian siete (7) requerimientos, haciendo distinción del nivel en que repercuten.

Requerimiento # 1

"Para la educación común, nivel Inicial, detectar cuales son las características mas distintivas entre las escuelas privadas o estatales, que se ubican en ámbitos tanto rurales como urbanos. Con lo cual, se debe encontrar características principales de estos cuatro grandes grupos que son las escuelas Estatal – Urbano, Privada – Urbano. Estatal – Rural, Privada - Rural. Además, de las diferencias y similitudes entre los grupos, investigar cual de ellos prevalece en cantidad de establecimiento y producción educativa a nivel nacional."

Requerimiento # 2

"Para la educación común, nivel Primario/EGB, detectar cuales son las características mas distintivas entre las escuelas privadas o estatales, que se ubican en ámbitos tanto rurales como urbanos. Con lo cual, se debe encontrar características principales de estos cuatro grandes grupos que son las escuelas Estatal - Urbano, Privada - Urbano. Estatal - Rural, Privada - Rural. Además, de las diferencias y similitudes entre los grupos, investigar cual de ellos prevalece en cantidad de establecimiento y producción educativa a nivel nacional."

Requerimiento # 3

"Investigar como es el comportamiento de la planta funcional en el nivel Primario/EGB para cada establecimiento educativo de los distintos ámbitos y sectores. Establecer cual de ellos prevalece en la mayoría de los establecimientos."

Requerimiento # 4

"Para la educación común, nivel Medio/Polimodal, detectar cuales son las modalidades con mayor cantidad de matriculados y egresados en el año de estudio. Determinar también el ámbito y el sector para dichas modalidades."

Requerimiento # 5

"Investigar como es el comportamiento de la planta funcional en el nivel Medio/Polimodal para cada establecimiento educativo en los distintos ámbitos y sectores. Establecer cual de ellos prevalece en la mayoría de los establecimientos."

Requerimiento # 6

"Para la educación común, nivel Superior No Universitario, detectar cual es el tipo de formación con mayor cantidad de matriculados y egresados en el año de estudio. Determinar también el sector predominante para los distintos tipos de formación."

Requerimiento # 7

"*Investigar como es el comportamiento de la planta funcional en el nivel Superior No Universitario para cada los establecimientos educativos en los distintos ámbitos y sectores. Establecer cual de ellos prevalece en la mayoría de los establecimientos y cual es su ubicación geográfica. También estimar las edades promedio de los estudiantes del nivel en cuestión."*

Criterios de éxito en el proceso de minería de datos

Aquí se definen los criterios para un resultado exitoso en términos técnicos. Si estos factores críticos de éxitos (FCE) aparecen durante el proyecto, aumenta la probabilidad de conseguir los requerimientos planteados anteriormente.

Objetivo	FCE	Recurso Afectado
Utilización de todos los datos provistos por las bases.	Contar con información completa y confiable.	DINIECE
Utilizar el software seleccionado para la modelación y ejecución.	La facilidad y la flexibilidad del la aplicación sea la comentada por su tutorial.	Autor
Obtener conclusiones señalando un posible accionar en el ámbito educativo.	Análisis de los resultados obtenidos por parte del experto en educación.	Tutor – Experto en educación.

Tabla 3. Factores críticos de éxitos

Plan del proyecto

Este proyecto se divide en cuatro (4) entregas respetando la metodología CRISP DM.

La primera entrega, como se puede observar en el figura 1, se correspondo con la desarrollada en esta fase. Se puede detectar que todas sus secciones principales fueron completadas siguiendo la metodología CRISP DM. En resumen, como se pudo observar, esta fase tiene como objetivo una descripción completa del marco en el que se realiza el proyecto.

Para la segunda entrega se informa como se realiza el proceso de comprensión y preparación de los datos para su futura aplicación en el modelo. En estas fases se fijan que tipo de variables son necesarias para la investigación, cuales de ellas tienen datos coherentes, y en caso de no ser coherentes, como se realiza su transformación y limpieza para su utilización. Tener presente que esta entrega abarca dos fases del la metodología.

En la tercera entrega, se documentan las técnicas utilizadas en el modelado, el proceso de construcción del modelo junto con el setéo de los parámetros, y su evaluación previa a la obtención de resultados.

Finalmente en la última entrega, se informan los resultados más relevantes obtenidos del modelo. Con los conocimientos del experto en educación se analizan dichos resultados con el propósito de obtener reglas y patrones generales sobre educativo nacional. La dinámica del proceso descrito puede visualizarse en la figura 1.

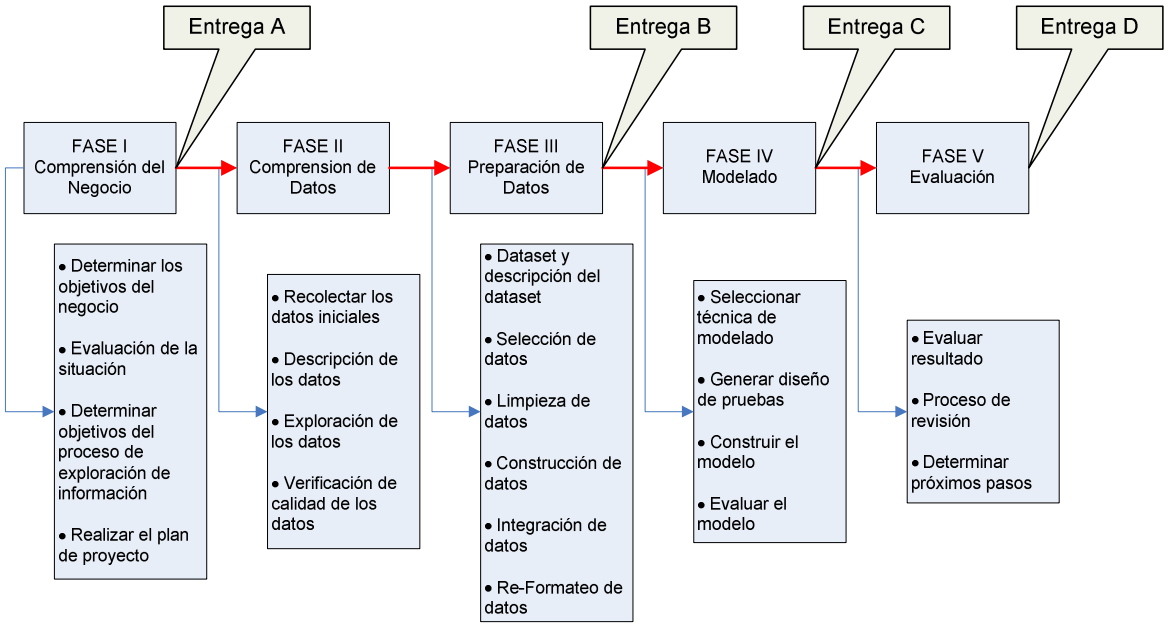

Figura 1. Dinámica del Proceso de Minería de Datos - CRISP DM.

Para la realización de cada entrega se estima un período de ochenta (80) horas, totalizando unas trescientas veinte (320) horas para completar el proyecto final. Una vez recibida la "Entrega A" del proyecto, las restantes tres se entregarán una por semana. Con lo cual, al término de las tres semanas posteriores a partir de la primera entrega, el proyecto estará listo para su corrección.

3.2. Fase II: Comprensión de los datos

Esta fase involucra las tareas que hacen a la comprensión de los datos, para la cual, los datos se deben describir, coleccionar, organizar, verificar y limpiar antes de realizar el análisis de los mismos. Esto puede consumir mucho tiempo y es crítico para el proyecto de exploración de información.

La Fase II se muestra como una fase genérica para todos los requerimientos y objetivos planteados sobre la educación común nacional.

Recolección de los datos iniciales

La tarea en esta sección es adquirir o acceder a los datos definidos en el plan de recursos del proyecto. Esta recolección inicial incluye datos que se cargan, si es necesario, para la comprensión de los mismos. Es importante destacar que si los datos se adquieren de múltiples fuentes o bases, la integración de los mismos es una tarea que se realiza en la posterior fase de preparación de datos.

El dataset es la base de datos correspondiente al Relevamiento Anual del Año 2005 (RA_2005). El mismo se encuentra comprimido en la página Web de la DINIECE y es de uso público. Los archivos dentro de esta carpeta tiene una extensión .db (.database), con lo cual se la transforma en archivos .xls. Generalizar el tipo de archivo que se utiliza para las bases de datos facilita el armado de los modelos tanto para las entradas como para sus salidas.

De más esta decir que la gran ventaja de este proyecto es a la alta disponibilidad de los datos para el análisis. No existen patentes ni restricciones en la utilización de los mismos, ni tampoco es necesaria la integración de los datos con otras bases. Las encuestas realizadas por el organismo se adjuntan con la información complementaria, con lo cual al compararlas con las bases de datos, no se detectan inconvenientes de ausencia de información.

Además de las encuestas, los archivos complementarios contienen un glosario reducido, el cual fue ampliado para este proyecto y se encuentra en el anexo. A su vez, se encuentra un documento que indica el alcance y las limitaciones de cada base expuesta, ya que como se comentó anteriormente, existen otras encuestas realizadas para otros fines que no forman parte del proyecto.

Descripción de los datos

Una de las dificultades que se presentan en el proyecto es la forma en que fueron volcados los datos a las bases luego de los RA. Estos se encuentran en diferentes tablas formadas por campos. Algunos de ellos están presentes en mas de una tabla, facilitando la relación entre estas. El nombre de cada tabla se debe al tipo de información que se guarda en ellas.

Para conseguir resultados que cumplan con los requerimientos, es necesaria la utilización de más de una tabla a la vez. A su vez, hay tablas genéricas que involucran a todos los niveles, con lo cual, se debe tener un amplio conocimiento de sus contenidos a fin de decidir correctamente cuales participan de la resolución del requerimiento a analizar. Afortunadamente, existe un campo que se repite en todas las tablas (ID_RA) y es por medio de él que se unen sus contenidos para lograr los resultados deseados.

Como se ha mencionado, resulta de vital importancia comprender los campos de las tablas y los posibles valores que estos contienen. A su vez, la comprensión lleva a combinarlos y relacionarlos con otros, obteniendo campos calculados o indirectos que ayudan a extraer resultados mas claros a la hora de exponerlos. Esta tarea es una de las bases fundamentales de un proyecto de minería de datos, ya que una comprensión equivocada desde su concepción, lleva indudablemente a resultados errados. No solo eso, sino que al no comprender la totalidad de los campos, algunos de gran importancia pueden quedar fuera del análisis, disminuyendo las posibilidades de alcanzar los objetivos planteados.

Ayudados por la información que presenta la DINIECE, se pudo distinguir y comprender la totalidad de los datos que se encuentran en las tablas del RA del Año 2005. El relevamiento tiene un total de once (11) tablas que abarcan todos los niveles de la educación común de nuestro país. Estas tablas son:

1. CAR2005: Presenta información sobre la planta funcional de los establecimientos en cada uno de los niveles. También muestra numéricamente los cargos y las horas desarrolladas por cada uno de ellos.
2. EDI2005: Muestra la cantidad de alumnos que se ubican en los diferentes rangos de edades que se señalan en cada campo, para todos los establecimientos del nivel Inicial.
3. EDMP2005: Muestra la cantidad de alumnos que se ubican en los diferentes rangos de edades que se señalan en cada campo, para todos los establecimientos del nivel Medio/Polimodal. Los datos se encuentran divididos por año o grado escolar correspondiente.

4. EDPE2005: Muestra la cantidad de alumnos que se ubican en los diferentes rangos de edades que se señalan en cada campo, para todos los establecimientos del nivel Primario/EGB. Los datos se encuentran divididos por año o grado escolar correspondiente.
5. EDS2005: Muestra la cantidad de alumnos que se ubican en los diferentes rangos de edades que se señalan en cada campo, para todos los establecimientos del nivel Superior No Universitario.
6. EMP2005: Presenta la cantidad de alumnos egresados en cada establecimiento, según la modalidad elegida en el nivel Medio/Polimodal.
7. ESNU2005: Presenta la cantidad de alumnos egresados en cada establecimiento, según la carrera elegida en el nivel Superior No Universitario. También muestra el tipo de formación del establecimiento en cuestión.
8. MAE2005: Presenta información sobre todos los establecimientos nacionales. Muestra los niveles que abarca, el ámbito y sector al que pertenecen, la provincia y el departamento donde se ubican
9. MAT2005: Presenta la cantidad de alumnos matriculados y repetidos de cada año, para cada establecimiento, según todos los niveles educativos. También muestra el tipo de sección del establecimiento en cuestión.
10. MMP2005: Presenta la cantidad de alumnos matriculados de cada año para cada establecimiento, según la modalidad elegida en el nivel Medio/Polimodal.
11. MSNU2005: Presenta la cantidad de alumnos matriculados de cada año para cada establecimiento, según la carrera elegida en el nivel Superior No Universitario. También muestra el tipo de formación del establecimiento en cuestión.

Se decide tomar el año 2005, ya que es el último relevamiento liberado por este organismo. Sin embargo, también existen relevamientos realizados añas anteriores, con las mismas tablas y campos presentes en este análisis.

De las once (11) tablas señaladas no se utilizan dos (2) de ellas, que son EDPE2005 y EDMP2005. Esto se debe a que los datos de los registros están divididos por año o grado escolar, haciendo imposible combinar estos datos con otras tablas. Sí, en cambio, se aceptan las tablas EDI2005 y EDS2005, ya que estas no están categorizadas por año o grado escolar, sino que es un único registro por establecimiento educativo.

Debido a la gran cantidad de campos que se obtienen del total de las tablas, y adelantando que se realizan modificaciones en la mayoría de ellos, se decide que la explicación de cada uno de ellos se realiza en la FASE III del proyecto. Por el momento, es importante tener presente la información que brinda cada tabla, basándose en la breve introducción expuesta.

Para más información, las características de los campos de las tablas de la DINIECE, se muestran en la tabla 2 del anexo.

Exploración de los datos

En esta sección se realiza un análisis de los datos a los que se aplican los modelos que contempla la Minería de Datos. Dicha investigación preliminar es de utilidad para conocer cuales son los datos que se utilizan, y para mostrar una información adicional sobre estos, permitiendo así delimitar el escenario donde se desarrollan las futuras tendencias y patrones.

Para comenzar el análisis de los datos difundidos, se expone la siguiente tabla. En ella se observa la cantidad de alumnos por nivel/ciclo de enseñanza según división político-territorial. Como se comenta anteriormente los datos que se toman para este análisis son los que pertenecen a la educación común.

		Nivel / Ciclo de enseñanza						
División	Total	Inicial	EGB 1 y 2 / Primario		EGB 3	Polimodal / Medio		Superior No
Político-Territorial			EGB 1 y 2	Primario		Polimodal	Medio	Universitario
Total País	**9.890.037**	**1.324.529**	**4.526.121**	**157.842**	**1.826.419**	**1.143.672**	**402.320**	**509.134**
Buenos Aires	3.580.005	577.379	1.577.421	-	756.708	546.440	758	121.299
Partidos del Conurbano	2.179.310	329.222	973.595	-	462.935	349.542	86	63.930
Buenos Aires Resto	1.400.695	248.157	603.826	-	293.773	196.898	672	57.369
Catamarca	109.000	9.636	53.772	-	23.239	15.645	399	6.309
Chaco	322.848	32.121	169.605	8.778	43.756	26.826	25.068	16.694
Chubut	118.347	14.943	56.208	-	26.354	16.653	-	4.189
Ciudad de Buenos Aires	634.346	90.281	226.326	36.523	271	353	192.013	88.579
Córdoba	811.148	101.269	361.642	-	180.697	106.012	681	60.847
Corrientes	293.942	33.303	158.364	1.668	50.130	24.077	12.528	13.872
Entre Ríos	317.554	38.734	156.638	-	67.998	20.290	21.804	12.090
Formosa	170.482	16.422	93.186	-	34.946	18.043	3.844	4.041
Jujuy	206.586	22.317	95.175	11.072	11.153	7.318	44.199	15.352
La Pampa	74.111	6.997	35.579	-	17.938	11.383	38	2.176
La Rioja	95.076	12.744	46.810	-	19.949	11.221	-	4.352
Mendoza	424.891	43.899	200.772	-	95.621	57.489	14	27.096
Misiones	314.390	31.705	178.872	-	61.037	30.039	62	12.675
Neuquén	159.183	18.287	459	82.466	421	216	46.456	10.878
Río Negro	169.842	20.683	80.776	12.373	422	-	47.499	8.089
Salta	360.704	34.744	177.668	-	77.623	46.844	1.084	22.741
San Juan	164.403	18.684	85.931	-	34.216	20.832	47	4.693
San Luis	101.286	13.531	52.924	-	21.456	12.048	-	1.327
Santa Cruz	63.203	9.012	31.112	-	13.400	6.697	1.954	1.028
Santa Fe	742.119	101.612	342.312	-	161.071	95.892	870	40.362
Santiago del Estero	247.523	31.722	138.346	4.962	38.760	22.952	177	10.604
Tierra del Fuego	37.012	5.438	14.912	-	8.753	5.227	-	2.682
Tucumán	372.036	39.066	191.311	-	80.500	41.175	2.825	17.159

Tabla 4. Alumnos por tipo de educación según división político- territorial.

Observando los totales, se puede distinguir claramente que las principales provincias que desarrollan una amplia actividad educativa son Buenos Aires, como principal exponente, seguida por Córdoba, Santa Fe y la Ciudad de Buenos Aires. Esta tabla también es muy representativa para determinar la evolución del alumnado a lo largo del ciclo educativo.

Dado que el nivel Superior No Universitario no es obligatorio y puede ser sustituido por una carrera universitaria, puede no ser muy significativo. Sin embargo los otros niveles son obligatorios, y además deben tener precedencia con niveles posteriores. Del nivel Inicial al Primario/EGB, se comprueba un gran salto en cada una de las provincias. Luego este número cae fuertemente al ingresar en los estudios medios.

Si bien es importante conocer cual es la producción educativa del año 2005, este análisis tiene como principal protagonista a los establecimientos educativos y es en ellos donde se debe poner el foco de atención. A continuación se presenta una tabla similar a la anterior, que muestra la distribución de las unidades educativas por nivel/ciclo de enseñanza según división político – territorial.

División Político-Territorial	Nivel / Ciclo de enseñanza								
	Inicial	EGB 1 y 2 / Primario			EGB 3	Polimodal / Medio			Superior no Universitario
		EGB 1y2	EGB 1 y 2 / Primario	Primario		Polimodal	Medio / Polimodal	Medio	
Total País	**16.298**	**18.329**	**3.514**	**353**	**15.062**	**5.663**	**292**	**1.063**	**1.870**
Buenos Aires	4.374	5.952	-	-	5.532	2.095	119	4	495
Partidos del Conurbano	2.254	2.450	-	-	2.453	1.141	56	1	205
Buenos Aires Resto	2.120	3.502	-	-	3.079	954	63	3	290
Catamarca	44	454	-	-	424	92	-	1	17
Chaco	383	283	722	-	220	105	33	37	42
Chubut	189	238	-	-	204	93	2	-	24
Ciudad de Buenos Aires	678	24	873	2	1	-	1	477	256
Córdoba	1.755	2.146	-	-	745	743	3	-	199
Corrientes	848	576	357	1	203	121	26	38	45
Entre Ríos	1.174	1.312	-	-	519	180	5	163	83
Formosa	103	493	-	-	462	72	24	1	32
Jujuy	439	91	324	-	84	34	16	67	20
La Pampa	85	210	-	-	108	83	1	-	18
La Rioja	190	378	-	-	313	77	-	1	35
Mendoza	792	826	-	-	1.082	311	1	-	74
Misiones	769	909	-	-	681	196	2	-	63
Neuquén	270	3	-	349	5	4	3	104	28
Río Negro	262	21	361	-	9	-	-	144	31
Salta	804	808	-	-	917	238	4	2	60
San Juan	343	396	-	-	279	112	2	1	32
San Luis	190	358	-	-	184	84	-	-	6
Santa Cruz	74	101	-	-	93	39	-	9	2
Santa Fe	1.302	1.594	-	1	2.142	614	14	6	179
Santiago del Estero	500	362	875	-	256	144	10	-	50
Tierra del Fuego	44	50	-	-	33	26	-	-	7
Tucumán	686	744	2	-	566	200	26	8	72

Tabla 5. Unidades educativas por tipo de educación según división político- territorial.

Como es de esperar la cantidad de establecimientos que existen a nivel nacional aumentan del nivel Inicial hacia el Primario, y es en este último donde se observa la mayor cantidad de datos. Luego descienden hacia los demás niveles, evidenciando un comportamiento que acompaña a la tendencia de alumnos. Sin embargo, se puede observar que el ranking de mayor cantidad de establecimientos por provincia, no es exactamente el mismo que por cantidad de alumnos. Las provincias con mayor cantidad de unidades educativas son Buenos Aires, Córdoba, Santa Fe y Entre Ríos. En un quinto puesto se encuentra la Capital, mostrando que la densidad de estudiantes en establecimientos es la mayor entre todas las provincias, mientras que la de Entre Ríos es la menor.

El porcentaje de mujeres en los niveles Inicial, Primario /EGB 1, 2, 3 y Polimodal/Medio, no merece una mención especial, ya que presenta un comportamiento similar en todos los años, grados o ciclos correspondientes. Este valor se encuentra oscilando el 50 % intuitivo.

Nivel Inicial	Total		Ciclo					
			1°		2°		3°	
	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres
Total País	**1.324.529**	**49,6**	**206.903**	**50,1**	**414.737**	**49,7**	**702.889**	**49,3**

Tabla 6. Alumnos y porcentaje de mujeres del nivel de enseñanza Inicial.

Nivel Primario / EGB 1 y 2	Total	Ciclo y año de estudio					
		1er. Ciclo			2° Ciclo		
		1°	2°	3°	4°	5°	6°
Alumnos	**4.597.404**	**824.542**	**781.103**	**776.954**	**765.551**	**741.013**	**708.241**
% mujeres	**48,9**	**48,2**	**48,5**	**48,9**	**49**	**49,3**	**49,7**

Tabla 7. Alumnos y porcentaje de mujeres del nivel de enseñanza Primario / EGB 1 y 2.

Nivel EGB 3	Año de estudio							
	Total		7°		8°		9°	
	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres
Total País	**2.098.453**	**50,5**	**727.403**	**49,6**	**744.970**	**50,2**	**626.080**	**51,9**

Tabla 8. Alumnos y porcentaje de mujeres del nivel de enseñanza EGB 3.

Nivel Polimodal / Medio		Polimodal / Medio								Nivel Medio
		Año de estudio								
		1°		2°		3°		4°		Año de Estudio
	Total	1° Polimodal	3° Medio	2° Polimodal	4° Medio	3° Polimodal	5° Medio	4° Polimodal	6° Medio	7°
Alumnos	**1.360.174**	**482.200**	**81.788**	**361.757**	**67.926**	**297.522**	**55.764**	**2.193**	**10.965**	**59**
% mujeres	**53,2**	**52,4**	**49,5**	**54,4**	**50,8**	**56,2**	**52,3**	**26,2**	**23,8**	**30,5**

Tabla 9. Alumnos y porcentaje de mujeres del nivel de enseñanza Polimodal/Medio.

Ahora bien, cuando se observa al nivel Superior No Universitario este comportamiento se modifica. La causa lógica de este cambio es la división de los estudios en distintas carreras. Como se puede observar en la siguiente tabla, las carreras relacionadas con la tecnología y ciencias aplicadas son elegidas mayormente por hombres. Mientras que las ciencias humanas, de salud, básicas y sociales son frecuentadas por mujeres. A nivel general, existen mayor cantidad de mujeres que comienzan una carrera terciaria. Se estima que se debe a que los hombres optan por seguir una carrera universitaria o a trabajar una vez terminado el ciclo Medio/Polimodal.

Nivel SNU	Total		Ciencias Aplicadas y Tecnología		Ciencias Básicas		Ciencias de la Salud		Ciencias Humanas		Ciencias Sociales	
	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres	Alumnos	% mujeres
Total País	**488.896**	**68,5**	**65.674**	**44,4**	**34.610**	**70,2**	**34.730**	**75,1**	**223.757**	**79,6**	**130.125**	**59,6**

Tabla 10. Alumnos y porcentaje de mujeres del nivel de enseñanza Superior no Universitario.

Si se habla sobre la doble escolaridad en el nivel Inicial, se observa que no tiene repercusión en el ámbito nacional, ya que la cantidad de alumnos que asisten a escuelas con doble jornada es igual a 29.184, lo que corresponde a un 2,2% del total de los alumnos de

este nivel. Las provincias donde existen mayor cantidad de establecimientos y, por ende, alumnos que participan de esta modalidad son la Ciudad de Buenos Aires y Buenos Aires, como se puede apreciar en el siguiente gráfico.

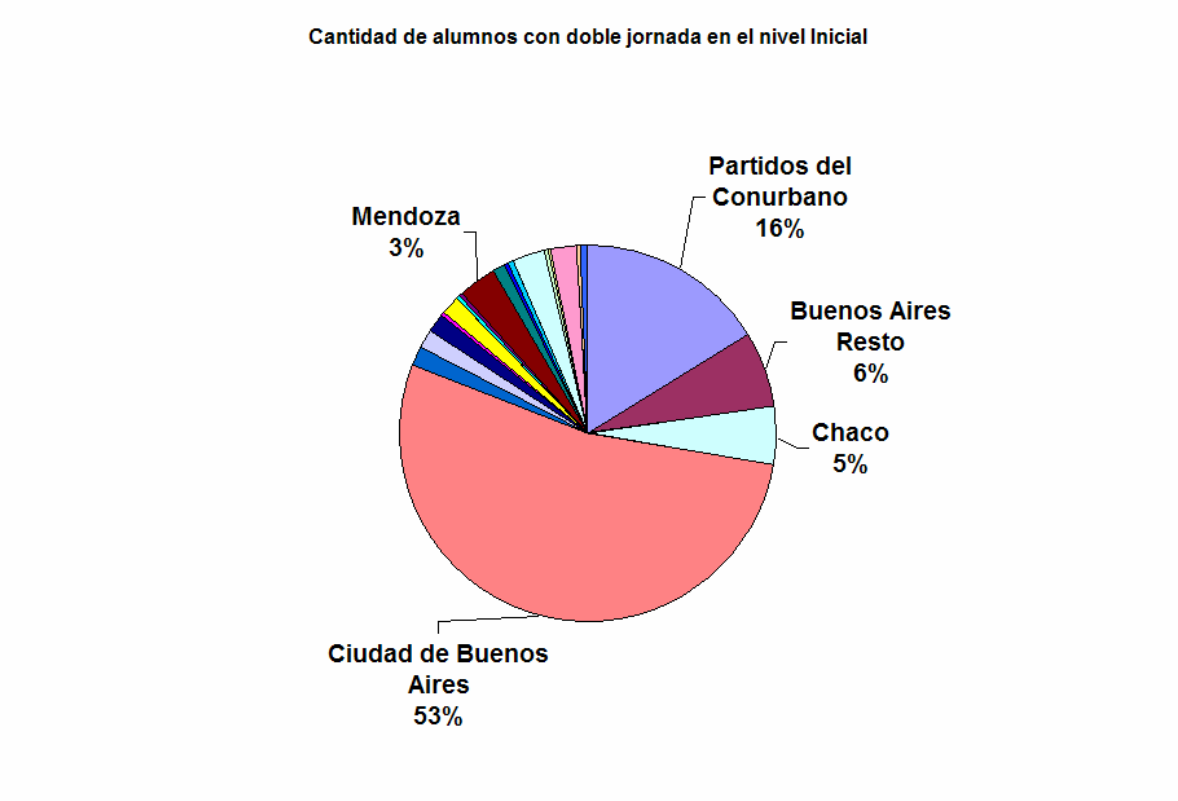

Figura 2. Cantidad de alumnos con doble jornada en el nivel Inicial.

En el caso del nivel Primario/EGB 1 y 2, la cantidad de alumnos con doble escolaridad aumenta significativamente con respecto a las del nivel Inicial. El total de alumnos que forman parte de esta modalidad es igual a 252.376, esto representa un 5,5 % del total de alumnos a nivel nacional. Las provincias con mayor cantidad de alumnos dentro de esta modalidad son las nombradas anteriormente, junto con Tucumán.

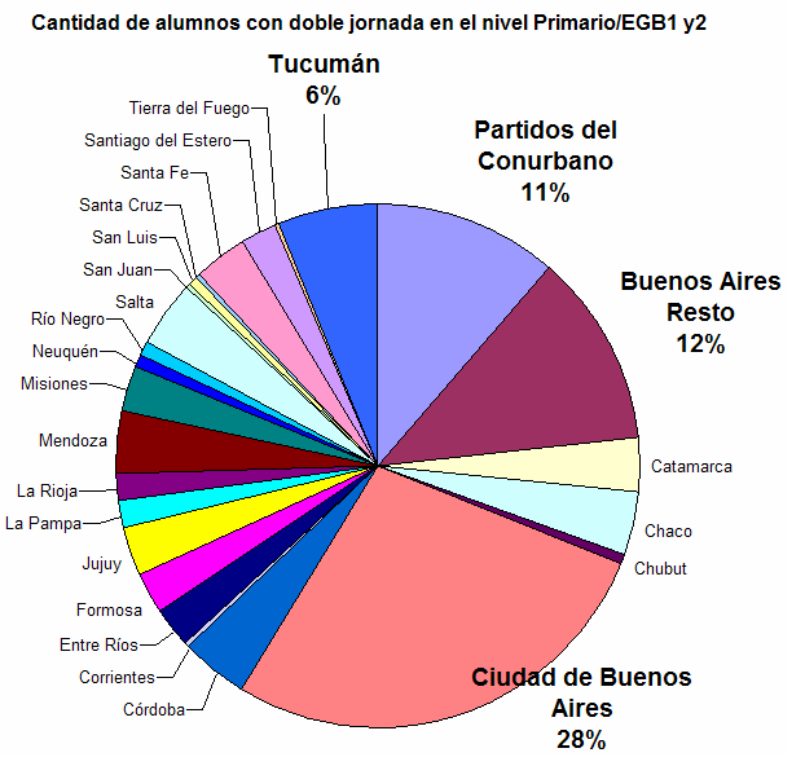

Figura 3. Cantidad de alumnos con doble jornada en el nivel Primario, EGB 1 y 2.

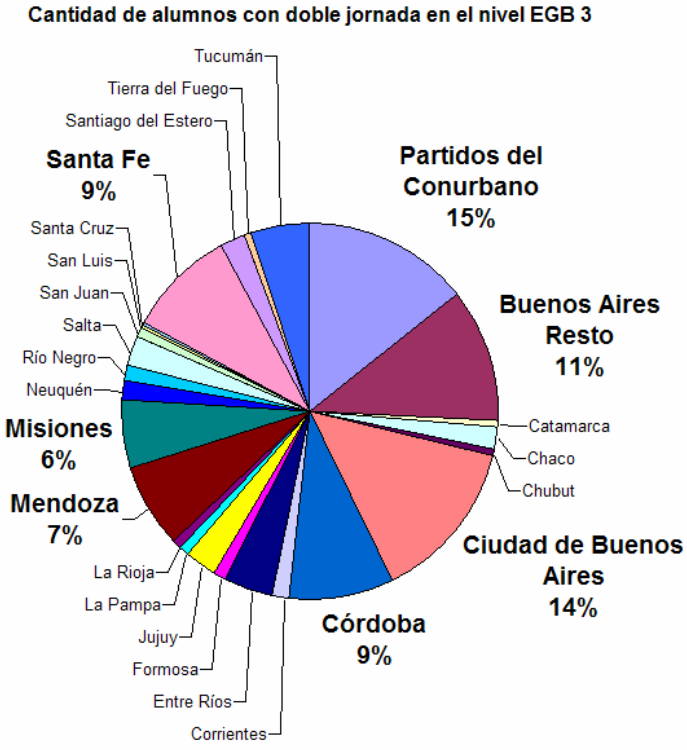

Figura 4. Cantidad de alumnos con doble jornada en el nivel EGB 3.

Observando el gráfico anterior (figura 4), las provincias que tienen mayor cantidad de alumnos con doble jornada no cambian, aunque solamente aumentan los porcentajes de las ciudades con mayores cantidades de estudiantes y establecimientos educativos. El porcentaje que representa a estos alumnos aumenta a un 6,2 % de un total de 4.597.747 alumnos.

Finalmente para escuelas Polimodales y Medias, la provincia de Córdoba es el principal exponente de alumnos con doble escolaridad, luego le sigue Buenos Aires (con sus divisiones), la Ciudad de Buenos Aires, Santa Fe y Mendoza. Sobre un total de 1.360.174 alumnos, el 7,1 % asiste a estas escuelas.

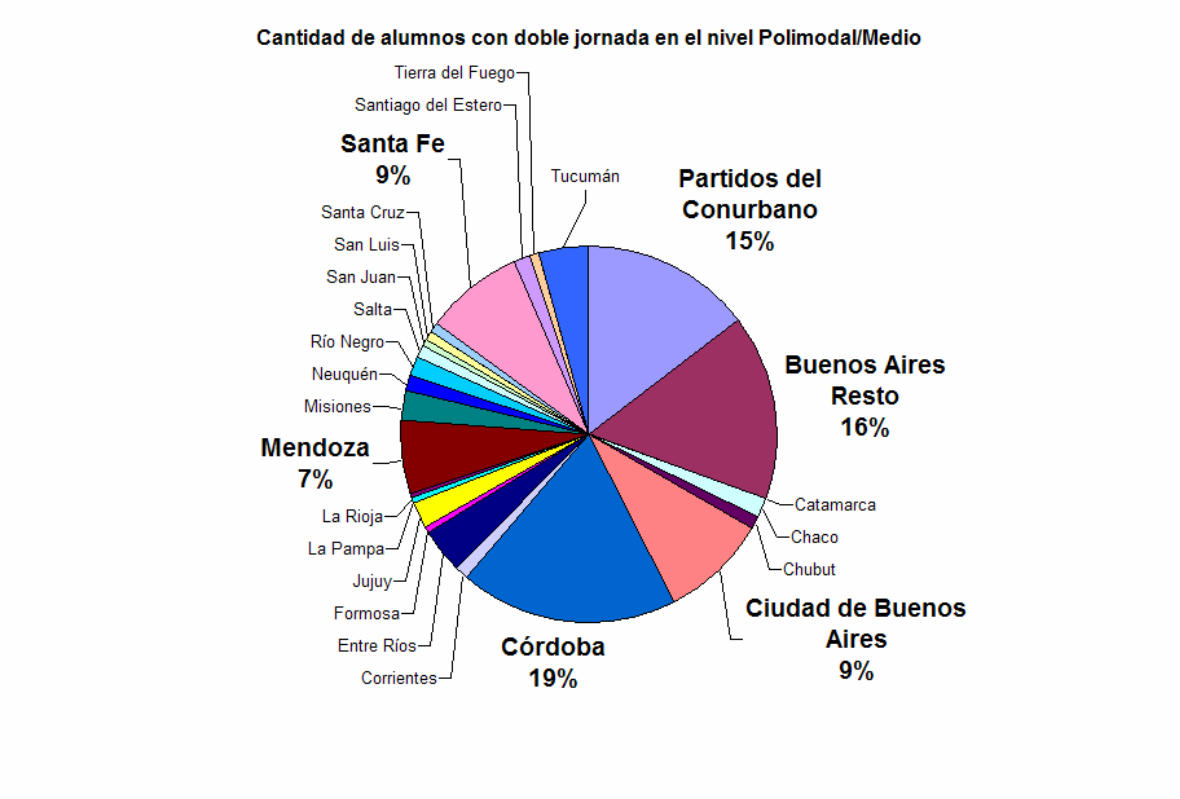

Figura 5. Cantidad de alumnos con doble jornada en el nivel Medio/Polimodal.

El nivel SNU, como es de esperar, no posee doble escolaridad.

Verificación de la calidad de los datos

Según datos brindados por la DINIECE, se pueden observar los porcentajes de participación de los establecimientos en el Relevamiento Anual 2005 según provincias.

División Político Territorial	Inicial	EGB 1y2/Primario	EGB 3	Medio/Polimodal	Superior no Universitario
Total País	**98,2**	**98,1**	**97,1**	**96,4**	**93,3**
Buenos Aires	95,5	96,3	95,4	94,2	86,9
Catamarca	97,7	99,6	98,8	98,9	94,1
Chaco	100,0	100,0	92,3	100,0	97,6
Chubut	88,4	95,0	94,1	88,4	75,0
Ciudad de Buenos Aires	99,7	99,6	100,0	99,4	98,0
Córdoba	100,0	100,0	100,0	97,0	100,0
Corrientes	100,0	100,0	100,0	100,0	100,0
Entre Rios	100,0	100,0	100,0	100,0	100,0
Formosa	100,0	100,0	100,0	100,0	100,0
Jujuy	99,5	99,8	98,8	99,2	95,0
La Pampa	100,0	100,0	100,0	100,0	94,4
La Rioja	99,5	100,0	99,7	98,7	97,1
Mendoza	99,7	91,4	99,9	100,0	100,0
Misiones	99,7	99,6	99,7	98,5	85,0
Neuquén	100,0	100,0	100,0	100,0	100,0
Rio Negro	100,0	100,0	100,0	100,0	100,0
Salta	99,9	99,8	99,8	100,0	98,3
San Juan	99,7	99,7	98,6	97,4	87,5
San Luis	88,4	90,2	90,2	92,9	42,9
Santa Cruz	100,0	100,0	100,0	100,0	100,0
Santa Fe	97,5	96,4	95,4	91,5	88,8
Santiago del Estero	100,0	99,4	93,8	93,4	98,0
Tierra del Fuego	100,0	100,0	100,0	100,0	85,7
Tucumán	99,4	99,2	98,4	95,3	93,1

Tabla 11. Porcentaje de cobertura del Relevamiento Anual 2005.

La base presenta escasos registros con errores cuya depuración se detalla en la Fase III. Con lo cual, se puede afirmar con seguridad que, debido al alcance del relevamiento y a la presentación de los datos, se cumple con los requisitos para lograr los objetivos planteados.

3.3. Fase III: Preparación de los datos

Esta fase involucra las tareas que hacen a la preparación de los datos, es decir seleccionar los datos, limpiarlos, estructurarlos integrarlos y definir el formato final de los mismos. A su vez, en caso de ser necesario como lo es en este proyecto, la construcción de nuevos datos obtenidos o calculados a partir de los datos ya existentes también se debe documentar en esta fase.

A diferencia de la Fase I y II, a partir de la Fase III en adelante, se desarrollan las restantes para cada requerimiento en particular. Esto ayuda a que la comprensión de la resolución de cada requerimiento sea clara.

Solo tres de las nueve tablas seleccionadas forman parte del análisis de todos los requerimientos, con lo cual solo estas son explicadas de manera general. Estas tablas son MAE2005, CAR2005 y MAT2005. Las secciones de la Fase III se repiten sistemáticamente para cada requerimiento. Se recomienda, remitirse a la tabla 1 del anexo donde se encuentran las características de los campos de las tablas de la DINIECE.

TABLA MAE2005

Selección de los datos

En esta sección se deciden que datos son usados para el análisis. El criterio de selección aplicado debe ser lo suficientemente amplio para permitir incluir datos de relevancia en función de los objetivos del proyecto, como así también mantener las normas de calidad y requerimientos técnicos (límites de volumen o tipos de datos). Es de hacer notar que esta selección cubre tanto la cantidad de atributos (o columnas) como de registros (o filas).La salida de este paso son las listas de datos a incluir y excluir, con las razones que avalan estas decisiones.

La tabla MAE2005 presenta información sobre todos los establecimientos nacionales. Muestra los niveles que abarca, el ámbito y sector al que pertenecen, la provincia y el departamento donde se ubica.

Los campos INICIAL, EGB12, MEDIO, PRIMARIO, SNU, EGB3 y POLIMODAL, contienen valores de tipo marca o *flag*, con lo cual son eliminados del análisis, ya que el nivel se encuentra explícito en el resto de las tablas. Cada establecimiento tiene una cruz en los campos donde se enseña conocimientos del nivel educativo correspondiente. El campo

DEPARTAMENTO, responde, valga la redundancia, al departamento o partido en donde se ubica el establecimiento. Dado que muchos registros de dicho campo se encuentran vacíos y que su aporte no es de importancia para este proyecto, se decide eliminarlo. Finalmente el campo LEY no se encuentra explicado en ninguno de los documentos asociados a las bases de datos, con lo cual también queda excluido del análisis.

Por consecuencia, los campos que se seleccionan de esta tabla son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo. El nombre del campo se fundamenta por ser es el ID de la escuelas que participan en todos los Relevamientos Anuales (RA) realizados desde el 2000.
- **AMBITO:** Se refiere a la ubicación geográfica de los establecimientos educativos Sus valores son:

    - Urbano
    - Rural

- **SECTOR:** Alude a la responsabilidad de la gestión de los servicios educativos. La gestión puede ser:

    - Estatal.
    - Privada.

- **PROVINCIA:** Responde a la ubicación geográfica del establecimiento educativo. Los valores corresponden a las veinticuatro (24) provincias de nuestro país. Estos son:

    - BUENOS AIRES
    - CATAMARCA
    - CHACO
    - CHUBUT
    - CIUDAD DE BUENOS AIRES
    - CORDOBA
    - CORRIENTES
    - ENTRE RIOS
    - FORMOSA
    - JUJUY
    - LA PAMPA
    - LA RIOJA
    - MENDOZA
    - MISIONES
    - NEUQUEN
    - RIO NEGRO
    - SALTA
    - SAN JUAN
    - SAN LUIS
    - SANTA CRUZ
    - SANTA FE
    - SANTIAGO DEL ESTERO
    - TIERRA DEL FUEGO
    - TUCUMAN

- **SEDE:** Es el lugar donde cumple sus funciones la máxima autoridad del establecimiento como responsable pedagógico y/o administrativo. La sede puede no tener alumnos. El anexo es la sección o grupo de secciones que depende administrativa y/o pedagógicamente de un establecimiento sede y funciona en distintos lugares geográficos. En caso que el establecimiento imparta educación en un anexo el campo SEDE tendrá como valor un "2". Con lo cual, los valores posibles son:

    - 1
    - 2

Limpieza de datos

El objetivo es optimizar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar selección de subconjuntos limpios de los datos, la inserción de valores por defecto convenientes o aplicación de técnicas de estimación de datos perdidos. Como resultado de esta sección se deben describir que decisiones se tomaron y que acciones se tomarán para solucionar los problemas de calidad de datos que se informaron durante la tarea de "Verificación de calidad de los datos" de la fase "Comprensión de los datos".

Dado que en la tabla MAE2005, no existen registros incompletos en los campos seleccionados, no se realiza ningún tipo de limpieza.

Construcción e integración de datos

Esta tarea de construcción incluye las operaciones de preparación y construcción de datos, como así también la producción de atributos derivados, completando con los nuevos registros o los valores transformados con los atributos existentes. Mientras que en la tarea de integración se aplican métodos que combinan información de múltiples tablas o archivos para crear nuevos registros o valores.

En esta sección la tabla MAE2005 tampoco requiere ninguna acción, ya que no se agregan nuevos campos o atributos a partir de campos existentes de esta tabla o de otras relacionadas.

Formato de los datos

Las transformaciones de estructuras se refieren a las modificaciones principalmente sintácticas realizadas a los datos que no cambian su significado, pero podría requerirse por la herramienta del modelo.

Algunas herramientas tienen requisitos en el orden de los atributos, como por ejemplo el primer campo es un único identificador para cada registro o el último sea el campo del resultado total del modelo a predecir. En función de lo expuesto en el párrafo anterior, también puede ser necesario cambiar el orden de los registros en el dataset. Hay herramientas que requieren que estén ordenados conforme al valor del atributo. Adicionalmente, existen herramientas que pueden requerir alguna tarea extra, como es por ejemplo separar los campos con comas o punto y coma o limitar la longuitos de algunos campos a un tamaño determinado.

Afortunadamente la aplicación utilizada no requiere ningún formato especial de los datos de la tabla MAE2005, ni de ninguna de las tablas utilizadas. Los formatos de los datos se setéan directamente en la aplicación, lo que será explicado en la Fase IV. Por lo tanto, esta sección no estará presente en el análisis de las tablas posteriores.

TABLA CAR2005

Selección de los datos

Esta tabla presenta información sobre la planta funcional de los establecimientos en cada uno de los niveles. También muestra numéricamente los cargos y las horas desarrolladas por cada uno de ellos.

En dicha tabla existen los campos de nombre CARGO( ) con valores en "( )" del uno (1) al cinco (5). Estos valores se corresponden a los diferentes tipos de cargos docentes para todos los niveles. Esta distinción de cargos no se encuentra expresada en ninguno de los documentos asociados a las bases de datos, con lo cual dicho dato queda excluido del análisis. Lo mismo ocurre con el campo HORA( ) con valores en "( )" del uno (1) al dos (2), los cuales también se desconocen sus características.

Por consecuencia, los campos que se seleccionan de esta tabla son:

- **ID_RA**: Número de identificación otorgado al establecimiento sede y/o anexo

- **NIVEL**: Los niveles de enseñanza son los tramos en que se estructura el sistema educativo formal. Se corresponden con las necesidades individuales de las etapas del proceso psico-físico-evolutivo articulado en la del desarrollo psico-físico-social y cultural. Los valores de este campo, los cuales fueron nombrados anteriormente, son:

    - Inicial
    - Primario/EGB
    - Medio/Polimodal
    - Superior No Universitario.

- **POF**: Se denomina POF a la planta funcional, que es el conjunto de cargos y horas cátedra asignados legal y presupuestariamente al establecimiento, estén éstos cubiertos o sin cubrir, independientemente de que quienes los ocupen estén en uso de licencia, comisión de servicio o tareas pasivas. Para los establecimientos privados, también incluye las horas y cargos no subvencionados o extracurriculares. La POF tiene dos valores que informan si los establecimientos están *dentro* de lo presupuestado en horas, cargos y módulo o *fuera*.

    - dentro
    - fuera

Limpieza de datos

En la tabla MAE2005, no existen registros incompletos en los campos seleccionados, con lo cual, no se realiza ningún tipo de limpieza.

Construcción e integración de datos

Para esta tabla existen algunas modificaciones que dependen del nivel al que se refiera el requerimiento en cuestión. El conflicto se genera en el campo "NIVEL" de la tabla, ya que algunos de los valores de este campo en otras tablas son diferentes.

Los niveles educativos que afecta este error producido en el procesamiento de datos del RA, son el Medio/Polimodal y el Primario/EGB. En las tablas que contienen información sobre estos niveles se realiza una distinción entre las escuelas que siguen el plan Primario o EGB y Medio o Polimodal. Con lo cual, al querer fundir tablas por sus niveles, se genera una incompatibilidad de valores.

La solución que se propone es crear dos tablas adicionales que tendrán los mismos campos seleccionados en la tabla CAR2005, y cuyos registros se obtienen a partir de filtrar el campo NIVEL. Con lo cual, la tabla CARPE2005 tendrá solamente los registros de las escuelas Primario/EGB, mientras que CARMP2005 tendrá los registros de las escuelas Medio/Polimodal. Al vincular las escuelas, el dato del nivel no será necesario por haber sido filtrado anteriormente, con lo cual se elimina y se toma al "ID_RA" como campo de conexión entre las tablas. Lo importante es saber de la creación de las tablas CARPE2005 y CARPE2005, ya que en la fase de modelado, se verá funcionalmente el cambio propuesto.

TABLA MAT2005

Selección de los datos

La tabla MAT2005 presenta la cantidad de alumnos matriculados y repetidos en cada año, para cada establecimiento, para todos los niveles educativos. También muestra el tipo de sección del establecimiento en cuestión.

Se toman todos los campos que conforman la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo
- **NIVEL:** Todos los valores correspondientes a los cuatro niveles, aunque separados según el plan educativo del establecimiento. Los valores de este campo son:

    - Inicial
    - Primario
    - EGB
    - Medio
    - Polimodal
    - Superior No Universitario.

- **TIPO _ SE:** Grupo escolar formado por alumnos que cursan en el mismo espacio, al mismo tiempo y con el mismo docente o equipo de docentes. Pueden estar cursando el mismo o diferentes grado. Los valores posibles son los siguientes:

    - I → **Sección Independiente (I)**: Las actividades de enseñanza corresponden a un mismo grado o año.
    - M → **Secciones Múltiples (M)**: Las actividades de enseñanza corresponden a varios años de estudio.

- **ALU( ):** Los valores en "( )" van de uno (1) a nueve (9), y corresponden a la cantidad de alumnos matriculados en dicho año o grado de enseñanza para un determinado tipo de sección y nivel de enseñanza.
- **REP( ):** Los valores en "( )" van de uno (1) a nueve (9), y corresponden a la cantidad de alumnos repetidos en dicho año o grado de enseñanza para un determinado tipo de sección y nivel de enseñanza.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros responden a los campos ALU() y REP(), comentados anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga "0" como resultado, y no "vació" o "null".

Segundo, se modifica el nombre de los campos ALU() y REP() para identificar mas fácilmente cual es su contenido sin necesidad de remitirse a su definición. Los nombres respectivos son MAT_NIV_TIPO_() y REP_NIV_TIPO_() (dentro de "()" hay un valor numérico cuyo significado corresponde al año o grado de enseñanza).

Construcción e integración de datos

En cuanto a la construcción, se crean dos nuevos campos a partir de los seleccionados en esta tabla. El primero es la sumatoria de los campos MAT_NIV_TIPO_() del uno (1) al nueve (9), con lo cual, siguiendo la nomenclatura se denomina MAT_NIV_TIPO. Este campo muestra, como es de esperar, los matriculados anuales en cada nivel para cada establecimiento, o sea, ID_RA.

El segundo es similar al anterior, pero aplicado a los campos REP_NIV_TIPO_() del uno (1) al nueve (9). Por lo tanto, la sumatoria de estos forma el campo REP_NIV_TIPO, correspondiente a la cantidad de alumnos repetidos durante el año en curso, 2005 en este caso, para cada establecimiento o ID_RA.

Para la integración de los datos con otras tablas, se puede anticipar que existe una dificultad observando los valores del campo NIVEL. Estos no se presentan de la forma en que se dividen los niveles en el proyecto, sino cada uno por separado dependiendo del plan que siga dicha institución. A fin de solucionar este problema, se crean cuatro nuevas tablas, con los mismos campos base e indirectos (calculados) de MAT2005, filtrando el campo NIVEL según los valores de la tabla MAE2005. Por lo tanto, las cuatro nuevas tablas son:

- MATI2005 → con NIVEL = Inicial
- MATPE2005 → con NIVEL = Primario/EGB
- MATMP2005 → con NIVEL = Medio/Polimodal
- MATSNU2005 → con NIVEL = Superior No Universitario

Al ingresar en el análisis de la Fase IV se podrá comprender la razón funcional de su creación.

Hasta aquí, se ha desarrollado la Fase III para las tablas que forman parte de todos los requerimientos. Con las mismas secciones denotadas para cada tabla, se realiza el análisis de los requerimientos, en donde se encuentran las seis (6) tablas restantes divididas de acuerdo a su necesidad.

Requerimiento # 1

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MATI2005 → ya fue analizada
- CAR2005 → ya fue analizada
- EDI2005

Por lo tanto, se completa el análisis para dicho requerimiento realizando la Fase III para EDI2005.

TABLA EDI2005

Selección de los datos

La tabla EDI2005 muestra la cantidad de alumnos que se ubican en los diferentes rangos de edades que se señalan en cada campo para todos los establecimientos con un nivel de educación Inicial.

Se toman todos los campos que conforma la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo
- **NIVEL:** El único valor es = Inicial

- **ED( ):** Los valores en "( )" van de uno (1) a seis (6) y corresponden a las edades que tienen los alumnos que asisten a dicho nivel. El dato muestra la cantidad de alumnos que tienen la edad señalada por este campo.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros responden a los campos ED( ), comentados anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga "0" como resultado, y no "vació" o "null".

Construcción e integración de datos

En cuanto a la construcción, se crea un nuevo campo a partir de los seleccionados en esta tabla. Este es el promedio de los campos ED() del uno (1) al seis (6), con lo cual, siguiendo la nomenclatura se denomina EDAD_PROM. Este campo muestra, como es de esperar, las edades promedio del nivel inicial para cada establecimiento, o sea, ID_RA.

Requerimiento # 2

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MATPE2005 → ya fue analizada
- CARPE2005 → ya fue analizada

Por lo tanto, considerando que los cambios realizados en estas tablas son válidos, la Fase III se encuentra completa para este requerimiento.

Requerimiento # 3

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MATPE2005 → ya fue analizada
- CARPE2005 → ya fue analizada

Por lo tanto, considerando que los cambios realizados en estas tablas son válidos, la Fase III se encuentra completa para este requerimiento.

Requerimiento # 4

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MMP2005
- EMP2005

Por lo tanto, se completa el análisis para dicho requerimiento desarrollando la Fase III para MMP2005 y EMP2005.

TABLA MMP2005

Selección de los datos

Presenta la cantidad de alumnos matriculados en cada establecimiento, según la modalidad elegida en el nivel Medio/Polimodal.

Se toman todos los campos que conforman la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo
- **NIVEL:** Los valores son:

  - Medio
  - Polimodal

- **MODALIDAD:** Responde las distintas modalidades que se enseñan en los establecimientos correspondientes a dicho nivel. Los valores son:

  - Agropecuaria
  - Artística
  - Bachiller
  - Ciclo básico
  - Ciencias Naturales
  - Comercial
  - Comunicación, Artes y Diseño
  - Economía y Gestión de las Organizaciones
  - Humanidades y Ciencias Sociales
  - Otros
  - Producción de Bienes y Servicios
  - Técnica

- **MAT():** Los valores en "()" van de uno (1) a siete (7), y corresponden a la cantidad de alumnos matriculados en dicho año o grado de enseñanza. En el caso del nivel Medio/Polimodal se diferencian los matriculados por año según su modalidad.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros se encuentran en los campos MAT() comentados anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga "0" como resultado, y no "vació" o "null".

Segundo, se modifica el nombre de los campos MAT() para identificar mas fácilmente cual es su contenido sin necesidad de remitirse a su definición. Los nombres serán MAT_NIV_MOD_() (dentro de "()" hay un valor numérico cuyo significado corresponde al año o grado de enseñanza).

Construcción e integración de datos

En cuanto a la construcción, se crea un nuevo campo a partir de los seleccionados. Este es la sumatoria de los campos MAT_NIV_MOD_() del uno (1) a siete (7), con lo cual, siguiendo la nomenclatura se denomina MAT_NIV_MOD. Este campo muestra, como es de esperar, los matriculados anuales en cada nivel para cada establecimiento, o sea, ID_RA, según su modalidad.

TABLA EMP2005

Selección de los datos

Presenta la cantidad de alumnos egresados en cada establecimiento, según la modalidad elegida en el nivel Medio/Polimodal.

Se toman todos los campos que conforman la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo.
- **NIVEL:** Ídem MMP2005.
- **MODALIDAD:** Ídem MMP2005.
- **EGRESADOS:** Correspondiente a la cantidad de alumnos egresados del establecimiento en cuestión. En el caso del nivel Medio/Polimodal se diferencian los egresados según su modalidad.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros se encuentran en el campo EGRESADOS comentado anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga "0" como resultado, y no "vació" o "null".

Segundo, se modifica el nombre del campos EGRESADOS para identificar mas fácilmente cual es su contenido sin necesidad de remitirse a su definición. El nombre será EGRE_NIV_MOD.

Construcción e integración de datos

En esta sección la tabla EMP2005 no requiere ninguna acción, ya que no se agregan nuevos campos o atributos a partir de campos existentes de esta tabla o de otras relacionadas.

Requerimiento # 5

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MATMP2005 → ya fue analizada
- CARMP2005 → ya fue analizada

Por lo tanto, considerando que los cambios realizados en estas tablas son válidos, la Fase III se encuentra completa para este requerimiento.

Requerimiento # 6

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MSNU2005
- ESNU2005

Por lo tanto, se completa el análisis para dicho requerimiento desarrollando la Fase III para MSNU2005 y ESNU2005.

TABLA MSNU2005

Selección de los datos

Presenta la cantidad de alumnos matriculados en cada establecimiento, según la carrera elegida en el nivel Superior no Universitario. También muestra el tipo de formación del establecimiento en cuestión.

Se toman todos los campos que conforman la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo.
- **NIVEL:** El único valor es = Superior No Universitario.
- **CARRERA:** Es la carrera que se enseña en el establecimiento correspondiente a dicho nivel. Este campo posee ciento quince (115) valores que se exponen en la tabla 3 del anexo.
- **TIPOFORMAC:** Dependiendo de las carreras que se enseñen en un establecimiento se obtiene el tipo de formación de la misma. Los posibles valores son:

    - Exclusivamente Técnico – Profesional
    - Exclusivamente Docente
    - Ambos tipos de formación

- **MATRICULA:** Correspondiente a la cantidad de alumnos matriculados del año en curso. En el caso del nivel Superior No Universitario se diferencian los matriculados según su carrera.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros se encuentran en el campo MATRICULA comentado anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga “0” como resultado, y no “vació” o “null”.

Segundo, se modifica el nombre del campo MATRICULA para identificar mas fácilmente cual es su contenido sin necesidad de remitirse a su definición. El nombre será MAT_CAR.

Construcción e integración de datos

Esta sección no requiere ninguna acción para la tabla MSNU2005, ya que no se agregan nuevos campos o atributos a partir de campos existentes de esta tabla o de otras relacionadas.

TABLA ESNU2005

Selección de los datos

Presenta la cantidad de alumnos egresados en cada establecimiento, según la carrera elegida en el nivel Superior no Universitario. También muestra el tipo de formación del establecimiento en cuestión.

Se toman todos los campos que conforman la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo.
- **NIVEL:** Ídem MSNU2005.
- **CARRERA:** Ídem MSNU2005.
- **TIPOFORMAC:** Ídem MSNU2005.
- **EGRESADOS:** Correspondiente a la cantidad de alumnos egresados del año en curso. En el caso del nivel Superior No Universitario se diferencian los egresados según su carrera.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros se encuentran en el campo EGRESADOS comentado anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga “0” como resultado, y no “vació” o “null”.

Segundo, se modifica el nombre del campo EGRESADOS para identificar mas fácilmente cual es su contenido sin necesidad de remitirse a su definición. El nombre será EGRE_CAR.

Construcción e integración de datos

Esta sección no requiere ninguna acción para la tabla ESNU2005, ya que no se agregan nuevos campos o atributos a partir de campos existentes de esta tabla o de otras relacionadas.

Requerimiento # 7

Para dicho requerimiento se utilizan las siguientes tablas:

- MAE2005 → ya fue analizada
- MATSNU2005 → ya fue analizada
- CAR2005 → ya fue analizada
- EDS2005

Por lo tanto, se completa el análisis para dicho requerimiento desarrollando la Fase III para EDS2005.

TABLA EDS2005

Selección de los datos

Muestra la cantidad de alumnos que se ubican en los diferentes rangos de edades que se señalan en cada campo, para todos los establecimientos del nivel Superior no Universitario.

Se toman todos los campos que conforman la tabla. Estos son:

- **ID_RA:** Número de identificación otorgado al establecimiento sede y/o anexo.
- **NIVEL:** El único valor es = Superior no Universitario
- **ED( ):** Los valores en "()" van de diecisiete (17) a veintinueve (29), con un rango continuo de treinta (30) a treinta y cuatro (34), de treinta y cinco (35) a treinta y nueve (39), y luego cuarenta (40). Los números dentro de "()" corresponden a las edades que tienen los alumnos que asisten a dicho nivel. El dato muestra la cantidad de alumnos que tienen la edad señalada por este campo.

Limpieza de datos

La primera corrección que se realiza es completar con ceros (0) los registros de tipo numérico que se encuentren vacíos. Estos registros responden a los campos ED(), comentados anteriormente. Esto se realiza para que una vez corrida la aplicación se obtenga "0" como resultado, y no "vació" o "null".

Construcción e integración de datos

En cuanto a la construcción, se crea un nuevo campo a partir de los seleccionados en esta tabla. Este es el promedio de los campos ED(), con lo cual, siguiendo la nomenclatura se denomina EDAD_PROM. Este campo muestra, como es de esperar, las edades promedio del nivel Superior No Universitario para cada establecimiento, o sea, ID_RA.

3.4. Fase IV: Modelado

El arte del trabajo especializado del proceso de Exploración de Información toma lugar en esta fase. Aquí se puede intentar probar unas hipótesis específicas o aplicar métodos que permitan el descubrimiento de información de forma automatizada. Además, se deben interpretan los resultados de análisis realizados en esta fase en el contexto planteado en la Fase I del proyecto.

Como ya se ha anticipado, la resolución de esta fase será metodológicamente similar a la de la anterior. En ella se diseña un modelo que responde a cada uno de los requerimientos, logrando un total de siete (7) modelos diferentes. Se recuerda también, que para la resolución de los mismos se utilizan algoritmos de inducción y clusterización, conocidos en la aplicación como CHAID Y KOHONEN respectivamente.

El software seleccionado cuenta con las siguientes ventajas:

- Gran flexibilidad para la carga de datos (inputs).
- Alta facilidad de uso.
- Posibilidad de fundir datos/tablas por campos claves de referencia.
- Efectividad al obtener algoritmos de inducción y de agrupamiento.
- Posibilidad de crear reglas de decisión, sesgando la salida según el nivel de confianza de la misma.
- Rapidez en la ejecución

Sus desventajas mas significativas son:

- Salidas poco claras para personas que no habitúan aplicar minería de datos.
- Bajo potencial para la obtención de gráficos o diagramas.

Para mas detalles operativos de la herramienta de modelización, remitirse al anexo, donde se explican los conceptos básicos del programa.

En la Fase IV se seleccionan y se aplican diversas técnicas de modelado, como así también, se determinan los valores de los parámetros y variables de calibración. Para esta tarea generalmente se puede contar con más de una técnica. Algunas de ellas pueden tener requerimientos específicos en cuanto a la configuración de los datos, lo cual puede plantear volver a la fase de preparación de los datos para realizar modificaciones.

Las dos primeras secciones de esta fase y la última, se desarrollan de manera genérica dado que no hay distinción entre las técnicas de modelado y en el diseño de pruebas implementado para cada requerimiento. La construcción del modelo se divide según los requerimientos ya que obviamente, se buscan objetivos diferentes. Se recomienda también que se vuelva sobre los requerimientos, para comprender la razón de la modelización.

Selección de técnicas de modelado

Dado que ya se ha seleccionado la herramienta de modelado o aplicación, esta presenta diversos modelos que se pueden observar en la paleta de modelos que se encuentra en el anexo. Como se adelantara, los modelos que se utilizan son algoritmos de inducción y clusterización.

Para los algoritmos de inducción, la aplicación presenta tres diferentes modelos, que son C.5, CRT y CHAID. Como es habitual en un proceso de Minería de Datos, el analista no conoce operativamente que tipo de algoritmo se reproduce al ejecutar el modelo, sin embargo, puede evaluar la efectividad de los resultados obtenidos. En este modelo se sigue dicha secuencia, que consta en correr todos los modelos inductivos, y de acuerdo al resultado obtenido, se selecciona uno de ellos. Esto se realiza con el primer requerimiento, y los modelos elegidos se extienden a los restantes.

Se anticipa que los resultados obtenidos de la ejecución de los modelos son reglas de decisión logradas por medio de una opción que brinda la herramienta. La gran utilidad de esta opción es la posibilidad de sesgar los resultados de acuerdo al nivel de confianza que presente la misma. Para este análisis el nivel de confianza se setéa en 75 puntos.

Al observar los resultados de cada algoritmo se detecta lo siguiente:

- C.5 → El algoritmo tiene un alto poder de discriminación entre los datos, siendo incapaz de agrupar patrones o comportamientos similares. Consecuentemente, la cantidad de reglas que entrega por corrida superan, en la mayoría de los casos, las cincuenta (50) complicando el análisis. Generalmente se obtienen reglas sustentados por una baja cantidad de registros y con niveles de confianza altos. El principal campo por el cual discrimina esta herramienta es por PROVINCIA.

- CRT → Este algoritmo, a diferencia del anterior, genera gran cantidad de agrupaciones brindando información de confianza reducida. Como es de esperar, las reglas de decisión son pocas y con niveles de confianza cercanos al sesgo señalado. Los principales campos por los cuales esta herramienta agrupa son los que poseen solamente dos valores, mas conocidos como campos de tipo marca o *flag*.

- CHAID → El poder de discriminación de este algoritmo se califica como alto. Sin embargo, a diferencia del C.5, este logra agrupar comportamientos similares en una sola regla de decisión manteniendo el nivel de confianza por arriba del sesgo. Esto facilita el análisis y ayuda a detectar patrones que abarcan simultáneamente a diversos valores de campos, como por ejemplo provincias, sectores, ámbitos, carreras, etc..

Considerando lo expuesto para los modelos inductivos, se utiliza el algoritmo CHAID. En caso que este algoritmo no presente resultados coherentes, se ejecuta el algoritmo C.5 debido a su alto poder de discriminación. Se podrá observar en las figuras de cada modelo que finalizan con los modelos señalados. Los íconos correspondientes a estos algoritmos se muestran en el anexo, junto a una breve explicación.

Continuando con los algoritmos de agrupación, conglomeración o, de su traducción del inglés, clusterización, la aplicación presenta a dos herramientas. Estas son KOHONEN y K-MEANS. Para detectar las fortalezas de cada mecanismo, se ejecutan ambos y de acuerdo a sus resultados, se selecciona uno de ellos.

La efectividad de ambos algoritmos a la hora de obtener resultados es similar. Solamente para tablas con gran cantidad de datos, se observa que el tiempo de ejecución de KOHONEN supera a K-MEANS. Sin embargo, la principal diferencia entre ellas, por la cual se descarta K-MEANS, se debe a la necesidad de setear en los parámetros de ejecución la cantidad de grupos o conglomerados que deseo obtener como salida. K-MEANS, por lo tanto, fuerza a los datos a agruparse en la cantidad seleccionada. En cambio, KOHONEN genera los grupos de acuerdo a los datos y a un parámetro de optimización. La función del parámetro de optimización es realizar una ejecución rápida o una en la que prevalezca la memoria del algoritmo. Como es de esperar, la primera opción trae como resultados una menor cantidad de grupos. Este parámetro esta presente en ambas técnicas y en la sección correspondiente a la parametrización de los modelos se muestra como se setéan los mismos.

Generación de diseño pruebas

Antes de generar el modelo se debe desarrollar un procedimiento o mecanismo para probar la calidad y validez de los modelos y del dataset. Para ello, la herramienta de modelización y ejecución facilita algunas herramientas que se encuentran en la paleta de resultados ubicada en el anexo. Para realizar el diseño de pruebas se selecciona al nodo Calidad y nodo Tabla. Estos son actividades que se aplican para obtener información del dataset que circula por el modelo. Las funciones básicas con sus íconos asociados son:

Nodo Calidad 

El nodo Calidad informa sobre la calidad de los datos buscando valores perdidos o vacíos. El nodo puede tener en cuenta las definiciones vacías o tratar valores vacíos o en blanco. De todas maneras en este proyecto los valores vacíos de los campos numéricos fueron completados con ceros (0), como se explicó en la limpieza de datos de la Fase III. Solamente los valores tipo texto incompletos aparecerán vacíos. Esta herramienta se aplica a cada tabla que se carga en un nodo origen como también en la fundición de estos, a fin de controlar que la calidad de los datos no se distorsione con el armado del modelo. Esta herramienta sirve también como ayuda para encontrar soluciones a mensajes de error que levanta la aplicación al realizar la ejecución total del modelo.

Nodo Tabla 

El nodo Tabla permite crear una tabla a partir de los datos que ingresan, permitiendo mostrarla en pantalla o exportarla en un archivo. Esto es útil en cualquier momento en que se necesite examinar sus valores de los datos o exportarlos en un formato fácilmente legible. En este proyecto, su utilización es posterior al nodo Calidad, con lo cual, ante un error visualizado en la calidad del dataset que circula por el modelo, se genera la salida en pantalla de la tabla cargada para examinar cual es el error en cuestión. Los errores generalmente ocurren al fundir tablas y no al cargar datos orígenes, y es allí donde esta herramienta tiene su mayor funcionalidad.

En la figura 6 se muestra ambos nodos aplicados a una tabla origen (MAE2005) con su correspondiente salida en pantalla.

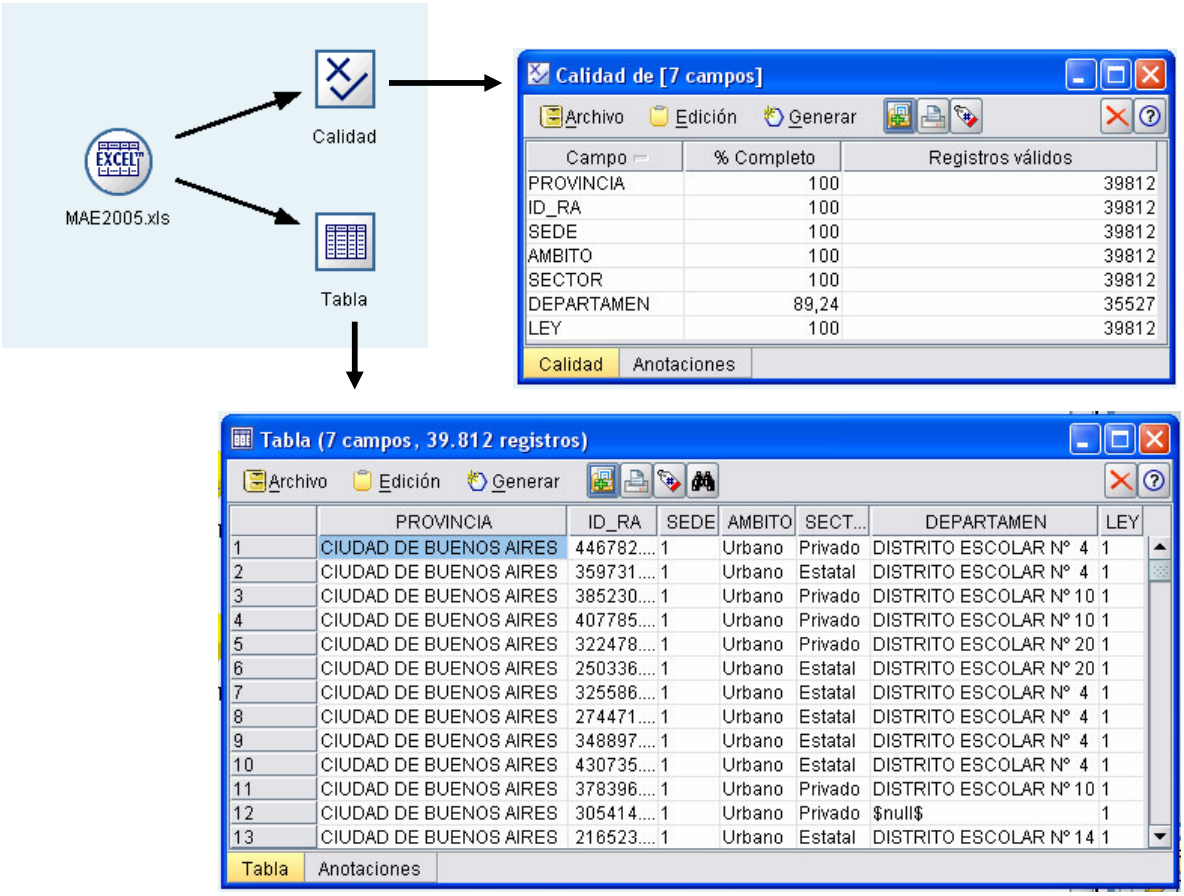

Figura 6. Salidas de los nodos Calidad y Tabla a partir de la nodo origen de la tabla MAE2005.

Como se puede observar, en la salida del nodo Calidad todos los campos se encuentran completos menos el DEPARTAMENTO, con el porcentaje y la cantidad de registros válidos que se indica. Al aplicar el nodo Tabla se observa que la causa de la ausencia de datos son los registros incompletos del campo, pudiendo visualizar uno de ellos en el registro número 12. Observar que la cantidad de campos y registros que señala la salida del nodo Tabla es coherente con lo que muestra la salida del nodo Calidad.

Esta es la forma como se realiza la prueba la calidad y validez del modelo y dataset, a medida que se construye el mismo. Adicionalmente, en caso de no detectar una anomalía con el diseño de pruebas descrito, el software levanta mensajes de error indicando el origen del mismo.

Construcción del modelo

Como se mencionó anteriormente, a partir de esta sección se explica cada modelo haciendo referencia al requerimiento correspondiente. Pero antes de comenzar con la construcción de los mismos, y con la finalidad de que se comprenda su armado y distribución, se describen los nodos que se utilizan de forma general dado en la mayoría de modelos.

Los nodos que se utilizan de manera genérica en todos los modelos:

- Nodo Origen (importar a Excel)
- Nodo Fundir
- Nodo Tipo
- Nodo Filtro
- Nodo Salida (exportar a Excel)
- Nodo C.5
- Nodo CHAID
- Nodo KOHONEN

Nodo Origen (importar a Excel)

Este nodo ya fue visualizado en la sección de diseño de pruebas. Como bien explica su nombre, este nodo se encarga de tomar las tablas provenientes del exterior para su utilización en la aplicación. Con este campo comienza la construcción de cualquier modelo. A diferencia de los otros nodos de importación de datos, este toma archivos con extensión .xls y no otra. Con lo cual, es compatible con la extensión que poseen las tablas que se comentan en fases anteriores. Al conectar el nodo con la tabla externa, este toma automáticamente el nombre de la misma.

Nodo Fundir

La función de un nodo Fundir es tomar varios registros de entrada para crear un registro de salida que contenga todos o algunos de los campos de entrada. Se trata de una operación útil cuando se desean fusionar datos de diferentes orígenes. Existen dos modos de fusionar datos:

- Fusionar por orden: concatena registros correspondientes procedentes de todos los orígenes en el orden de entrada hasta vaciar el origen de datos más pequeño. Si se usa esta opción, es importante haber ordenado previamente los datos con un nodo Ordenar.

- Fusionar usando un campo clave: concatena datos (como el ID de cliente) con lo cual se debe especificar cómo relacionar los registros procedentes de un origen de datos con los procedentes de otros. Ofrece varias posibilidades de unión, incluidas la unión interior, la exterior, la exterior parcial y la anti-unión.

En este proyecto se usa la fusión por campo clave, y como es de esperar, el ID_RA será muchas veces este campo.

Nodo Tipo

Las propiedades del campo se pueden especificar en un nodo Origen o en un nodo Tipo independiente. La funcionalidad es similar en ambos nodos y se usa para describir características de los datos en un campo determinado. Los tipos que se pueden encontrar en la aplicación son:

- Rango → Números continuos dentro de un rango.
- Discreto → Números no continuos.
- Marca → Flag o campo con 2 valores posibles.
- Conjunto → Texto o numérico con valores repetidos.
- Conjunto ordenado → Texto o numérico con valores repetidos ordenados.
- Sin tipo → No siguen un patrón

Nodo Filtro

Los nodos Filtro tienen tres funciones:

- Filtrar o descartar campos de registros que pasan por ellos. Por ejemplo, como investigador médico, es posible que no esté interesado en el nivel de potasio (datos de nivel de campo) de los pacientes (datos de nivel de registro); por ello, puede filtrar el campo K (potasio).

- Cambiar el nombre de los campos.
- Establecer correspondencias de campos entre un nodo de origen y otro.

En este proyecto solamente se utilizan las dos primeras funciones.

Nodo Salida (exportar a Excel) 

El nodo de exportación Excel ofrece los datos resultantes en formato de Excel (.xls). Si se desea, se puede elegir iniciar automáticamente Excel y abrir el archivo exportado cuando se ejecuta el nodo.

Los nodos que representan a los algoritmos comentados al comienzo de la Fase IV, no requieren explicación y son los que se muestran a continuación.

Terminado con los nodos utilizados, se comienza con el modelado de cada requerimiento describiendo paralelamente con la configuración y ajustes de los parámetros de cada nodo y la descripción de la construcción propiamente dicha.

Tener presente que la mayoría de los requerimientos se resuelven con varios modelos, dado que los datos al igual que las técnicas seleccionadas no permiten obtener resultados con una sola corrida. Con lo cual, las figuras que se exponen a continuación sobre la estructura de los modelos, son muchas veces el análisis de una parte de los datos aplicando una técnica especifica (inducción / clusterización + inducción) y no la totalidad del modelo. Se comenta al comenzar cada requerimiento en cuantas partes se divide el modelado del mismo.

Requerimiento # 1

Este requerimiento, correspondiente al nivel Inicial, se lo estudia en dos partes ya que se utilizaron dos modelos de análisis diferentes para la obtención de reglas de decisión. Uno se realiza aplicando solamente inducción, mientras que en el segundo se aplica clusterización y sobre los grupos formados, inducción. Para el modelo de inducción se obtuvieron reglas seleccionando las variables ÁMBITO, y posteriormente, SECTOR como salidas, con lo cual este modelo se divide en dos evaluaciones diferentes.

A su vez, cada uno de los tres (3) modelos (inducción ámbito, inducción sector, clusterización + inducción) se corren sobre dos (2) tablas fundidas (INI y INI_(CONPOF)) que se obtienen a partir de las tablas iniciales señaladas en la Fase III para este requerimiento. Si bien algunos campos se repiten en ambas tablas, otros no están presentes simultáneamente. Esto se debe a que la presencia de ciertos campos en forma combinada distorsiona el análisis de las reglas de decisión. En resumen, este requerimiento es resuelto por tres (3) modelos aplicados a dos (2) tablas, totalizando seis (6) ejecuciones, y por ende, seis (6) salidas diferentes.

La tabla INI esta formada por los campos de las tablas MAE2005 y MATI2005, con lo cual el campo clave por el cual se funden las tablas es el ID_RA. El modelo que se observa en la figura 7, crea la tabla INI y obtiene resultados por inducción.

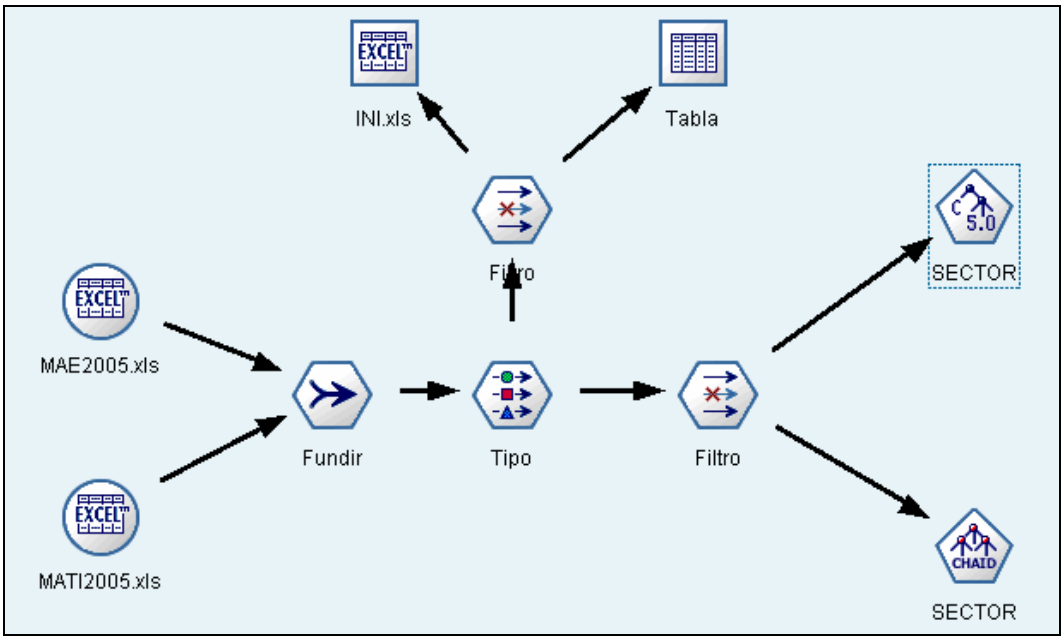

Figura 7. Modelo inductivo de la tabla INI, con el SECTOR como salida.

En el modelo se puede observa a los nodos Origen que importan las dos tablas a Excel para fundirlas en un nodo Fundir. El ajuste de los parámetros de nodo Fundir se pueden observar en la figuras 8 y 9. En la primera, correspondiente a la solapa de entradas, se puede observar que aparecen las dos tablas que conforman la nueva. Mientras que en la segunda, correspondiente a la solapa de fundir, se puede observar que el método de fusión optado es por campos claves y el seleccionado es el ID_RA, o ID del establecimiento.

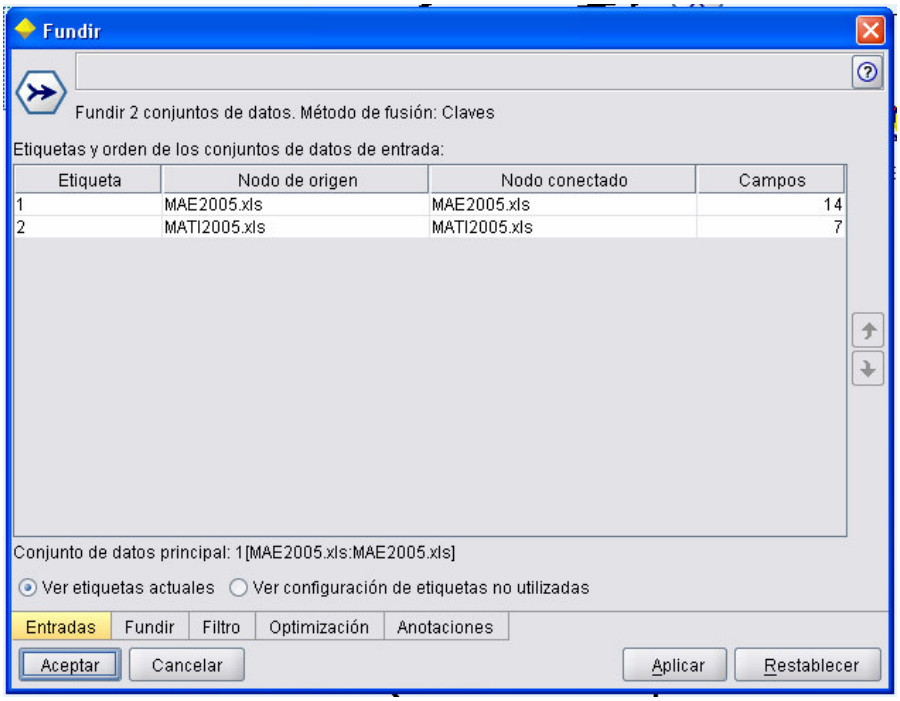

Figura 8. Setéo de los parámetros de fundición. Nodo Fundir, solapa Entradas.

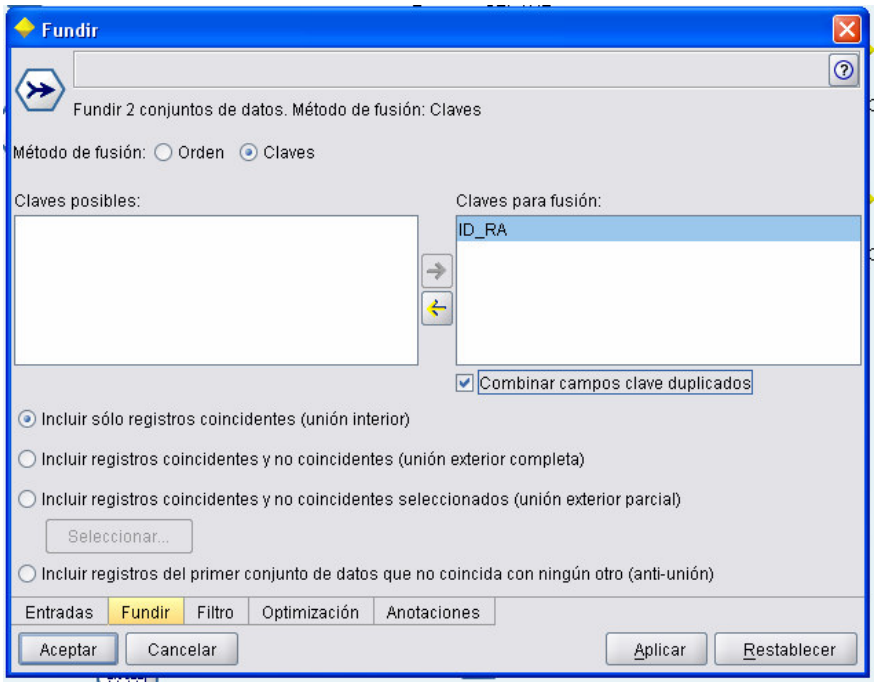

Figura 9. Setéo de los parámetros de fundición. Nodo Fundir, solapa Fundir.

Si se vuelve al modelo, se observa que se le entrega a los campos una tipificación para que sean identificados por la aplicación. Esto es un procedimiento obligatorio para la ejecución del modelo. Para lograrlo se utiliza el nodo Tipo, el cual también setéa el campo que se requiere como salida. En la figura 10 se muestra que el campo SECTOR es seteado como

tal. Posteriormente se cambia a ÁMBITO, que se encuentra un registro por encima del campo SECTOR. A su vez, se pueden apreciar los distintos tipos de valores que toman los campos en cuestión.

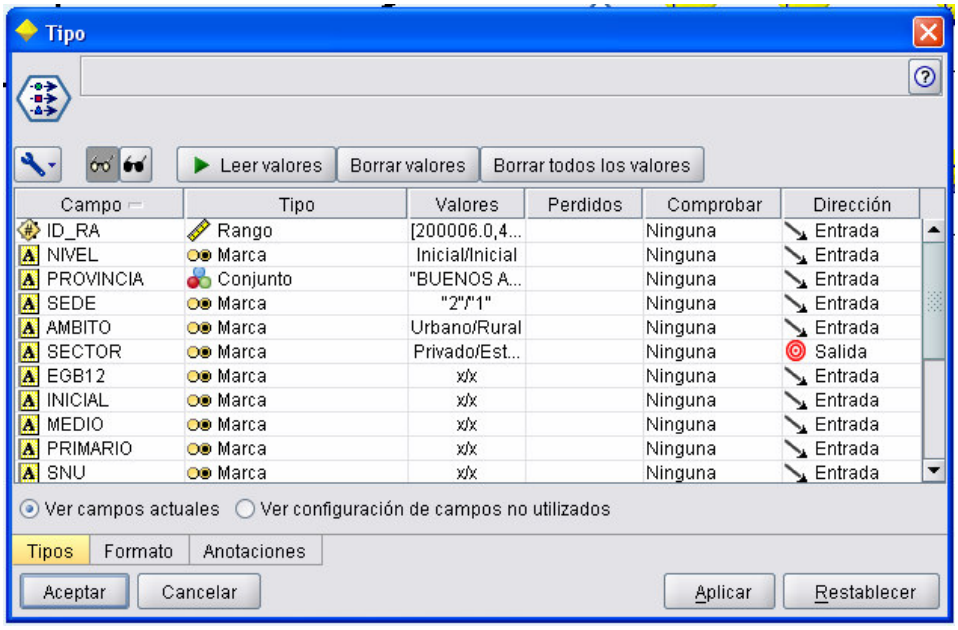

Figura 10. Setéo de los parámetros de tipificación. Nodo Tipo, solapa Tipos.

Posteriormente se filtran los campos que se deciden dejar fuera del requerimiento, con lo cual el nodo Filtro los elimina del análisis de la forma que se muestra en la figura 11.

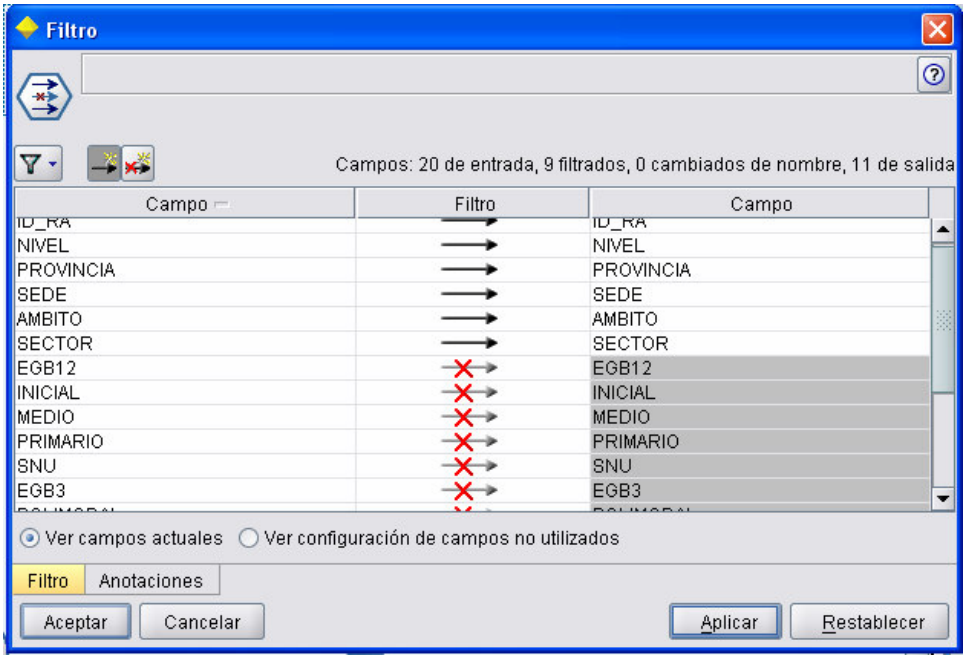

Figura 11. Setéo de los parámetros de filtro, solapa Filtro.

Finalmente la nueva tabla (fundida) queda formada por los siguientes campos:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- TIPO _ SE
- MAT_TIPO
- MAT_TIPO_3
- MAT_TIPO_4
- MAT_TIPO_5

Las salidas del modelo explicado son obtenidas por C.5 y CHAID, los cuales no requieren comentarios sobre el setéo de sus parámetros. En la parte superior del modelo se puede observar que la nueva tabla se exporta a un Excel bajo el nombre de INI, como fue denominado desde su concepción. Esta tabla se utiliza para correr el algoritmo de clusterización + inducción, ahorrando así, el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla INI es el que se muestra en la figura 12.

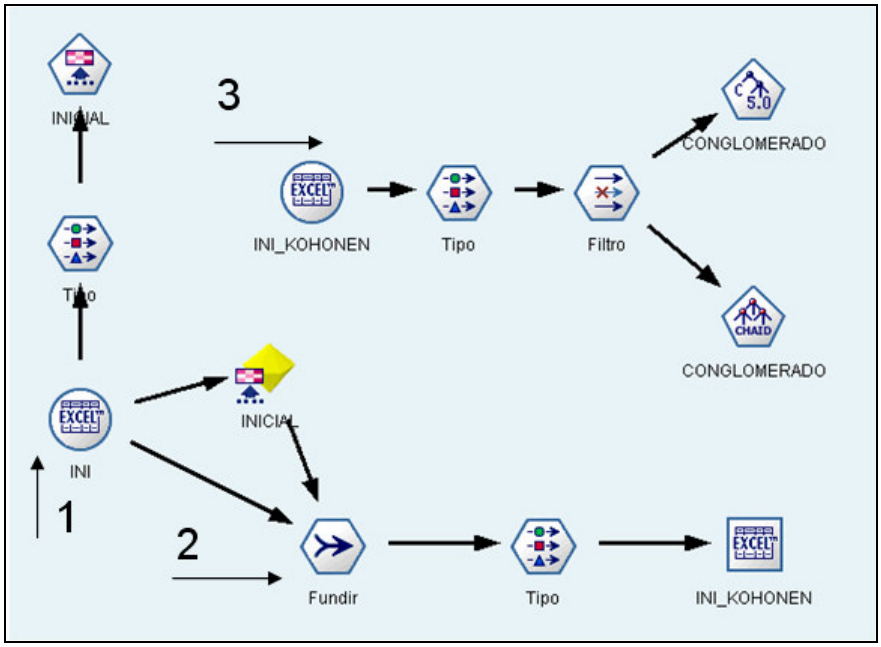

Figura 12. Modelo de clusterización + inducción de la tabla INI.

El modelo comienza, como es habitual con el nodo Origen de nombre INI correspondiente a la tabla exportada del modelo anterior. Si se observa solamente la rama que se dirige hacia arriba (Rama 1), la información pasa por un nodo Tipo para luego ingresar al nodo

KOHONEN. Se recuerda que antes de ejecutar el modelo KOHONEN se debe ajustar el parámetro de optimización de la forma que se muestra en la figura 13.

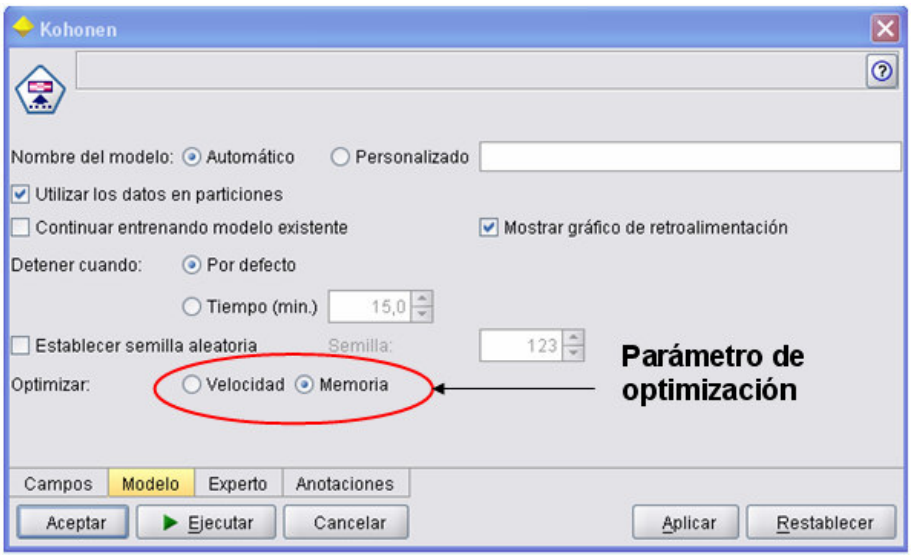

Figura 13. Ajustes de parámetro de optimización del nodo KOHONEN.

De la ejecución de este modelo, se obtiene como salida el icono amarillo que se observa en la figura 12, que básicamente es una nueva tabla con los atributos adquiridos en la ejecución. En este caso, al correr el modelo de clusterización, se suman dos campos adicionales. Estos se corresponden a coordenadas ("x" e "y"), que indican la ubicación del conglomerado en el plano. Sin embargo, para que esta información esté disponible se la exporta a una tabla de Excel (Rama 2), con el nombre de INI_KOHONEN. La nomenclatura es correcta dado que posee los campos aportados por el modelo KOHONEN, fundidos con la tabla INI. En esta tabla se concatenan ambas coordenadas para identificar al conglomerado con una sola variable, esto es

Si x = 0 y si y = 0 → el conglomerado será = grupo 00.

De mas está decir, que el nombre de este nuevo campo es CONGLOMERADO. Finamente, INI_KOHONEN se importa por un nuevo nodo Origen (Rama 3) y se dirige a las técnicas inductivas. Estas se encuentran seteadas para tomar como salida al campo CONGLOMERADO. Con esto se concluye la modelización de los datos correspondientes para la tabla INI.

El modelo inductivo de la tabla INI_(CONPOF) es muy similar al anterior, ya que su principal diferencia radica en las tablas utilizadas, que son EDI2005, CAR2005 y MAE2005. En la figura 14 se puede observar su estructura.

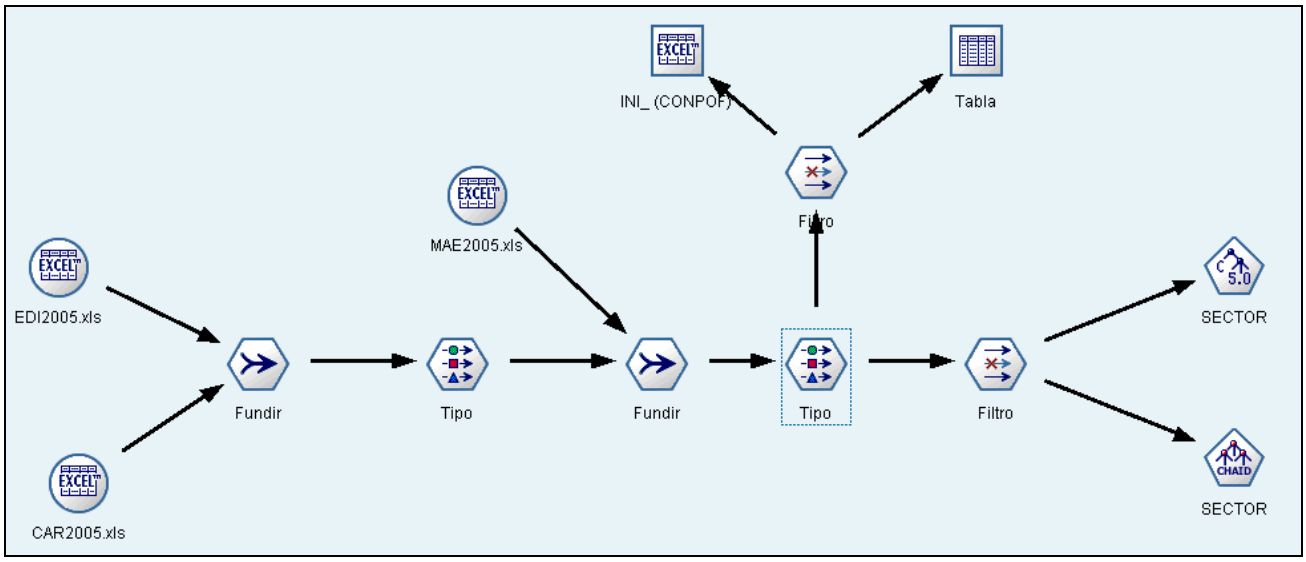

Figura 14. Modelo inductivo de la tabla INI_(CONPOF), con el SECTOR como salida.

La razón por la cual MAE2005 no se funde en el mismo nodo que las dos restantes, es que no posee el campo NIVEL. La primera fundición se realiza por el ID_RA y por NIVEL, lo que es coherente ya que para un establecimiento de la tabla CAR2005 pueden visualizarse varios niveles y por ende diferentes registros, generando conflictos al fundir. Por lo tanto, en el setéo de la primera fundición los campos claves son los que se muestran en la figura 15, mientras que el de la segunda se muestra en la figura 16. El ajuste de parámetros de los nodos restantes es el mismo que se realiza para el primer modelo inductivo del requerimiento.

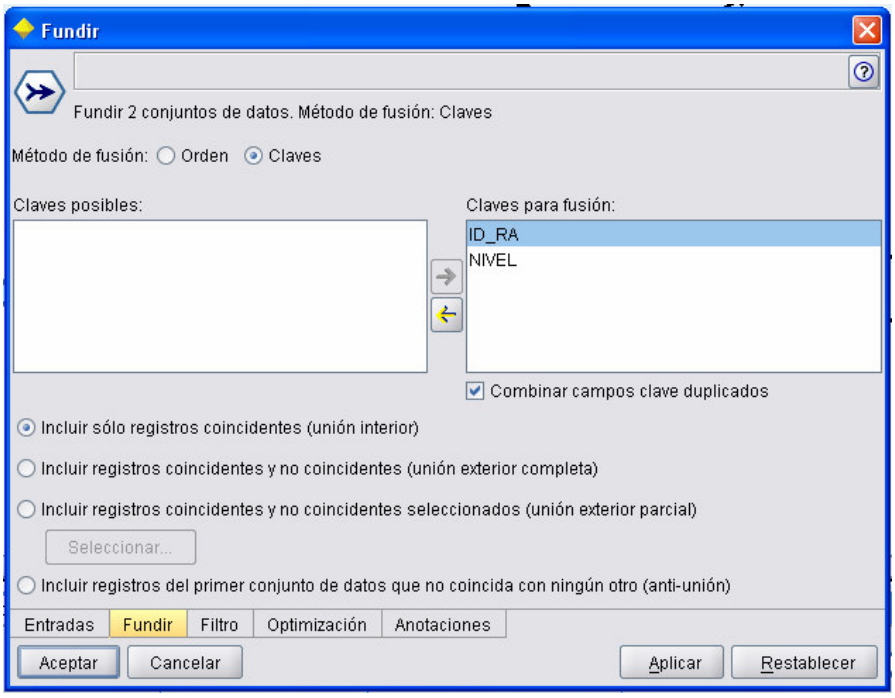

Figura 15. Setéo de los parámetros de fundición. Nodo Fundir, solapa Fundir.

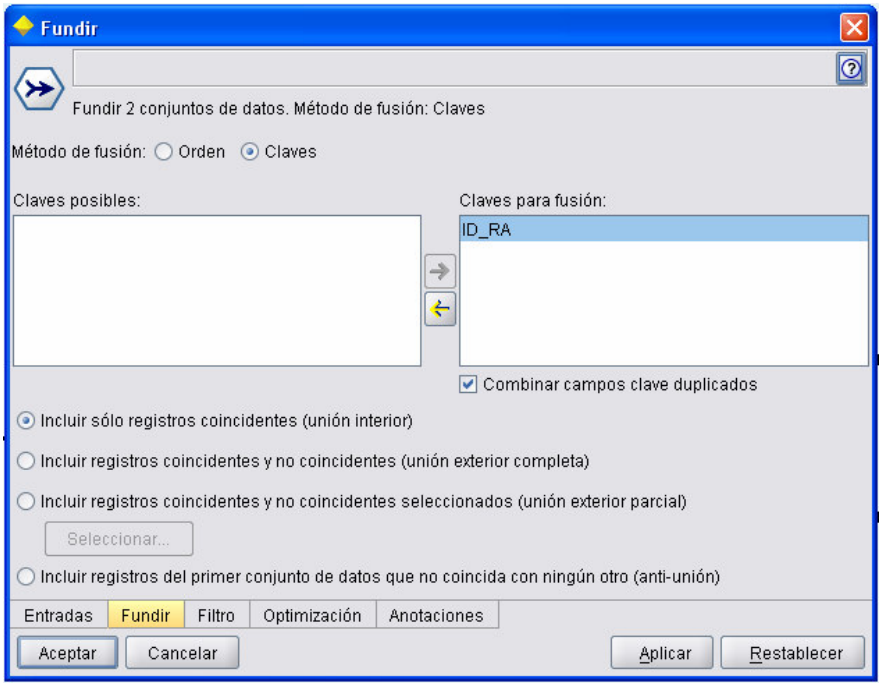

Figura 16. Setéo de los parámetros de fundición. Nodo Fundir, solapa Fundir.

La tabla INI_(CONPOF) contiene, valga la renuncia, el campo “POF”. Este campo, como se describe en la FASE III, muestra si los establecimientos están *dentro* de lo presupuestado en horas, cargos y módulo o *fuera*. Los campos presentes en esta tabla son:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- DEPARTAMENTO
- EDAD_PROM
- POF

Las salidas del modelo explicado son obtenidas por C.5 y CHAID, los cuales no requieren comentarios sobre el setéo de sus parámetros. En la parte superior del modelo se puede observar que la nueva tabla se exporta a un Excel bajo el nombre de INI_(CONPOF), como fue denominado desde su concepción. Esta tabla se utiliza para correr el algoritmo de clusterización + inducción, ahorrando así, el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla INI_(CONPOF) es el que se muestra en la figura 17.

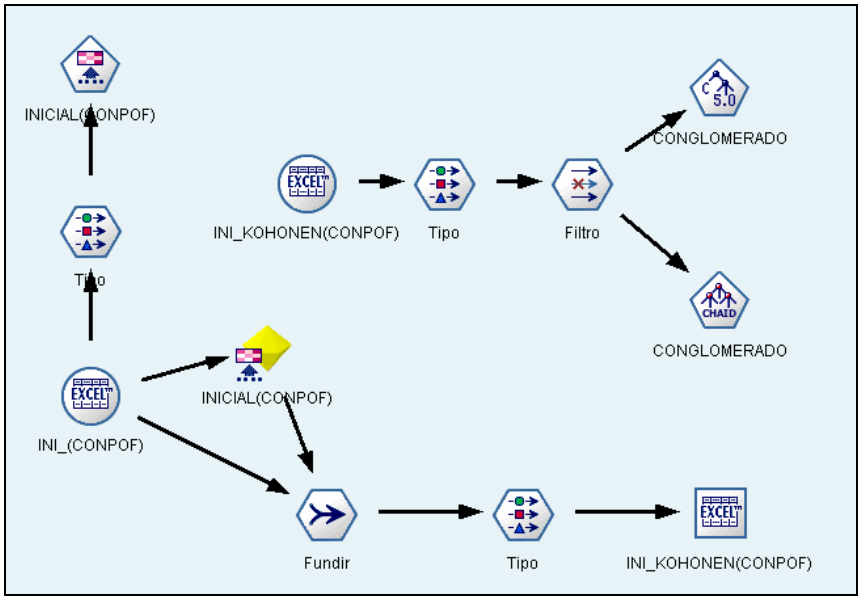

Figura 17. Modelo de clusterización + inducción de la tabla INI_(CONPOF).

El mismo es idéntico al primer modelo de clusterización + inducción del requerimiento, cambiando las tablas de los nodos orígenes. El ajuste de parámetros de los nodos restantes es el mismo que se realiza en el primer modelo de clusterización + inducción del requerimiento. Con esto se concluye la modelización de los datos correspondientes a la tabla INI_(CONPOF) y a requerimiento.

Requerimiento # 2

Este requerimiento, correspondiente al nivel Primario/EGB, se lo estudia utilizando dos modelos de análisis diferentes para la obtención de reglas de decisión. Uno se realiza aplicando solamente inducción, mientras que en el segundo se aplica clusterización y sobre los grupos formados, inducción. Para el modelo de inducción se obtuvieron reglas seleccionando las variables ÁMBITO, y posteriormente, SECTOR como salidas, con lo cual este modelo se divide en dos evaluaciones diferentes.

Cada uno de los tres (3) modelos (inducción ámbito, inducción sector, clusterización + inducción) se corren sobre una (1) tabla fundida (PEGB_(CONPOF)) que se obtiene a partir de las tablas iniciales señaladas en la Fase III para este requerimiento. En resumen, este requerimiento es resuelto por tres (3) modelos aplicados a una (1) tabla, totalizando tres (3) ejecuciones, y por ende, tres (3) salidas diferentes.

La tabla PEGB_(CONPOF) esta formada por los campos de las tablas MAE2005, MATPE2005 y CARPE2005, con lo cual el campo clave por el cual se funden las tablas es el ID_RA. El modelo que se observa en la figura 18, crea la tabla PEGB_(CONPOF) y obtiene resultados por inducción.

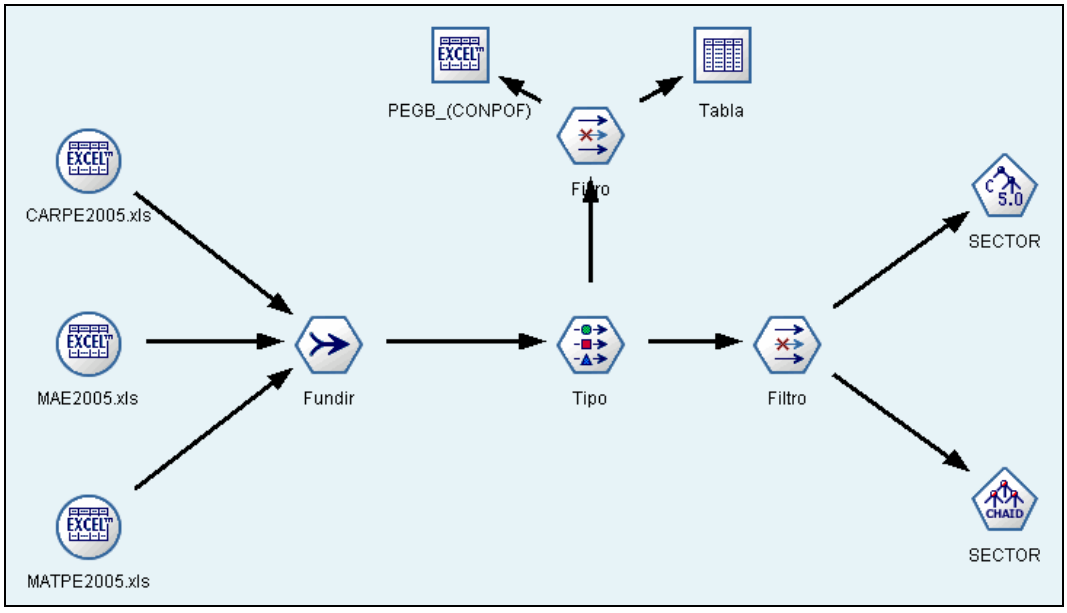

Figura 18. Modelo inductivo de la tabla PEGB_(CONPOF), con el SECTOR como salida.

El ajuste de parámetros es el mismo que se realiza en el primer modelo inductivo del requerimiento # 1. Los campos que forman parte de la nueva tabla (fundida) son los siguientes:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- DEPARTAMENTO
- TIPO_SE
- MAT_NIV_TIPO
- MAT_NIV_TIPO_(1-9)
- REP_NIV_TIPO
- REP_NIV_TIPO_(1-9)
- POF

Las salidas del modelo explicado son obtenidas por C.5 y CHAID. En la parte superior del modelo se puede observar como la nueva tabla se exporta a un Excel bajo el nombre de PEGB_(CONPOF). Esta tabla se utiliza para correr el algoritmo de clusterización + inducción, ahorrando así, el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla PEGB_(CONPOF) es la que se muestra en la figura 19.

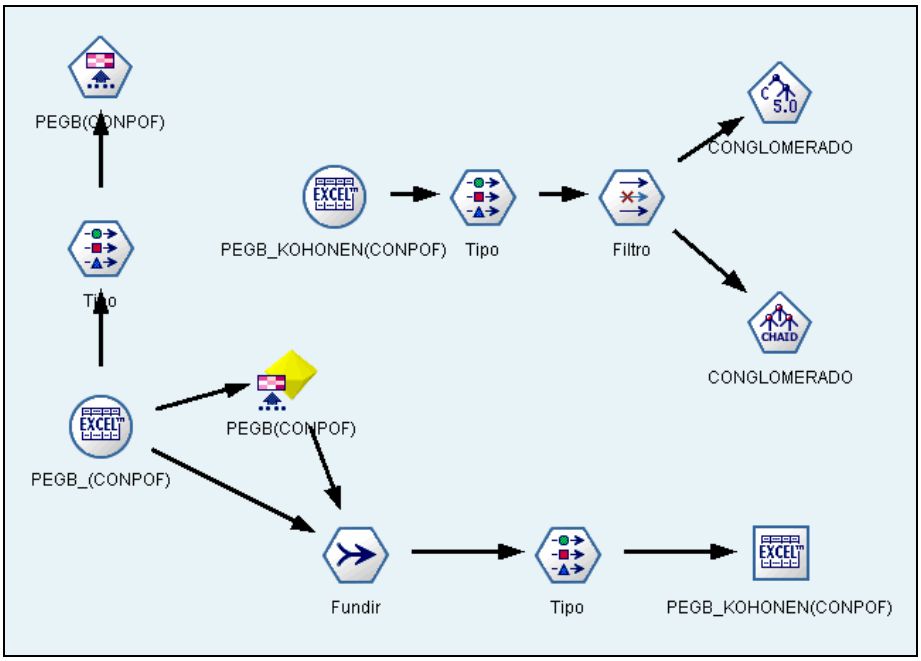

Figura 19. Modelo de clusterización + inducción de la tabla PEGB_(CONPOF).

El mismo es idéntico al primer modelo de clusterización + inducción del requerimiento # 1, cambiando las tablas de los nodos orígenes. El ajuste de parámetros de los nodos restantes es el mismo que se realiza en el primer modelo de clusterización + inducción del requerimiento # 1. Con esto se concluye la modelización de los datos correspondientes a la tabla PEGB_(CONPOF) y al requerimiento.

Requerimiento # 3

Este requerimiento, correspondiente al nivel Primario/EGB, se lo estudia a partir de los resultados obtenidos de los dos modelos de análisis del requerimiento # 2. El requerimiento se focaliza en el comportamiento de la POF de este nivel. Dado que los dos modelos anteriores contienen esa información, el resultado de dicho requerimiento se obtendrá de la ejecución de estos. Con lo cual, obtenidas las reglas de decisión para el requerimiento # 2, se buscan las que resuelvan el requerimiento # 3.

Requerimiento # 4

Este requerimiento, correspondiente al nivel Medio/Polimodal, se lo estudia utilizando dos modelos de análisis diferentes para la obtención de las reglas de decisión. Uno se realiza aplicando solamente inducción, mientras que en el segundo se aplica clusterización y sobre los grupos formados, inducción. Para el modelo de inducción se obtuvieron reglas seleccionando las variables ÁMBITO, y posteriormente, SECTOR como salidas, con lo cual este modelo se divide en dos evaluaciones diferentes.

En cada uno de los tres (3) modelos (inducción ámbito, inducción sector, clusterización + inducción) se corren sobre una (1) tabla fundida (MP) que se obtienen a partir de las tablas iniciales señaladas en la Fase III para este requerimiento. En resumen, este requerimiento es resuelto por tres (3) modelos aplicados a una (1) tabla, totalizando tres (3) ejecuciones, y por ende, tres (3) salidas diferentes.

La tabla MP esta formada por los campos de las tablas MAE2005, MMP2005 y EMP2005. El modelo que se observa en la figura 20, crea la tabla MP y obtiene resultados por inducción.

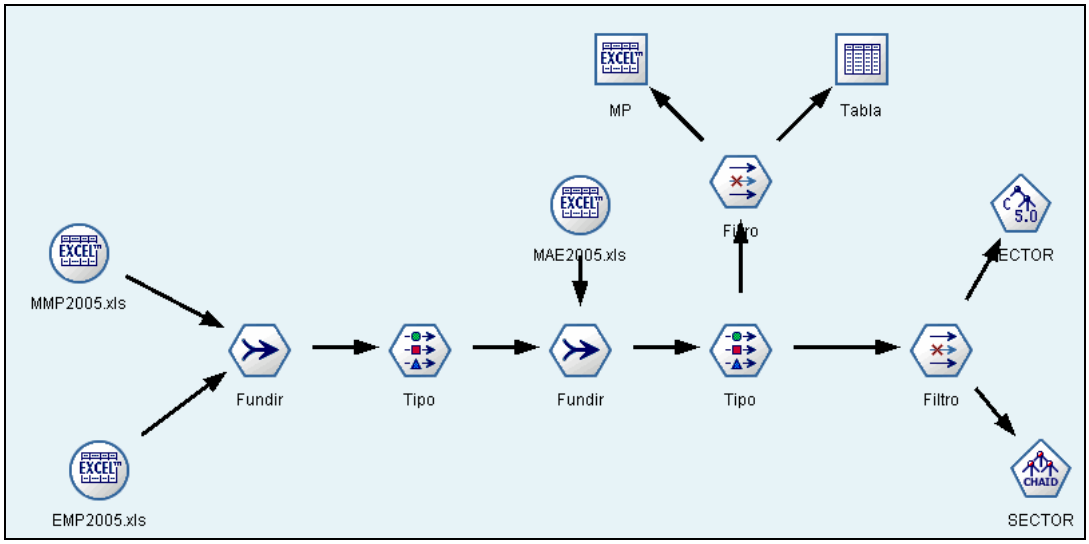

Figura 20. Modelo inductivo de la tabla MP, con el SECTOR como salida.

La razón por la cual MAE2005 no se funde en el mismo nodo que las dos restantes, es que esta no posee el campo NIVEL ni MODALIDAD. La primera fundición se realiza por el ID_RA, NIVEL y MODALIDAD. Por lo tanto, para el setéo de la primera fundición, los campos claves son los que se muestran en la figura 21. Mientras que para la segunda, el ajuste se muestra en la figura 22. Los ajustes de los parámetros de los nodos restantes son los mismos que se realizan en el primer modelo inductivo del requerimiento # 1.

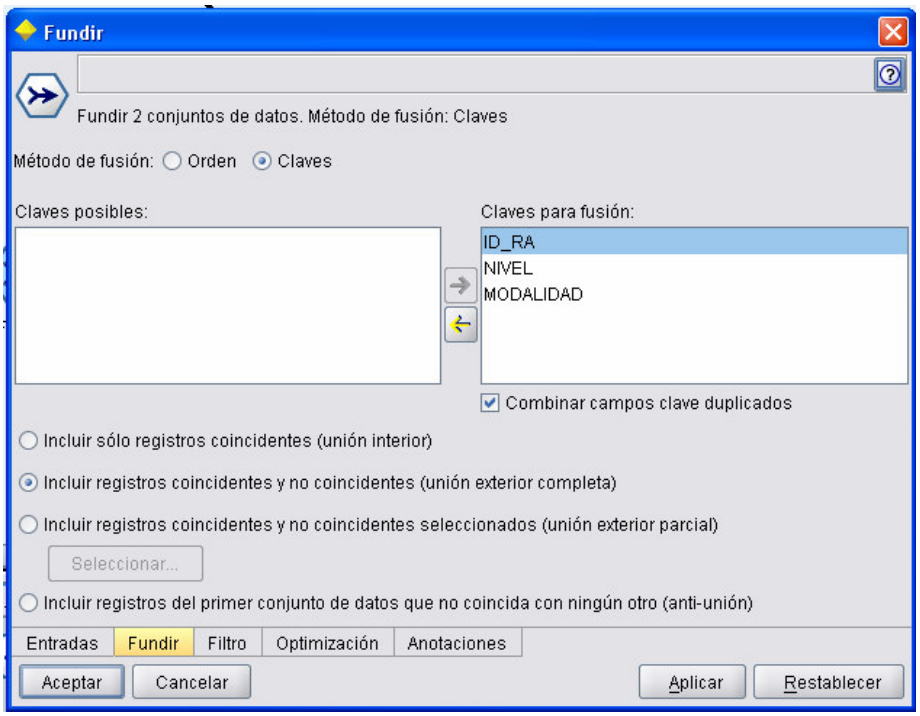

Figura 21. Setéo de los parámetros de fundición. Nodo Fundir, solapa Fundir.

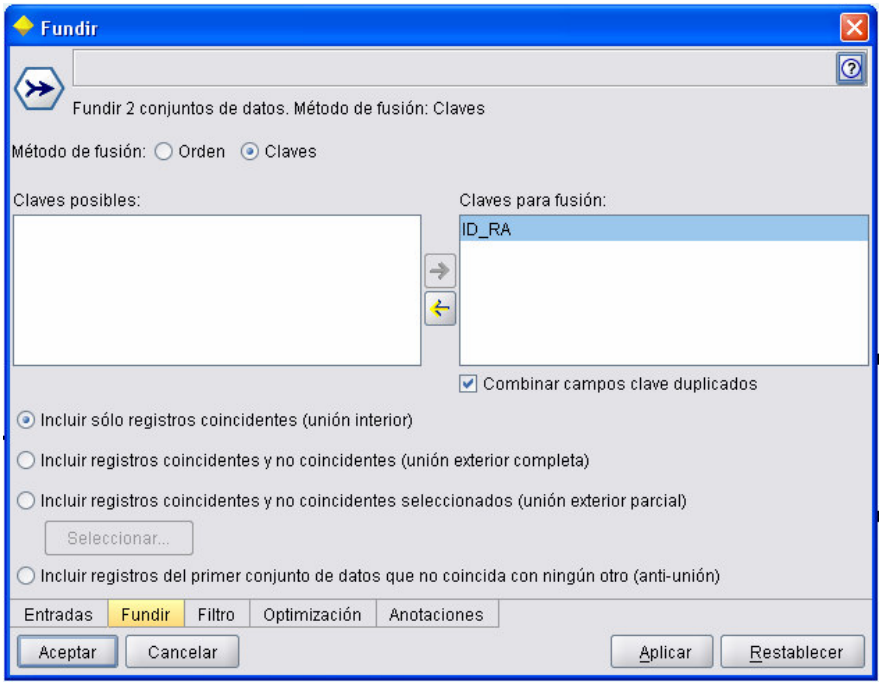

Figura 22. Setéo de los parámetros de fundición. Nodo Fundir, solapa Fundir.

Los campos que forman la nueva tabla (fundida) son:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- DEPARTAMENTO
- MODALIDAD
- EGRE_NIV_MOD
- MAT_NIV_MOD

Las salidas del modelo explicado son obtenidas por C.5 y CHAID. En la parte superior del modelo se puede observar que la nueva tabla se exporta a un Excel bajo el nombre de MP, como fue denominada desde su concepción. Esta tabla se utiliza para correr el algoritmo de clusterización + inducción, ahorrando así, el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla MP es la que se muestra en la figura 23.

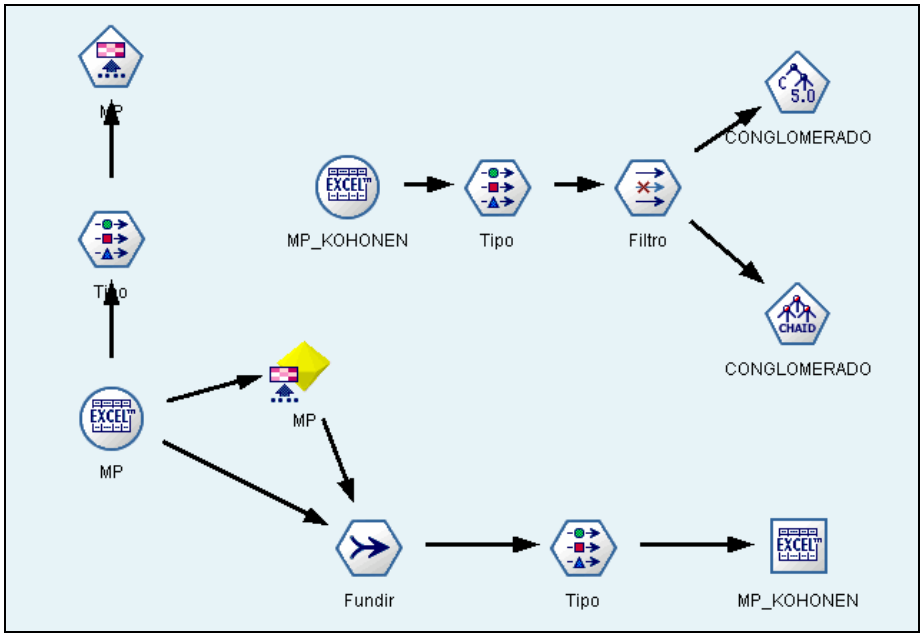

Figura 23. Modelo de clusterización + inducción de la tabla MP.

El mismo es idéntico al modelo de clusterización + inducción del requerimiento # 1, cambiando las tablas de los nodos orígenes. El ajuste de parámetros de los nodos restantes es el mismo que se realiza en el modelo de clusterización + inducción del requerimiento # 1. Con esto se concluye la modelización de los datos correspondientes a la tabla MP y al requerimiento.

Requerimiento # 5

Este requerimiento, correspondiente al nivel Medio/Polimodal, se lo estudia utilizando dos modelos de análisis diferentes para la obtención de las reglas de decisión. Uno se realiza aplicando solamente inducción, mientras que en el segundo se aplica clusterización y sobre los grupos formados, inducción. Para el modelo de inducción se obtuvieron reglas seleccionando a las variables ÁMBITO, y posteriormente, SECTOR como salidas, con lo cual este modelo se divide en dos evaluaciones diferentes.

En cada uno de los tres (3) modelos (inducción ámbito, inducción sector, clusterización + inducción) se corren sobre una (1) tabla fundida (MP_(CONPOF)) que se obtienen a partir de las tablas iniciales señaladas en la Fase III para este requerimiento. En resumen, este requerimiento es resuelto por tres (3) modelos aplicados a una (1) tabla, totalizando tres (3) ejecuciones, y por ende, tres (3) salidas diferentes.

La tabla MP_(CONPOF) esta formada por los campos de las tablas MAE2005, CARMP2005 y MATMP2005. Con lo cual el campo clave por el cual se funden las tablas es el ID_RA. El modelo que se observa en la figura 24, crea la tabla MP_(CONPOF) y obtiene resultados por inducción.

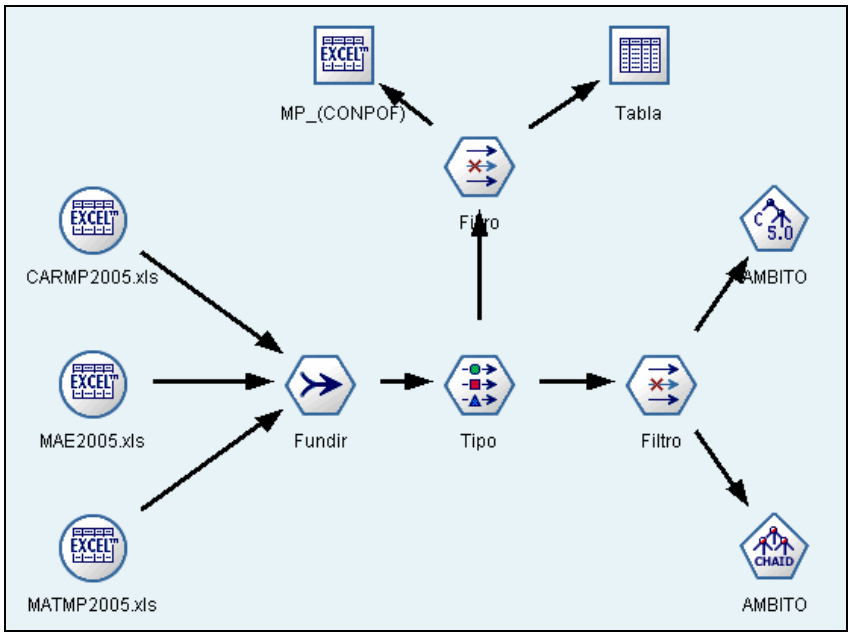

Figura 24. Modelo inductivo de la tabla MP_(CONPOF), con el AMBITO como salida.

El ajuste de parámetros es el mismo que se realiza en el modelo inductivo del requerimiento # 1. Los campos que conforman la nueva tabla (fundida) son:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- DEPARTAMENTO
- TIPO_SE
- MAT_NIV
- REP_NIV
- POF

Las salidas del modelo explicado son obtenidas por C.5 y CHAID, los cuales no requieren comentarios sobre el setéo de sus parámetros. En la parte superior del modelo se puede observar que la nueva tabla se exporta a un Excel bajo el nombre de MP_(CONPOF), como fue denominada desde su concepción. Esta tabla se utiliza para correr el algoritmo de clusterización + inducción, ahorrando así, el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla MP_(CONPOF) es la que se muestra en la figura 25.

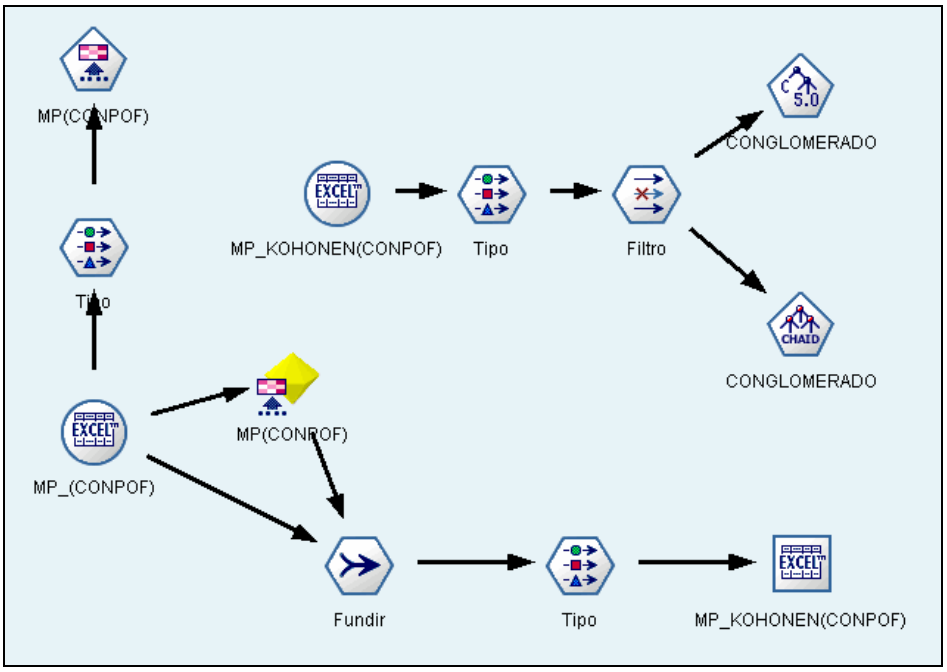

Figura 25. Modelo de clusterización + inducción de la tabla MP_(CONPOF).

El mismo es idéntico al modelo de clusterización + inducción del requerimiento # 1, cambiando las tablas de los nodos orígenes. El ajuste de los parámetros de los nodos restantes es el mismo que se realiza en el primer modelo de clusterización + inducción del requerimiento # 1. Con esto se concluye la modelización de los datos correspondientes a la tabla MP_(CONPOF) y al requerimiento.

Requerimiento # 6

Este requerimiento, correspondiente al nivel Superior No Universitario, se lo estudia utilizando dos modelos de análisis diferentes para la obtención de reglas de decisión. Uno se realiza aplicando solamente inducción, mientras que en el segundo se aplica clusterización y sobre los grupos formados, inducción. Para el modelo de inducción se obtuvieron reglas seleccionando a las variables ÁMBITO, y posteriormente, SECTOR como salidas, con lo cual este modelo se divide en dos evaluaciones diferentes.

En cada uno de los tres (3) modelos (inducción ámbito, inducción sector, clusterización + inducción) se corren sobre una (1) tabla fundida (SNU) que se obtienen a partir de las tablas iniciales señaladas en la Fase III para este requerimiento. En resumen, este requerimiento es resuelto por tres (3) modelos aplicados a una (1) tabla, totalizando tres (3) ejecuciones, y por ende, tres (3) salidas diferentes.

La tabla SNU esta formada por los campos de las tablas MAE2005, ESNU2005 y MSNU2005. El modelo que se observa en la figura 26, crea la tabla SNU y obtiene resultados obtenidos por inducción.

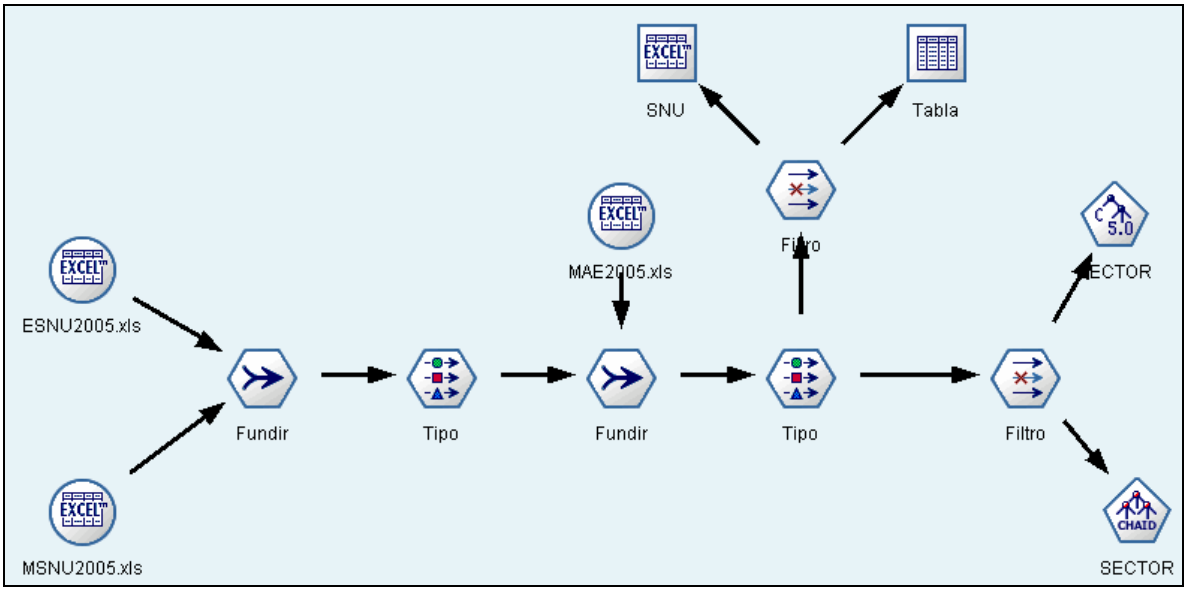

Figura 26. Modelo inductivo de la tabla SNU, con el SECTOR como salida.

La razón por la cual MAE2005 no se funde en el mismo nodo que las dos restantes, es que no posee el campo NIVEL ni CARERRA. La primera fundición se realiza por el ID_RA, NIVEL y CARRERA. El ajuste de parámetros para el modelo inductivo es el mismo que se realiza en el requerimiento # 4, cambiando el campo MODALIDAD por CARRERA como campos claves.

Luego, los campos que forman parte de la nueva tabla (fundida) son los siguientes:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- DEPARTAMENTO
- CARRERA
- TIPOFORMAC
- EGRE_CAR
- MAT_CAR

Las salidas del modelo explicado son obtenidas por C.5 y CHAID, los cuales no requieren comentarios sobre el setéo de sus parámetros. En la parte superior del modelo se puede observar que la nueva tabla se exporta a un Excel bajo el nombre de SNU, como fue denominada desde su concepción. Esta tabla se utiliza para correr el algoritmo de

clusterización + inducción, ahorrando el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla SNU es la que se muestra en la figura 27.

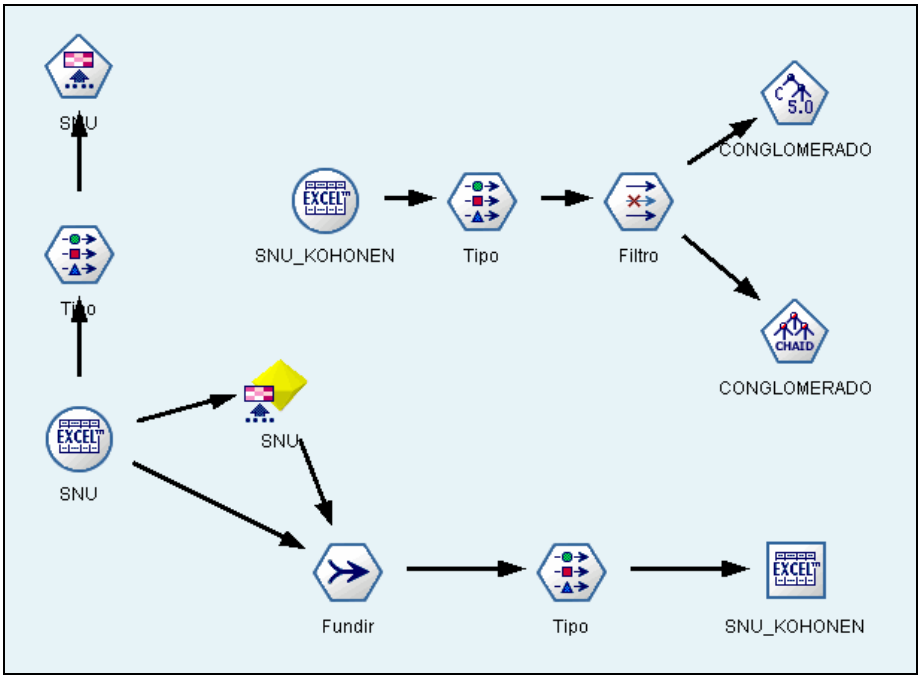

Figura 27. Modelo de clusterización + inducción de la tabla SNU.

El mismo es idéntico al modelo clusterización + inducción del requerimiento # 1, cambiando las tablas de los nodos orígenes. El ajuste de los parámetros de los nodos restantes es el mismo que se realiza en el primer modelo de clusterización + inducción del requerimiento # 1. Con esto se concluye la modelización de los datos correspondientes a la tabla SNU y al requerimiento.

Requerimiento # 7

Este requerimiento, correspondiente al nivel Superior No Universitario, se lo estudia utilizando dos modelos de análisis diferentes para la obtención de reglas de decisión. Uno se realiza aplicando solamente inducción, mientras que en el segundo se aplica clusterización y sobre los grupos formados, inducción. Para el modelo de inducción se obtuvieron reglas seleccionando a las variables ÁMBITO, y posteriormente, SECTOR como salidas, con lo cual este modelo se divide en dos evaluaciones diferentes.

En cada uno de los tres (3) modelos (inducción ámbito, inducción sector, clusterización + inducción) se corren sobre una (1) tabla fundida (SNU_(CONPOF)) que se obtienen a partir de las tablas iniciales señaladas en la Fase III para este requerimiento. En resumen, este

requerimiento es resuelto por tres (3) modelos aplicados a una (1) tabla, totalizando tres (3) ejecuciones, y por ende, tres (3) salidas diferentes.

La tabla SNU_(CONPOF) esta formada por los campos de las tablas MAE2005, CAR2005, EDS2005 y MATSNU2005. El modelo que se observa en la figura 28, crea la tabla SNU_(CONPOF) y obtiene resultados por inducción.

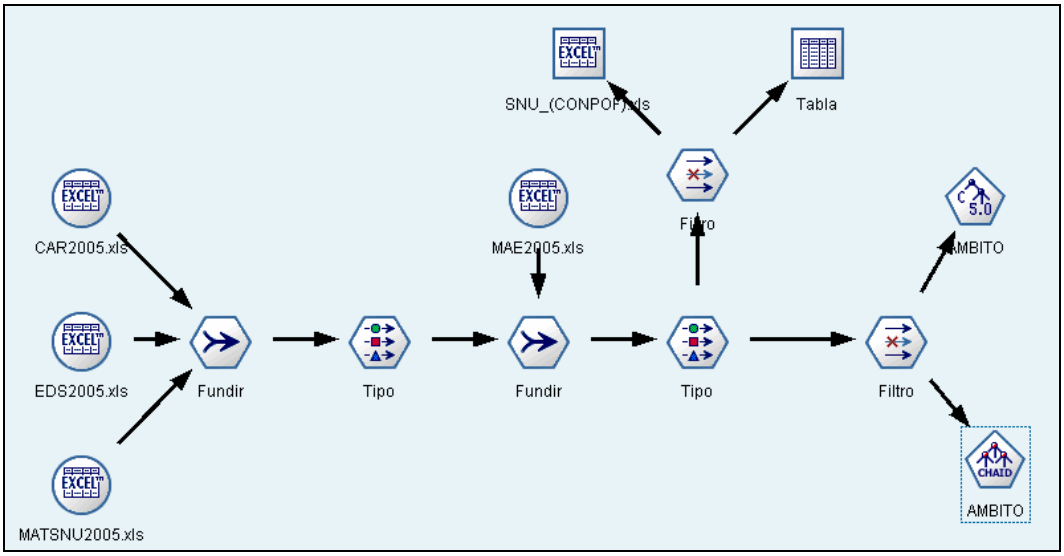

Figura 28. Modelo inductivo de la tabla SNU_(CONPOF), con el AMBITO como salida.

La razón por la cual MAE2005 no se funde en el mismo nodo que las dos restantes, es que no posee el campo NIVEL. La primera fundición se realiza por el ID_RA, NIVEL. El ajuste de parámetros es el mismo que se realiza para segundo modelo inductivo del requerimiento # 1.

Luego, los campos que forman parte de la nueva tabla (fundida) son los siguientes:

- ID_RA
- NIVEL
- PROVINCIA
- SEDE
- AMBITO
- SECTOR
- DEPARTAMENTO
- MAT_NIV
- EDAD_PROM
- POF

Las salidas del modelo explicado son obtenidas por C.5 y CHAID, los cuales no requieren comentarios sobre el setéo de sus parámetros. En la parte superior del modelo se puede observar que la nueva tabla se exporta a un Excel bajo el nombre de SNU_(CONPOF), como fue denominada desde su concepción. Esta tabla se utiliza para correr el algoritmo de clusterización + inducción, ahorrando el desarrollo anterior. Con lo cual, la estructura del modelo de clusterización + inducción para la tabla SNU_(CONPOF) es la que se muestra en la figura 29.

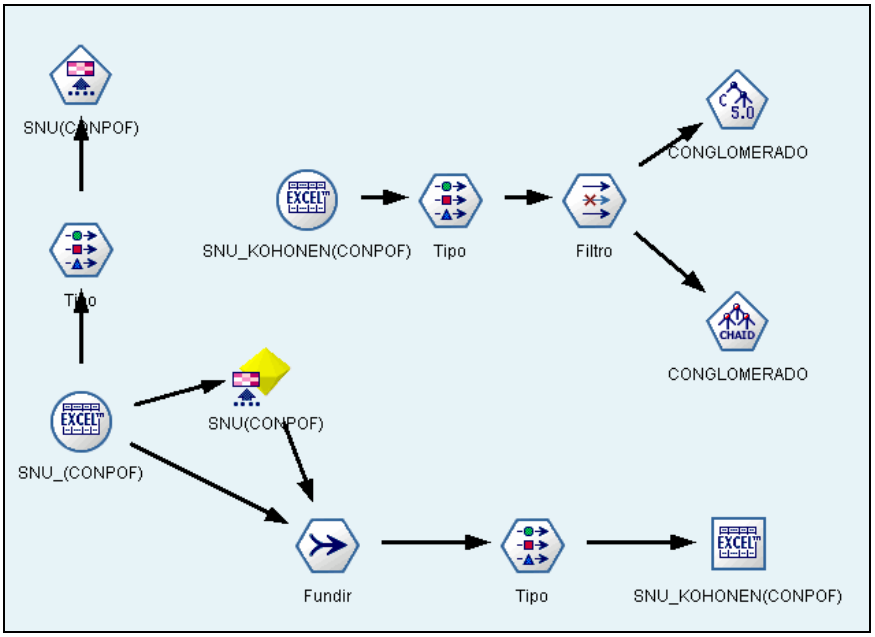

Figura 29. Modelo de clusterización + inducción de la tabla SNU_(CONPOF).

El mismo es idéntico al modelo clusterización + inducción del requerimiento # 1, cambiando las tablas de los nodos orígenes. El ajuste de los parámetros de los nodos restantes es el mismo que se realiza en el primer modelo de clusterización + inducción del requerimiento # 1. Con esto se concluye la modelización de los datos correspondientes a la tabla SNU_(CONPOF) y al requerimiento.

Evaluación del modelo

La evaluación que se puede hacer sobre los modelos anteriores es que estos responden, con éxito, a lo planteado por cada requerimiento. No solo eso, sino que todas las tablas que difunde la DINIECE en el Relevamiento Anual del año 2005, fueron utilizadas en los modelos creados. Con lo cual, estos contienen todos los datos existentes.

En cuanto a los parámetros, si bien los ajustes realizados son coherentes, se podrían setear de diferente manera, entregando diferentes resultados. Es importante aclara que en muchos de los nodos existen diversos parámetros para “jugar” con el modelo, sobre todo en las técnicas de salida (KOHONEN, CHAID, C.5, KMEANS). Pero también es importante saber que se buscó una homogeneidad en los ajustes de cada nodo, ya que de lo contrario, las condiciones en las que se hubieran obtenido los resultados de los requerimientos, serían poco claras.

3.5. Fase V: Resultados

Evaluación de los resultados

Evaluación respecto a los factores de éxito.

De acuerdo a lo planteado en la Fase I sobre los factores críticos de éxito mas significativos del proyecto, los resultados obtenidos y su posterior análisis reflejan en que medida fueron cumplidos. Con respecto a la utilización de todos los datos provistos por las bases, el resultado es favorable. Si bien la información difundida por la DINIECE presenta algunos registros vacíos, con formatos diferentes o campos que no se conoce su significado, se pudieron utilizar los campos disponibles para analizar todos los requerimientos y plantear una resolución para cada uno de ellos.

En cuanto a la herramienta de modelado y ejecución, si bien en toda aplicación existe un periodo de aprendizaje, las virtudes denotadas son reales y fueron de gran ayuda en la resolución de los requerimientos. Sin ir mas lejos, en la creación de los modelos se puede observa la simplicidad a la hora de su construcción y setéo de los parámetros.

Finalmente, la tarea del experto en educación, si bien hasta el momento no puede ser evaluada, se adelanta que fue un pilar fundamental para la comprensión de los comportamientos que se observan en la siguiente sección.

Evaluación de los resultados

Se observan, a continuación, los resultados obtenidos para cada requerimiento. Se recuerda que sus conclusiones se encuentran en la última sección del documento. Todas las reglas de decisión que se obtienen de la aplicación muestran dos valores numéricos que se pueden observar junto a cada una de ellas. El primero de ellos es el “peso”, esto es, la cantidad de registros que conforman la regla. El segundo valor es el “nivel de confianza” de la regla. En la figura 30 se puede observar la ubicación de estos valores.

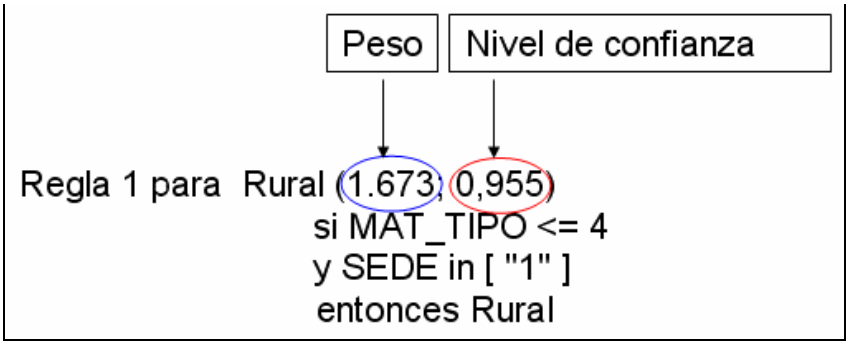

Figura 30. Formato de salidas de las reglas de decisión.

Para sesgar las salidas, se exige que las reglas obtenidas tengan una confianza mayor al 75%.

Requerimiento # 1

Este requerimiento posee seis (6) ejecuciones y, por ende, seis (6) salidas.

Se aplica el algoritmo de inducción conocido como CHAID a la tabla INI y se busca como variable de salida el ÁMBITO de las escuelas iniciales. Esta tabla contiene una cantidad de 20.000 registros. Las reglas son:

1. Como se puede observar, la cota en la que se encuentra la mayor cantidad de escuelas rurales se muestra en las cuatro primeras reglas (0 a 12 alumnos). Las reglas 1 y 2, se complementan en cuanto a la sede. Se puede observar una mayor cantidad de escuelas con una única sede, lo cual es lógico para el ámbito rural. Por otra parte, las reglas 3, 4 y 5 también se complementan, aunque esta vez por el tipo de sección y por provincias. En este caso se puede detectar una mayor cantidad de escuelas rurales con un tipo de sección múltiple, con lo cual, los alumnos comparten las actividades de enseñanza correspondientes a varios años de estudio, independientemente de su edad.

```
Regla 1 para  Rural (1.673; 0,955)
        si MAT_TIPO <= 4
        y SEDE in [ "1" ]
        entonces Rural
Regla 2 para  Rural (256; 0,879)
        si MAT_TIPO <= 4
        y SEDE in [ "2" ]
        entonces Rural
```

Regla 3 para Rural (367; 0,752)
si MAT_TIPO > 4
y MAT_TIPO <= 12
y TIPO_SE in ["I"]
entonces Rural
Regla 4 para Rural (1.171; 0,948)
si MAT_TIPO > 4
y MAT_TIPO <= 12
y TIPO_SE in ["M"]
y PROVINCIA in ["BUENOS AIRES" "CHACO" "CHUBUT" "CORDOBA" "CORRIENTES" "JUJUY" "LA PAMPA" "MENDOZA" "MISIONES" "SALTA" "SAN JUAN" "SAN LUIS"]
entonces Rural
Regla 5 para Rural (566; 0,857)
si MAT_TIPO > 4
y MAT_TIPO <= 12
y TIPO_SE in ["M"]
y PROVINCIA in ["CATAMARCA" "CIUDAD DE BUENOS AIRES" "ENTRE RIOS" "FORMOSA" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SANTA CRUZ" "SANTA FE" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO"]
entonces Rural

2. Desde doce (12) matriculados totales hasta diecinueve (19), solamente las escuelas del sector estatal de las provincias que se exponen a continuación, siguen perteneciendo al ámbito rural. Por otro lado, las escuelas privadas pertenecen indefectiblemente al ámbito urbano. Las siguientes reglas ratifican lo dicho.

Regla 6 para Rural (385; 0,875)
si MAT_TIPO > 12
y MAT_TIPO <= 19
y SECTOR in ["Estatal"]
y PROVINCIA in ["CATAMARCA" "CHACO" "LA RIOJA" "MENDOZA" "SAN JUAN" "SAN LUIS" "SANTIAGO DEL ESTERO"]
entonces Rural
Regla 7 para Rural (557; 0,948)
si MAT_TIPO > 12
y MAT_TIPO <= 19
y SECTOR in ["Estatal"]
y PROVINCIA in ["CORRIENTES" "JUJUY" "MISIONES" "SALTA" "SANTA CRUZ" "TUCUMAN"]
entonces Rural
Regla 1 para Urbano (232; 0,918)
si MAT_TIPO > 12
y MAT_TIPO <= 19
y SECTOR in ["Privado"]
entonces Urbano

3. El ámbito es incierto para escuelas donde el número de matriculados totales es mayor que diecinueve (19) y menor que veintisiete (27). Solo en pocas provincias se pueden detectar pequeñas diferencias, que fundamentan la formación de las siguientes reglas. Es claro que el peso de estas reglas no es significativamente alto con respecto a la cantidad de registros que contiene la tabla. Es por eso que, en este intervalo de matriculados, no se tiene un panorama claro del ámbito escolar a nivel nacional.

Regla 8 para Rural (247; 0,769)
si MAT_TIPO > 19
y MAT_TIPO <= 27
y PROVINCIA in ["CATAMARCA" "CORRIENTES" "SALTA" "SAN JUAN"]
entonces Rural
Regla 9 para Rural (285; 0,863)
si MAT_TIPO > 19
y MAT_TIPO <= 27
y PROVINCIA in ["SANTA CRUZ" "SANTIAGO DEL ESTERO" "TUCUMAN"]
entonces Rural
Regla 2 para Urbano (376; 0,763)
si MAT_TIPO > 19
y MAT_TIPO <= 27
y PROVINCIA in ["BUENOS AIRES" "CIUDAD DE BUENOS AIRES" "TIERRA DEL FUEGO"]
y SECTOR in ["Estatal"]
entonces Urbano
Regla 3 para Urbano (276; 0,978)
si MAT_TIPO > 19
y MAT_TIPO <= 27
y PROVINCIA in ["BUENOS AIRES" "CIUDAD DE BUENOS AIRES" "TIERRA DEL FUEGO"]
y SECTOR in ["Privado"]
entonces Urbano

Para el intervalo de matriculados que va desde veintiocho (28) hasta cincuenta y ocho (58), tampoco se puede detectar una tendencia nacional definida. Si bien se estima que escuelas con mas de veintiocho (28) alumnos en total son establecimientos ubicados en ciudades, se observa que los pesos de las reglas que avalan dicha tendencia, no son significativos en comparación a los que se muestran a continuación (regla 4, 5 y 6). Es por esto que a partir de cincuenta y ocho (58) matriculados totales se descarta toda posibilidad de encontrar una escuela rural en cualquier sector (privado o estatal). Por lo tanto, en las tres reglas restantes se extraen solamente conclusiones del ámbito urbano.

4. A partir de cincuenta y ocho (58) matriculados hasta ochenta y uno (81), se puede detectar una cantidad pareja de establecimientos privados y estatales. Ahora bien, dentro del sector estatal, a nivel nacional, las escuelas con un tipo de sección independiente prevalecen ampliamente por sobre las de enseñanza múltiple.

```
Regla 11 para  Urbano (940; 0,862)
        si MAT_TIPO > 58
        y MAT_TIPO <= 81
        y SECTOR in [ "Estatal" ]
        y TIPO_SE in [ "I" ]
        entonces Urbano
Regla 12 para  Urbano (257; 0,957)
        si MAT_TIPO > 58
        y MAT_TIPO <= 81
        y SECTOR in [ "Estatal" ]
        y TIPO_SE in [ "M" ]
        entonces Urbano
Regla 13 para  Urbano (786; 0,991)
        si MAT_TIPO > 58
        y MAT_TIPO <= 81
        y SECTOR in [ "Privado" ]
        entonces Urbano
```

5. Para mas de ochenta y un (81) alumnos matriculados, las escuelas estatales duplican a las privadas.

```
Regla 14 para  Urbano (1.333; 0,954)
        si MAT_TIPO > 81
        y MAT_TIPO <= 111
        y SECTOR in [ "Estatal" ]
        entonces Urbano
Regla 15 para  Urbano (670; 0,994)
        si MAT_TIPO > 81
        y MAT_TIPO <= 111
        y SECTOR in [ "Privado" ]
```

6. Y es alrededor de los ciento once (111) alumnos matriculados donde se encuentra la media de las escuelas urbanas para el nivel inicial. Las reglas solo hacen una diferenciación por provincias, pero esto se debe solamente a distintos niveles de confianza que ellas presentan.

Regla 16 para Urbano (1.739; 0,998)
si MAT_TIPO > 111
y PROVINCIA in ["BUENOS AIRES" "CATAMARCA" "CHACO" "CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "CORRIENTES" "ENTRE RIOS" "FORMOSA" "JUJUY" "LA PAMPA" "MISIONES" "NEUQUEN" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "TUCUMAN"]
y MAT_TIPO <= 168
entonces Urbano
Regla 17 para Urbano (1.858; 1,0)
si MAT_TIPO > 111
y PROVINCIA in ["BUENOS AIRES" "CATAMARCA" "CHACO" "CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "CORRIENTES" "ENTRE RIOS" "FORMOSA" "JUJUY" "LA PAMPA" "MISIONES" "NEUQUEN" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "TUCUMAN"]
y MAT_TIPO > 168
entonces Urbano
Regla 18 para Urbano (383; 0,979)
si MAT_TIPO > 111
y PROVINCIA in ["LA RIOJA" "MENDOZA" "RIO NEGRO" "SALTA" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO"]
entonces Urbano

Al cambiar la salida del mismo modelo al SECTOR los resultados encontrados son:

1. Las escuelas rurales a nivel nacional pertenecen al sector estatal. El tipo de sección es múltiple, aunque se detecta un grupo menor de establecimientos con sección independiente.

Regla 1 para Estatal (1.932; 0,965)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y SEDE in ["1"]
entonces Estatal
Regla 2 para Estatal (284; 1,0)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y SEDE in ["2"]
entonces Estatal
Regla 3 para Estatal (362; 0,945)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y SEDE in ["1"]
y PROVINCIA in ["BUENOS AIRES" "FORMOSA"]
entonces Estatal

Regla 4 para Estatal (2.083; 0,999)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y SEDE in ["1"]
y PROVINCIA in ["CHACO" "CHUBUT" "CORDOBA" "CORRIENTES" "JUJUY" "LA PAMPA" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO" "TUCUMAN"]
entonces Estatal

Regla 5 para Estatal (1.708; 0,986)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y SEDE in ["1"]
y PROVINCIA in ["ENTRE RIOS" "MENDOZA" "MISIONES" "SALTA" "SANTA FE"]
entonces Estatal

Regla 6 para Estatal (771; 0,997)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y SEDE in ["2"]
entonces Estatal

2. Las escuelas estatales tienen una gran supremacía en el ámbito urbano en todas las provincias, salvo en la Ciudad de Buenos Aires. Solamente en esta provincia, las escuelas estatales del nivel inicial son superadas levemente por las privadas. La razón por la que se forman varias reglas con distintas provincias, es que estas se agrupan por iguales valores de otras variables. También el nivel de confianza, como se explicó otras veces, divide a reglas que se complementan y que marcan un único patrón o comportamiento.

Regla 7 para Estatal (187; 0,813)
si AMBITO in ["Urbano"]
y PROVINCIA in ["BUENOS AIRES" "TIERRA DEL FUEGO" "TUCUMAN"]
y MAT_TIPO > 58
y MAT_TIPO <= 111
y TIPO_SE in ["M"]
y MAT_TIPO > 81
entonces Estatal

Regla 8 para Estatal (337; 0,837)
si AMBITO in ["Urbano"]
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "FORMOSA" "LA PAMPA" "SAN LUIS"]
y SEDE in ["1"]
y MAT_TIPO > 81
entonces Estatal

Regla 9 para Estatal (332; 0,991)
si AMBITO in ["Urbano"]
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "FORMOSA" "LA PAMPA" "SAN LUIS"]
y SEDE in ["2"]
entonces Estatal
Regla 11 para Estatal (2.525; 0,75)
si AMBITO in ["Urbano"]
y PROVINCIA in ["CORDOBA" "CORRIENTES" "LA RIOJA" "MENDOZA" "MISIONES" "NEUQUEN" "SAN JUAN" "SANTIAGO DEL ESTERO"]
y TIPO_SE in ["I"]
entonces Estatal
Regla 12 para Estatal (297; 0,852)
si AMBITO in ["Urbano"]
y PROVINCIA in ["CORDOBA" "CORRIENTES" "LA RIOJA" "MENDOZA" "MISIONES" "NEUQUEN" "SAN JUAN" "SANTIAGO DEL ESTERO"]
y TIPO_SE in ["M"]
entonces Estatal
Regla 13 para Estatal (221; 0,95)
si AMBITO in ["Urbano"]
y PROVINCIA in ["ENTRE RIOS" "JUJUY" "RIO NEGRO" "SALTA" "SANTA CRUZ" "SANTA FE"]
y SEDE in ["2"]
entonces Estatal

3. La excepción en Capital.

Regla 1 para Privado (319; 0,831)
si AMBITO in ["Urbano"]
y PROVINCIA in ["CIUDAD DE BUENOS AIRES"]
y SEDE in ["1"]
y MAT_TIPO <= 81
entonces Privado
Regla 10 para Estatal (216; 1,0)
si AMBITO in ["Urbano"]
y PROVINCIA in ["CIUDAD DE BUENOS AIRES"]
y SEDE in ["2"]
entonces Estatal

Al aplicar el algoritmo de clusterización junto con el de inducción a la tabla INI, con el CONGLOMERADO como salida, los resultados que se obtienen son:

1. El tipo de sección independiente prevalece fuertemente en las escuelas estatales urbanas. Esta tendencia esta avalada por la regla número 4 del análisis netamente inductivo con la variable ÁMBITO como salida, pero, en este análisis, se ve una

diferencia mas determinante debido al peso de las reglas. Se puede ver que la regla de los grupos 22 y 32 difiere en el tipo de sección con la regla del grupo 00, duplicando en cantidad de registros a este último. También, por medio de la regla del grupo 32, se detecta, para este ámbito y sector, una mayoría de escuelas con una sola sede educativa.

```
Reglas grupo para 00 - contiene 1 regla(s)
        Regla 1 para grupo 00 (2.111; 1,0)
                si SECTOR in [ "Estatal" ]
                y AMBITO in [ "Urbano" ]
                y TIPO_SE in [ "M" ]
                entonces grupo 00
Reglas para grupo 22 - contiene 1 regla(s)
        Regla 1 para grupo 22 (627; 1,0)
                si SECTOR in [ "Estatal" ]
                y AMBITO in [ "Urbano" ]
                y TIPO_SE in [ "I" ]
                y SEDE in [ "2" ]
                y MAT_TIPO <= 111
                entonces grupo 22
Reglas para grupo 32 - contiene 1 regla(s)
        Regla 1 para grupo 32 (5.562; 1,0)
                si SECTOR in [ "Estatal" ]
                y AMBITO in [ "Urbano" ]
                y TIPO_SE in [ "I" ]
                y SEDE in [ "1" ]
                entonces grupo 32
```

2. Las escuelas estatales rurales también se clusterizan por tipo de sección, pero en este caso, la enseñanza múltiple supera a la independiente de acuerdo al peso que se observa en las siguientes reglas. Esta tendencia también se presentó en ambos análisis del modelo inductivo. Un dato interesante es que se observa la cota superior de cincuenta y ocho (58) matriculados totales, anteriormente mencionada para las escuelas rurales. Lógicamente, también se detecta una mayoría de escuelas rurales estatales con una sola sede.

```
Reglas para grupo 02 - contiene 1 regla(s)
        Regla 1 para  grupo 02 (4.107; 1,0)
                si SECTOR in [ "Estatal" ]
                y AMBITO in [ "Rural" ]
                y TIPO_SE in [ "M" ]
                y SEDE in [ "1" ]
                entonces grupo 02
Reglas para grupo 12 - contiene 2 regla(s)
```

```
Regla 1 para  grupo 12 (1.946; 1,0)
        si SECTOR in [ "Estatal" ]
        y AMBITO in [ "Rural" ]
        y TIPO_SE in [ "I" ]
        y MAT_TIPO <= 58
        entonces grupo 12
Regla 2 para  grupo 12 (769; 1,0)
        si SECTOR in [ "Estatal" ]
        y AMBITO in [ "Rural" ]
        y TIPO_SE in [ "M" ]
        y SEDE in [ "2" ]
        entonces grupo 12
```

3. Finalmente se ve como las escuelas privadas con tipo de sección independiente son aproximadamente el triple que las de sección múltiple.

```
Reglas para grupo 10 - contiene 3 regla(s)
Regla 1 para  grupo 10 (118; 0,924)
        si SECTOR in [ "Privado" ]
        y TIPO_SE in [ "M" ]
        y MAT_TIPO <= 42
        y MAT_TIPO > 12
        y MAT_TIPO <= 19
        entonces 10
Regla 2 para  grupo 10 (478; 0,973)
        si SECTOR in [ "Privado" ]
        y TIPO_SE in [ "M" ]
        y MAT_TIPO <= 42
        y MAT_TIPO > 12
        y MAT_TIPO > 19
        entonces grupo 10
Regla 3 para  grupo 10 (338; 0,997)
        si SECTOR in [ "Privado" ]
        y TIPO_SE in [ "M" ]
        y MAT_TIPO > 42
        entonces grupo 10
Reglas para grupo 30 - contiene 1 regla(s)
Regla 1 para  grupo 30 (3.639; 1,0)
        si SECTOR in [ "Privado" ]
        y TIPO_SE in [ "I" ]
        entonces grupo 30
```

La tabla INI_(CONPOF), la cual contiene 15.900 registros aproximadamente, posee, valga la renuncia, el campo “POF”. Cabe destacar que para el análisis de las siguientes reglas, se obtuvo por medio de cálculos que la edad promedio para el nivel inicial es de 4,42 años (4

años y 5 meses). Este dato ayuda al análisis del campo EDAD_PROM. Se comienza aplicando el modelo CHAID, netamente inductivo, y obteniendo como salida a la variable ÁMBITO. La única regla obtenida es la siguiente:

1. Como se obtuvo en reglas anteriores, las escuelas privadas de este nivel se ubican únicamente en ciudades. En estas escuelas se detecta una edad promedio superior a la nacional. A su vez, se observa una mayor cantidad de establecimientos privados que se ajustan al presupuesto designado (POF dentro), sin dejar de tener en cuenta el significativo grupo de escuelas que no lo hacen (POF fuera).

```
Regla 7 para  Urbano (2.742; 0,985)
        si SECTOR in [ "Privado" ]
        y EDAD_PROM <= 4.45
        y POF in [ "dentro" ]
        entonces Urbano
Regla 8 para  Urbano (723; 0,997)
        si SECTOR in [ "Privado" ]
        y EDAD_PROM <= 4.45
        y POF in [ "fuera" ]
        entonces Urbano
Regla 9 para  Urbano (967; 0,97)
        si SECTOR in [ "Privado" ]
        y EDAD_PROM > 4.45
        y EDAD_PROM <= 4.8
        entonces Urbano
Regla 10 para  Urbano (390; 0,936)
        si SECTOR in [ "Privado" ]
        y EDAD_PROM > 4.8
        entonces Urbano
```

Si se ubica al SECTOR como salida, las reglas son las siguientes:

1. El intervalo de edades con mayor cantidad de registros en este sector es el que indica la regla 5. Este intervalo es el mismo que se muestra en la única regla de la corrida anterior. Todas estas escuelas poseen una POF fuera.

```
Regla 2 para  Privado (159; 0,912)
        si AMBITO in [ "Urbano" ]
        y EDAD_PROM > 3.98473282442748
        y EDAD_PROM <= 4.06569343065693
        y POF in [ "fuera" ]
        entonces Privado
```

Regla 4 para Privado (480; 0,879)
si AMBITO in ["Urbano"]
y EDAD_PROM > 4.06569343065693
y EDAD_PROM <= 4.45
y POF in ["fuera"]
entonces Privado
Regla 5 para Privado (206; 0,85)
si AMBITO in ["Urbano"]
y EDAD_PROM > 4.45
y EDAD_PROM <= 4.63559322033898
y POF in ["fuera"]
entonces Privado

2. Las escuelas rurales estatales se encuentran con un promedio de edad superior a la media nacional.

Regla 1 para Estatal (752; 0,955)
si AMBITO in ["Rural"]
y SEDE in ["1"]
y EDAD_PROM <= 4.27586206896552
entonces Estatal
Regla 2 para Estatal (3.036; 0,979)
si AMBITO in ["Rural"]
y SEDE in ["1"]
y EDAD_PROM > 4.27586206896552
entonces Estatal
Regla 3 para Estatal (736; 1,0)
si AMBITO in ["Rural"]
y SEDE in ["2"]
entonces Estatal

Finalmente, al aplicar el modelo de clusterización + inducción se encontraron los siguientes resultados:

1. Las escuelas estatales urbanas a nivel nacional se ajustan al presupuesto en horas, cargos y módulo que les otorga el gobierno o la institución destinada a tal fin. Esto se ve claramente, debido a la amplia superioridad de escuelas que cumplen con la regla del grupo 00, opuesta en la POF con la regla del grupo 10. No es novedad que dichas escuelas presenten una sola sede educativa.

```
Reglas para grupo 00 - contiene 1 regla(s)
        Regla 1 para  00 (5.966; 1,0)
                si AMBITO in [ "Urbano" ]
                y SECTOR in [ "Estatal" ]
                y SEDE in [ "1" ]
                y POF in [ "dentro" ]
                entonces grupo 00
Reglas para grupo 10 - contiene 1 regla(s)
        Regla 1 para  10 (159; 1,0)
                si AMBITO in [ "Urbano" ]
                y SECTOR in [ "Estatal" ]
                y SEDE in [ "1" ]
                y POF in [ "fuera" ]
                entonces grupo 10
```

2. En el aspecto anteriormente analizado, las escuelas privadas urbanas tiene un comportamiento similar. Sin embargo, existe un nicho más importante y significativo de escuelas con un presupuesto *fuera* del estimado. Se puede decir que un cuarto de los establecimientos educativos nacionales sobrepasan dicho presupuesto. Esta tendencia también se expone en la primera regla netamente inductiva, y son las siguientes reglas la que la refuerzan:

```
Reglas para grupo 20 - contiene 1 regla(s)
        Regla 1 para  20 (997; 1,0)
                si AMBITO in [ "Urbano" ]
                y SECTOR in [ "Privado" ]
                y POF in [ "fuera" ]
                entonces grupo 20
Reglas para grupo 30 - contiene 1 regla(s)
        Regla 1 para  30 (3.728; 1,0)
                si AMBITO in [ "Urbano" ]
                y SECTOR in [ "Privado" ]
                y POF in [ "dentro" ]
                entonces grupo 30
```

3. En las escuelas rurales se detecta una tendencia en cuanto a la edad promedio de los alumnos del nivel inicial. Como se puede observar dentro del grupo 32, la regla 2 tiene casi el triple de peso que la regla 1. Con lo cual, la edad promedio de los alumnos estará por encima de 4,3 años. Analizando la tabla con mas detalle se vio que existe, a nivel nacional, un nivel bajo de niños con 3 años de edad o menos, que asiste a clase. Este nivel de alumnos se cuadriplica luego de 2 años biológicos, donde ingresan, sí o sí, a sala de 5 años (tener presente que muchos alumnos ingresan a sala de 5 teniendo menos de 5 años). Por lo tanto, del total de niños de

zonas rurales, más de la mitad comienza su formación en sala de 5 años. Esta tendencia también se señala en la segunda regla inductiva con el SECTOR como salida.

Reglas para grupo 32 - contiene 2 regla(s)
Regla 1 para 32 (752; 0,955)
si AMBITO in ["Rural"]
y SEDE in ["1"]
y EDAD_PROM <= 4.27586206896552
entonces grupo 32
Regla 2 para grupo 32 (3.036; 0,979)
si AMBITO in ["Rural"]
y SEDE in ["1"]
y EDAD_PROM > 4.27586206896552
entonces grupo 32

Requerimiento # 2

La tabla PEGB_(CONPOF) contiene 35.500 registros aproximadamente. Se comienza con el análisis netamente inductivo con el modelo CHAID, y se coloca como salida al campo SECTOR. Se detectan veintisiete (27) reglas para el sector estatal y cuatro (4) para el privado. Por defecto, la cantidad de escuelas primarias estatales superan a las privadas. Sin embargo, se debe tener en cuenta que las veintisiete (27) reglas no describen veintisiete (27) diferentes comportamientos, sino que muchas de ellas se dividen por no compartir iguales niveles de confianza. Esto ocurre generalmente cuando el modelo divide a un mismo comportamiento por provincias, y al verificar con mas detalle las reglas, se observa que todas las provincias lo poseen.

1. El primer comportamiento que se detecta se refiere a las escuelas del ámbito rural. Antes que nada, se aclara que no se observan reglas que relacionen a escuelas de este ámbito con el sector privado, con lo cual, todas serán estatales. Del análisis se obtiene que las escuelas de sección múltiple superan ampliamente a las de tipo independiente a nivel nacional. En estas (las múltiple) se puede detectar un nivel de repetidos muy bajo, ya que la cota que divide reglas complementarias es de 0 repetidos (regla 10 y 11). Los matriculados también son bajos debido a que la cota es de 15 alumnos y las reglas que hacen esta distinción no difieren fuertemente en el peso que poseen. Como se explicó anteriormente, esta tendencia agrupa varias reglas por su división por provincias. La única provincia que no aparece en las reglas por no tener escuelas rurales es la Ciudad de Buenos Aires.

Regla 9 para Estatal (2.304; 0,998)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y PROVINCIA in ["BUENOS AIRES" "CORDOBA" "NEUQUEN" "SALTA" "SANTA FE" "TUCUMAN"]
y MAT_NIV_TIPO <= 15
entonces Estatal
Regla 10 para Estatal (903; 0,984)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y PROVINCIA in ["BUENOS AIRES" "CORDOBA" "NEUQUEN" "SALTA" "SANTA FE" "TUCUMAN"]
y MAT_NIV_TIPO > 15
y REP_NIV_TIPO <= 0
entonces Estatal
Regla 11 para Estatal (1.760; 0,997)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y PROVINCIA in ["BUENOS AIRES" "CORDOBA" "NEUQUEN" "SALTA" "SANTA FE" "TUCUMAN"]
y MAT_NIV_TIPO > 15
y REP_NIV_TIPO > 0
entonces Estatal
Regla 12 para Estatal (6.244; 0,999)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "CORRIENTES" "FORMOSA" "JUJUY" "LA PAMPA" "LA RIOJA" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO"]
entonces Estatal
Regla 13 para Estatal (1.766; 0,986)
si AMBITO in ["Rural"]
y TIPO_SE in ["M"]
y PROVINCIA in ["ENTRE RIOS" "MENDOZA" "MISIONES" "SANTA CRUZ"]
entonces Estatal

2. La cantidad de escuelas estatales rurales con tipo de sección independiente es menor que la múltiple, ya que el peso de las reglas lo demuestra en cada una de las provincias. Un patrón que se observa para escuelas de sección independiente es el aumento de la cota divisoria de alumnos repetidos como también de alumnos matriculados. La posible razón por la cual la cota de alumnos repetidos aumenta se debe a una mayor cantidad de matriculados. En esta tendencia, también se encuentran todas las provincias menos la Capital Federal.

Regla 1 para Estatal (546; 0,985)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y PROVINCIA in ["BUENOS AIRES" "CHACO" "CORDOBA" "MENDOZA" "SALTA" "SANTIAGO DEL ESTERO" "TUCUMAN"]
y REP_NIV_TIPO <= 6
y MAT_NIV_TIPO <= 43
entonces Estatal
Regla 2 para Estatal (441; 0,925)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y PROVINCIA in ["BUENOS AIRES" "CHACO" "CORDOBA" "MENDOZA" "SALTA" "SANTIAGO DEL ESTERO" "TUCUMAN"]
y REP_NIV_TIPO <= 6
y MAT_NIV_TIPO > 43
entonces Estatal
Regla 3 para Estatal (1.049; 0,985)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y PROVINCIA in ["BUENOS AIRES" "CHACO" "CORDOBA" "MENDOZA" "SALTA" "SANTIAGO DEL ESTERO" "TUCUMAN"]
y REP_NIV_TIPO > 6
entonces Estatal
Regla 4 para Estatal (992; 0,994)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y PROVINCIA in ["CATAMARCA" "CHUBUT" "FORMOSA" "JUJUY" "LA PAMPA" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO"]
entonces Estatal
Regla 5 para Estatal (259; 0,988)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y PROVINCIA in ["CORRIENTES" "ENTRE RIOS" "LA RIOJA" "MISIONES" "NEUQUEN"]
y MAT_NIV_TIPO <= 27
entonces Estatal
Regla 6 para Estatal (314; 0,876)
si AMBITO in ["Rural"]
y TIPO_SE in ["I"]
y PROVINCIA in ["CORRIENTES" "ENTRE RIOS" "LA RIOJA" "MISIONES" "NEUQUEN"]
y MAT_NIV_TIPO > 27
y REP_NIV_TIPO <= 6
entonces Estatal

```
Regla 7 para  Estatal (576; 0,951)
        si AMBITO in [ "Rural" ]
        y TIPO_SE in [ "I" ]
        y PROVINCIA in [ "CORRIENTES" "ENTRE RIOS" "LA RIOJA" "MISIONES" "NEUQUEN" ]
        y MAT_NIV_TIPO > 27
        y REP_NIV_TIPO > 6
        entonces Estatal
Regla 8 para  Estatal (356; 0,888)
        si AMBITO in [ "Rural" ]
        y TIPO_SE in [ "I" ]
        y PROVINCIA in [ "SANTA FE" ]
        entonces Estatal
```

Las reglas restantes de este modelo pertenecen al ámbito urbano. Según se puede detectar, la cantidad de alumnos repetidos totales de un establecimiento es el dato más influyente a la hora de decidir si la escuela es estatal o privada. No solo eso, sino que, dentro de un mismo sector, las reglas se dividen según la cantidad de repetidos. Por supuesto que los campos restantes varían sus rangos o valores, pero es indudable que los repetidos totales tienen un peso importante en la división de las reglas. Es por eso que, para este ámbito, se comienza el análisis con las reglas que poseen menor cantidad de repetidos.

3. Para el rango de cero (0) a seis (6) repetidos totales anuales, las escuelas privadas superan a las estatales. La cantidad de matriculados totales parece no ser un factor que afecte a la cantidad de repetidos en el sector privado. Las siguientes reglas avalan este comportamiento, ya que muestran altos pesos para una cantidad mayor de ciento cuarenta y nueve (149) matriculados en varias provincias.

```
Regla 1 para  Privado (716; 0,802)
        si AMBITO in [ "Urbano" ]
        y REP_NIV_TIPO <= 3
        y MAT_NIV_TIPO > 149
        y PROVINCIA in [ "BUENOS AIRES" "CATAMARCA" "ENTRE RIOS" "MENDOZA" "SALTA" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO" "TUCUMAN" ]
        y REP_NIV_TIPO <= 0
        entonces Privado
Regla 2 para  Privado (602; 0,934)
        si AMBITO in [ "Urbano" ]
        y REP_NIV_TIPO <= 3
        y MAT_NIV_TIPO > 149
        y PROVINCIA in [ "BUENOS AIRES" "CATAMARCA" "ENTRE RIOS" "MENDOZA" "SALTA" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO" "TUCUMAN" ]
        y REP_NIV_TIPO > 0
        entonces Privado
```

Regla 3 para Privado (638; 0,782)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO <= 3
y MAT_NIV_TIPO > 149
y PROVINCIA in ["CORDOBA" "SANTA FE" "SANTIAGO DEL ESTERO"]
entonces Privado
Regla 4 para Privado (456; 0,781)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 3
y REP_NIV_TIPO <= 6
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "LA RIOJA" "SAN LUIS" "SANTA CRUZ" "TUCUMAN"]
entonces Privado

4. Mientras que en el sector estatal, si bien se observan algunas reglas con el rango de repetidos nombrado anteriormente, estas no poseen un peso importante teniendo en cuenta la amplia cantidad de provincias que las conforman. A su vez, este rango de repetidos esta acompañado por una cantidad de matriculados ampliamente menor al que muestran las privadas. Con lo cual en este sector, sí existe algún tipo de relación positiva entre repetidos y matriculados, aunque no se puede decir que se comporte con igual proporción que la privada.

Regla 14 para Estatal (421; 0,922)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO <= 3
y MAT_NIV_TIPO <= 27
y PROVINCIA in ["BUENOS AIRES" "CATAMARCA" "CHACO" "CHUBUT" "CORDOBA" "ENTRE RIOS" "FORMOSA" "JUJUY" "LA RIOJA" "MENDOZA" "MISIONES" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO"]
entonces Estatal
Regla 15 para Estatal (365; 0,841)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO <= 3
y MAT_NIV_TIPO > 27
y MAT_NIV_TIPO <= 71
y PROVINCIA in ["BUENOS AIRES" "CHACO" "CORDOBA" "FORMOSA" "JUJUY" "RIO NEGRO" "SANTA CRUZ"]
entonces Estatal
Regla 16 para Estatal (95; 0,884)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 3
y REP_NIV_TIPO <= 6
y PROVINCIA in ["CATAMARCA" "CHACO" "CIUDAD DE BUENOS AIRES" "CORDOBA" "CORRIENTES" "ENTRE RIOS" "FORMOSA" "JUJUY" "LA PAMPA" "MENDOZA"

"MISIONES" "NEUQUEN" "RIO NEGRO" "SALTA" "SAN JUAN" "SANTA FE" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO"]
y MAT_NIV_TIPO <= 43
entonces Estatal

5. De seis (6) a diez (10) repetidos es imposible distinguir un sector de otro sin hacer referencia a una provincia. Ahora bien, a partir de diez (10) repetidos en adelante las escuelas son sin duda, estatales. Primero, por que no se detectan reglas con este rango para el sector privado; segundo, debido a los altos pesos que muestran las reglas del sector estatal.

Regla 18 para Estatal (174; 0,793)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 10
y REP_NIV_TIPO <= 20
y PROVINCIA in ["BUENOS AIRES" "SANTA CRUZ"]
y MAT_NIV_TIPO <= 247
entonces Estatal
Regla 19 para Estatal (665; 0,931)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 10
y REP_NIV_TIPO <= 20
y PROVINCIA in ["CATAMARCA" "CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "FORMOSA" "JUJUY" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "TIERRA DEL FUEGO"]
y MAT_NIV_TIPO <= 588
entonces Estatal
Regla 20 para Estatal (70; 0,8)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 10
y REP_NIV_TIPO <= 20
y PROVINCIA in ["CATAMARCA" "CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "FORMOSA" "JUJUY" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "TIERRA DEL FUEGO"]
y MAT_NIV_TIPO > 588
entonces Estatal
Regla 21 para Estatal (482; 0,795)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 10
y REP_NIV_TIPO <= 20
y PROVINCIA in ["CHACO" "ENTRE RIOS" "LA PAMPA" "MENDOZA" "MISIONES" "SALTA"]
entonces Estatal

Regla 22 para Estatal (326; 0,853)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 20
y REP_NIV_TIPO <= 43
y PROVINCIA in ["BUENOS AIRES" "SANTA CRUZ" "SANTIAGO DEL ESTERO"]
y MAT_NIV_TIPO <= 383
entonces Estatal
Regla 23 para Estatal (896; 0,974)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 20
y REP_NIV_TIPO <= 43
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "FORMOSA" "LA PAMPA" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "TIERRA DEL FUEGO"]
entonces Estatal
Regla 24 para Estatal (1.153; 0,871)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 20
y REP_NIV_TIPO <= 43
y PROVINCIA in ["CORRIENTES" "ENTRE RIOS" "JUJUY" "MENDOZA" "MISIONES" "SALTA" "SANTA FE" "TUCUMAN"]
y MAT_NIV_TIPO > 71
entonces Estatal
Regla 25 para Estatal (203; 0,911)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 43
y MAT_NIV_TIPO <= 247
entonces Estatal
Regla 26 para Estatal (973; 0,987)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 43
y MAT_NIV_TIPO > 247
y ID_RA <= 290277
entonces Estatal
Regla 27 para Estatal (2.208; 0,97)
si AMBITO in ["Urbano"]
y REP_NIV_TIPO > 43
y MAT_NIV_TIPO > 247
y ID_RA > 290277
entonces Estatal

En cuanto al mismo modelo con el ÁMBITO como salida, se observa una estructura de salida similar a la que se obtiene anteriormente. Las reglas obtenidas, como se mencionara, están regidas por la cantidad de alumnos repetidos. En este caso, el campo que divide a las reglas de decisión es MAT_NIV_TIPO, esto es, la cantidad total de alumnos matriculados. Sin embargo es interesante mostrar los distintos intervalos ya que pueden presentarse

comportamientos diferentes en cada uno de ellos. Se comienza estudiando las reglas para los niveles mas bajo de matriculados y detectando el ámbito correspondiente de cada una de ellas.

1. Como es de esperar, para una cantidad de matriculados que no supere los cuarenta y tres (43) alumnos, el colegio se ubica en zonas rurales. Aparece también el campo TIPO_SE, señalando que la mayoría de estas escuelas tienen un tipo de sección múltiple. Entre otros comportamientos se puede observar la ausencia de la Ciudad de Buenos Aires cuando la cantidad de matriculados es mayor a quince (15); un peso máximo en la regla 10 que señala que el promedio de matriculados se encuentra entre quince (15) y veintisiete (27) alumnos.

Regla 1 para Rural (787; 0,97)
si MAT_NIV_TIPO <= 7
y PROVINCIA in ["BUENOS AIRES" "JUJUY"]
entonces Rural
Regla 2 para Rural (1.837; 0,989)
si MAT_NIV_TIPO <= 7
y PROVINCIA in ["CATAMARCA" "CHACO" "CORRIENTES" "FORMOSA" "LA PAMPA" "LA RIOJA" "MENDOZA" "MISIONES" "SALTA" "SAN JUAN" "SANTIAGO DEL ESTERO" "TUCUMAN"]
entonces Rural
Regla 3 para Rural (108; 0,759)
si MAT_NIV_TIPO <= 7
y PROVINCIA in ["CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "ENTRE RIOS" "NEUQUEN" "RIO NEGRO" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "TIERRA DEL FUEGO"]
y REP_NIV_TIPO <= 3
y ID_RA <= 230078
entonces Rural
Regla 4 para Rural (789; 0,887)
si MAT_NIV_TIPO <= 7
y PROVINCIA in ["CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "ENTRE RIOS" "NEUQUEN" "RIO NEGRO" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "TIERRA DEL FUEGO"]
y REP_NIV_TIPO <= 3
y ID_RA > 230078
entonces Rural
Regla 5 para Rural (786; 0,95)
si MAT_NIV_TIPO > 7
y MAT_NIV_TIPO <= 15
y TIPO_SE in ["M"]
y PROVINCIA in ["BUENOS AIRES" "CORRIENTES" "NEUQUEN" "SALTA" "SAN LUIS"]

entonces Rural
Regla 6 para Rural (829; 0,99)
si MAT_NIV_TIPO > 7
y MAT_NIV_TIPO <= 15
y TIPO_SE in ["M"]
y PROVINCIA in ["CATAMARCA" "CHACO" "FORMOSA" "JUJUY" "LA RIOJA" "SAN JUAN" "SANTIAGO DEL ESTERO" "TUCUMAN"]
entonces Rural
Regla 7 para Rural (1.180; 0,896)
si MAT_NIV_TIPO > 7
y MAT_NIV_TIPO <= 15
y TIPO_SE in ["M"]
y PROVINCIA in ["CHUBUT" "CIUDAD DE BUENOS AIRES" "CORDOBA" "ENTRE RIOS" "LA PAMPA" "MENDOZA" "MISIONES" "RIO NEGRO" "SANTA CRUZ" "SANTA FE" "TIERRA DEL FUEGO"]
entonces Rural
Regla 8 para Rural (834; 0,827)
si MAT_NIV_TIPO > 15
y MAT_NIV_TIPO <= 27
y PROVINCIA in ["BUENOS AIRES" "ENTRE RIOS" "JUJUY" "MENDOZA" "RIO NEGRO" "SANTA CRUZ" "TIERRA DEL FUEGO"]
entonces Rural
Regla 9 para Rural (528; 0,975)
si MAT_NIV_TIPO > 15
y MAT_NIV_TIPO <= 27
y PROVINCIA in ["CATAMARCA" "CORRIENTES" "FORMOSA" "LA RIOJA"]
entonces Rural
Regla 10 para Rural (2.044; 0,943)
si MAT_NIV_TIPO > 15
y MAT_NIV_TIPO <= 27
y PROVINCIA in ["CHACO" "CHUBUT" "CORDOBA" "LA PAMPA" "MISIONES" "NEUQUEN" "SALTA" "SAN JUAN" "SAN LUIS" "SANTA FE" "SANTIAGO DEL ESTERO" "TUCUMAN"]
entonces Rural
Regla 11 para Rural (595; 0,929)
si MAT_NIV_TIPO > 27
y MAT_NIV_TIPO <= 43
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "CORRIENTES" "LA PAMPA" "LA RIOJA" "NEUQUEN" "SALTA" "SAN JUAN" "SAN LUIS" "SANTIAGO DEL ESTERO" "TUCUMAN"]
y ID_RA <= 321371
entonces Rural

Regla 12 para Rural (723; 0,972)
si MAT_NIV_TIPO > 27
y MAT_NIV_TIPO <= 43
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "CORRIENTES" "LA PAMPA" "LA RIOJA" "NEUQUEN" "SALTA" "SAN JUAN" "SAN LUIS" "SANTIAGO DEL ESTERO" "TUCUMAN"]
y ID_RA > 321371
y ID_RA <= 469713
entonces Rural

Regla 13 para Rural (147; 0,925)
si MAT_NIV_TIPO > 27
y MAT_NIV_TIPO <= 43
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "CORRIENTES" "LA PAMPA" "LA RIOJA" "NEUQUEN" "SALTA" "SAN JUAN" "SAN LUIS" "SANTIAGO DEL ESTERO" "TUCUMAN"]
y ID_RA > 321371
y ID_RA > 469713
entonces Rural

Regla 14 para Rural (1.194; 0,876)
si MAT_NIV_TIPO > 27
y MAT_NIV_TIPO <= 43
y PROVINCIA in ["CORDOBA" "ENTRE RIOS" "FORMOSA" "JUJUY" "MENDOZA" "MISIONES" "SANTA CRUZ" "SANTA FE"]
y REP_NIV_TIPO <= 10
entonces Rural

Regla 15 para Rural (83; 0,771)
si MAT_NIV_TIPO > 27
y MAT_NIV_TIPO <= 43
y PROVINCIA in ["CORDOBA" "ENTRE RIOS" "FORMOSA" "JUJUY" "MENDOZA" "MISIONES" "SANTA CRUZ" "SANTA FE"]
y REP_NIV_TIPO > 10
entonces Rural

2. Desde los cuarenta y tres (43) matriculados hasta los ciento cuarenta y nueve (149) se encuentra el rango donde se mezclan ambos ámbitos. Solo las provincias de Catamarca, Misiones y Santiago del Estero están presentes en todas las reglas de ámbito rural. Por otro lado, las provincias que conforman todas las reglas del ámbito urbano son Tierra del Fuego, Río Negro y la Ciudad de Buenos Aires. La presencia de la Capital Federal dentro de este grupo no es extraño, aunque si lo es la de las dos primeras. A continuación, solo se muestran las reglas mas significativas para este comportamiento.

```
        Regla 19 para  Rural (632; 0,965)
                si MAT_NIV_TIPO > 43
                y MAT_NIV_TIPO <= 71
                y PROVINCIA in [ "CORRIENTES" "LA PAMPA" "MISIONES" "NEUQUEN"
"SALTA" "SANTA CRUZ" "SANTIAGO DEL ESTERO" ]
                y REP_NIV_TIPO > 0
                y REP_NIV_TIPO <= 10
                entonces Rural
Regla 21 para  Rural (462; 0,825)
                si MAT_NIV_TIPO > 71
                y MAT_NIV_TIPO <= 149
                y SECTOR in [ "Estatal" ]
                y PROVINCIA in [ "CATAMARCA" "CORRIENTES" "SALTA" "SAN JUAN"
"SANTIAGO DEL ESTERO" ]
                entonces Rural
        Regla 22 para  Rural (454; 0,894)
                si MAT_NIV_TIPO > 71
                y MAT_NIV_TIPO <= 149
                y SECTOR in [ "Estatal" ]
                y PROVINCIA in [ "MISIONES" "TUCUMAN" ]
                entonces Rural
        Regla 1 para  Urbano (386; 1,0)
                si MAT_NIV_TIPO > 15
                y MAT_NIV_TIPO <= 27
                y PROVINCIA in [ "CIUDAD DE BUENOS AIRES" ]
                entonces Urbano
        Regla 2 para  Urbano (390; 1,0)
                si MAT_NIV_TIPO > 27
                y MAT_NIV_TIPO <= 43
                y PROVINCIA in [ "CIUDAD DE BUENOS AIRES" "TIERRA DEL FUEGO" ]
                entonces Urbano
        Regla 3 para  Urbano (561; 0,966)
                si MAT_NIV_TIPO > 43
                y MAT_NIV_TIPO <= 71
                y PROVINCIA in [ "CIUDAD DE BUENOS AIRES" "RIO NEGRO" "TIERRA DEL
FUEGO" ]
                entonces Urbano
```

3. Para un rango que pertenece al citado en la regla 6 (71 – 149), en el único caso donde “gana” el ámbito urbano es cuando las escuelas son privadas. El campo de los alumnos repetidos divide la regla en 2, ya que como se puede ver, sus valores son complementarios. El mayor peso en esta división se encuentra en el menor número de repetidos, esto es, menor a tres.

Regla 5 para Urbano (745; 0,966)
si MAT_NIV_TIPO > 71
y MAT_NIV_TIPO <= 149
y SECTOR in ["Privado"]
y REP_NIV_TIPO <= 3
entonces Urbano
Regla 6 para Urbano (278; 0,878)
si MAT_NIV_TIPO > 71
y MAT_NIV_TIPO <= 149
y SECTOR in ["Privado"]
y REP_NIV_TIPO > 3
entonces Urbano

4. Finalmente, a partir de doscientos cuarenta y siete (247) matriculados en adelante, cualquier escuela pertenece a zonas urbanas. Si bien aparece información adicional como el sector o la cantidad de repetidos, se observa que las reglas que forman parte de los distintos rangos de matriculados son complementarias en cuanto a la provincia. Con lo cual, juntas completan esta tendencia a nivel nacional. El máximo peso se detecta en la última regla de decisión (regla 22), donde se encuentran la mayor cantidad de escuelas de todo el país.

Regla 10 para Urbano (743; 0,942)
si MAT_NIV_TIPO > 247
y MAT_NIV_TIPO <= 383
y PROVINCIA in ["BUENOS AIRES" "CORDOBA" "ENTRE RIOS" "NEUQUEN"]
y SECTOR in ["Estatal"]
entonces Urbano
Regla 11 para Urbano (588; 0,995)
si MAT_NIV_TIPO > 247
y MAT_NIV_TIPO <= 383
y PROVINCIA in ["BUENOS AIRES" "CORDOBA" "ENTRE RIOS" "NEUQUEN"]
y SECTOR in ["Privado"]
entonces Urbano
Regla 12 para Urbano (169; 0,976)
si MAT_NIV_TIPO > 247
y MAT_NIV_TIPO <= 383
y PROVINCIA in ["CHACO" "CORRIENTES" "FORMOSA" "JUJUY" "LA PAMPA" "LA RIOJA" "MENDOZA" "SAN JUAN"]
y REP_NIV_TIPO <= 6
entonces Urbano
Regla 13 para Urbano (564; 0,817)
si MAT_NIV_TIPO > 247
y MAT_NIV_TIPO <= 383
y PROVINCIA in ["CHACO" "CORRIENTES" "FORMOSA" "JUJUY" "LA PAMPA" "LA RIOJA" "MENDOZA" "SAN JUAN"]

y REP_NIV_TIPO > 6
entonces Urbano
Regla 14 para Urbano (564; 0,931)
si MAT_NIV_TIPO > 247
y MAT_NIV_TIPO <= 383
y PROVINCIA in ["CHUBUT" "RIO NEGRO" "SAN LUIS" "SANTA FE" "TIERRA DEL FUEGO"]
entonces Urbano
Regla 15 para Urbano (479; 1,0)
si MAT_NIV_TIPO > 247
y MAT_NIV_TIPO <= 383
y PROVINCIA in ["CIUDAD DE BUENOS AIRES" "SANTA CRUZ"]
entonces Urbano
Regla 16 para Urbano (829; 0,993)
si MAT_NIV_TIPO > 383
y MAT_NIV_TIPO <= 588
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "CIUDAD DE BUENOS AIRES" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO"]
y SECTOR in ["Estatal"]
entonces Urbano
Regla 17 para Urbano (662; 1,0)
si MAT_NIV_TIPO > 383
y MAT_NIV_TIPO <= 588
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "CIUDAD DE BUENOS AIRES" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO"]
y SECTOR in ["Privado"]
entonces Urbano
Regla 18 para Urbano (370; 0,9)
si MAT_NIV_TIPO > 383
y MAT_NIV_TIPO <= 588
y PROVINCIA in ["CATAMARCA" "MISIONES" "SALTA" "TUCUMAN"]
entonces Urbano
Regla 19 para Urbano (496; 0,946)
si MAT_NIV_TIPO > 383
y MAT_NIV_TIPO <= 588
y PROVINCIA in ["CHACO" "CORRIENTES" "JUJUY" "MENDOZA" "SANTIAGO DEL ESTERO"]
entonces Urbano
Regla 20 para Urbano (387; 1,0)
si MAT_NIV_TIPO > 383
y MAT_NIV_TIPO <= 588
y PROVINCIA in ["CORDOBA" "ENTRE RIOS" "FORMOSA" "LA PAMPA" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SANTA FE"]
y REP_NIV_TIPO <= 10
entonces Urbano
Regla 21 para Urbano (792; 0,973)
si MAT_NIV_TIPO > 383

y MAT_NIV_TIPO <= 588
y PROVINCIA in ["CORDOBA" "ENTRE RIOS" "FORMOSA" "LA PAMPA" "LA RIOJA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SANTA FE"]
y REP_NIV_TIPO > 10
entonces Urbano

Regla 22 para Urbano (3.541; 0,998)
si MAT_NIV_TIPO > 588
entonces Urbano

Requerimiento # 3

Se recuerda que este requerimiento se resuelve con el modelo planteado para el requerimiento anterior ya que la tabla analizada posee información sobre la POF. La única corrida de las tres que brinda resultados que aplican al requerimiento es la que se realiza con KOHONEN + CHAID. Con lo cual, los resultados obtenidos son:

1. Las escuelas estatales urbanas tienen un tipo de sección independiente para los dos niveles que se muestran en la tabla (Primario / EGB). A su vez, tiene una POF dentro de lo presupuestado en la mayoría de los establecimientos.

Reglas para grupo 02 - contiene 1 regla(s)
Regla 1 para grupo 02 (8.673; 1,0)
si TIPO_SE in ["I"]
y AMBITO in ["Urbano"]
y SECTOR in ["Estatal"]
y POF in ["dentro"]
y NIVEL in ["EGB"]
entonces grupo 02

Reglas para grupo 12 - contiene 1 regla(s)
Regla 1 para grupo 12 (997; 1,0)
si TIPO_SE in ["I"]
y AMBITO in ["Urbano"]
y SECTOR in ["Estatal"]
y POF in ["dentro"]
y NIVEL in ["Primario"]
entonces grupo 12

Reglas para grupo 11 - contiene 2 regla(s)
Regla 1 para grupo 11 (951; 1,0)
si TIPO_SE in ["I"]
y AMBITO in ["Urbano"]
y SECTOR in ["Estatal"]
y POF in ["fuera"]
y MAT_NIV_TIPO <= 588
entonces grupo 11

Regla 2 para grupo 11 (127; 0,945)
si TIPO_SE in ["I"]
y AMBITO in ["Urbano"]
y SECTOR in ["Estatal"]
y POF in ["fuera"]
y MAT_NIV_TIPO > 588
entonces grupo 11

2. Finalmente, en las escuelas privadas no se obtiene información acerca de su planta funcional (POF), mientras que el tipo de sección es individual y su ámbito, como es de esperar, es urbano.

Reglas para grupo 00 - contiene 1 regla(s)
Regla 1 para grupo 00 (6.307; 1,0)
si TIPO_SE in ["I"]
y AMBITO in ["Urbano"]
y SECTOR in ["Privado"]
entonces grupo 00

Requerimiento # 4

Al pedirle a la aplicación que obtuviese reglas de decisión aplicando el modelo inductivo CHAID a la tabla MP, esta levanta un mensaje de error que informa no poder crearlas debido a la complejidad de los datos. Es por ello que no se pueden obtener reglas colocando al sector y al ámbito como salidas. Al intentar crearlas con el modelo C.5, presenta el mismo error.

Luego, se aplica clusterización e inducción (KOHONEN + CHAID) a la tabla MP, y se observa que al ordenar los datos en grupos, facilita que la aplicación comprenda los datos y logre obtener reglas de decisión.

1. Se detectan dos reglas de decisión muy significativas y simples, que involucran a dos modalidades.

Reglas para grupo 00 - contiene 1 regla(s)
Regla 1 para grupo 00 (2.859; 1,0)
si MODALIDAD in ["Economía y Gestión de las Organizaciones"]
entonces grupo 00
Reglas para grupo 02 - contiene 1 regla(s)
Regla 1 para grupo 02 (2.455; 1,0)
si MODALIDAD in ["Humanidades y Ciencias Sociales"]
entonces grupo 02

Luego se encuentran reglas de decisión con menos peso que van agrupando cada una de las modalidades restantes, diferenciándolas de acuerdo al sector que pertenezcan mas que al ámbito o de alguna otra variable. Algunos ejemplos de las reglas más significativas son.

2. Se observa que existe una mayor cantidad de escuelas estatales que enseñan Ciencias Naturales, que privadas.

Reglas para grupo 20 - contiene 4 regla(s)
Regla 1 para grupo 20 (59; 0,814)
si MODALIDAD in ["Ciencias Naturales"]
y SECTOR in ["Privado"]
y MAT_NIV_MOD <= 25
entonces grupo 20
Regla 2 para grupo 20 (727; 1,0)
si MODALIDAD in ["Ciencias Naturales"]
y SECTOR in ["Privado"]
y MAT_NIV_MOD > 25
entonces grupo 20
Reglas para grupo 30 - contiene 4 regla(s)
Regla 1 para grupo 30 (23; 0,826)
si MODALIDAD in ["Ciencias Naturales"]
y SECTOR in ["estatal"]
y MAT_NIV_MOD <= 25
entonces grupo 30
Regla 2 para grupo 30 (1.052; 1,0)
si MODALIDAD in ["Ciencias Naturales"]
y SECTOR in ["estatal"]
y MAT_NIV_MOD > 25
entonces grupo 30

3. Las reglas 3 y 4 del grupo 30 y la regla 4 del grupo 20 tienen un análisis similar que la anterior conclusión, aunque en este caso aumenta la diferencia entre la cantidad de escuelas estatales y las privadas.

Regla 4 para grupo 20 (269; 1,0)
si MODALIDAD in ["Producción de Bienes y Servicios"]
y SECTOR in ["Privado"]
y MAT_NIV_MOD > 25
entonces grupo 20
Regla 3 para grupo 30 (52; 0,885)
si MODALIDAD in ["Producción de Bienes y Servicios"]
y SECTOR in ["estatal"]
y MAT_NIV_MOD <= 25
entonces grupo 30

Regla 4 para grupo 30 (991; 1,0)
si MODALIDAD in ["Producción de Bienes y Servicios"]
y SECTOR in ["estatal"]
y MAT_NIV_MOD > 25
entonces grupo 30

4. Las estatales de las modalidades "Agropecuaria", "Ciclo básico" y "Otros" poseen mas cantidad de alumnos matriculados que las privadas, ya que el peso es alto considerando que solo contabiliza una cantidad de matriculados mayor a doscientos setenta y ocho (278) alumnos.

Reglas para grupo 32 - contiene 5 regla(s)
Regla 1 para grupo 32 (139; 0,914)
si MODALIDAD in ["Agropecuaria" "Ciclo basico" "Otros"]
y SECTOR in ["estatal"]
y MAT_NIV_MOD > 278
entonces grupo 32

Reglas para grupo 22 - contiene 4 regla(s)
Regla 1 para grupo 22 (224; 1,0)
si MODALIDAD in ["Agropecuaria" "Ciclo basico" "Otros"]
y SECTOR in ["Privado"]
entonces grupo 22

5. Superan en cantidad de establecimientos, pero no en matriculados, las escuelas estatales a las privadas en las modalidades "Artística", "Bachiller" y "Técnica".

Regla 2 para grupo 22 (344; 1,0)
si MODALIDAD in ["Artística" "Bachiller" "Técnica"]
y SECTOR in ["Privado"]
y MAT_NIV_MOD > 25
y MAT_NIV_MOD <= 278
entonces grupo 22
Regla 3 para grupo 22 (61; 0,951)
si MODALIDAD in ["Artística" "Bachiller" "Técnica"]
y SECTOR in ["Privado"]
y MAT_NIV_MOD > 25
y MAT_NIV_MOD > 278
entonces grupo 22
Regla 2 para grupo 32 (123; 0,911)
si MODALIDAD in ["Artística" "Bachiller" "Técnica"]
y SECTOR in ["estatal"]
y MODALIDAD in ["Artística" "Técnica"]
y MAT_NIV_MOD > 186
entonces grupo 32

Regla 3 para grupo 32 (646; 1,0)
si MODALIDAD in ["Artística" "Bachiller" "Técnica"]
y SECTOR in ["estatal"]
y MODALIDAD in ["Bachiller"]
entonces grupo 32

Requerimiento # 5

De los tres (3) modelos aplicados a la tabla MP_(CONPOF), totalizando tres (3) ejecuciones, y por ende, tres (3) salidas. Las conclusiones que se obtienen son siete (7), y las mismas se logran colocando al SECTOR como salida del modelo:

1. Las escuelas del sector estatal y ámbito rural de todas las provincias menos de la Ciudad de Buenos Aires (no tiene rurales), Córdoba, Santa Fe, Corrientes, Misiones y Rió Negro tiene una cantidad total de alumnos repetidos menor o igual a tres. La poca cantidad de matriculados en escuelas rurales influye en cierta medida en esta tendencia.

Regla 1 para estatal (146; 0,788)
si REP_NIV <= 3
y AMBITO in ["rural"]
y PROVINCIA in ["BUENOS AIRES" "LA PAMPA" "LA RIOJA" "MENDOZA" "TUCUMAN"]
entonces estatal
Regla 2 para estatal (90; 1,0)
si REP_NIV <= 3
y AMBITO in ["rural"]
y PROVINCIA in ["CATAMARCA" "CHUBUT" "FORMOSA" "NEUQUEN" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO"]
entonces estatal
Regla 3 para estatal (158; 0,93)
si REP_NIV <= 3
y AMBITO in ["rural"]
y PROVINCIA in ["CHACO" "ENTRE RIOS" "JUJUY" "SALTA" "SANTIAGO DEL ESTERO"]
entonces estatal

2. Debido al peso que poseen estas reglas muy similares, se puede inducir que en todas las provincias las escuelas que posean una cantidad de alumnos repetidos mayores que veinticuatro (24) serán de sector estatal.

Regla 11 para estatal (269; 0,792)
si REP_NIV > 24
y REP_NIV <= 48
y PROVINCIA in ["BUENOS AIRES" "CIUDAD DE BUENOS AIRES" "NEUQUEN" "RIO NEGRO"]
y NIVEL in ["Polimodal"]
entonces estatal
Regla 12 para estatal (232; 1,0)
si REP_NIV > 24
y REP_NIV <= 48
y PROVINCIA in ["CATAMARCA" "CHACO" "CHUBUT" "ENTRE RIOS" "FORMOSA" "JUJUY" "LA PAMPA" "LA RIOJA" "MENDOZA" "MISIONES" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO" "TUCUMAN"]
entonces estatal
Regla 13 para estatal (168; 0,952)
si REP_NIV > 24
y REP_NIV <= 48
y PROVINCIA in ["CORDOBA" "CORRIENTES" "SALTA" "SANTA FE"]
y MAT_NIV > 92
entonces estatal
Regla 14 para estatal (793; 0,961)
si REP_NIV > 48
y MAT_NIV > 151
entonces estatal

3. La provincia de Bs. As., Chubut, Córdoba y Jujuy, cuyas escuelas pertenezcan al ámbito urbano y tengan una cantidad de repetidos anuales menores o iguales a tres (3) alumnos, con una cantidad total de matriculados mayor a cuarenta y ocho (48), serán privadas.

Regla 1 para Privado (144; 0,854)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "CORDOBA" "JUJUY"]
y MAT_NIV > 48
y MAT_NIV <= 71
entonces Privado
Regla 2 para Privado (292; 0,932)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "CORDOBA" "JUJUY"]
y MAT_NIV > 71
y MAT_NIV <= 115
y POF in ["dentro"]
entonces Privado

Regla 3 para Privado (81; 1,0)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "CORDOBA" "JUJUY"]
y MAT_NIV > 71
y MAT_NIV <= 115
y POF in ["fuera"]
entonces Privado
Regla 4 para Privado (411; 0,908)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["BUENOS AIRES" "CHUBUT" "CORDOBA" "JUJUY"]
y MAT_NIV > 115
y MAT_NIV <= 315
entonces Privado

4. En las provincias de Catamarca, Chaco, Entre Ríos, Misiones, Salta y Santiago del Estero, se verifica la misma tendencia que la regla anterior, pero en menor medida. Se ha analizado cada una de las provincias que constituyen esta regla, filtrando por menos de tres alumnos repetidos para el ámbito urbano, y se concluye que las privadas superan en cantidad a las estatales en un 30 %, aproximadamente.

Regla 5 para Privado (228; 0,763)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["CATAMARCA" "CHACO" "ENTRE RIOS" "MISIONES" "SALTA" "SANTIAGO DEL ESTERO"]
y MAT_NIV > 48
y MAT_NIV <= 151
entonces Privado

5. En Capital se mantiene la misma tendencia.

Regla 6 para Privado (195; 1,0)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["CIUDAD DE BUENOS AIRES"]
y MAT_NIV <= 477
entonces Privado
Regla 7 para Privado (18; 0,889)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["CIUDAD DE BUENOS AIRES"]
y MAT_NIV > 477
entonces Privado

Regla 8 para Privado (318; 0,767)

6. Las provincias que se muestran a continuación manifiestan la misma tendencia, aunque con un peso menor. Esto es debido a que la regla abarca muchas provincias, pero pocos son los registros que se contabilizan.

Regla 8 para Privado (318; 0,767)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["LA PAMPA" "MENDOZA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "TIERRA DEL FUEGO" "TUCUMAN"]
y POF in ["dentro"]
y REP_NIV <= 1
entonces Privado
Regla 9 para Privado (132; 0,977)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["LA PAMPA" "MENDOZA" "NEUQUEN" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "SANTA FE" "TIERRA DEL FUEGO" "TUCUMAN"]
y POF in ["fuera"]
entonces Privado

7. Esta regla muestra las únicas provincias que no siguen la tendencia de las cuatro últimas reglas.

Regla 4 para estatal (32; 0,938)
si REP_NIV <= 3
y AMBITO in ["Urbano"]
y PROVINCIA in ["CORRIENTES" "FORMOSA" "LA RIOJA"]
y REP_NIV > 1
entonces estatal

Ahora se aplica el mismo algoritmo (CHAID) a la misma tabla, pero esta vez se deja a la variable ÁMBITO como salida del modelo. Con lo cual, las siguientes cuatro reglas muestran comportamientos dependientes de la ubicación de las escuelas.

1. Se detecta que las escuelas con una cantidad total de matriculados menor a cuarenta y ocho (48), que pertenecen al sector estatal y al nivel Polimodal, son rurales. Con lo cual, muestra que habrá pocas escuelas estatales - urbanas con de cuarenta y ocho (48) alumnos matriculados.

Se analizó la misma regla en la tabla modificando en nivel a Medio, y se pudo ver que el ámbito es más parejo. Un detalle que se puede observar es que en estas escuelas no existen repetidos.

Reglas para rural - contiene 1 regla(s)
Regla 1 para rural (155; 0,845)
si MAT_NIV <= 48
y SECTOR in ["estatal"]
y NIVEL in ["Polimodal"]
y REP_NIV <= 0
entonces rural

Entonces como conclusión, las escuelas urbanas tienen más de cuarenta y ocho (48) matriculados. Esta afirmación me permite introducir las tres reglas restantes.

2. La segunda regla engloba a todas las escuelas del sector privado. Se puede observar que en mas de una regla aparecen variables con valores complementarios (REP_NIV, POF). Esto se debe a que el modelo forma reglas de distinto peso y niveles de confianzas, pero que en definitiva muestran una única tendencia o patrón. Luego, aquí se muestra que las escuelas privadas independientemente de alguna otra variable son urbanas.

Reglas para Urbano - contiene 16 regla(s)
Regla 1 para Urbano (181; 0,901)
si MAT_NIV <= 48
y SECTOR in ["Privado"]
y PROVINCIA in ["BUENOS AIRES" "CATAMARCA" "CHACO" "CHUBUT" "CIUDAD DE BUENOS AIRES" "ENTRE RIOS" "FORMOSA" "JUJUY" "MENDOZA" "MISIONES" "NEUQUEN" "RIO NEGRO" "SALTA" "SAN JUAN" "SAN LUIS" "SANTA CRUZ" "TUCUMAN"]
entonces Urbano
Regla 2 para Urbano (195; 0,908)
si MAT_NIV > 48
y MAT_NIV <= 71
y SECTOR in ["Privado"]
y REP_NIV <= 1
entonces Urbano
Regla 3 para Urbano (223; 0,785)
si MAT_NIV > 48
y MAT_NIV <= 71
y SECTOR in ["Privado"]
y REP_NIV > 1
entonces Urbano

Regla 5 para Urbano (420; 0,931)
si MAT_NIV > 71
y MAT_NIV <= 92
y SECTOR in ["Privado"]
y POF in ["dentro"]
entonces Urbano
Regla 6 para Urbano (94; 0,989)
si MAT_NIV > 71
y MAT_NIV <= 92
y SECTOR in ["Privado"]
y POF in ["fuera"]
entonces Urbano
Regla 8 para Urbano (505; 0,974)
si MAT_NIV > 92
y MAT_NIV <= 115
y SECTOR in ["Privado"]
entonces Urbano
Regla 10 para Urbano (417; 0,988)
si MAT_NIV > 115
y MAT_NIV <= 151
y SECTOR in ["Privado"]
entonces Urbano
Regla 12 para Urbano (967; 0,99)
si MAT_NIV > 151
y MAT_NIV <= 237
y SECTOR in ["Privado"]
entonces Urbano
Regla 14 para Urbano (412; 1,0)
si MAT_NIV > 237
y MAT_NIV <= 315
y SECTOR in ["Privado"]
entonces Urbano

3. En cuanto a las escuelas estatales urbanas, se puede decir que son escuelas de setenta y un (71) matriculados totales en adelante. Se observa también una distinción por provincias, pero la realidad es que las provincias que no aparecen en esta regla, muestran grupos muy parejos y reducidos para ambos ámbitos.

Regla 4 para Urbano (191; 0,764)
si MAT_NIV > 71
y MAT_NIV <= 92
y SECTOR in ["estatal"]
y PROVINCIA in ["BUENOS AIRES" "CORDOBA" "CORRIENTES" "ENTRE RIOS" "FORMOSA" "LA PAMPA" "MISIONES" "RIO NEGRO" "SAN JUAN" "SAN LUIS" "SANTA FE" "SANTIAGO DEL ESTERO" "TIERRA DEL FUEGO"]
entonces Urbano

Regla 9 para Urbano (388; 0,879)
si MAT_NIV > 115
y MAT_NIV <= 151
y SECTOR in ["estatal"]
entonces Urbano
Regla 11 para Urbano (638; 0,948)
si MAT_NIV > 151
y MAT_NIV <= 237
y SECTOR in ["estatal"]
entonces Urbano

4. Las últimas dos reglas son mas que claras. No hay escuelas rurales con más de trescientos quince (315) matriculados en ambos niveles, sin distinción de sector.

Regla 15 para Urbano (556; 0,993)
si MAT_NIV > 315
y NIVEL in ["Medio"]
entonces Urbano
Regla 16 para Urbano (1.043; 1,0)
si MAT_NIV > 315
y NIVEL in ["Polimodal"]
entonces Urbano

Finalmente, al aplicar el algoritmo de KOHONEN junto con CHAID a la misma tabla de 8.000 registros se obtienen las siguientes reglas:

1. Las dos primeras reglas pertenecen a los grupos con mayor cantidad de registros. La regla del grupo 00, no precisa explicación y es la única que aparece en este conglomerado. Mientras que la regla 1 para el grupo 02 es la base de las reglas de su grupo, ya que se mantienen las variables y solo se modifican los valores de otros campos. Con lo cual, se observa que en la Ciudad de Buenos Aires, Neuquén y Río Negro no prevalecen las escuelas del nivel Polimodal, sino del Medio. La variable de los matriculados no importa realmente ya que hay otra regla que complementa hacia abajo, esto es desde noventa y dos (92) hasta cero (0).

Reglas para grupo 00 - contiene 1 regla(s)
Regla 1 para 00 (1.092; 1,0)
si AMBITO in ["Urbano"]
y NIVEL in ["Medio"]
y POF in ["dentro"]
entonces grupo 00

Reglas para grupo 02 - contiene 5 regla(s)
Regla 1 para grupo 02 (115; 0,878)
si AMBITO in ["Urbano"]
y NIVEL in ["Polimodal"]
y SECTOR in ["estatal"]
entonces grupo 02
Regla 3 para 02 (1.106; 0,989)
si AMBITO in ["Urbano"]
y NIVEL in ["Polimodal"]
y SECTOR in ["estatal"]
y MAT_NIV > 92
y PROVINCIA in ["BUENOS AIRES" "CATAMARCA" "CHACO" "ENTRE RIOS" "FORMOSA" "JUJUY" "MISIONES" "SANTA CRUZ" "TIERRA DEL FUEGO"]
entonces 02
Regla 4 para grupo 02 (825; 0,943)
si AMBITO in ["Urbano"]
y NIVEL in ["Polimodal"]
y SECTOR in ["estatal"]
y MAT_NIV > 92
y PROVINCIA in ["CHUBUT" "CORDOBA" "LA PAMPA" "LA RIOJA" "MENDOZA" "SALTA" "SAN JUAN" "SAN LUIS" "SANTIAGO DEL ESTERO" "TUCUMAN"]
entonces grupo 02
Regla 5 para grupo 02 (271; 0,978)
si AMBITO in ["Urbano"]
y NIVEL in ["Polimodal"]
y SECTOR in ["estatal"]
y MAT_NIV > 92
y PROVINCIA in ["CORRIENTES" "SANTA FE"]
entonces grupo 02

2. La regla del grupo 30 muestra que gran cantidad de establecimientos privados tienen un POF “dentro”. Sin embargo hay un nicho abalado por la regla del grupo 20 que muestra que hay establecimientos privados con POF “fuera”. Estos establecimientos superan en cantidad a las escuelas estatales – urbanas, de nivel Polimodal y con un POF también “fuera”.

Reglas para grupo 30 - contiene 1 regla(s)
Regla 1 para grupo 30 (2.316; 1,0)
si AMBITO in ["Urbano"]
y NIVEL in ["Polimodal"]
y SECTOR in ["Privado"]
y POF in ["dentro"]
entonces grupo 30

Reglas para grupo 20 - contiene 1 regla(s)

Regla 1 para grupo 20 (684; 1,0)
si AMBITO in ["Urbano"]
y NIVEL in ["Polimodal"]
y SECTOR in ["Privado"]
y POF in ["fuera"]
entonces grupo 20

3. Algunas escuelas del sector estatal y del ámbito rural, poseen dos sedes, como se puede observar en la siguiente regla. También se pueden ver ciento ochenta y seis escuelas privadas en el ámbito rural.

Reglas para grupo 22 - contiene 1 regla(s)
Regla 1 para grupo 22 (82; 0,878)
si AMBITO in ["rural"]
y SECTOR in ["estatal"]
y SEDE in ["2"]
entonces grupo 22
Reglas para grupo 32 - contiene 1 regla(s)
Regla 1 para grupo 32 (757; 0,974)
si AMBITO in ["rural"]
y SECTOR in ["estatal"]
y SEDE in ["1"]
entonces grupo 32
Reglas para grupo 31 - contiene 1 regla(s)
Regla 1 para grupo 31 (186; 0,871)
si AMBITO in ["rural"]
y SECTOR in ["Privado"]
entonces grupo 31

Requerimiento # 6

Al aplicar el algoritmo de inducción conocido como CHAID a la tabla SNU, colocando al ÁMBITO como salida, se obtiene como único resultado el ámbito urbano. Esto se debe a que el modelo no puede hallar reglas sobre regiones rurales, ya que todos los institutos que dictan estudios superiores no universitarios se localizan en zonas urbanas.

Cuando se desea obtener resultados aplicando el mismo modelo con el SECTOR como salida, el programa no permite procesar una respuesta ya que las condiciones de entrada son muy complejas para formar reglas de decisión. Se estima que esto es debido a la diversa lista de valores que posee el campo CARRERA. Es por ello que se pasa directamente al modelo de clusterización complementado con el de inducción. La tabla SNU tiene una cantidad aproximada de 5.000 registros. Los grupos que se pueden logran presentan las siguientes reglas:

1. La división de los grupos se da claramente por el tipo de formación. Se estima que debido a la gran diversidad de carreras, el modelo no puede generar reglas que hagan referencia a este campo. Sí, en cambio, se distingue a los establecimientos según su sector, y es en este nivel únicamente donde el privado supera levemente en cantidad de institutos al estatal. Esta mayoría se da fuertemente en escuelas de tipo técnico – profesional, pero no así en escuelas con formación docente. Para escuelas que tienen ambos tipos de formación el resultado es similar para ambos sectores.

```
Reglas para grupo 00 - contiene 1 regla(s)
        Regla 1 para  00 (1.006; 1,0)
                si TIPOFORMAC in [ "Exclusivamente Docente" ]
                y SECTOR in [ "Privado" ]
                entonces grupo 00
Reglas para grupo 02 - contiene 1 regla(s)
        Regla 1 para  grupo 02 (1.632; 1,0)
                si TIPOFORMAC in [ "Exclusivamente Técnico - Profesional" ]
                y SECTOR in [ "Privado" ]
                entonces grupo 02
Reglas para grupo 11 - contiene 1 regla(s)
        Regla 1 para  grupo 11 (91; 1,0)
                si TIPOFORMAC in [ "Ambos tipos de formación" ]
                y SECTOR in [ "Privado" ]
                entonces grupo 11
Reglas para grupo 22 - contiene 1 regla(s)
        Regla 1 para grupo 22 (128; 1,0)
                si TIPOFORMAC in [ "Ambos tipos de formación" ]
                y SECTOR in [ "Estatal" ]
                entonces grupo 22
Reglas para grupo 30 - contiene 1 regla(s)
        Regla 1 para  grupo 30 (1.904; 1,0)
                si TIPOFORMAC in [ "Exclusivamente Docente" ]
                y SECTOR in [ "Estatal" ]
                entonces grupo 30
Reglas para 32 - contiene 1 regla(s)
        Regla 1 para  grupo 32 (894; 1,0)
                si TIPOFORMAC in [ "Exclusivamente Técnico - Profesional" ]
                y SECTOR in [ "Estatal" ]
                entonces grupo 32
```

Requerimiento # 7

La tabla SNU_(CONPOF) posee solamente 2.000 registros y por lo tanto las reglas que se muestran a continuación tienen un peso menor en relación a las de otros niveles. Como ocurre con la tabla anterior, al aplicar el algoritmo de inducción colocando al ÁMBITO como salida, se obtiene como única respuesta el ámbito urbano ya que casi no existen escuelas en zonas rurales.

En cuanto al modelo inductivo con el SECTOR como salida, sí se logran resultados. Esto refuerza el supuesto que se mencionó anteriormente sobre la complejidad que trae al modelo la diferenciación de los registros por carrera. La regla que se obtiene es:

1. La regla con mayor cantidad de registros muestra una tendencia que relaciona a los establecimientos privados con una POF fuera de lo presupuestado. Las reglas que conforman dicha tendencia hacen una diferenciación por provincia, y muestran que en Catamarca, Formosa, La Rioja, San Luís, Santiago del Estero y Tierra del Fuego no existen escuelas privadas.

```
Regla 1 para  Privado (56; 0,786)
        si PROVINCIA in [ "BUENOS AIRES" "JUJUY" "SALTA" ]
        y POF in [ "fuera" ]
        entonces Privado
Regla 2 para  Privado (25; 0,84)
        si PROVINCIA in [ "CHACO" "CHUBUT" "CORRIENTES" "NEUQUEN" "SAN JUAN"]
        y POF in [ "fuera" ]
        entonces Privado
Regla 3 para  Privado (118; 0,907)
        si PROVINCIA in [ "CIUDAD DE BUENOS AIRES" ]
        y MAT_NIV <= 115
        entonces Privado
Regla 4 para  Privado (28; 0,929)
        si PROVINCIA in [ "CIUDAD DE BUENOS AIRES" ]
        y MAT_NIV > 115
        y POF in [ "fuera" ]
        entonces Privado
Regla 5 para  Privado (68; 0,971)
        si PROVINCIA in [ "CORDOBA" "ENTRE RIOS" "LA PAMPA" "RIO NEGRO" ]
        y POF in [ "fuera" ]
        y MAT_NIV <= 636
        entonces Privado
```

```
Regla 6 para  Privado (90; 0,878)
        si PROVINCIA in [ "MENDOZA" "MISIONES" "SANTA FE" "TUCUMAN" ]
        y MAT_NIV <= 61
        entonces Privado
Regla 7 para  Privado (39; 1,0)
        si PROVINCIA in [ "MENDOZA" "MISIONES" "SANTA FE" "TUCUMAN" ]
        y MAT_NIV > 61
        y MAT_NIV <= 413
        y POF in [ "fuera" ]
        y MAT_NIV <= 275
        entonces Privado
Regla 8 para  Privado (5; 0,8)
        si PROVINCIA in [ "MENDOZA" "MISIONES" "SANTA FE" "TUCUMAN" ]
        y MAT_NIV > 61
        y MAT_NIV <= 413
        y POF in [ "fuera" ]
        y MAT_NIV > 275
        entonces Privado
```

Por el lado del sector estatal no existe ninguna regla que muestre un patrón significativo para ser mencionado, dado que el peso que muestran es bajo.

```
Regla 1 para  Estatal (69; 0,826)
        si PROVINCIA in ["CHACO" "CHUBUT" "CORRIENTES" "NEUQUEN" "SAN JUAN"]
        y POF in [ "dentro" ]
        y MAT_NIV > 206
        entonces Estatal
Regla 2 para  Estatal (47; 0,915)
        si PROVINCIA in ["LA RIOJA" "SAN LUIS" "SANTA CRUZ" "TIERRA DEL FUEGO"]
        entonces Estatal
```

Al aplicar clusterización (KOHONEN) junto con inducción (CHAID) para obtener reglas de decisión de los grupos, los resultados obtenidos son:

1. La mayoría de los establecimientos superiores no universitarios pertenecen al sector privado, con un gasto en horas, cargos y módulo que se encuentra dentro de lo presupuestado. A su vez, este tipo de institutos se concentran en una sola sede. En las escuelas estatales se mantiene la misma tendencia con relación a la POF y a la cantidad de sedes educativas, aunque en menor escala.

Reglas para grupo 02 - contiene 1 regla(s)
Regla 1 para grupo 02 (1.002; 0,995)
si POF in ["dentro"]
y SEDE in ["1"]
y SECTOR in ["Privado"]
entonces grupo 02
Reglas para grupo 30 - contiene 1 regla(s)
Regla 1 para grupo 30 (785; 1,0)
si POF in ["dentro"]
y SEDE in ["1"]
y SECTOR in ["Estatal"]
entonces grupo 30

2. Se detecta que la POF excede lo presupuestado principalmente en los institutos privados dado que se analiza la tabla a fin de encontrar el sector para el valor "fuera". Esta regla enseña divisiones de poco interés, dado que el número de matriculados no tiene relación con la POF del establecimiento. Estas divisiones se deben a los diferentes intervalos de confianza que abarcan las tres reglas del mismo grupo.

Reglas para grupo 22 - contiene 3 regla(s)
Regla 1 para grupo 22 (31; 0,903)
si POF in ["fuera"]
y MAT_NIV <= 40
entonces grupo 22
Regla 2 para grupo 22 (196; 0,995)
si POF in ["fuera"]
y MAT_NIV > 40
y MAT_NIV <= 635
entonces grupo 22
Regla 3 para grupo 22 (24; 0,833)
si POF in ["fuera"]
y MAT_NIV > 635
entonces grupo 22

Revisión del proceso

En esta sección se analiza como se satisfizo la recolección y análisis de los resultados logrados. Se debe realizar una completa revisión de los datos para determinar si existe algún factor importante o tarea que se ha pasado por alto en algún momento. También el objetivo de esta tarea es asegurar la calidad de los modelos generados.

En relación a los objetivos, los datos responden aceptablemente a los requerimientos planteados. Sin embargo, existen otras bases que se exponen en la página Web de la DINIECE llamadas Operativo Nacional de Evaluación de la Calidad (ONE) y Censos de Docentes. Estas tienen gran cantidad de información, aunque son de dimensiones menores que los Relevamientos Anuales (RA). Desafortunadamente no se utilizaron dado que no brindaban información relacionada con los requerimientos planteados. Con lo cual, la selección, depuración y manejo de datos en las fases correspondientes se considera exitosa.

En cuanto a los modelos, también puede considerarse como aceptable su rendimiento y calidad. Esto se debe principalmente a la facilidad de manejo y a la efectividad de la aplicación. También cabe aclarar, que se trabajó de la forma mas completa para cubrir todos los requerimientos, configurando los parámetros de forma tal que no condicionen fuertemente las salidas.

Se cree que los resultados gráficos son relevantes en un análisis, sobre todo para personas que no están habituadas a observar resultados expuestos como reglas de decisión. Por lo tanto, dentro de las conclusiones se adjuntan algunas imágenes interesantes que complementan las reglas obtenidas.

4. CONCLUSIONES

En resumen del trabajo que se desarrollara en las diferentes secciones, a continuación se enuncian conclusiones sobre los resultados obtenidos en el ámbito educativo nacional como también algunas apreciaciones sobre realización del proyecto. La principal causa por la cual se añaden conclusiones sobre aspectos ajenos a la educación, se debe a que el trabajo en cuestión forma parte del proyecto final de una carrera universitaria. Con lo cual, es pertinente mencionar conclusiones de acuerdo al aprendizaje obtenido al margen de los objetivos planteados.

En la primera parte se habla del aprendizaje logrado a partir del abordaje a las diversas áreas, técnicas y metodologías. Esta contiene a las conclusiones de la Minería de Datos con CRISP DM como metodología, los datos seleccionados, las herramientas de modelización y ejecución. La segunda parte contiene a las conclusiones obtenidas del trabajo propiamente dicho sobre la educación nacional. Finalmente, en la última subsección se comentan posibles líneas para futuras investigaciones, esto es, oportunidades detectadas durante la realización del trabajo que podría complementar y hasta mejorar la investigación.

4.1. Sobre el aprendizaje

La Minería de Datos basada en la metodología CRISP DM, se aplica satisfactoriamente en este proyecto. Sin embargo, existen herramientas dentro de la sección de comprensión del negocio (Fase I) como por ejemplo, organigramas de la empresa, mapas estáticos del negocio, análisis FODA, mapa de condicionamientos de los objetivos, etc., que no fueron utilizadas debido a la información disponible. En un análisis del negocio enfocado a corporaciones, donde gran parte de los datos están disponibles, todas las herramientas citadas y otras tantas son totalmente válidas y necesarias. Sin duda un análisis de estas características lleva a que los resultados obtenidos sean estudiados con mayor conocimiento y profundidad.

A su vez, el “negocio” analizado en este proyecto imposibilita implementar acciones (Fase VI) que modifiquen los patrones negativos de forma eficiente como en una corporación. Esto es real, ya que es el gobierno el principal responsable de la administración y desarrollo educativo. Sin embargo, esta fase puede ser suplantada por otras actividades que ayuden a su futura implementación, como ser la publicación del trabajo en medios públicos o la presentación del mismo en conferencias de investigación educativa.

En definitiva, se demuestra que CRISP DM, planteando requerimientos a partir de bases de datos para obtener resultados por medio de algoritmos, sirve como una técnica eficiente y complementaria a la estadística, para el análisis de datos.

En cuanto a los datos y la aplicación seleccionada, ya se ha comentado algo en la revisión del proceso. Las tablas, si bien son completas en la mayoría de los registros, fallan al no tener campos que conecten el relevamiento con otras investigaciones educativas. Este aspecto forma parte de las futuras líneas de investigación. Con respecto a la aplicación se ha trabajo de la forma mas completa posible para cubrir todos los requerimientos, aunque algunas funcionalidades adicionales no fueron tenidas en cuenta a la hora de la modelización. Sin embargo, se ha comprendido como configurar los datos a partir de las actividades planteadas en las Fases II y III, y se ha aprendido a modelar con una aplicación desconocida al comienzo del proyecto, logrando un alto rendimiento y calidad.

4.2. Sobre la educación

De los resultados del requerimiento # 1 correspondiente al nivel Inicial, se detectan comportamientos distintivos y otros similares a cada sector y ámbito. Como conclusión compartida por todas las escuelas se señala que estas poseen una sola sede educativa. No es frecuente que los establecimientos de educación inicial, a menos que pertenezca a escuelas con ciclo básico completo (Inicial, Primaria/EGB y Medio), tengan un anexo adicional a la sede central.

Los establecimientos privados se localizan solamente en áreas urbanas y la cantidad de establecimientos y alumnos prevalecen con respecto a las estatales solamente en la Ciudad de Buenos Aires. Este comportamiento se fundamenta en la rentabilidad económica que brinda el nivel en la Capital. Estas escuelas son las únicas que presentan un nicho con una planta funcional fuera de lo presupuestado, ya que por medio del pago de los alumnos afrontan las horas y cargos extras. El tipo de sección elegido por estas es independiente, con lo cual, aquí también el aspecto económico tiene su relevancia, ya que al tener una enseñanza que se divide en 3 años se percibirá un mayor beneficio económico comparado al que se obtiene en menos cantidad años por la misma educación.

Para las escuelas rurales el tipo de sección que prevalece es el múltiple, debido a la escasez de establecimientos y alumnos. A su vez, que la edad promedio de los alumnos este por encima del promedio del país, es otro factor que repercute en el tipo de sección seleccionada. Esto se debe fundamentalmente a la lejanía de las escuelas de donde habita la población y también a la preferencia por mantener a los niños en sus hogares debido al bajo

nivel infraestructura y confort de las escuelas rurales. Por lo tanto, del total de matriculados en zonas rurales, más de la mitad comienza su formación en sala de 5 años que es obligatoria.

En la figura 31 se observa la ausencia de escuelas privadas en el ámbito rural y la inscripción masiva en salas de 5 años de este ámbito, señalada en el párrafo anterior.

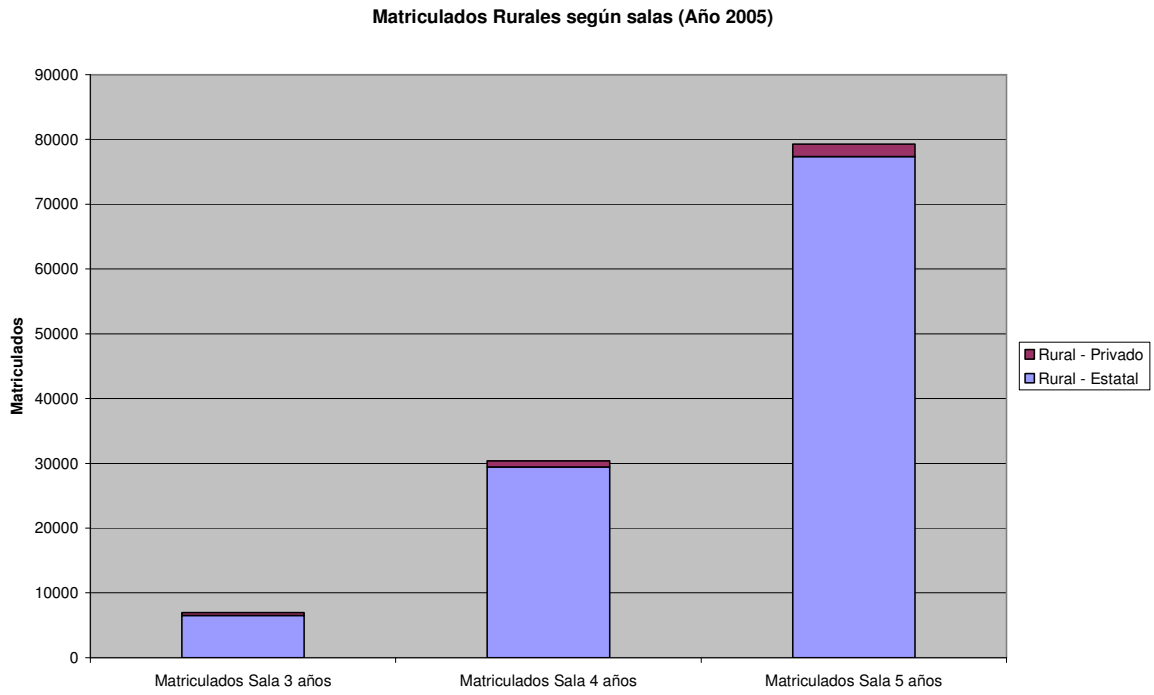

Figura 31. Matriculados rurales del nivel Inicial, según salas (Año 2005)

Las estatales urbanas poseen la mayor cantidad de establecimientos y alumnos del nivel. Estas presentan una tendencia similar a la rural sobre los matriculados en salas de 5 años, pero en menor proporción. La figura 32 muestra esta tendencia.

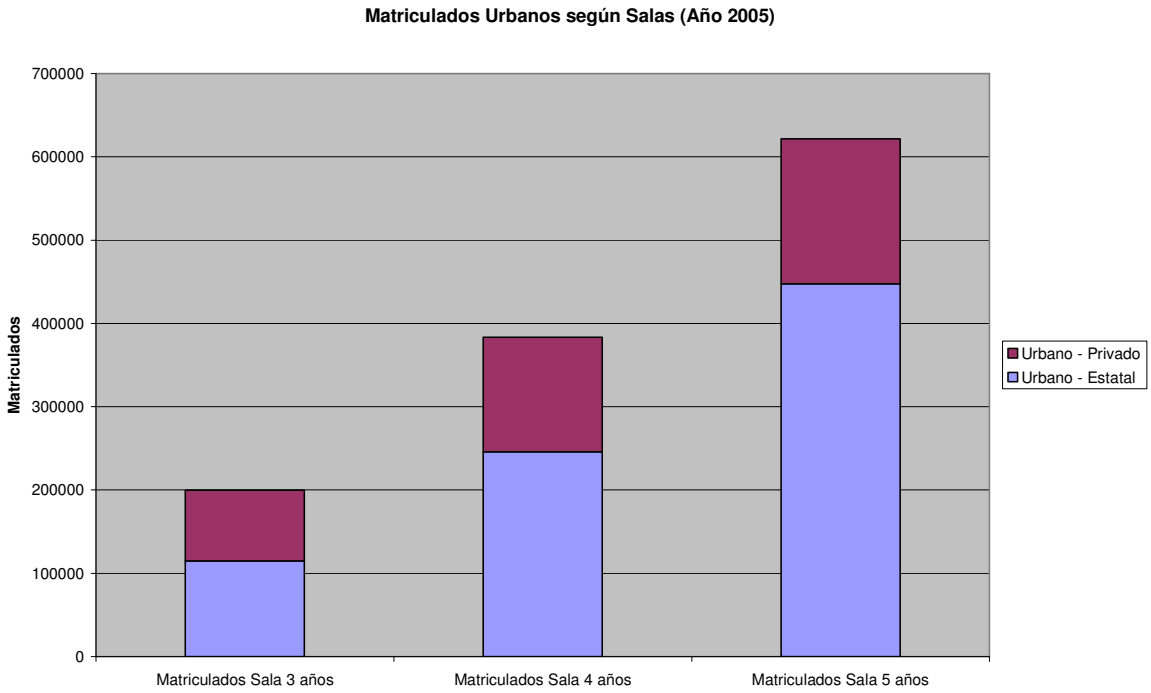

Figura 32. Matriculados urbanos del nivel Inicial, según salas (Año 2005).

La educación del nivel Primario y EGB, presenta algunos comportamientos interesantes al mezclar los sectores y ámbitos de los establecimientos. Estos patrones responden a los requerimientos # 2 y # 3.

Para el sector privado no se encuentran tendencias en el ámbito rural. Mientras que para el sector estatal, las provincias de Catamarca, Misiones y Santiago del Estero están presentes en todos los resultados relacionados con este ámbito. Esto se debe a que existen pocas zonas urbanas o ciudades mas allá de sus capitales. El tipo de sección múltiple es una constante en este ámbito por las mismas razones mencionadas en el nivel Inicial.

En el ámbito urbano Tierra del Fuego, Río Negro y la Ciudad de Buenos Aires aparecen en todas sus reglas. La presencia de la Capital Federal dentro de este grupo no es extraño, aunque si lo es la de las otras dos. Según comenta el experto, este comportamiento se debe a la baja población en los ámbitos rurales, fundamentada en el clima que poseen dichas regiones. Con respecto a las estatales de este ámbito, si bien algunas reglas muestran niveles bajos de alumnos repetidos, estas se relacionan proporcionalmente con la cantidad de alumnos matriculados. A su vez, tienen una POF dentro de lo presupuestado en la mayoría de los establecimientos, ya que es complejo ampliar las actividades, cuando la paga de los profesores proviene del estado.

En cuanto a las privadas urbanas, el principal resultado detectado es la baja cantidad de alumnos repetidos. En escuelas con mas de ciento cuarenta y nueve (149) matriculados, el rango de repetidos va de cero (0) a seis (6). Claramente, la retención de alumnos beneficia económicamente a las instituciones privadas, como también ayuda a mantener una imagen de escuelas con pocos alumnos aplazados. Sin duda esta tendencia muestra la alta permeabilidad que existe en las escuelas privadas con los alumnos que no tiene los conocimientos que se demandan en cada año.

Al margen de los resultados, la figura 33 muestra una disminución en la cantidad de alumnos que comienzan la escuela Primaria/ EGB. No se logra explicar este comportamiento, por lo que debe ser monitoreado en los relevamientos de años posteriores.

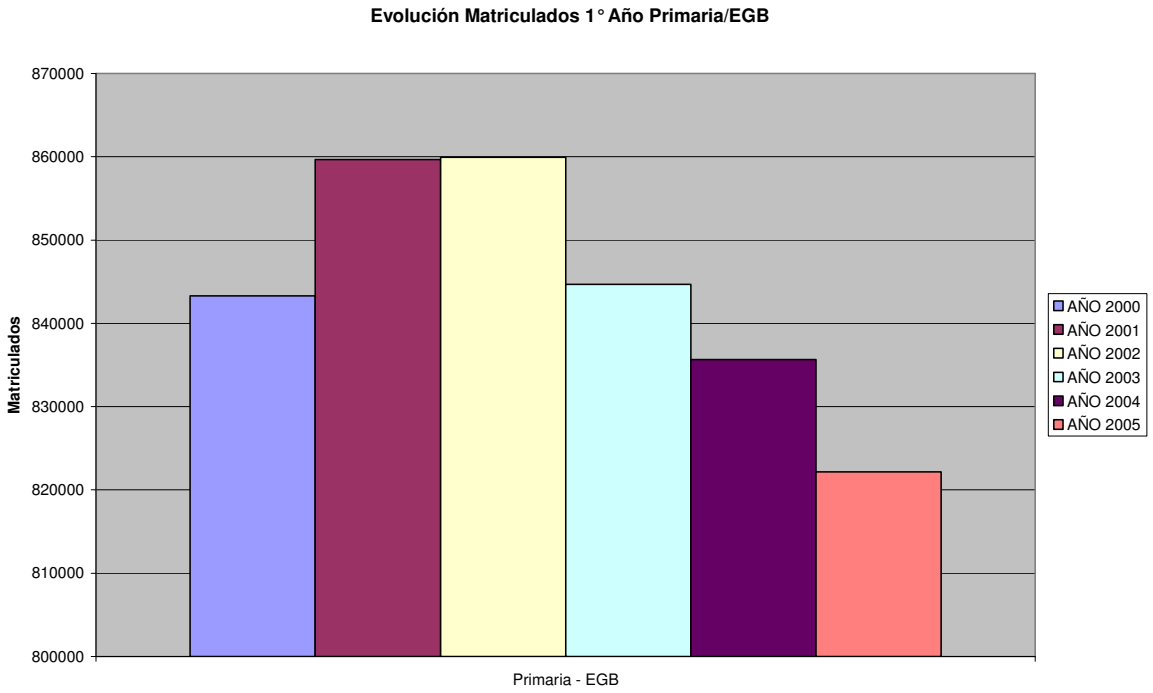

Figura 33. Evolución de matriculados del 1º año del nivel Primario/EGB.

En el nivel Medio/Polimodal, analizado por los requerimientos # 4 y # 5, se detectan dos modalidades que prevalecen del resto. Hasta el año 2000, el bachiller era la modalidad de enseñanza mas frecuente de las escuelas medias. Luego de los cambios en los planes educativos, algunas escuelas se abocaron a la enseñanza de modalidades mas específicas. En siguientes figuras, así como también en las reglas obtenidas, se puede observar la migración del alumnado a orientaciones humanas y económicas. La figura 34 expone la evolución de los egresados según cada modalidad. En ella se detecta a simple vista el comportamiento señalado sobre los egresados de bachilleratos.

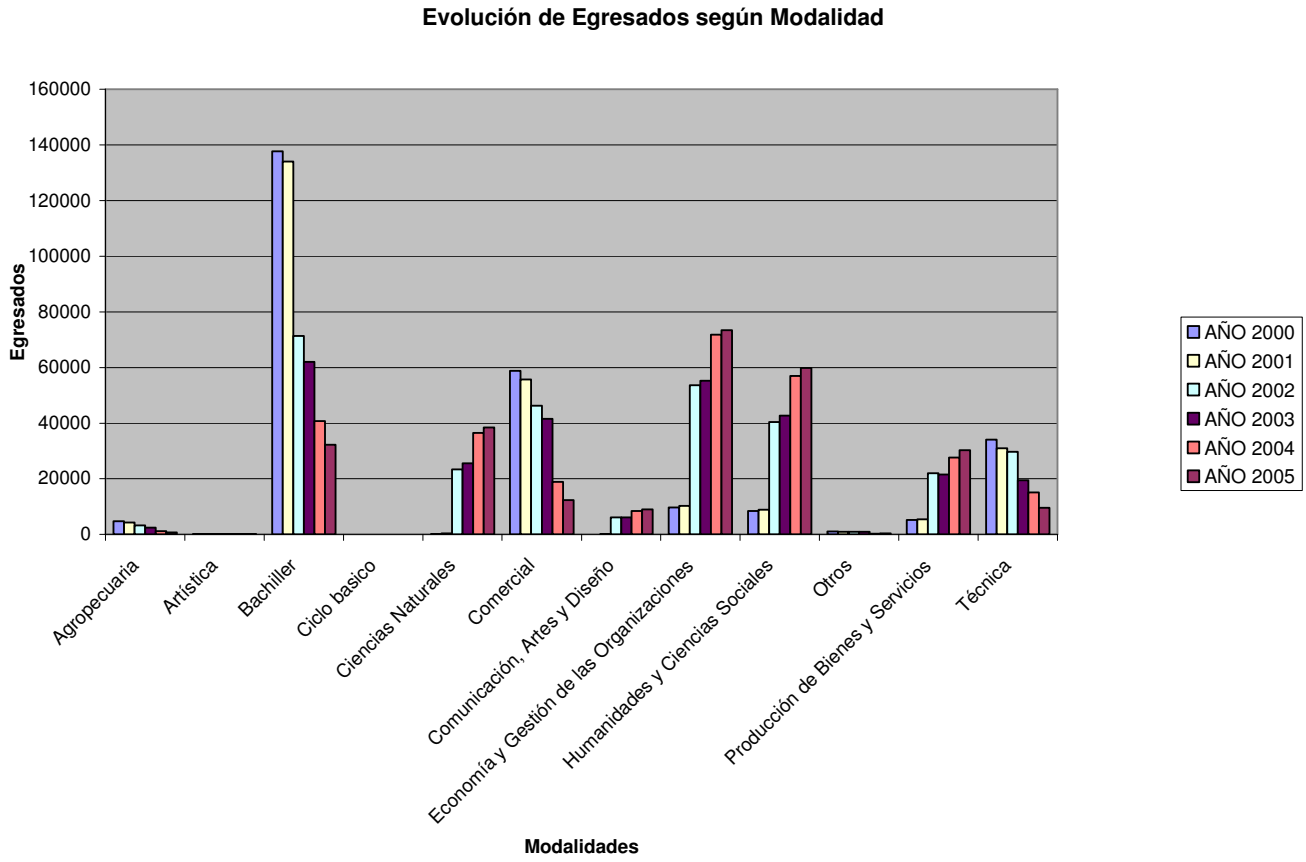

Figura 34. Evolución de egresados del nivel Medio/ Polimodal, según la modalidad.

La figura 35 sobre la evolución de los matriculados del nivel según modalidad, muestra a que modalidades migran los alumnos; estas son Economía y Gestión de las Organizaciones y Humanidades y Ciencias Sociales. En menor escala también aumentan los matriculados en las escuelas de Ciencias Naturales, Producción de Bienes y Servicios, y, muy levemente en las escuelas Técnicas.

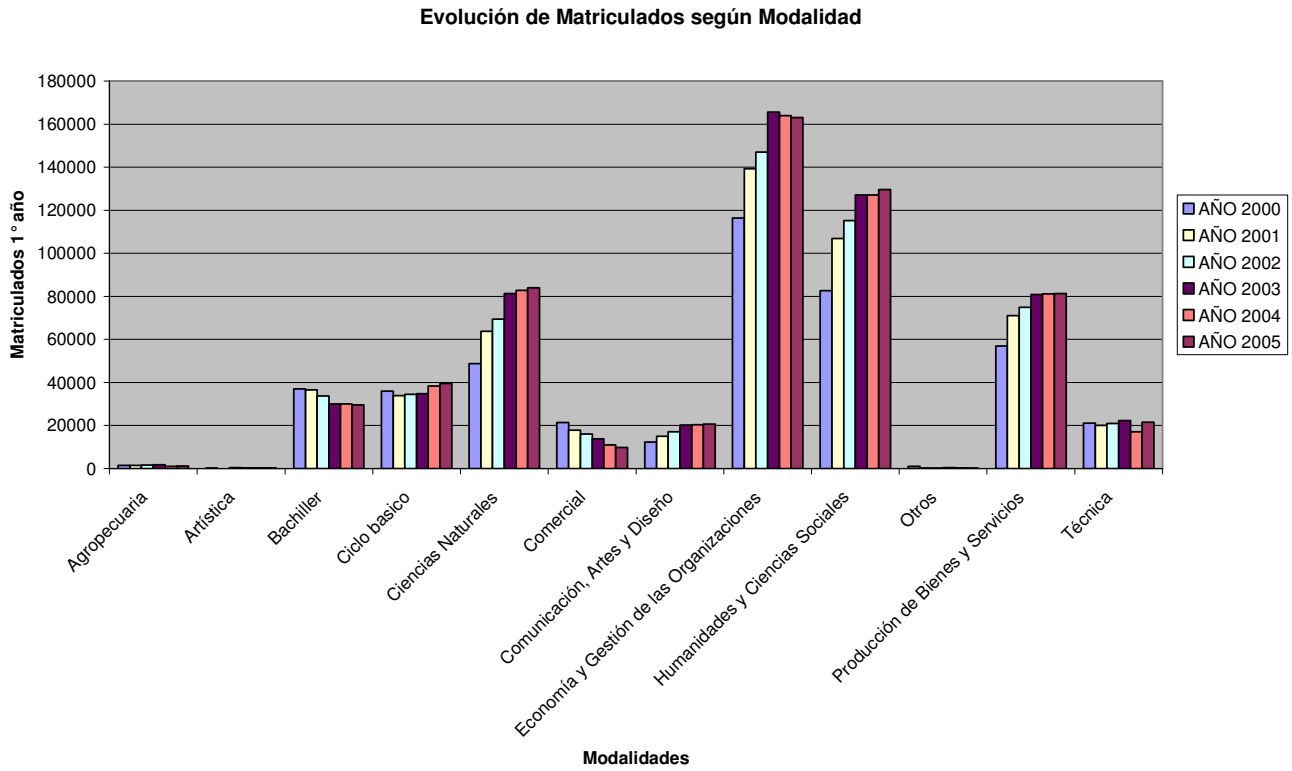

Figura 35. Evolución de matriculados del nivel Medio/ Polimodal, según la modalidad.

Se puede agrega que existe un nicho de escuelas privadas con una POF fuera de lo presupuestado. Esta tendencia que se encuentra en otros niveles también esta presente en el este, y se debe a que poseen mas facilidades económicas para afrontar los cargos y horas extras.

En cuanto al requerimiento # 6 del nivel Superior no Universitario, las figuras 36 y 37 avalan la tendencia detectada sobre la falta de establecimientos en zonas rurales. Si bien las figuras se refieren a alumnos matriculados, la cantidad de establecimientos es directamente proporcional a dicha variable. Como ya se ha mencionado, la educación terciaria no forma parte de los estudios obligatorios, con lo cual, en zonas rurales difícilmente se localicen establecimientos educativos, aumentando consecuentemente la migración de alumnos a ciudades. A su vez, son los establecimientos que pertenecen al sector privado los que prevalecen en este nivel. Con lo cual, los intereses económicos de los institutos hacen que los mismos se ubiquen en centro urbanos con mayor cantidad de habitantes.

La figura 36 demuestra el predominio privado en las escuelas técnicas, mientras que en la figura 37 se detecta la baja cantidad de matriculados rurales en relación a los urbanos.

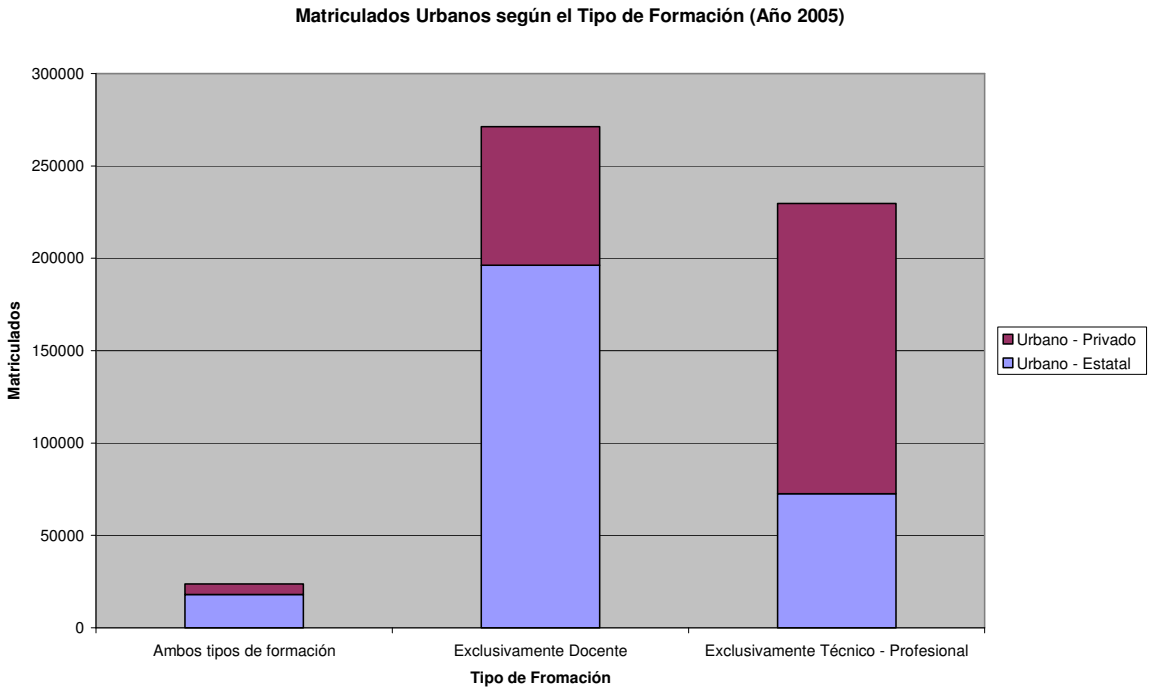

Figura 36. Matriculados urbanos del nivel SNU, según el tipo de formación (Año 2005).

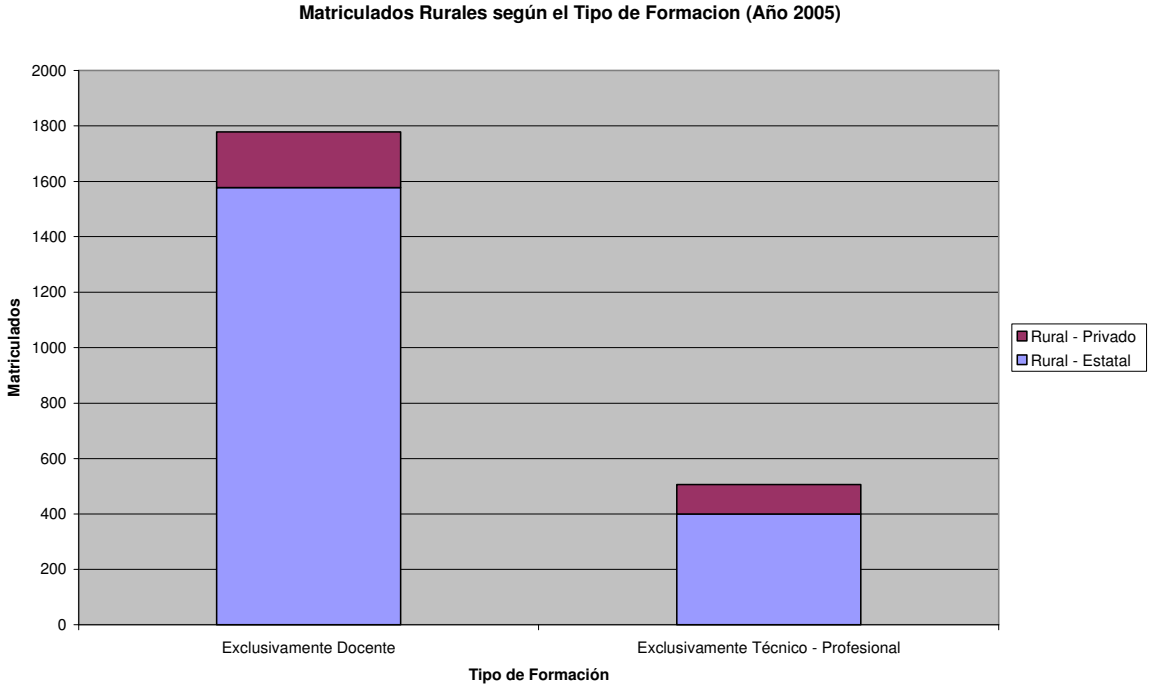

Figura 37. Matriculados rurales del nivel SNU, según el tipo de formación (Año 2005).

Con respecto al requerimiento # 7 relacionado también a este nivel, es lógico que prevalezca las escuelas con planta funcional "fuera" sean privadas. El gobierno no desea pagar por horas, cargos o módulos extras a los presupuestados para la carrera. Mientras que

las privadas, pueden que aumenten las horas de enseñanza de una carrera, no así los años de la misma, para captar mayor cantidad de alumnos.

Los estudiantes perciben los mejores salarios al seguir una carrera de formación técnica. Como se observa en la exploración de datos, son mayormente hombres los que ingresan en carreras técnicas, deseando conseguir trabajo a corto plazo sin realizar una carrera universitaria. Las mujeres, que prevalecen en carreras docentes y humanas, son influenciadas por los mejores salarios del sector mercado técnico, detectando una migración de alumnas hacia este tipo de formación. La figura 38 muestra la tendencia señalada, que claramente se profundiza en los últimos años del relevamiento (2004 y 2005).

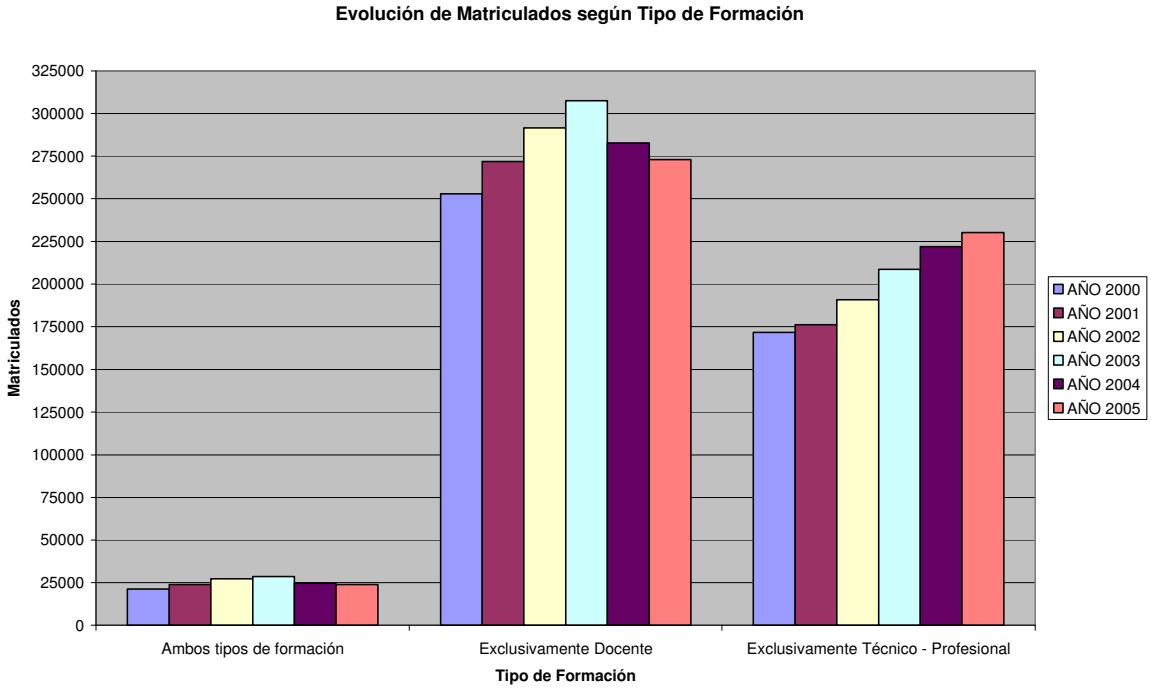

Figura 38. Evolución de matriculados del nivel SNU, según el tipo de formación.

4.3. Sobre futuras líneas de investigación

Uno de las grandes falencias en cuanto a los datos, como ya se mencionó, es la falta de conexión entre los tres relevamientos realizados por la DINIECE (Operativo Nacional de Evaluación de la Calidad (ONE), Censos de Docentes, Relevamientos Anuales (RA)). La técnica de modelización de fundir tablas, que se emplea en los modelos expuestos, no se podría utilizar para analizar el total de los datos. Sin embargo, este error es fácilmente corregible, ya que por ejemplo, en el Censo de Docentes además de un ID por docente se podría agregar el campo del ID del establecimiento correspondiente (ID_RA). Esto

ayudaría a conectar tablas y por ende saber si los resultados destacados anteriormente se sustentan en patrones de las tablas de los docentes. Sin duda, la posible sinergia lograda a través de la unificación de información de las bases de conocimiento, es fundamental para futuras investigaciones.

Con respecto a la aplicación, algunas funcionalidades adicionales de la misma no fueron tenidas en cuenta durante este proyecto. La causa es principalmente la capacidad de alcanzar los objetivos sin la necesidad de crear modelos demasiados sofisticados y complejos, que dificulten la obtención de resultados y la comprensión de los mismos. No debe olvidarse que existe un setéo de parámetros para cada nodo, aumentando las diversas posibilidades de modelización.

La aplicación posee algoritmos o técnicas de procesamiento mas complejas que no tampoco se han mencionado. Estas pueden brindar resultados interesantes, partiendo de modelos sin duda más complejos. Entre estas técnicas se encuentran los modelos de reglas de asociación utilizando algoritmos A priori, reglas de asociación de secuencia, redes bayesianas, etc.

REFERENCIAS

Britos, P., Hossian, A., García-Martinez, R. y Sierra, E. 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Editorial Nueva Librería. ISBN 987-1104-30-8

DINIECE. 2006. *Promoción del Uso de Estadísticas Educativas en Investigación*. http://diniece.me.gov.ar/concurso2006/basededatos.html. Página vigente al 26/02/07.

Filmus, D. (1999) *La descentralización educativa en argentina: elementos para el análisis de un proceso abierto*. Coloquio Regional sobre Descentralización de la Educación en América Central, Cuba y República Dominicana (1997 Nov. 3-5: San José).-CLAD; Países Bajos. Gobierno Nacional; Costa Rica. Ministerio de Planificación Nacional y Política Económica; Costa Rica. Ministerio de Educación Pública.

Tedesco, J. C. (2003) *Los pilares de la educación del futuro.* IIPE. Instituto Internacional de Planeamiento de la Educación.

Tiramonti, G. (2003) *Información para la gestión: Un estudio de caso en tres jurisdicciones de la Argentina.* FLACSO.

ANEXO

Metodología CRISP DM

Fase I - Comprensión del negocio

Esta primera fase se focaliza en entender los objetivos del negocio y los requerimientos desde la perspectiva del negocio, para incluir estos conocimientos dentro de la definición del problema de Exploración de Información y el diseño del plan preliminar para lograr los objetivos.

Fase II - Comprensión de los datos:

La fase de comprensión de los datos comienza con la recolección inicial de datos y prosigue con actividades que apuntan a la familiarización de los datos, identificar problemas de calidad y detectar relaciones interesantes entre los mismos que permitan generar hipótesis sobre información oculta.

Fase III - Preparación de los datos:

Esta fase contempla un conjunto de actividades destinadas a la construcción del dataset a partir de los datos iniciales. Esta fase implica múltiples tareas que pueden desarrollarse al mismo tiempo sin un orden estricto. Estas tareas incluyen la selección, limpieza y transformación de tablas, registros y atributos para poder ingresarlas en la herramienta de modelado.

Fase IV - Modelado:

En esta fase se seleccionarán y aplicarán varias técnicas de modelado, como así también, opcionalmente, se podrán determinar los valores de los parámetros y variables de calibración. Para esta tarea generalmente se puede contar con más de una técnica que realice la misma función. Algunas técnicas pueden tener requerimientos específicos en cuanto a la conformación de los datos, lo cual puede hacer que se deba volver a la fase de preparación de los datos para realizar alguna adecuación.

Fase V - Evaluación:

En esta fase se verifica que los resultados obtenidos en la fase de modelado sean de alta calidad desde la perspectiva del análisis de datos. Antes de realizar la implementación del modelo, es importante evaluar que los resultados del modelado y revisar los pasos realizados para la construcción del modelo y verificar que estos sean apropiados en función de los objetivos del negocio.

Fase VI - Implementación:

El final del proyecto, generalmente, no se encuentra en la creación del modelo. En la mayoría de los casos el analista de datos debe generar un informe final para ser presentado al cliente o en caso de que el propio cliente desee realizar el informe final se deberá transferir a este el conocimiento para que este pueda hacer una correcta interpretación de los datos.

A continuación, en la Tabla i, se describen cada una de las fases con sus tareas genéricas:

Comprensión del negocio	Comprensión de los datos	Preparación de los datos	Modelado	Evaluación	Implementación
Determinar los objetivos del negocio: ▪ *Escenario actual* ▪ *Objetivos del negocio* ▪ *Factores críticos de éxito*	**Recolectar los datos Iniciales:** ▪ *Reporte inicial de la colección de los datos*	▪ *Dataset* ▪ *Descripción del Dataset*	**Seleccionar técnica de modelado:** ▪ *Técnica de modelado* ▪ *Supuestos de modelado*	**Evaluar Resultado:** ▪ *Evaluación de los resultados del proceso de Exploración de Información respecto de los factores críticos del éxito* ▪ *Aprobación del Modelo*	**Plan de Implementación:** ▪ *Plan de Implementación*
Evaluación de la situación: ▪ *Inventario de Recursos* ▪ *Requisitos, supuestos y requerimientos* ▪ *Riesgos y Contingencias* ▪ *Terminologías* ▪ *Costos y beneficios*	**Descripción de los datos:** ▪ *Reporte de Descripción de los Datos*	**Seleccionar los datos:** ▪ *Inclusión/ Exclusión de Datos*	**Generar diseño de pruebas:** ▪ *Diseño de Pruebas*	**Proceso de revisión:** ▪ *Revisión del Proceso*	**Plan de monitoreo y mantenimiento:** ▪ *Plan de Monitoreo y Mantenimiento*
Determinar Objetivos del Proceso de Exploración de Información: ▪ *Metas del Proceso de Exploración de Información* ▪ *Criterios de Éxito del Proceso de Exploración de Información*	**Exploración de los datos:** ▪ *Reporte de Exploración de datos*	**Limpiar los datos:** ▪ *Reporte de limpieza de datos*	**Construir el Modelo:** ▪ *Configuración de parámetros* ▪ *Modelo* ▪ *Descripción del Modelo*	**Determinar Próximos pasos:** ▪ *Lista de Posibles Acciones* ▪ *Decisión*	**Armado del Informe Final:** ▪ *Informe Final* ▪ *Presentación Final*

Comprensión del negocio	Comprensión de los datos	Preparación de los datos	Modelado	Evaluación	Implementación
Realizar el Plan del Proyecto: ▪ *Plan de Proyecto* ▪ *Validación inicial de técnicas y herramientas*	**Verificación de calidad de los datos:** ▪ *Reporte de Calidad de Datos*	**Construcción de datos:** ▪ *Atributos derivados* ▪ *Generación de Registros*	**Evaluar el modelo:** ▪ *Evaluar el Modelo* ▪ *Revisión de la Configuración de Parámetros*		**Revisión del proyecto:** ▪ *Documentación de la experiencia*
		Integrar los datos: ▪ *Unificación de los Datos*			
		Formato de los datos: ▪ *Reformatear los Datos*			

Tabla i. Fases de la metodología CRISP DM.

Glosario

Ámbito: Se refiere a la ubicación geográfica que se encuentra el establecimiento educativo. Sus valores son Urbano o Rural.

Edad Promedio: Correspondiente a la edad promedio de los alumnos de un establecimiento. Se muestra un campo con la edad promedio del total de los alumnos (ED_PROM) y otros que muestran la edad promedio de cada uno de los años o grados escolares del establecimiento (ED_PROM_(Nro de año o grado)). Estos campos no son datos sino que fueron calculados a partir de otros.

Egresados: Correspondiente a la cantidad de alumnos egresados del año en curso. El campo se visualiza con el nombre de EGRE_NIV para el total de egresados por nivel. Este campo está presente en los niveles MP y SNU solamente. Al igual que la división de los matriculados en el nivel MP según modalidades, los egresados siguen la misma nomenclatura denominándose EGRE_NIV_MOD. Para el nivel SNU el campo es EGRE_NIV_CAR dividiendo los registros por carreras.

Establecimiento Educativo: Es la unidad donde se organiza la oferta educativa, cuya creación o autorización se registra bajo un acto administrativo –ley, decreto, resolución o disposición. Existe una autoridad máxima como responsable pedagógico y/o administrativo, con una planta orgánico-funcional asignada, para impartir educación a un grupo de alumnos. El establecimiento constituye la unidad organizacional que contiene en su interior a la/s unidad/es educativa/s, las cuales forman parte del establecimiento y se corresponden con cada uno de los niveles de enseñanza para los cuales se imparte educación. Dicha educación puede darse en el mismo lugar físico donde se encuentra el responsable pedagógico y/o administrativo, fuera del mismo, o en forma combinada, independientemente de la organización y modalidad de prestación (presencial o a distancia). Para la educación formal, el establecimiento puede organizar el servicio en una o más unidades educativas.

Matrícula: Corresponde a la matrícula inicial y corresponde a la cantidad de alumnos matriculados según situación al 30 de abril del año en curso. El campo que muestra los matriculados anuales en cada nivel se presenta bajo el nombre de MAT_NIV, mientras que si se hace una distinción por año el campo lleva el nombre de MAT_NIV_1, por ejemplo, para primer año escolar. Para el nivel PEGB los matriculados se dividen por Tipo de Sección, es por ello que el nombre del campo es MAT_NIV_TIPO para el total de matriculados del establecimiento, agregándose un número al final del campo que corresponde al grado o año escolar. Las tablas del nivel MP presentan como campo dato a los matriculados divididos por modalidades, con lo cual se diferenció dicho campo bajo el nombre de MAT_NIV_MOD para el total de los alumnos y, para cada año en particular, se le agrega un número al final del campo que corresponde al grado o año escolar, como se comentó anteriormente. Finalmente, en cuanto al nivel SNU el campo dato de los matriculados se divide por carrera, con lo cual el nombre del mismo es MAT_NIV_CAR.

Nivel: Los niveles de enseñanza son los tramos en que se estructura el sistema educativo formal. Se corresponden con las necesidades individuales de las etapas del proceso psico-físico-evolutivo articulado en la del desarrollo psico-físico-social y cultural. Los niveles son:

- **Nivel Inicial (INI)**: Tiene por objeto la socialización y educación temprana y asistencia adecuada, que garantice la calidad de los resultados en todas las etapas de aprendizaje. En la Educación Común se orienta a niños/as de 45 días a 5 años de edad, siendo para aquellos de 5 años obligatorio.
- **Nivel PRIMARIO/EGB 1, 2 y 3 (PEGB)**: Tiene por objeto la adquisición de competencias básicas, la apropiación de conocimientos elementales y comunes, imprescindibles para toda la población. La Educación PRIMARIA/ EGB 1 y 2 es obligatoria. El EGB 3 es el ciclo que continúa a partir de EGB 2 y es obligatorio.
- **Nivel MEDIO/POLIMODAL (MP)**: Profundiza el conocimiento en un conjunto de saberes según orientaciones científicas, técnicas, humanísticas, sociales etc. Para Educación Común, tiene una duración de 3 años como mínimo después del cumplimiento de la Educación General Básica o de la Primaria. Trayectos Técnicos Profesionales y/o Itinerarios Formativos son una oferta complementaria integrada a polimodal. En ello los alumnos reciben, además del título de base de la orientación polimodal cursada, una o más certificaciones. Los datos aparecen segmentados por el campo MODALIDAD, que, como su nombre lo indica, es la modalidad que se enseña en dicho establecimiento.
- **Nivel Superior (S)**: Es la formación académica de grado para el ejercicio de la docencia, el desempeño técnico, profesional, artístico o el conocimiento y la investigación científico-tecnológica a través de instituciones no universitarias (SNU) y universitarias (SU). Para el nivel SNU, que es el que se estudia en este trabajo, se segmenta por CARRERA. Dependiendo de las carreras que se enseñen en un establecimiento, esto es el Tipo de Formación (TIPOFORMAC), los establecimientos pueden ser "Exclusivamente Técnico – Profesional", "Exclusivamente Docente" o "Ambos tipos de formación".

Planta funcional (POF): Es el conjunto de cargos y horas cátedra asignados legal y presupuestariamente al establecimiento, estén éstos cubiertos o sin cubrir, independientemente de que quienes los ocupen estén en uso de licencia, comisión de servicio o tareas pasivas. Para los establecimientos privados, también incluye las horas y cargos no subvencionados o extracurriculares. La POF tiene 2 valores, "Dentro" o "Fuera". Con lo cual, este campo nos informa si los establecimientos están *dentro* de lo presupuestado en horas, cargos y módulo o *fuera*.

Repetidos: Correspondiente a la cantidad de alumnos repetidos durante el año en curso. El campo se visualiza con el nombre de REP_NIV, haciendo también una división para los alumnos repetidos de los diferentes años de los niveles con el nombre de REP_NIV_(Nro de año o grado). Este campo tiene dependencia con los campos de matriculados, a fin de que las divisiones que en este existan se visualizarán también en los campos de los alumnos repetidos.

Sector de Gestión: Alude a la responsabilidad de la gestión de los servicios educativos. La gestión puede ser Estatal o Privada. El campo correspondiente lleva el nombre de SECTOR. Los valores posibles son:

- **Estatal**: establecimientos administrados directamente por el Estado.
- **Privada**: establecimientos administrados por instituciones o personas particulares que pueden ser o no subvencionados por el Estado.

Sede: Es el lugar donde cumple sus funciones la máxima autoridad del establecimiento como responsable pedagógico y/o administrativo. La sede puede no tener alumnos. El anexo es la sección o grupo de secciones que depende administrativa y/o pedagógicamente de un establecimiento sede y funciona en distintos lugares geográficos. En caso que el establecimiento imparta educación en un anexo el campo SEDE tendrá como valor un "2".

Tipos de Educación: Son las diferentes formas en que se organiza la educación formal en función de la población a la que se dirige, definida a partir de la edad de los alumnos, de sus necesidades educativas, o de sus inquietudes o motivaciones. Cada uno de los tipos de educación cuenta con una organización curricular específicamente diseñada, con modalidades pedagógicas particulares y una articulación interna en niveles de complejidad creciente. Los Tipos de Educación son: Común, Especial, Adultos y Artística.

- **Educación Común**: Se dirige a la educación de la mayor parte de la población, para la adquisición de los conocimientos, las destrezas y las capacidades que la estructura del sistema educativo prevé en los plazos preestablecidos y en las edades teóricas previstas. Los contenidos apuntan a la formación general y homogénea, permitiendo la especialización a medida que el alumno avanza en la complejidad y en los niveles educativos. Contiene los siguientes niveles: Inicial, Primario/EGB1 y 2, EGB3, Medio/polimodal y Superior Universitario y Superior No Universitario (de formación docente y de formación técnico/profesional).
- **Educación Especial**: Se dirige a las personas cuyos procesos de aprendizaje se ven dificultados por motivos de origen psico-físico y social, por lo cual, requieren atención educativa particular, ya sea de manera transitoria o permanente. Sus estrategias de enseñanza se caracterizan por una alta flexibilidad y por tanto variabilidad (sistemas diferentes de organización de contenidos, de evaluación y de acreditación), definidas a partir de la problemática específica que presenten los sujetos. Contiene los siguientes niveles: Inicial, Primario/EGB1 y 2, EGB3, Medio/polimodal.
- **Educación de Adultos**: Son los procesos educativos organizados por los cuales jóvenes y adultos mejoran sus capacidades técnicas y profesionales, desarrollan sus habilidades o enriquecen sus conocimientos con los propósitos de completar un nivel de educación formal, adquirir o actualizar conocimiento y habilidades en un área específica. Contiene los siguientes niveles: Primario/EGB1 y 2, EGB3, Medio/polimodal.
- **Educación Artística**: Responde a la necesidad de aquellas personas que a partir de diferentes motivaciones, inquietudes e iniciativas demandan una educación en los diversos campos del arte. Contiene los siguientes niveles: Inicial, Primario/EGB1 y 2, EGB3, Medio/polimodal y Superior Universitario y Superior No Universitario (de formación docente y de formación técnico/profesional).

Tipos de designación: Situación legal administrativa bajo la cual el docente desempeña sus actividades educativas. Se refiere a cada uno de los tipos de puestos de trabajo con que cuenta un establecimiento educativo, que tiene asignada una partida presupuestaria y un conjunto de tareas a desempeñar por una persona. Los tipos y número de cargos, horas cátedra y módulos están vinculados con la matriz curricular y las dimensiones del establecimiento. Las categorías son designación por cargo, designación por hora cátedra y designación por módulo.

- **Cargo**: es el puesto de trabajo definido en función de una determinada carga horaria (organizada de acuerdo a horas reloj) y de determinadas tareas a desarrollar.
- **Hora cátedra**: es la unidad mínima de tiempo (40-50 minutos) para desarrollar actividades de enseñanza–aprendizaje en un establecimiento educativo. La hora cátedra constituye la unidad de medida más frecuente para contratación del personal docente en los niveles EGB3, Medio/polimodal y Superior no universitario (aunque puede verificarse también en otros tipos de educación y ciclo/nivel). Las horas cátedra se destinan principalmente al dictado de clases, pero también pueden ser dedicadas a capacitación, actividades de extensión, investigación u otras. El conjunto de horas cátedra aprobado, así como la distribución de las mismas por materia se fundamenta en el plan de estudios respectivo. La distribución de dichas horas por docente es competencia del establecimiento educativo.
- **Módulo**: es similar a la hora cátedra, pero su duración es de 60 minutos.

Tipos de Sección: Grupo escolar formado por alumnos que cursan en el mismo espacio, al mismo tiempo y con el mismo docente o equipo de docentes. Pueden estar cursando el mismo o diferentes grado. El nombre del campo es TIPO _ SE y los valores son:

- **Sección Independiente (I)**: Las actividades de enseñanza corresponden a un mismo grado o año. Incluyen las independientes de recuperación (enseñanza personalizada), independientes mixtas (diferentes modalidades) e independientes semipresenciales.
- **Secciones Múltiples (M)**: Las actividades de enseñanza corresponden a varios años de estudio. Incluyen las múltiples de recuperación, múltiples de multinivel y múltiples semipresenciales.

Características de los campos obtenidos de las tablas de DINIECE

Tabla	Campo	Tipo de dato	Observaciones de la tabla
CAR2005	ID_RA	Numérico	Presenta información sobre la planta funcional de los establecimientos en cada uno de los niveles. También muestra numéricamente los cargos y las horas desarrolladas por cada uno de ellos.
	NIVEL	Texto	
	POF	Texto	
	CARGO1	Numérico	
	CARGO2	Numérico	
	CARGO3	Numérico	
	CARGO4	Numérico	
	CARGO5	Numérico	
	HORA1	Numérico	
	HORA2	Numérico	
EDI2005	ID_RA	Numérico	Muestra la cantidad de alumnos que se ubican en los diferentes rango de edades que se señalan en cada campo, para todos los establecimientos del nivel Inicial.
	NIVEL	Texto	
	ED2	Numérico	
	ED3	Numérico	
	ED4	Numérico	
	ED5	Numérico	
	ED6	Numérico	
EDPE2005	ID_RA	Numérico	Muestra la cantidad de alumnos que se ubican en los diferentes rango de edades que se señalan en cada campo, para todos los establecimientos del nivel Primario/EGB. Los datos se encuentran divididos por ano o grado escolar correspondiente.
	NIVEL	Texto	
	ANIO_ES	Numérico	
	ED5	Numérico	
	ED6	Numérico	
	ED7	Numérico	
	ED8	Numérico	
	ED9	Numérico	
	ED10	Numérico	
	ED11	Numérico	
	ED12	Numérico	
	ED13	Numérico	
	ED14	Numérico	
	ED15	Numérico	
	ED16	Numérico	
	ED17	Numérico	
	ED18	Numérico	

EDMP2005	ID_RA	Número	Muestra la cantidad de alumnos que se ubican en los diferentes rango de edades que se señalan en cada campo, para todos los establecimientos del nivel Medio/Polimodal. Los datos se encuentran divididos por año o grado escolar correspondiente.
	NIVEL	Texto	
	ANIO_ES	Número	
	ED11	Número	
	ED12	Número	
	ED13	Número	
	ED14	Número	
	ED15	Número	
	ED16	Número	
	ED17	Número	
	ED18	Número	
	ED19	Número	
	ED20_24	Número	
	ED25	Número	
EDS2005	ID_RA	Número	Muestra la cantidad de alumnos que se ubican en los diferentes rango de edades que se señalan en cada campo, para todos los establecimientos del nivel Superior no Universitario.
	NIVEL	Texto	
	ED17	Número	
	ED18	Número	
	ED19	Número	
	ED20	Número	
	ED21	Número	
	ED22	Número	
	ED23	Número	
	ED24	Número	
	ED25	Número	
	ED26	Número	
	ED27	Número	
	ED28	Número	
	ED29	Número	
	ED30_34	Número	
	ED35_39	Número	
	ED40	Número	
EMP2005	ID_RA	Número	Presenta la cantidad de alumnos egresados en cada establecimiento, según la modalidad elegida en el nivel Medio/Polimodal.
	NIVEL	Texto	
	MODALIDAD	Texto	
	EGRESADOS	Número	
ESNU2005	ID_RA	Número	Presenta la cantidad de alumnos egresados en cada establecimiento, según la carrera elegida en el nivel Superior no Universitario. También muestra el tipo de formación del establecimiento en cuestión.
	NIVEL	Texto	
	CARRERA	Texto	
	TIPOFORMAC	Texto	
	EGRESADOS	Número	
MAE2005	PROVINCIA	Texto	Presenta información sobre todos los establecimientos nacionales. Muestra los niveles que abarca, el ámbito y sector al que pertenecen, y la provincia y el departamento donde se ubica.
	ID_RA	Número	
	SEDE	Número	
	AMBITO	Texto	
	SECTOR	Texto	
	EGB12	Marca	
	INICIAL	Marca	
	MEDIO	Marca	
	PRIMARIO	Marca	
	SNU	Marca	
	EGB3	Marca	
	POLIMODAL	Marca	
	DEPARTAMEN	Texto	
	LEY	Número	

MAT2005	ID_RA	Número	Presenta la cantidad de alumnos matriculados y repetidos en cada año, para cada establecimiento, para todos los niveles educativos. También muestra el tipo de sección del establecimiento en cuestión.
	NIVEL	Texto	
	TIPO_SE	Texto	
	ALU1	Número	
	ALU2	Número	
	ALU3	Número	
	ALU4	Número	
	ALU5	Número	
	ALU6	Número	
	ALU7	Número	
	ALU8	Número	
	ALU9	Número	
	ALUAP	Número	
	REP1	Número	
	REP2	Número	
	REP3	Número	
	REP4	Número	
	REP5	Número	
	REP6	Número	
	REP7	Número	
	REP8	Número	
	REP9	Número	
	REPAP	Número	
MMP2005	ID_RA	Número	Presenta la cantidad de alumnos matriculados en cada ano para cada establecimiento, según la modalidad elegida en el nivel Medio/Polimodal.
	NIVEL	Texto	
	MODALIDAD	Texto	
	MAT1	Número	
	MAT2	Número	
	MAT3	Número	
	MAT4	Número	
	MAT5	Número	
	MAT6	Número	
	MAT7	Número	
MSNU2005	ID_RA	Número	Presenta la cantidad de alumnos matriculados en cada año para cada establecimiento, según la carrera elegida en el nivel Superior no Universitario. También muestra el tipo de formación del establecimiento en cuestión.
	NIVEL	Texto	
	CARRERA	Texto	
	TIPOFORMAC	Texto	
	MATRICULA	Número	

Tabla ii. Características de los campos obtenidos de las tablas de DINIECE.

Valores del campo CARRERA

CARRERA
ACTUALIZACIÒN PEDAGÒGICA
ADMINISTRACION
AGROPECUARIA
ANESTESIOLOGIA
ARCHIVOLOGIA
ARTE
ARTES AUDIOVISUALES
ARTES GRAFICAS
ARTES PLASTICAS/VISUALES
ORGANIZACIÓN Y GESTIÓN
ENFERMERIA
BIBLIOTECOLOGIA
BIOLOGIA
CALIGRAFIA Y ESTENOGRAFIA
CANTO
CASTELLANO
TURISMO
EDUCACION PRIMARIA/ EGB
CERAMICA
CIENCIAS DE LA EDUCACION
RADIOLOGIA Y AFINES
CIENCIAS NATURALES
CIENCIAS POLITICAS
EDUCACION NIVEL INICIAL
CIENCIAS SOCIALES
LABORATORIO
COMERCIALIZACIÓN
INSTRUMENTACION QUIRURGICA
COMERCIO EXTERIOR
COMPUTACION
IDIOMA INGLES
COMUNICACIÓN SOCIAL
COMUNICACIONES
DIBUJO Y DISEÑO
CONSTRUCCIONES
CONTABILIDAD
CONTROL DE PROCESOS
LENGUA
HISTORIA
COOPERATIVISMO
DANZA
DECORACION
MUSICA
DEPORTES
INSTRUMENTO
TECNOLOGIA
DERECHO
DIBUJO
MATEMATICA
DIRECCION CORAL
DIRECCION ORQUESTAL
ECONOMIA
EDUCACIÓN DEL ADULTO
HOTELERÍA Y GASTRONOMÍA
EDUCACION ESPECIAL
EDUCACION FISICA
INDUSTRIA DE LA ALIMENTACION
ELECTRICIDAD
ELECTRONICA
ESCENOGRAFIA
SISTEMAS
ESCULTURA
INFORMATICA
ESTADISTICA
FILOSOFIA
SEGURIDAD INDUSTRIAL
FISICA
RELACIONES PUBLICAS
ODONTOLOGIA
FONOAUDIOLOGIA
FORMACION ETICA Y CIUDADANA
FOTOGRAFIA
GEOGRAFIA
GRABADO
GRAFOLOGÍA
HEMOTERAPIA E INMUNOHEMATOLOGIA
HIGIENE
PSICOLOGIA
IDIOMA ALEMAN
SERVICIO SOCIAL
IDIOMA FRANCES
IDIOMA ITALIANO
TEOLOGÏA
IDIOMA PORTUGUES
IMPUESTOS
INGENIERIA FORESTAL
PSICOPEDAGOGIA
LETRAS
PEDAGOGIA
LITERATURA
MECANICA
MEDIO AMBIENTE/ ECOLOGÍA
MINAS
OTRAS
TEATRO
MUSEOLOGIA
SECRETARIADO
OTRAS ESPECIALIDADES
ROBOTICA
NAVEGACION
QUIMICA
NUTRICION
SANIDAD
SEGURIDAD PUBLICA
OBSTETRICIA
OPTICA
VETERINARIA
RELACIONES HUMANAS
OTROS
PINTURA
PRODUCCION E INSTALACIONES INDUSTRIALES
PSICOMOTRICIDAD
RELACIONES LABORALES
TERAPIA OCUPACIONAL
VITRAL

Tablas iii. Tipos de carreras.

Herramienta de modelización y ejecución de algoritmos

Paleta de nodos

- Orígenes. Nodos utilizados para introducir datos.
- Operaciones con registros. Nodos utilizados para operaciones con registros de datos como la selección, la fusión y la adición.
- Operaciones con campos. Nodos utilizados para operaciones con campos de datos como el filtrado, la derivación de campos nuevos y la determinación del tipo de datos de campos dados.
- Gráficos. Nodos utilizados para visualizar los datos antes y después del modelado. Entre ellos se incluyen gráficos, histogramas, nodos de malla y diagramas de evaluación.
- Modelado. Nodos que representan los potentes algoritmos de modelado, tales como las redes neuronales, los árboles de decisión, los algoritmos de conglomerados y las secuencias de datos.
- Resultado. Nodos utilizados con el fin de producir una variedad de resultados para los datos, los gráficos y los resultados de los modelos.

Conceptos básicos

Los nodos de origen permiten importar los datos almacenados en distintos formatos, entre los que se incluyen archivos planos, Microsoft Excel y bases de datos (.db). También puede generar datos sintéticos mediante el

Conceptos básicos de las operaciones con campos

Una vez que haya realizado una exploración de datos inicial, es posible que tenga que seleccionar, limpiar o construir datos para preparar el análisis. La paleta Operaciones con campos contiene muchos nodos útiles para esta transformación y preparación.

Por ejemplo, con un nodo Derivar se puede crear un atributo que no se representa en los datos en la actualidad. También se puede utilizar un nodo Intervalos para volver a codificar automáticamente valores de campos en análisis objetivos. Comprobará que con frecuencia utiliza un nodo Tipo, lo cual se debe a que permite asignar tipos de datos, valores y papeles de modelado para cada campo del conjunto de datos. Estas operaciones son útiles para gestionar valores perdidos y modelado posterior en la ruta.

Conceptos básicos de las operaciones con registros

Los nodos de operaciones con registros se usan para realizar cambios en los datos a nivel de registro. Estas operaciones son importantes durante las fases Comprensión de los datos y Preparación de los datos del análisis porque permiten adaptar los datos a las necesidades particulares de su negocio.

Por ejemplo, según los resultados de la auditoría de datos realizada con el nodo Auditar datos (paleta Resultado), puede que desee fusionar los registros de las compras realizadas por los clientes durante los últimos tres meses. Con el nodo Fundir, puede fusionar registros basándose en los valores de un campo clave, como el ID de cliente.F. Así mismo, puede descubrir que es imposible administrar una base de datos con información sobre visitas al sitio Web con más de un millón de registros. Con el nodo Muestrear, puede seleccionar un subconjunto de datos para utilizarlo en el modelado.

Conceptos básicos de los modelos generados

Los modelos generados son el resultado de la tarea de modelado de datos. Siempre que se ejecuta correctamente un nodo de modelado, se crea un nodo de modelo generado. Los modelos generados contienen información sobre el modelo creado y ofrecen un mecanismo para utilizar dicho modelo para generar pronósticos y facilitar la tarea de una minería más profunda de los datos.

Cuando se crean dichos modelos, se colocan en la paleta de modelos generados. Se pueden seleccionar y examinar para ver detalles del modelo. Los modelos generados que no sean modelos de reglas sin refinar se pueden colocar en la ruta para generar pronósticos o para llevar a cabo un análisis más detallado de sus propiedades.

Se puede identificar el tipo de un nodo de modelo generado por su icono:

Icono	Tipo de nodo	Icono	Tipo de nodo
	Red neuronal		Red de Kohonen
	Modelo de árbol C5.0		Ecuación de regresión lineal
	Conjunto de reglas		Modelos de *K*-Medias
	Ecuación de regresión logística		Modelo de árbol C&R
	Ecuación de PCA/Factorial		Conjunto de secuencias
	Modelo A priori		Modelo CARMA
	Árbol QUEST		Modelo de extracción de texto
	Modelo de selección de características		Modelo de detección de anomalías
	Modelos sin refinar, como los modelos GRI y CEMI (sólo paleta de modelos generados)		Árbol CHAID

Conceptos básicos sobre nodos de resultados

Los nodos de resultados ofrecen los medios para obtener información acerca de los datos y los modelos. También proporcionan un mecanismo para exportar datos en varios formatos y poder interactuar con otras herramientas de software.