Identificación de Patrones de Comportamiento de Oficios Judiciales en el Gabinete de Procesamiento

Trabajo Final de Grado
Ingeniería de Sistemas

INSTITUTO UNIVERSITARIO AERONÁUTICO

Alumna: GIUBBANI, CINTIA ANAHÍ

Tutor: Mgter. CIOLLI, MARÍA ELENA

AÑO: 2016

Declaración de derechos de autor

Declaro en forma libre y voluntaria que la presente investigación y elaboración del Trabajo Final de Grado: "Identificación de Patrones de Comportamiento de Oficios Judiciales en Gabinete de Telecomunicaciones" como así también las expresiones vertidas en la misma son de autoría propia realizada en base a recopilación bibliográfica y consultas en internet debidamente citadas.

Debido al tipo de información que aquí se describe, se solicita la confidencialidad de los datos.

Dedicatoria

A mi compañero de vida.

Agradecimientos

A mi familia por la paciencia, la comprensión y el apoyo.

Resumen

La Información, como objeto de análisis de investigaciones judiciales, y la minería de datos, como herramienta para extraer conocimiento, encuentran su punto de fusión en la búsqueda de patrones de comportamiento que permitan definir estrategias y metodologías de investigación. Por medio de la investigación, se pretende ayudar a los operadores, a descubrir patrones de comportamiento e identificar tendencias en los casos judiciales que ingresan día a día, observando características o atributos que influyen en las investigaciones judiciales y que permitan la definición de un planteo táctico.

Hoy en día, la información necesaria, se obtiene de los datos recolectados a través de los oficios solicitados por los distintos organismos Judiciales. Evaluando dichos datos, se vislumbra la necesidad de estudiar el comportamiento de los operadores, quienes, se encargan de concretar los pedidos realizados en los oficios. El análisis de dicho rol, permitirá al encargado anticipar y determinar el accionar de la oficina.

El presente estudio, se ocupa de explorar los datos disponibles, construir modelos, validarlos, estudiarlos y exponer los resultados a disposición de quien esté encargado del análisis estratégico de la oficina.

El uso de computadoras, permiten llevar a cabo algoritmos complejos, analizando un elevado volumen de datos, que la mente humana no podría procesar por si sola. La información es confrontada entre sí, evaluada y ponderada según su eficiencia.

Dicho proceso, tiene como resultado la identificación de patrones de comportamiento, relevantes y significativos, tanto de los oficios judiciales que ingresan de las diferentes entidades, como de las aptitudes de cada uno de los operadores, que permiten anticipar la manera de actuar del encargado desde diferentes perspectivas.

En este trabajo permite demostrar como las herramientas informáticas pueden ser protagonistas en la toma de decisiones de un encargado, cualquiera sea el ámbito de trabajo.

Poder contar con herramientas que estudien específicamente este rol (operador), puede ser el punto de partida para ayudar a un encargado a predecir estrategias válidas y sacar ventaja al momento de llevar adelante la resolución de oficios judiciales

Índice de contenidos

1. INTRODUCCIÓN

1.1. ANTECEDENTES

"El Análisis de la información obtenida de investigaciones, es el cerebro de la seguridad nacional. Se recurre a ello, tanto para guiar estratégicas, movimientos oportunos y lograr resolver casos en todo el mundo. Un análisis equivocado de la misma puede ser totalmente inútil para la resolución de casos delictivos."[1]

La ayuda que puede recibir un investigador ante la toma de decisiones en el análisis de casos puede resultar crucial en una exhaustiva investigación.

La minería de datos centrado en el análisis de información, a través de modelos validados, permite trabajar a nivel del conocimiento, con el fin de descubrir patrones, relaciones y asociaciones para la toma de decisiones en el análisis de las causas judiciales.

Dado a la experiencia personal de estar trabajando en el ámbito judicial, específicamente, en el Gabinete de Procesamiento, veo la necesidad de aplicar técnicas de minería de datos, capaces de brindar herramientas que permitan descubrir conocimiento, patrones e identificar tendencias en los casos judiciales que ingresan día a día. Estas particularidades serán claves en la definición de un planteo táctico. Desde el punto de vista judicial, se tiene acceso a datos e información de equipos móviles de distintas causas judiciales. Todo ello recolectado a través de los distintos oficios solicitados por Fiscalías de Instrucción, Juzgados y Unidades Judiciales, lo que se considera como la "materia prima" necesaria en la concreción de este proyecto.

1.2. HISTORIA Y CONCEPTOS CLAVES

1.2.1 Historia Universal

La Policía Judicial de Córdoba es una institución que tiene vida normativa desde el año 1.939. Es recién en el año 1.958 donde se le proporcionó estructura administrativa, y se creó, dentro del ámbito del Poder Judicial, el "Departamento de Policía Judicial", que se hizo depender del Tribunal Superior De Justicia, en lo administrativo, y funcionalmente quedó bajo la autoridad del Fiscal De La Cámara Del Crimen De Turno. Esta

legislación, sólo hizo referencia a gabinetes de especialidades técnicas auxiliares de la investigación penal (pericias automotrices, pericias físico- mecánicas, balística, grafo crítica, reconstrucción gráfica del rostro, huellas y rastros, planimetría legal, entre otras).

El 30 de abril de 1987, entró en vigencia una nueva "Constitución De La Provincia De Córdoba", en la cual se incorporó a la Policía Judicial, como una institución u organismo del Estado, y lo puso bajo la dirección del Fiscal General de la Provincia y todo esto dentro del Poder Judicial.

Siendo una institución novedosa dentro de la tradición judicial Argentina, la Policía Judicial de Córdoba fue pionera y señera en el país, y en esta calidad, fue observada por innumerables gobiernos de provincia y por partidos políticos que incluyeron el modelo Córdoba en sus plataformas.
Hoy, muchas provincias cuentan con estructuras similares a esta, gracias al valioso asesoramiento que recibieron de funcionarios con experiencia de la Policía Judicial de Córdoba.

En conclusión se puede decir que es una institución de carácter profesional técnico-científico, que colabora con la administración de justicia dentro del proceso penal en la investigación de los delitos de acción pública como órgano auxiliar del Ministerio Público Fiscal. Investiga, Impide, Individualiza y Reúne.

1.2.2. Gabinete de Procesamiento - Función

El Gabinete de Telecomunicaciones, creado en el año 2009 dentro del Departamento de Comportamiento Criminal de Policía Judicial, tiene como tarea gestionar y analizar toda la información que las empresas de telefonía almacenan derivadas de las comunicaciones de sus clientes y que son de interés en una investigación penal. Funciona como nexo entre las oficinas dependientes del Ministerio Público Fiscal, y las Gerencias legales de las empresas prestatarias de servicio telefónico en el país, centralizando así en una sola oficina todos los requerimientos judiciales que se hagan en materia de información telefónica. El personal del área se encarga de interpretar, analizar y transformar la información recabada para así poder traducirla en un lenguaje

claro para su utilización en el proceso penal, haciéndola entendible a cualquier persona que no sea idónea en la materia.

1.2.3. Sub – áreas del Gabinete

El Gabinete se encarga de coordinar y organizar administrativamente la gestión, procesamiento, sistematización, análisis y elaboración de informes con los datos que registran en sus sistemas de facturación las empresas prestatarias de servicio telefónico del país, y que pueden ser de vital importancia en la investigación de un hecho delictivo; por ello se divide en dos unidades:

- **Unidad Procesamiento de las Telecomunicaciones**: es la encargada de centralizar y gestionar ante las empresas prestatarias de servicios telefónicos la información relacionada a usuarios de los mismos cuando lo requiera un órgano judicial. Además, sistematiza la información receptada y confecciona informes para su remisión garantizando la comprensión de la misma.

- **Unidad Análisis de las Telecomunicaciones**: entiende en aquellos pedidos en los que, no sólo se solicita la información sobre una línea telefónica determinada (gestionada en la Unidad de Procesamiento de las Telecomunicaciones), sino que, además, se solicita la elaboración de un informe detallado sobre cuestiones que son de utilidad investigativa a la causa como, por ejemplo, cruces de comunicaciones entre dos o más líneas, la ubicación de líneas móviles utilizando el tráfico que éstas procesan a través de antenas, cambios de líneas producidas en un aparato, etc. Esta Unidad interpreta la información remitida por la empresa telefónica y la analiza buscando elementos útiles para la investigación penal conforme el criterio de búsqueda impuesto por quien entiende en la causa. Finalmente elabora un informe detallado y fácil de interpretar evitando el lenguaje técnico utilizado por las empresas en sus registros.

1.2.4. Principales Recursos Materiales y Tecnológicos

El Gabinete posee 8 computadores, y un plotter de última generación. Además, cuenta con un software denominado i2 Analyst´s Notebook, adquirido para representar mediante un entorno gráfico amigable, todas las conexiones existentes de una o más líneas telefónicas.

1.2.5 Recursos Humanos – El equipo de Trabajo

Cuenta con 9 operadores, entre ellos personal especializado en distintos tipos de delitos (homicidios, abusos, robos, etc.), Analistas, Ingenieros, y un encargado.
Entre los operadores, están los encargados de llevar adelante los entrecruzamientos de comunicaciones, mencionados en el apartado anterior.

1.3. SITUACIÓN PROBLEMÁTICA

Hoy en día, existen amplias razones para considerar que la información proveniente de los dispositivos móviles es de vital importancia en la investigación de delitos.

En abril de 2013, una estadística realizada por la Organización de las Naciones Unidas, relevó que existían seis mil millones de teléfonos móviles en cotejo con los siete mil millones de personas. Resulta casi asombroso el gran desarrollo de estos dispositivos y la inclusión casi infalible del celular en toda y cualquier actividad de las personas, ya sea de origen legal o criminal.

El análisis de los registros de **comunicaciones** relacionadas a **hechos criminale**s es fundamental para la conclusión de investigaciones.

En Córdoba, el Poder Judicial de la Provincia, fue uno de los pioneros en introducir el análisis de teléfonos en las investigaciones para la resolución de casos delictivos. Esto permitió a la Justicia de otras provincias, como Salta, Chubut, incursionar en el procesamiento de telecomunicaciones, mediante la guía brindada por Justicia Cordobesa.

Este tipo de análisis se ha realizado mediante herramientas estadísticas descriptivas básicas, considerando fundamentalmente relaciones primarias. Sin embargo, muchas veces la estadística descriptiva clásica no refleja la verdadera interrelación de las variables y por lo tanto, el problema real. Este contexto requiere un tratamiento más complejo que obliga a evolucionar en el análisis de información criminal.

Casi siempre, el tamaño de las **bases de datos** está establecido en aspectos como la capacidad y eficiencia de almacenamiento y no en el posterior uso o análisis de la información. Por esta razón, en muchos casos, los registros almacenados son demasiado grandes o complejos como para analizar y superan el alcance de la estadística.

Mediante las herramientas para el análisis de datos y de predicciones, se nos dan la oportunidad de gestionar y dar sentido a la información que llega, con la obtención de productos de inteligencia.

La **Minería de Datos (Data Mining)** es un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos. Busca proporcionar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas.

Ante casos de investigaciones criminales, la gran cantidad de información y de variables que se ven involucradas, justifica el uso de herramientas que permitan determinar relaciones multivariantes subyacentes.

En los juzgados de todo el mundo, la **evidencia científica en relación a la telefonía** está ganando **status de prueba foren**se de alto nivel, similar al fenómeno que fue el caso de las imágenes digitales de cámaras de vigilancia, archivos de correo electrónico y secuencias de comandos de Internet, seguido por las víctimas o sospechosos crímenes.

En este contexto, el objetivo de este trabajo es evaluar una implementación de minería de datos en el análisis de información criminal y comprobar su efectividad y valor agregado.

El encargado del área necesita conocer las características y tendencias de las causas que se encuentran registradas y aquellas que pueden llegar a ingresar a la oficina, lo cual le dará la posibilidad de anticiparse a situaciones que requieran mayor investigación, como lo son los secuestros virtuales, las desapariciones de persona, las causas con personas en riesgo de muerte, donde la información en lo que respecta a dispositivos telefónicos son claves.

Poder analizar, interpretar, estudiar, comprender y comunicar la información obtenida de las empresas telefónicas, con respecto a lo solicitado por la Justicia, puede ser la clave en el éxito o fracaso de una causa judicial.

El encargado, define una estrategia a priori, para ello utiliza los datos estadísticos recogidos de información proporcionada por la justicia e información externa (como la extraída de los medios de comunicación o de las demás áreas de investigación del Poder Judicial), para poder determinar el modo de actuación ante posibles causas graves o que pueden llegar a alterar el bienestar de la sociedad.

Lo que se destaca, a este punto, es la necesidad de enfocar, específicamente, el estudio de los datos disponibles hacia el comportamiento de los operadores. Ellos serán quienes darán respuestas a los pedidos realizados por las Fiscalías, y seguramente serán la clave para interpretar y anticipar el desempeño de la oficina. Poder contar con herramientas que estudien específicamente este rol (operador), puede ser el punto de partida para ayudar a un encargado a predecir estrategias válidas y sacar ventaja al momento de llevar adelante la resolución de oficios judiciales.

Se cuenta con estadísticas organizadas pero no se dispone de la lectura anticipada de lo que "puede o podría suceder si". Este trabajo intenta ayudar a superar esta dificultad, es decir "optimizar el proceso de planificación y distribución de expedientes y análisis de telecomunicaciones a los diferentes operarios, según el área de conocimientos de cada uno de ellos y los trabajos pendientes que puedan tener".

1.4. PROBLEMA

Se puede identificar la necesidad de estudiar las tendencias de las diferentes causas que ingresan al área, con el objetivo de identificar cuáles requieren mayor atención y análisis, y de esta forma distribuirla a los operadores correspondientes, según el nivel de conocimiento de cada uno (conocimiento en causas de homicidios, de robos, de secuestros, telecomunicaciones, analistas que realizan entrecruzamiento de comunicaciones, etc.).

Actualmente, para realizar un relevamiento a nivel general de los diferentes tipos de delitos, se lleva adelante un análisis estadístico básico, sin hacer un aprovechamiento exhaustivo de la información mediante el uso de técnicas o herramientas de Minería de Datos.

El área trabaja con diferentes entidades judiciales, como lo son las Unidades Judiciales, las Fiscalías de Instrucción y los Juzgados, todo en un ámbito provincial, pero también colabora en causas de Juzgados Federales de Córdoba como de otras provincias. Actualmente, se han incrementado en gran medida, los oficios judiciales que solicitan investigación por parte del área**.**

1.5. OBJETO DE ESTUDIO

El objeto de estudio del presente trabajo es el Gabinete de Procesamiento de las Telecomunicaciones apuntando específicamente a consolidar información estadística y hacerla disponible para la realización de análisis de datos para la toma de decisiones con respecto a la planificación de distribución de causas.

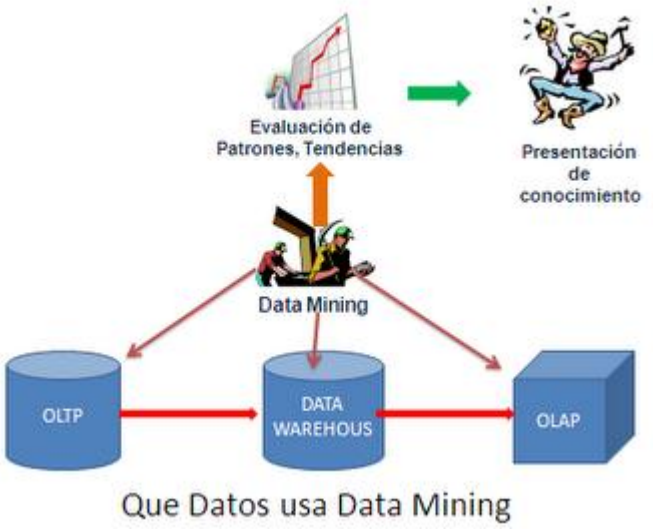
1.6. CAMPO DE ACCIÓN

"La Minería de Datos o Data Mining consiste en la extracción de información que reside de forma implícita en los datos. Se trata de información desconocida pero que puede resultar útil para efectuar algún proceso. Por lo tanto la minería de datos se ocupa de recabar, extraer, sondear, preparar y explorar los datos para sacar toda la información que ocultan". [5]

El objetivo general del proceso es **extraer información** de un conjunto de datos y **transformarla** en una estructura comprensible para su uso posterior, utilizando aspectos de bases de datos y gestión de datos, procesamiento previo, **métricas** y consideraciones de complejidad.

La minería de datos, así como el descubrimiento de conocimientos en los datos, permiten el desarrollo de estadísticas, aprendizajes automáticos y visualización de datos. Esta fusión de disciplinas se ocasiona por el significativo incremento del volumen de los datos, produciendo la necesidad de disponer de la mayor cantidad de elementos para establecer políticas de inteligencia criminal más ajustadas con base en los datos disponibles en los diferentes soportes.

Identificar variables claves y sus correlaciones, para descubrir patrones (y llegar a la creación de modelos abstractos) será la tarea primordial de este estudio.



1.7. OBJETIVOS

1.7.1. Objetivo General

Identificar patrones relevantes y significativos entre las diferentes causas que ingresan al área de las diferentes entidades judiciales, al igual que las aptitudes y conocimientos de cada uno de los operadores, con el fin de interpretar las tendencias y ayudar al encargado a la toma de decisiones en lo que respecta la planificación de tácticas a seguir y distribución de las causas a los operadores correspondientes.

El encargado del área desea resolver los siguientes requerimientos:

- Evolución de oficios:

  Seguimiento de los Expedientes comparando la cantidad de ingresos en los distintos meses del año y del año anterior, estudiando la evolución de los mismos, etc.

- Distribución de expedientes:

  Comparar los tipos de pedidos pendientes y resueltos, de cada entidad judicial, con el fin de indicar cuales son los Distritos que tiene mayor tendencias en determinadas causas (como lo son por ejemplo, los secuestros virtuales, que suelen darse en zonas donde existen Centros Penitenciaros, y por ende las causas suelen venir de determinadas Fiscalías que intervienen en el lugar), y visualizar cuales necesitan mayor atención y análisis.

- Desempeño de operadores:

  Comparar el desempeño de los distintos operadores y la evolución de dicho desempeño a lo largo del tiempo, con el objetivo de planificar quienes de ellos, serán los encargados de llevar adelante las causas que solicitan entrecruzamiento de comunicaciones.

1.7.2. Objetivos Específicos

Expresar los patrones identificados como modelo de los datos según las conclusiones a las que se arribe.

Contribuir en la interpretación anticipada de la elección del operador en función de:

- Las experiencias en distintas áreas de la justicia.
  - La rapidez con que interpreta la información recabada.
  - Experiencias con pedidos similares.
  - La entidad desde donde llega el oficio.
  - La urgencia y complejidad del pedido.
  - El momento de ingreso del pedido al área.
- Otros puntos de vista que puedan identificarse relevantes al momento de realizar el estudio.

Determinar tendencias en los diferentes tipos de causas, según la zona y ámbito en donde se desempeñan las entidades judiciales que envían los requerimientos.

1.8. IDEA A DEFENDER / PROPUESTA A JUSTIFICAR / SOLUCIÓN A COMPROBAR

1.8.1. Idea a defender

Conocer, identificar, analizar y comprender las tendencias y características de cada uno de los oficios que ingresan al área, como así también los patrones de trabajo de cada operador que integra el equipo de procesamiento de telecomunicaciones, con el fin de planificar y definir las estrategias correctas para llevar adelante investigaciones productivas.

1.8.2. Propuesta a Justificar

Existen pautas repetitivas en los tipos de oficios que ingresan, como así también en la forma de analizar, interpretar y dar respuesta de cada operador. Estas tendencias pueden ocurrir debido al ámbito de acción de cada entidad que realiza el pedido, a situaciones del momento en que ingresa el oficio al área y también a rasgos individuales de cada operador.

Se propone aplicar técnicas y herramientas de minería de datos sobre la información relevada que permita al encargado complementar el análisis actual con conclusiones de mayor valor agregado, que permita identificar patrones según los tipos de oficio.

1.8.3. Solución a comprobar

Identificar tendencias de pedidos realizados al área, como así también de cada uno de los operadores, lo cual permitiría definir estrategias que aumenten la probabilidad de concretar investigaciones de manera correcta, al menor tiempo posible y con la mayor información recabada.

1.9. DELIMITACIÓN DEL PROYECTO

Reconociendo en el operador un rol fundamental y determinante a la hora de llevar adelante una investigación, este proyecto se focalizará en el estudio del mismo y en las tendencias de los oficios ingresados. Relacionar las aptitudes, los conocimientos, el desempeño y particularidades de cada operador, permitirá al encargado poder definir de manera factible, oportuna y eficaz la mejor forma de llevar adelante un pedido de investigación.

El producto final de este proyecto será la obtención de reportes que muestren consultas sobre el contenido de los modelos evaluados. Los mismos servirán al encargado para observarlos desde la óptica que estime conveniente y le ayudarán a definir planteos tácticos.

1.10. APORTE TEÓRICO

El presente estudio, desde el punto de vista teórico, pretende enlazar la Minería de datos con el Análisis Criminal en el área en cuestión; permitiendo lograr estrategias para perfeccionar la capacidad de caracterizar con precisión, detectar y anticipar investigaciones sobre la base de un análisis exhaustivo de comportamiento predictivo y la planificación.

En esta primera aproximación, se buscará brindar un nuevo enfoque que permita avanzar hacia la implementación, a posterior, de técnicas de Inteligencia Artificial.
En una segunda etapa, se debería ampliar el trabajo de campo hacia el resto de las áreas de investigación criminal. Por el momento se intenta implementar estas técnicas en un área en particular, pero se considera que las demás áreas que se ven involucradas en una investigación están interrelacionadas.

Cabe destacar que no se puede automatizar las distintas situaciones que pueden darse para dar respuestas a una investigación, existen características ajenas a la planificación, como los rasgos psicológicos y conocimientos de cada operador, errores en la información enviada por las empresas, etc.

A pesar de todo, se considera que la aplicación de técnicas sistemáticas pueden proporcionar el medio requerido para planificar, programar y ayudar a la distribución de oficios para su posterior investigación.

1.11. APORTE PRÁCTICO

Desde un punto de vista práctico, este estudió colaborará con quien se desempeñe como encargado del Área de Telecomunicaciones de Policía Judicial. Aportará "conocimiento" que permitirá identificar patrones aplicables hacia la definición de estrategias de investigación. Es decir colaborará en la toma de decisiones.

Según las conclusiones a las que se arribe, se podrá construir una interfaz que sirva al encargado para realizar consultas e interactuar con el modelo obtenido. Dicha interfaz será un prototipo de lo que podría realizarse como definición final del proyecto, dado que la parte más importante en esta etapa inicial será la investigación y el descubrimiento o no de patrones de comportamiento, reglas de asociación, correlaciones, etc.

1.12. MÉTODOS Y MEDIOS DE INVESTIGACIÓN

A partir del planteo del problema se puede determinar claramente el tipo de investigación que se llevará adelante con el fin de lograr el objetivo y la respuesta a la pregunta propuesta.

El método seleccionado deberá servir para pronosticar el comportamiento del operador. Para llevar adelante esta tarea se deberá estudiar si existe relación y qué tan importante es entre las variables que definen las situaciones particulares de cada oficio y la manera en que el operador actúa en respuesta a ello.

Se vislumbra la necesidad de llevar adelante un estudio de tipo relacional, donde se determine la manera en cómo se relacionan entre sí las variables definidas o si es que no existe relación alguna que permita anticipar el proceder del operador.

Se busca pronosticar el valor de la variable objetivo (representa el actuar del operador) a partir del valor de las variables que la modifican.

Este tipo de investigación permite definir qué tan importante es la relación encontrada, es decir cuál es el grado de asociación.

Existe la necesidad del conocimiento del dominio del problema como estrategia esencial para llevar adelante este tipo de investigaciones. Habrá un aporte relacionado directamente a lo vivencial, la experiencia personal en el ambiente de trabajo, objeto de estudio de este proyecto.

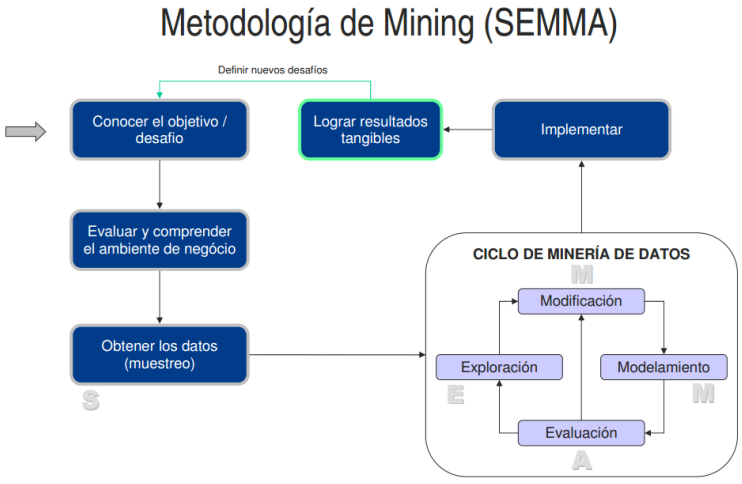
Se utilizarán también métodos empíricos que ayudarán a la recopilación de información, sobre todo a través de entrevistas con gente idónea en el tema.

1.13. MÉTODOS Y MEDIOS DE INGENIERÍA

Para elegir la metodología necesaria para la concreción del trabajo propuesto, se investigó, en primer lugar, aquellas que son las más utilizadas, teniendo en cuenta sus características fundamentales. Entre estas se pueden mencionar:

- SEMMA (Sample, Explore, Modify, Model, Assess)
- CRISP-DM (Cross Industry Standard Process for Data Mining)
- P3TQ (Product Place Price Time Quantity)

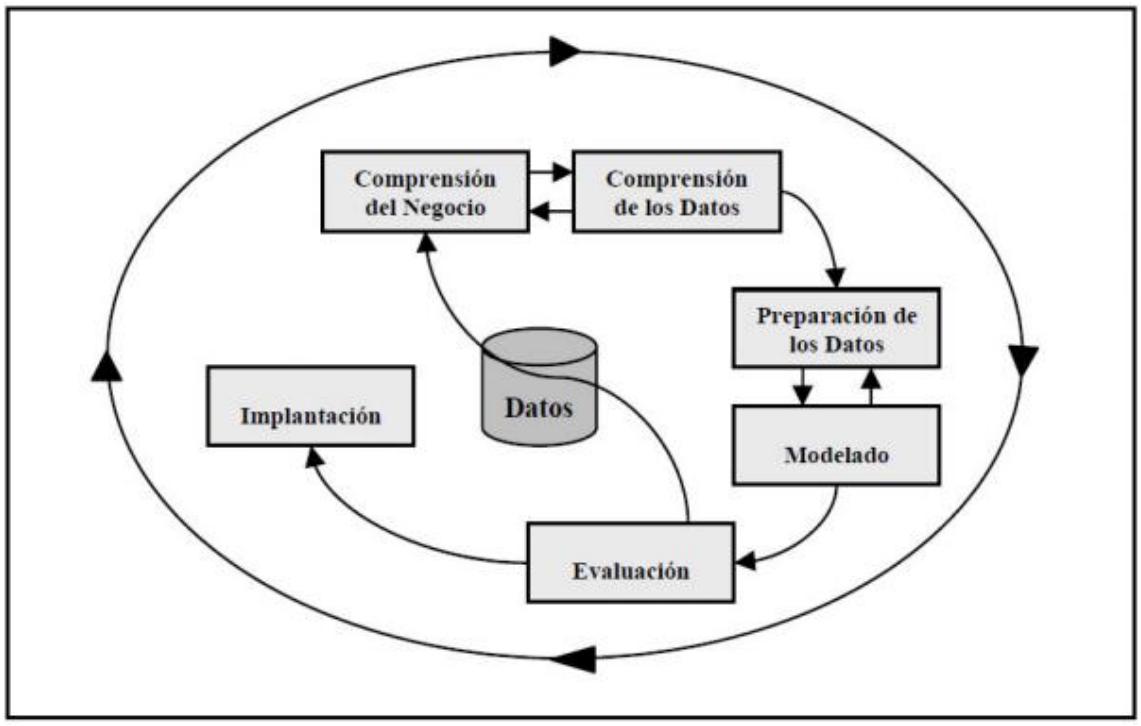
La metodología **SEMMA** se caracteriza, principalmente, por priorizar sus fases desde un punto de vista técnico, es decir, dando prioridad a las prácticas usadas para su implementación y obtención de resultados, más que en el análisis y comprensión del problema que se está abordando. Obteniendo una parte de la toda la población es como la metodología SEMMA comienza su trabajo, directamente va hacia la manipulación de datos de la Empresa, a la clasificación de variables e inmediatamente comenzar con el análisis de los mismos, con el fin de abreviar al máximo el problema que se pretende resolver. Durante su desarrollo se organizan nodos para cada una de las fases del proyecto. En esos nodos hay herramientas que llevan a cabo las tareas necesarias para poder avanzar. Se trata de una metodología en cascada pero a su vez es iterativa. Dentro de las herramientas usadas por SEMMA para la aplicación de la metodología usa productos creados por su propia Empresa "SAS". Se decide no hacer uso de dicha metodología, debido a que está ligada a un proveedor de software determinado y por su orientación técnica.



La metodología **P3TQ** o metodología **Catalyst** plantea la formulación de dos modelos, el Modelo de Negocio (MII) y el Modelo de Explotación de la Información (MIII). El primer modelo, proporciona una guía de pasos para identificar un problema de negocio u oportunidad y permite definir los requerimientos. El segundo proporciona una guía de pasos para la construcción y ejecución de modelos de minería de datos a partir del modelo de negocios. Debido a que dicha metodología se centra en el análisis de la cadena de valor organizacional definida por las relaciones entre precio/lugar/producto/tiempo/cantidad, no es considerada para su aplicación en este proyecto, dado que no se está investigando un problema u oportunidad en un negocio.

La metodología **CRISP-DM** es la más utilizada como guía en el desarrollo de proyectos de implementación de minería de datos. Incluye un modelo y una guía, estructurados en seis pasos: **Comprensión del negocio** (Objetivos y requerimientos desde una perspectiva no técnica), **Comprensión de los datos** (Familiarizarse con los datos teniendo presente los objetivos del negocio), **Preparación de los datos** (Obtener la vista minable o dataset), **Modelado** (Aplicar las técnicas de minería de datos a los dataset), **Evaluación** (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio) e **Implantación** (Despliegue, explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización). Se trata de un proceso en cascada. Generalmente, se trata el problema completo o dividido en sub problemas y se convierte en un ciclo de vida incremental. Esto permite ir construyendo modelos de minería de datos para cada uno de los sub problemas identificados. Cada fase es descompuesta en tareas y actividades. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas.

Dado que se trata de una metodología versátil, con posibilidad de adaptarse a cualquier tipo de proyecto, se considera una de la mejor opción para aplicar en este proyecto.



2. PRIMERA PARTE MARCO CONTEXTUAL

2.1. ENTORNO DEL OBJETO DE ESTUDIO

Actualmente, Argentina se vio afectada por una creciente ola de inseguridad caracterizada por un aumento en los índices delictivos y los niveles de violencia.

Los aparatos telefónicos son los protagonistas en la mayoría de estos delitos. Esto se debe a que, cuando se comete algún ilícito, dentro de la investigación, suele aparecer algún aparato celular, llamada o mensaje de texto. Ante esta situación, la Fiscalía suele oficiar a las compañías operadoras para que remitan información, ya sea para el seguimiento de comunicaciones telefónicas o para la recolección de datos de tráfico telefónico.

Cuando un celular es utilizado para cometer algún delito: como extorsionar a alguien (los comúnmente llamados secuestros virtuales), la fiscalía suele pedir el tráfico de comunicaciones de forma inmediata, aunque también puede utilizarse en casos en los que el delito es el robo del aparato.

También existen casos en los que el delito ha sido un homicidio, casos de violencia de género o desapariciones de persona, donde la información referida a los celulares, es para indicar la ubicación de una persona o conexión de esta con otras.

Es por ello que existe el Gabinete de telecomunicaciones, por donde se encaminan todos aquellos pedidos acerca de celulares, con el objetivo de prestar colaboración en causas delictivas. En la actualidad, los casos más comunes son las desapariciones de persona, donde se procede a oficiar a las empresas de telefonía de forma inmediata, con el objetivo de analizar las últimas comunicaciones de la persona desparecida, y tratar de identificar la zona de movimiento de la misma, mediante el análisis de las antenas que captaron las llamadas En aquellos casos de robo de celulares, se pide a las empresas los listados de comunicaciones, con el objetivo de identificar si los asaltantes se comunicaron con alguna otra

persona, y el número de IMEI del aparato sustraído, lo que permite observar si dicho celular está siendo operado con otra SIM CARD, y de este modo, tratar de llegar a la ubicación del mismo.

En este ambiente totalmente investigativo, la informática colabora aportando velocidad de procesamiento y capacidad de análisis brindando de esta manera un ambiente capaz de manipular datos, ordenarlos y presentarlos como información y conocimiento.

2.2. MI RELACIÓN CON EL ÁMBITO JUDICIAL

Mi incentivo para llevar adelante el presente trabajo surge de la experiencia laboral que me lleva a convivir día a día con las investigaciones referente a teléfonos, específicamente en el Análisis de telecomunicaciones. La manera más coherente de aplicar los conocimientos adquiridos durante los años de estudio es volcarlos en aquel ambiente que conozco y en el que me desempeño diariamente.

2.3. ANÁLISIS DE LOS PROBLEMAS OBSERVADOS

La idea surge del observar la incrementación de delitos en los que se ven involucrados los aparatos telefónicos, y de la necesidad de estudiar las causas para lograr un mejor análisis y la obtención de pruebas fehacientes.

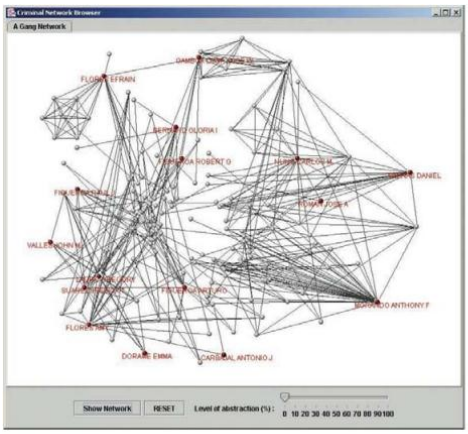
Debido a la diversidad de conocimientos de todos los operadores que trabajamos en la oficina, se deben distribuir las causas, según el tipo de hecho, al personal con mejor instrucción y de este modo optimizar tiempos y resultados.

2.4. ANTECEDENTES DE PROYECTOS SIMILARES

Se logró determinar la existencia de investigaciones realizadas por estudiantes de diferentes Universidades, en proyectos orientados a la Minería de Datos en el ámbito judicial, pero ninguno de ellos orientado específicamente a la Identificación de Patrones en los delitos de telecomunicaciones.

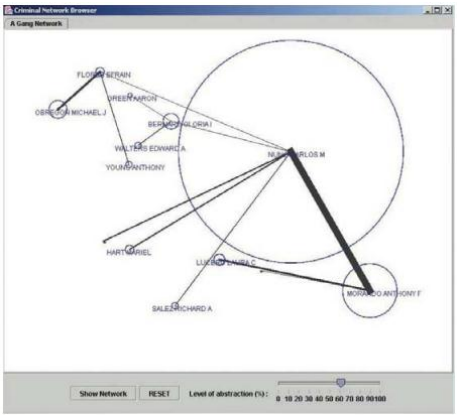
A nivel mundial, se pueden describir algunas aplicaciones de minería de datos en el análisis de información criminal:

- Proyecto COPLINK: El Proyecto COPLINK fue creado en el año 1997 en el Laboratorio de Inteligencia Artificial de la Universidad de Arizona, en Tucson, con el objetivo de servir de modelo para ser llevado a nivel nacional. Coplink está compuesto por dos sistemas integrados: Coplink Connect y Coplink Detect. El primero busca compartir información criminal entre distintos departamentos policiales, mediante un fácil acceso y una interface sencilla, integrando distintas fuentes de información. El segundo está diseñado para detectar de forma automática distintos tipos de asociaciones entre las bases de datos mediante técnicas de minería de datos.



Análisis de Redes Criminales: vínculos entre sospechosos [Coplink, 2004]

Entre otras aplicaciones Coplink provee Análisis de Redes Criminales, la cual consiste en: identificar las redes o bandas criminales, sus líderes o integrantes clave y como se relacionan entre sí. En primer lugar se utiliza la técnica de concept space para extraer relaciones de los sumarios policiales y construir una posible red de sospechosos. La fuerza del vínculo entre dos sospechosos se mide en base a la frecuencia de hechos en los que participaron ambos. Luego



se utiliza clustering jerárquico para partir la red en subgrupos y block modeling para identificar patrones de interacción entre los mismos. Finalmente se calcula el baricentro de cada subgrupo para determinar su miembro clave o líder.

- Proyecto OVER: El Proyecto OVER comenzó en el año 2000 en Reino Unido como una iniciativa conjunta de la Policía de West Midlands y el Centro de Sistemas de Adaptación y División de Psicología de la Universidad de Sunderland. El proyecto está enfocado en los casos de robo a domicilio particulares. Sus principales objetivos son: identificar los recursos críticos para establecer estrategias de prevención y detección más eficientes; proveer de fundamentos empíricos para el desarrollo de planes interdepartamentales orientados a la reducción del delito; identificar la información relevante a ser recolectada en el lugar del hecho, redundando en mejoras de eficiencia y reducción de tiempo del personal policial; alimentar al sistema tanto con información hard (información forense) como soft (información sobre la escena del delito); analizar la distribución espacio-temporal de los hechos y confirmar las suposiciones sobre tendencias y patrones.
- El Departamento de Policía de Ámsterdam utiliza el software de minería de datos DataDetective junto con Mapinfo para el análisis de registros criminales. Las principales técnicas empleadas son árboles de decisión y redes neuronales de backpropagation. Han unificado varias bases de datos policiales junto con información externa (clima, variables socioeconómicas y demográficas) en un único data warehouse.
- El Departamento de Policía de Nueva York inició en julio de 2005 el Real Time Crime Center. Este ambicioso proyecto tiene como objetivo conformar un enorme data warehouse y cruzar información de todo tipo mediante herramientas de inteligencia de negocios (como Repotnet 1.1 y Accurint Pro) de forma de detectar patrones de comportamiento y asociaciones antes desapercibidos.

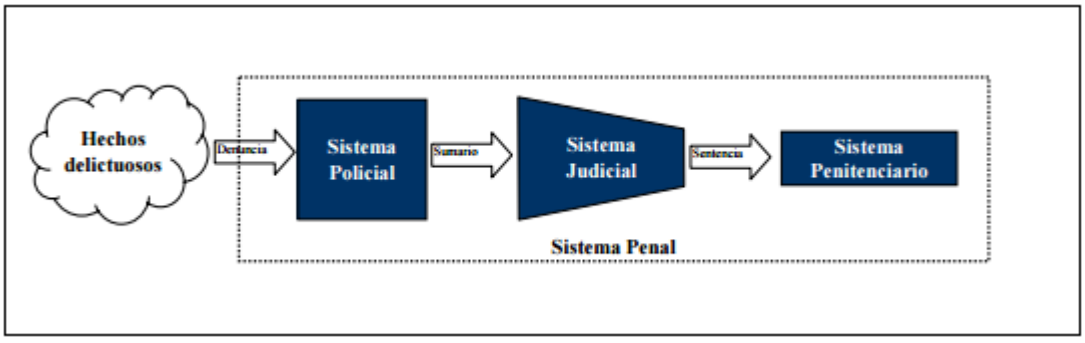
3. SEGUNDA PARTE MARCO TEÓRICO

3.1. MARCO TEÓRICO DEL OBJETO DE ESTUDIO

La Policía Judicial es una institución de carácter profesional técnico-científico, que colabora con la administración de justicia dentro del proceso penal en la investigación de los delitos de acción pública como órgano auxiliar del Ministerio Público Fiscal.

Es la encargada de Investigar, Individualizar, y reunir pruebas.

Se entiende por información criminal a toda aquella información resultante a partir de un presunto delito o de sus componentes (victima, victimario, propiedades, vehículos, etc.) que sea relevante para la toma de decisiones a posteriori. Ya sea en la prevención, detección y esclarecimiento del delito como en la persecución de delincuentes, la mejora de procesos judiciales y la creación de nuevas leyes. Según esta definición la mayor fuente de información criminal es el Sistema Penal, entendido como el conjunto de instituciones y procedimientos presentes en el proceso que transita un hecho delictuoso desde que es registrado por el Estado.



El estudio de la información de manera estratégica apunta sobre todo a ser una ayuda a la investigación judicial, ya sea a nivel provincial o nacional, donde el analizar una prueba que permita resolver un delito puede depender del ingenio de cada uno de los técnicos y especialistas, pero el resultado, es decir lograr recabar la prueba justa no es jamás una insignificancia.

Para llevar adelante el análisis mencionado, es oportuno recurrir a la ayuda de la informática, para la obtención de estadísticas. Lograr un centro de procesamiento de datos en la justicia provincial, donde se puedan realizar informes estadísticos que

permitan determinar la recurrencia de hechos en diferentes zonas, a fin de lograr adelantarse a los mismos. Es a partir de esto que se intentará llevar adelante la investigación planteada para poder obtener a través de la minería de datos, información clave para la resolución de los expedientes que ingresan.

Se hará hincapié en el uso de herramientas que permitan que los datos relevados durante un análisis dejen de ser solamente datos y se transformen en información y conocimiento útil al encargado del gabinete. Contar con información de apoyo para la toma de decisiones a la hora de distribuir las causas a los operadores, puede ser una ayuda importante y determinante en la obtención de la prueba.

Se considera a la minería de datos una herramienta efectiva a la hora de buscar una solución del problema propuesto. Esta técnica permite extraer conocimiento útil a partir de un importante volumen de datos.

Realizar pronósticos, detectar categorías, analizar indicadores, escenarios y calcular predicciones son actividades que sustentan la minería de datos.

Mediante dicha técnica, a través del agrupamiento de datos en clases, árboles de decisión y algoritmos de aprendizaje inductivos, se buscará detectar patrones ocultos y reglas que caractericen el perfil del operador que se debe asignar según el tipo de causa o hecho.

3.2. MARCO TEÓRICO DEL CAMPO DE ACCIÓN

Básicamente, el data mining surge para ayudarnos a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.
Para entender que es la Minería de Datos, es preciso expresar dos de las definiciones más importantes:

- Desde el punto de vista general, se puede mencionar la definición dada por Fayyad, Piatetsky–Shapiro y Smyth en 1996: *"La Minería de datos es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos".*
- Desde el punto de vista empresarial, Molina junto a otros autores, en el año 2001, la definen como: *"La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo a la toma de decisión".*

Se puede decir que, lo principal es el conocimiento del dominio del problema, saber que se desea hacer, el objetivo por el cual se está trabajando, la información disponible y donde se pueden encontrar los datos necesarios para ser trabajados.

Una vez que tenemos en cuenta el problema y la información disponible, se deben preparar los datos, investigarlos, generar modelos, validarlos e implementarlos, para poder ser utilizados cada vez que se necesiten.
Para llevar adelante este proceso, se escogió como herramienta **Microsoft Visual Studio 2012 - Business Intelligence Development**, que cuenta con la capacidad de realizar soluciones de inteligencia de negocio, y que además permite desarrollar proyectos de **Analysis Services, Integration Services, y Reporting Services**; Además se escogió el entorno integrado de **Microsoft SQL Server 2012,** que, al contar con dicha capacidad, permite completar todo el proceso planeado.

Se escogió Visual Studio 2012, ya que se puede acceder a él a partir de versiones libres y por contar con conocimientos previos.

Se representan los pasos a seguir en el proceso cíclico de Microsoft SQL Server en la figura 3.1.

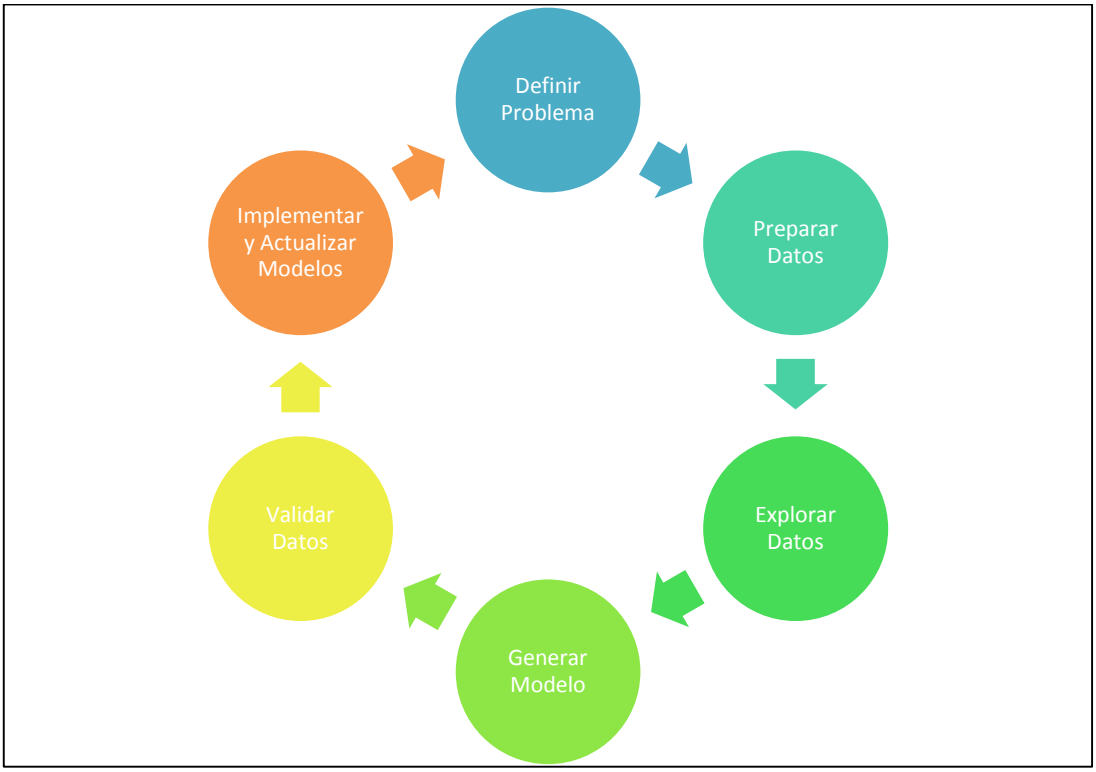
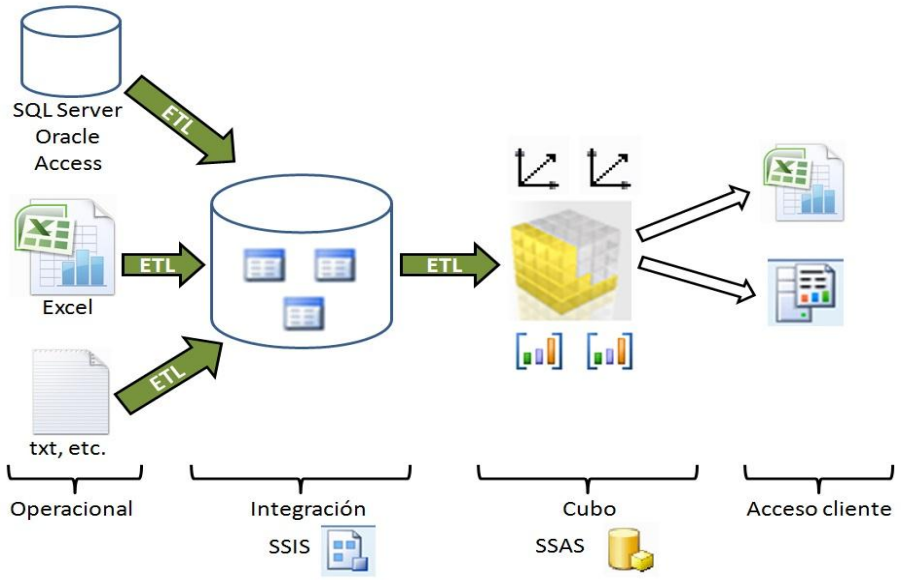


Fig. 3-1: Proceso cíclico Microsoft SQL Server 2012



1) Definir el problema: en este paso se trata de analizar los requisitos, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y de la delimitación de los objetivos concretos del proyecto de minería de datos.
2) Preparar los datos: El segundo paso del proceso de minería de datos consiste en consolidar y limpiar los datos identificados en el paso Definir el problema.

   Se deberá prestar atención a posibles inconsistencias de los datos. La limpieza de los mismos, no solamente implica quitar los datos no válidos, sino también buscar las correlaciones ocultas en los datos, identificar los orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis.
3) Explorar los datos*:* se deben conocer los datos para tomar las decisiones adecuadas. Al explorar los datos para conocer el problema, puede decidir si el conjunto de datos contiene datos defectuosos y buscar una estrategia para corregir los problemas. Se deberán limpiar los datos, seleccionarlos, integrarlos para obtener la vista minable o dataset.
4) Generar Modelos*:* este paso consiste en generar el modelo de minería de datos. Para ello se usaran los conocimientos adquiridos en el paso Explorar los datos para definir y crear los modelos, es decir, que se aplicarán las técnicas de minería de datos a los data-set obtenidos anteriormente. Se definirán las

columnas de entradas, el atributo que se intenta predecir y los parámetros que indicarán al algoritmo elegido cómo procesar los modelos.

5) Explorar y Validar los modelos*:* En este paso se debe verificar si los modelos son útiles a las necesidades, es decir comprobar su eficacia. Analysis Services proporciona herramientas que ayudan a separar los datos en conjuntos de datos de entrenamiento y pruebas, para que pueda evaluar con precisión el rendimiento de todos los modelos en los mismos datos.
6) Implementar y actualizar los modelos*:* en el último paso, el modelo ha sido implementado en producción, y se deberán presentar los patrones encontrados de manera útil al encargado de tomar las decisiones. Si fuera necesario se deberán corregir modelos, actualizarlos y volver a presentarlos.

4. TERCERA PARTE CONCRECIÓN DEL MODELO

Para llevar adelante el proyecto, se utilizó la mencionada metodología CRISP-DM. En los puntos a continuación, se exponen los pasos realizados en cada una de las fases y los resultados obtenidos.

4.1. TERCERA PARTE CONCRECIÓN DEL MODELO

Se llevó a cabo una reunión con el encargado del área, con el fin de plantear la idea propuesta.

A posterior, se realizaron reuniones específicamente para que cada uno de los involucrados, encargado y operadores, aportaran sus conocimientos, a fin de obtener una comprensión de lo que se intentaría hacer.

A continuación, se enumeran las tareas realizadas, que permitieron obtener un conocimiento completo del entorno, objetivos, herramientas y material disponible.

4.1.1. Determinar Objetivos del negocio

- *Background:* se puede decir que el contexto donde se va actuar es el ámbito judicial a nivel provincial, específicamente en el Gabinete de Telecomunicaciones. Allí se realizan tareas de análisis y de procesamiento de comunicaciones, donde existen roles definidos por los conocimientos de los operadores, según el tipo de causa con la que se esté trabajando.
- *Objetivos del Negocio:* el objetivo desde el punto de vista del negocio es lograr resolver las causas en el menor tiempo posible y con la mayor eficacia posible, ya que se tratan de causas judicializadas.
- *Criterios de éxito del negocio:* el negocio será exitoso si se lograr un nivel alto de resolución de causas de manera efectiva en el menor tiempo posible, y con la mejor investigación realizada.

4.1.2. Valoración de la Situación

- Inventario de recursos: Está la predisposición por parte de los operadores, de colaborar con la información necesaria para identificar variables, interpretar los

resultados y guiar la investigación para la obtención de resultados esperados, según el problema presentado. Además se cuenta con datos provenientes del sistema con el que cuenta el área.

- Requisitos, supuestos y restricciones:

Los supuestos para este proyecto son:

- Conocimiento del área por parte de los operadores.
- Establecer y llevar a cabo sistemáticamente las actividades que permitan cumplir con los objetivos de un proyecto en tiempos esperados.
- Alto grado de involucramiento de los operadores.

Las restricciones que posiblemente pueden surgir son:

- Falta de información precisa, debido a que no hay suficiente información formal documentada y que no se permite el acceso a la misma.

La documentación que se maneja hoy en día en el área es confidencial. Se cuenta con un sistema de registros de expedientes básico, donde las consultas se realizan de forma manual, y sin tener en cuentas indicadores, métricas ni conocimientos de los operadores, según las causas. Por otro lado, tenemos la información confidencial, ya que se trata de sábanas telefónicas, donde se la considera información privada y solos los operadores del área y personal de la justicia puede solicitarla.

Por esta razón, se considera que la falta de información formal y poco precisa existente es considerada como una posible restricción a la hora del relevamiento de información. No existen grandes limitaciones que afecten al proyecto, más allá de los conocimientos básicos de los usuarios, respecto de las herramientas informáticas. Existe suficiente compromiso, tiempo y recursos.

- Riesgos y contingencias:

RIESGO	PROBABILIDAD OCURRENCIA	CONTINGENCIA
No se cuenta con experiencia en proyectos de Minería de Datos	100%	Se cuenta con tutoriales acerca de Business Intelligence para adquirir experiencia.

Escasa cantidad de datos disponibles	60%	Las observaciones de los resultados obtenidos deberán ser llevadas a cabo en conjunto con el personal de la oficina, ya que son ellos quienes conocen el dominio del problema.
Escasa práctica en el uso de herramientas de minería de datos	70%	Antes de iniciar el proyecto se siguieron tutoriales y se realizó ejercitación necesaria para adquirir destreza.
Error al estimar el tiempo total del proyecto	50%	Si el tiempo excediera los límites académicos permitidos se buscará ayuda de terceros para reforzar la capacitación y poder así concluir el mismo.

Tabla 4.1: Riesgos y contingencias

- Costes y Beneficios*:* Por tratarse de un trabajo académico se utilizarán aplicaciones en versión trial, un posterior reconocimiento de los resultados podría derivar en la evaluación efectiva del costo total.

4.1.3. Determinar los objetivos de la minería de datos

- *Meta:* Ayudar al encargado a tomar la decisión más adecuada en su tarea de distribuir expedientes con análisis según la complejidad de la causa, el tipo de carátula y la mejor opción de operador, para su resolución en el menor tiempo posible, protegiendo legalmente la información, y realizando un trabajo con transparencia y confiabilidad.
- *Objetivos:* predecir, según la complejidad y tipo de causa, cuál es la mejor opción, en lo que respecta al operador. Y determinar, según la Unidad Judicial, que tipos de oficios son los que más ingresan.
- *Criterios de éxito:* el proyecto tendrá éxito si se revelan relaciones desconocidas hasta el momento en el comportamiento de los operadores que resulten útiles para el encargado en la preparación estrategias de distribución de oficios y resolución de los mismos.

4.1.4. Realizar el plan del proyecto

	Modo de	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
1		Compresión Negocio	15 días	lun 02/11/15	vie 20/11/15	
2		Compresión Datos	25 días	lun 23/11/15	vie 25/12/15	1
3		Preparación Datos	20 días	lun 28/12/15	vie 22/01/16	2
4		Modelado	12 días	lun 25/01/16	mar 09/02/1	3
5		Evaluación	15 días	mié 10/02/1	mar 01/03/1	4
6		Implantación	10 días	mié 02/03/1	mar 15/03/1	5

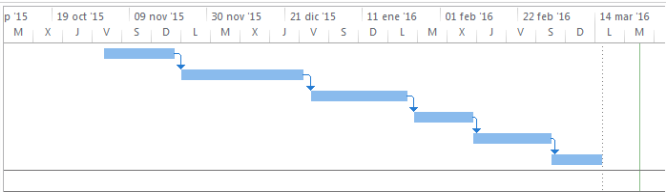


Fig.4-1: Plan del proyecto

4.2. FASE DE COMPRENSIÓN DE LOS DATOS

Los datos con los que se cuenta están registrados en planillas Excel. La recolección de los mismos se llevará a cabo a través de la carga on-line de los oficios y del análisis de los conocimientos de los operadores.

4.2.1. Recolección de datos iníciales

Se analizaron las planillas Excel correspondientes a los expedientes de los años 2014 y 2015, junto a la planilla de expedientes para análisis. Se considera que el análisis de dos años de trabajo, aporta comprensión acerca de la experiencia, solidez y factores que influyen en el rendimiento a posterior de los operadores.

La experiencia adquirida los trabajos constantes colabora con el perfeccionamiento permanente.

Los expedientes iniciados durante el transcurso del año 2014, de distintas dependencias judiciales fueron alrededor de 4200, repartidos entre 9 operadores, en cuyo caso solo en 21 expedientes, fueron solicitados entrecruzamiento de comunicaciones y análisis, distribuidos en solo 3 operadores.

En el año 2015, los expedientes iniciados hasta el mes de mayo, han sido 2000, distribuidos entre 8 operadores, de los cuales hasta el momento, solo en 15 se han solicitado análisis de las comunicaciones, repartidos en 3 operadores.

4.2.2. Descripción de los datos

Para poder describir los datos de manera comprensible se copia en la tabla 4.2 una muestra de la planilla de expedientes en formato Excel (los datos correspondientes a cada expediente son ficticios por tratarse de datos confidenciales).

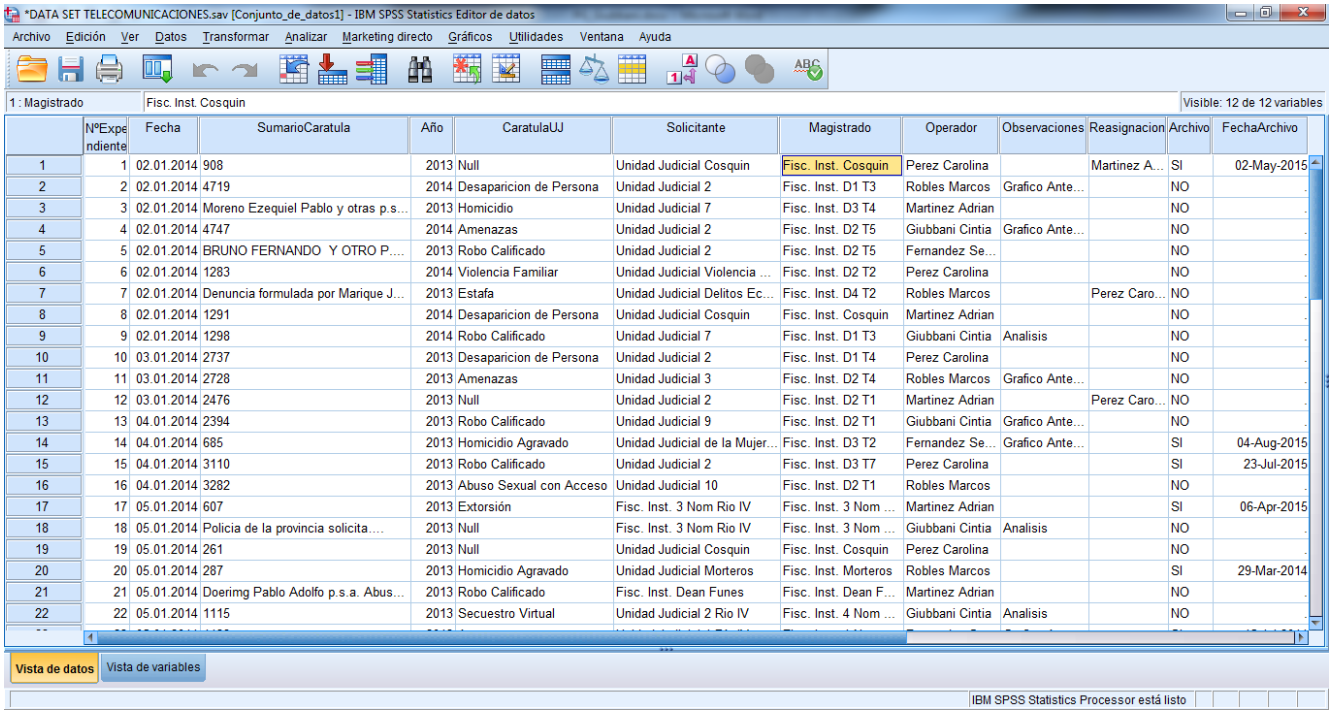
Nº Expendiente	Fecha	Sumario / Caratula	Año	Caratula UJ	Solicitante	Magistrado	Operador	Observaciones
1	02/01/2014 10:23	908	2013	Robo	Unidad Judicial Cosquin	Fisc. Inst. Cosquin	Perez Carolina	Analisis
2	02/01/2014 13:45	4719	2014	Desaparicion de Persona	Unidad Judicial 15	Fisc. Inst. D1 T3	Robles Marcos	
3	02/01/2014 13:22	Moreno Ezequiel Pablo y otras p.ss.aa Homicidio en ocasión de robo	2013	Homicio	Unidad Judicial 7	Fisc. Inst. D3 T4	Martinez Adrian	
4	02/01/2014 14:25	4747	2014	Amenaza	Unidad Judicial 2	Fisc. Inst. D2 T5	Giubbani Cintia	
5	02/01/2014 14:30	BRUNO FERNANDO Y OTRO P.SS.AA. ROBO, ETC.	2013	Robo	Unidad Judicial Huinca Renanco	Fisc. Inst. Hunica Renanco	Fernandez Sebastian	
6	02/01/2014 14:40	1283	2014	Violencia Familiar	Unidad Judicial Violencia Familiar	Fisc. Inst. D2 T2	Perez Carolina	Analisis
7	02/01/2014 15:00	Denuncia formulada por Marique Jorge Matias contra Agüero Diego	2013	Estafa	Unidad Judicial Delitos Economicos	Fisc. Inst. D4 T2	Robles Marcos	
8	02/01/2014 15:30	1291	2014	Desaparicion de Persona	Unidad Judicial Villa Maria	Fisc. Inst. Villa Maria	Martinez Adrian	
9	02/01/2014 16:31	1298	2014	Robo	Unidad Judicial Robo y Hurtos	Fisc. Inst. D1 T3	Giubbani Cintia	Analisis

Tabla 4.2: Muestra de la recolección de datos Planilla Expedientes 2014

La definición del tipo de variable, rangos permitidos, etc. se explica usando terminología del programa SPSS que es el que se utilizará para la preparación y exploración del dataset.

En una primera instancia se muestran los datos tal como están almacenados en su origen, posteriormente se analizarán las modificaciones que se consideren pertinentes según el objetivo del trabajo y las necesidades respectivas.

Ingresando el dataset a SPSS se obtuvo el resultado reflejado en la figura 4-2.



	NºExpediente	Fecha	SumarioCaratula	Año	CaratulaUJ	Solicitante	Magistrado	Operador	Observaciones	Reasignacion	Archivo	FechaArchivo
1	1	02.01.2014	908	2013	Null	Unidad Judicial Cosquin	Fisc. Inst. Cosquin	Perez Carolina		Martinez A...	SI	02-May-2015
2	2	02.01.2014	4719	2014	Desaparicion de Persona	Unidad Judicial 2	Fisc. Inst. D1 T3	Robles Marcos	Grafico Ante...		NO	
3	3	02.01.2014	Moreno Ezequiel Pablo y otras p.s...	2013	Homicidio	Unidad Judicial 7	Fisc. Inst. D3 T4	Martinez Adrian			NO	
4	4	02.01.2014	4747	2014	Amenazas	Unidad Judicial 2	Fisc. Inst. D2 T5	Giubbani Cintia	Grafico Ante...		NO	
5	5	02.01.2014	BRUNO FERNANDO Y OTRO P....	2013	Robo Calificado	Unidad Judicial 2	Fisc. Inst. D2 T5	Fernandez Se...			NO	
6	6	02.01.2014	1283	2014	Violencia Familiar	Unidad Judicial Violencia ...	Fisc. Inst. D2 T2	Perez Carolina			NO	
7	7	02.01.2014	Denuncia formulada por Marique J...	2013	Estafa	Unidad Judicial Delitos Ec...	Fisc. Inst. D4 T2	Robles Marcos		Perez Caro...	NO	
8	8	02.01.2014	1291	2014	Desaparicion de Persona	Unidad Judicial Cosquin	Fisc. Inst. Cosquin	Martinez Adrian			NO	
9	9	02.01.2014	1298	2014	Robo Calificado	Unidad Judicial 7	Fisc. Inst. D1 T3	Giubbani Cintia	Analisis		NO	
10	10	03.01.2014	2737	2013	Desaparicion de Persona	Unidad Judicial 2	Fisc. Inst. D1 T4	Perez Carolina			NO	
11	11	03.01.2014	2728	2013	Amenazas	Unidad Judicial 3	Fisc. Inst. D2 T4	Robles Marcos	Grafico Ante...		NO	
12	12	03.01.2014	2476	2013	Null	Unidad Judicial 2	Fisc. Inst. D2 T1	Martinez Adrian		Perez Caro...	NO	
13	13	04.01.2014	2394	2013	Robo Calificado	Unidad Judicial 9	Fisc. Inst. D2 T1	Giubbani Cintia	Grafico Ante...		NO	
14	14	04.01.2014	685	2013	Homicidio Agravado	Unidad Judicial de la Mujer...	Fisc. Inst. D3 T2	Fernandez Se...	Grafico Ante...		SI	04-Aug-2015
15	15	04.01.2014	3110	2013	Robo Calificado	Unidad Judicial 2	Fisc. Inst. D3 T7	Perez Carolina			SI	23-Jul-2015
16	16	04.01.2014	3282	2013	Abuso Sexual con Acceso	Unidad Judicial 10	Fisc. Inst. D2 T1	Robles Marcos			NO	
17	17	05.01.2014	607	2013	Extorsión	Fisc. Inst. 3 Nom Rio IV	Fisc. Inst. 3 Nom ...	Martinez Adrian			SI	06-Apr-2015
18	18	05.01.2014	Policia de la provincia solicita....	2013	Null	Fisc. Inst. 3 Nom Rio IV	Fisc. Inst. 3 Nom ...	Giubbani Cintia	Analisis		NO	
19	19	05.01.2014	261	2013	Null	Unidad Judicial Cosquin	Fisc. Inst. Cosquin	Perez Carolina			NO	
20	20	05.01.2014	287	2013	Homicidio Agravado	Unidad Judicial Morteros	Fisc. Inst. Morteros	Robles Marcos			SI	29-Mar-2014
21	21	05.01.2014	Doerimg Pablo Adolfo p.s.a. Abus...	2013	Robo Calificado	Fisc. Inst. Dean Funes	Fisc. Inst. Dean F...	Martinez Adrian			NO	
22	22	05.01.2014	1115	2013	Secuestro Virtual	Unidad Judicial 2 Rio IV	Fisc. Inst. 4 Nom ...	Giubbani Cintia	Analisis		NO	

Fig. 4-2: Dataset

La vista de las variables se recoge en la tabla 4.3.

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	NºExpendie...	Numérico	12	0	Nº Expediente	Ninguna	Ninguna	4	Derecha	Nominal	Entrada
2	Fecha	Fecha	20	0	Fecha	Ninguna	Ninguna	13	Derecha	Escala	Entrada
3	SumarioCar...	Cadena	67	0	Sumario / Carat...	Ninguna	Ninguna	29	Izquierda	Nominal	Entrada
4	Año	Numérico	12	0	Año	Ninguna	Ninguna	4	Derecha	Escala	Entrada
5	CaratulaUJ	Cadena	23	0	Caratula UJ	Ninguna	Ninguna	16	Izquierda	Nominal	Entrada
6	Solicitante	Cadena	34	0	Solicitante	Ninguna	Ninguna	24	Izquierda	Nominal	Entrada
7	Magistrado	Cadena	26	0	Magistrado	Ninguna	Ninguna	18	Izquierda	Nominal	Entrada
8	Operador	Cadena	19	0	Operador	Ninguna	Ninguna	14	Izquierda	Nominal	Entrada
9	Observacion...	Cadena	8	0	Observaciones	Ninguna	Ninguna	9	Izquierda	Nominal	Entrada

Tabla 4.3: Vista de metadatos

En esta etapa de estudio, los datos pueden ser llamados datos brutos.

Con respecto a la escala de medidas de variables, se pueden clasificar, las mismas, como cuantitativas discretas (NºExpedientes y Año,) y cualitativas nominales o categóricas (SumarioCaratula, Fecha, CaratulaUJ, Solicitante, Magistrado, Operador, Observaciones, Archivo, FechaArchivo).

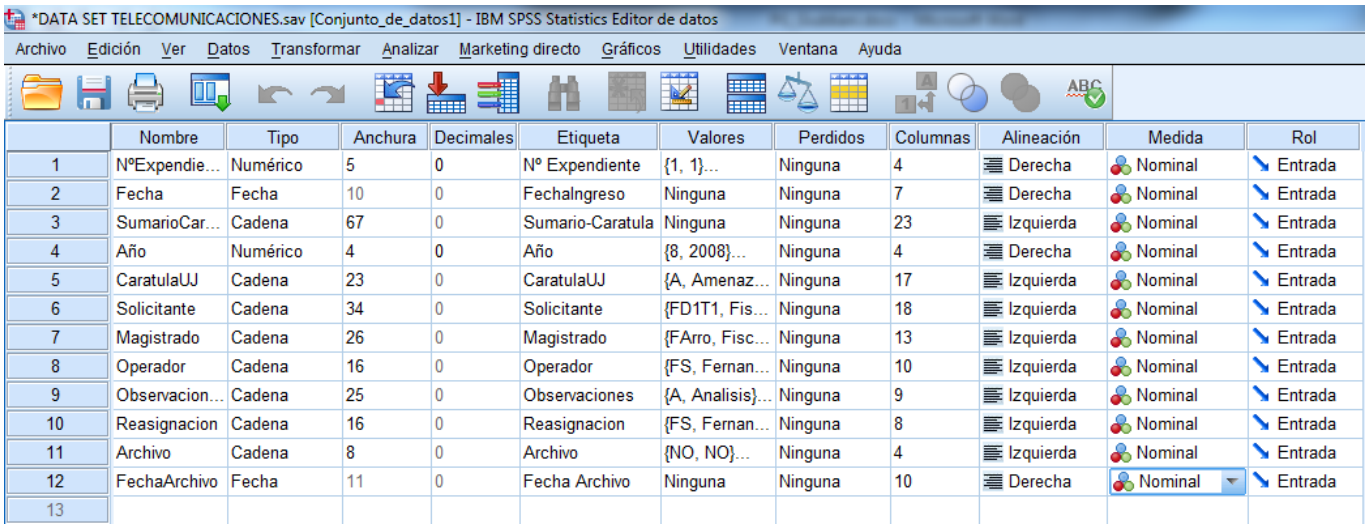
En lo que dimensionalidad se trata, podemos decir que estamos frente a datos multivariados, ya que las propiedades del dataset se miden en un conjunto específico de objetos.

Se consolidó todo en un único dataset. Se utilizó las funciones de SPSS, como Unir e Insertar Casos, que permitió obtener un dataset con 1044 registros.

La tabla resultante, posee una estructura de una fila por caso, es decir, que por cada variable oficio existe una fila con sus atributos que lo describen.

4.2.3. Reporte de exploración de los datos

Para llevar adelante la tarea de exploración de datos, primero se verificaron los tipos de datos y se definieron los rangos o valores permitidos. Esta es una tarea fundamental para poder aplicar las técnicas estadísticas adecuadas en cada caso. La vista de variables quedó redefinida como muestra la figura 4-3.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	NºExpendie...	Numérico	5	0	Nº Expendiente	{1, 1}...	Ninguna	4	Derecha	Nominal	Entrada
2	Fecha	Fecha	10	0	FechaIngreso	Ninguna	Ninguna	7	Derecha	Nominal	Entrada
3	SumarioCar...	Cadena	67	0	Sumario-Caratula	Ninguna	Ninguna	23	Izquierda	Nominal	Entrada
4	Año	Numérico	4	0	Año	{8, 2008}...	Ninguna	4	Derecha	Nominal	Entrada
5	CaratulaUJ	Cadena	23	0	CaratulaUJ	{A, Amenaz...	Ninguna	17	Izquierda	Nominal	Entrada
6	Solicitante	Cadena	34	0	Solicitante	{FD1T1, Fis...	Ninguna	18	Izquierda	Nominal	Entrada
7	Magistrado	Cadena	26	0	Magistrado	{FArro, Fisc...	Ninguna	13	Izquierda	Nominal	Entrada
8	Operador	Cadena	16	0	Operador	{FS, Fernan...	Ninguna	10	Izquierda	Nominal	Entrada
9	Observacion...	Cadena	25	0	Observaciones	{A, Analisis}...	Ninguna	9	Izquierda	Nominal	Entrada
10	Reasignacion	Cadena	16	0	Reasignacion	{FS, Fernan...	Ninguna	8	Izquierda	Nominal	Entrada
11	Archivo	Cadena	8	0	Archivo	{NO, NO}...	Ninguna	4	Izquierda	Nominal	Entrada
12	FechaArchivo	Fecha	11	0	Fecha Archivo	Ninguna	Ninguna	10	Derecha	Nominal	Entrada
13											

Fig. 4-3: Redefinición dataset

Todas las variables son ahora de tipo nominal, debido a que los atributos a los que hacen referencia son todos del tipo cualitativo. Explícitamente los valores posibles se expresan en la tabla 4.4.

Variable	Descripción	Valores Posible
NºExpendiente	Indica el N° de registro interno del oficio que ingresa	[1,6000]
Fecha	Fecha de ingreso del expediente	Ninguna
SumarioCaratula	N° o caratula que fue dado en la Unidad judicial	Ninguna
Año	Indica el año en que se inicia la causa	[2008, 2018]
CaratulaUJ	Indica el tipo de causa	DP (Desaparicion de Persona), SE (secuestro Extorsivo), SV (secuestro Virtual), H (Homicidio)
Solicitante	Dependencia donde se inicia el pedido	{FD1T1, Fiscalia D1 T1}...
Magistrado	Indicia el magistrado que controla la causa	{Farro(Fiscalia Arroyito), UJ1 (Unidad Judicial 1), UJ2 (Unidad Judicial 2)…
Operador	Persona encargada de llevar adelante el requerimiento	{FS (Fernandez Sebastian), GC (Giubbani Cintia), PC (Perez Carolina)
Observaciones	Pedidos realizados posteriormente al requerimiento inicial	A (Análisis), GA (Gráfico Antenas)
Reasignacion	Indica cuando un expediente es reasignado a otro operador, ya sea por unión con algún expediente existente o por experiencia del nuevo operario.	{FS (Fernandez Sebastian), GC (Giubbani Cintia), PC (Perez Carolina)}…
Archivo	Si el expediente se encuentra en estado archivado, esto se da cuando la causa se finaliza.	{SI, NO, PERDIDO}
FechaArchivo	Fecha en la que la causa ingresó al archivo	Ninguna

Tabla 4.4: Tabla de Datos

Como primera medida para la exploración de datos, se utilizó la exploración visual. Debido al tipo de variables cualitativas la seleccionada fue diagrama de barras.

Teniendo presente que el objetivo fundamental será predecir el tipo de causa se inicia la exploración por esa variable primordial.

Análisis de la variable CaratulaUJ

Estadísticos

CaratulaUJ

N	Válidos	1139
	Perdidos	0

Tabla de Frecuencia

CaratulaUJ

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Abuso Sexual con Acceso	9	,8	,8	,8
	Amenazas	214	18,8	18,8	19,6
	Amenazas Calificadas	23	2,0	2,0	21,6
	Denuncia Formulada	1	,1	,1	21,7
	Desaparcion de Personas	205	18,0	18,0	39,7
	Desobediencia a la Auto	4	,4	,4	40,0
	Encubrimiento	8	,7	,7	40,7
	Estafa	33	2,9	2,9	43,6
	Extorsión	33	2,9	2,9	46,5
	Homicidio	87	7,6	7,6	54,2
	Homicidio Agravado	37	3,2	3,2	57,4
	MED	18	1,6	1,6	59,0
	Null	290	25,5	25,5	84,5
	Robo calificado	4	,4	,4	84,8
	Robo Calificado	92	8,1	8,1	92,9
	Secuestro Virtual	57	5,0	5,0	97,9
	Violencia Familiar	24	2,1	2,1	100,0
	Total	1139	100,0	100,0	

Tabla 4.5: Distribución de frecuencias CaratulaUJ

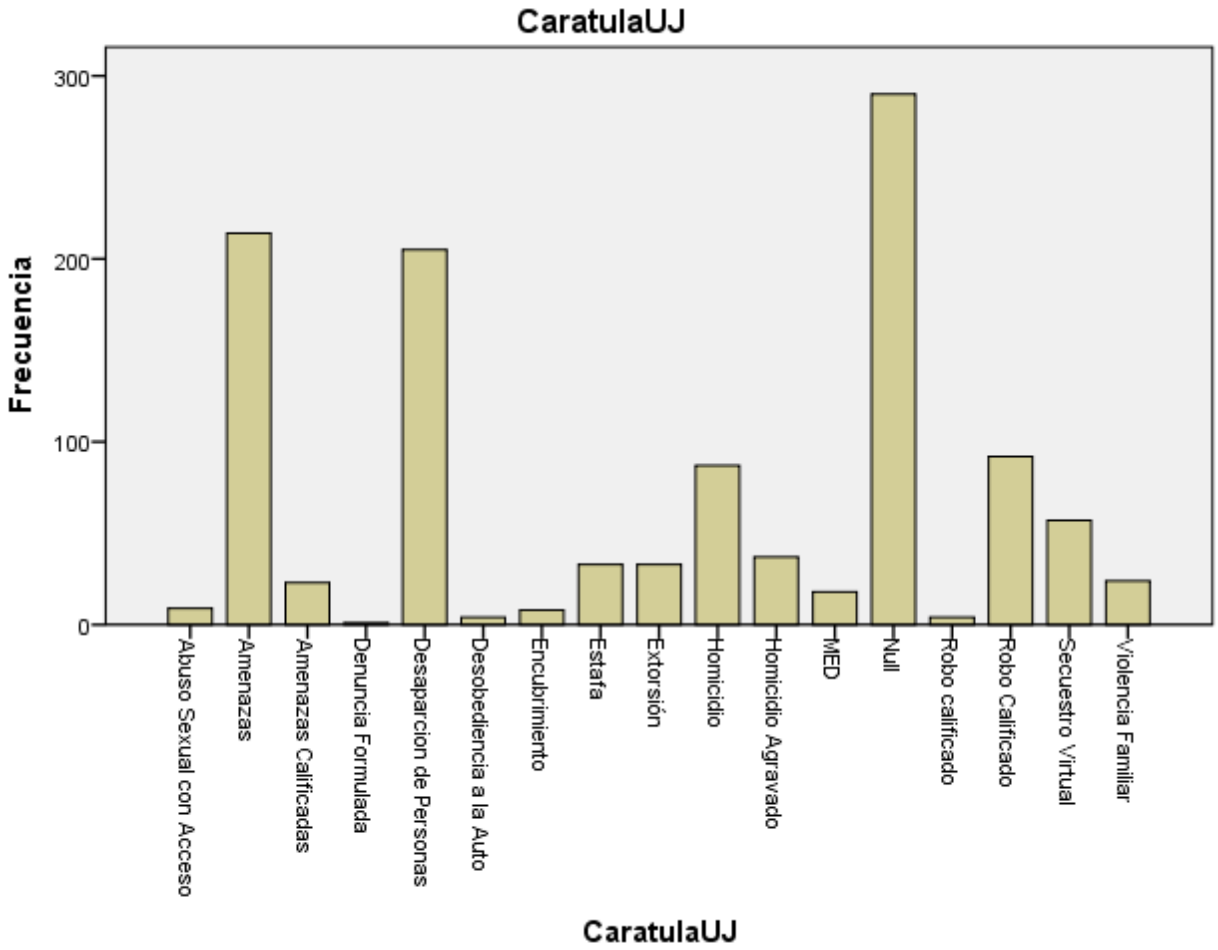


Fig. 4-4: Distribución de frecuencias CaratulaUJ

Se puede observar que existe una categoría de valores nulos, por lo que será preciso averiguar el porqué de esa información faltante, si se trata de datos perdidos, de campos que no deben considerarse, etc.

Se vislumbra que existe posiblemente un error en la registración del dato Robo Calificado dado que la tabla de frecuencias identifica tres categorías diferentes "Robo Calificado", "Robo calificado" y "Robo calificado por el…". Probablemente todas estén indicando el mismo tipo de dato "Robo Calificado". Lo mismo sucede con el dato "Amenaza" y "Amenazas", quizás haya habido una falta de criterio homogéneo en la manera de ingresar el dato.

Se nota también como particularidad que no se relevaron casos perdidos, es decir todos los registros han sido clasificados.

Se continúa el análisis por la variable Operador, uno de los indicadores claves en el estudio debido a la importancia que tiene en la resolución de oficios a corto tiempo.

Análisis de la variable Operador

Estadísticos

Operador

N	Válidos	1139
	Perdidos	0

Tabla de Frecuencia

Operador

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Fernandez Sebast	222	19,5	19,5	19,5
	Giubbani Cintia	229	20,1	20,1	39,6
	Martinez Adrian	229	20,1	20,1	59,7
	Perez Carolina	230	20,2	20,2	79,9
	Robles Marcos	229	20,1	20,1	100,0
	Total	1139	100,0	100,0	

Tabla 4.6: Distribución de frecuencias Operador

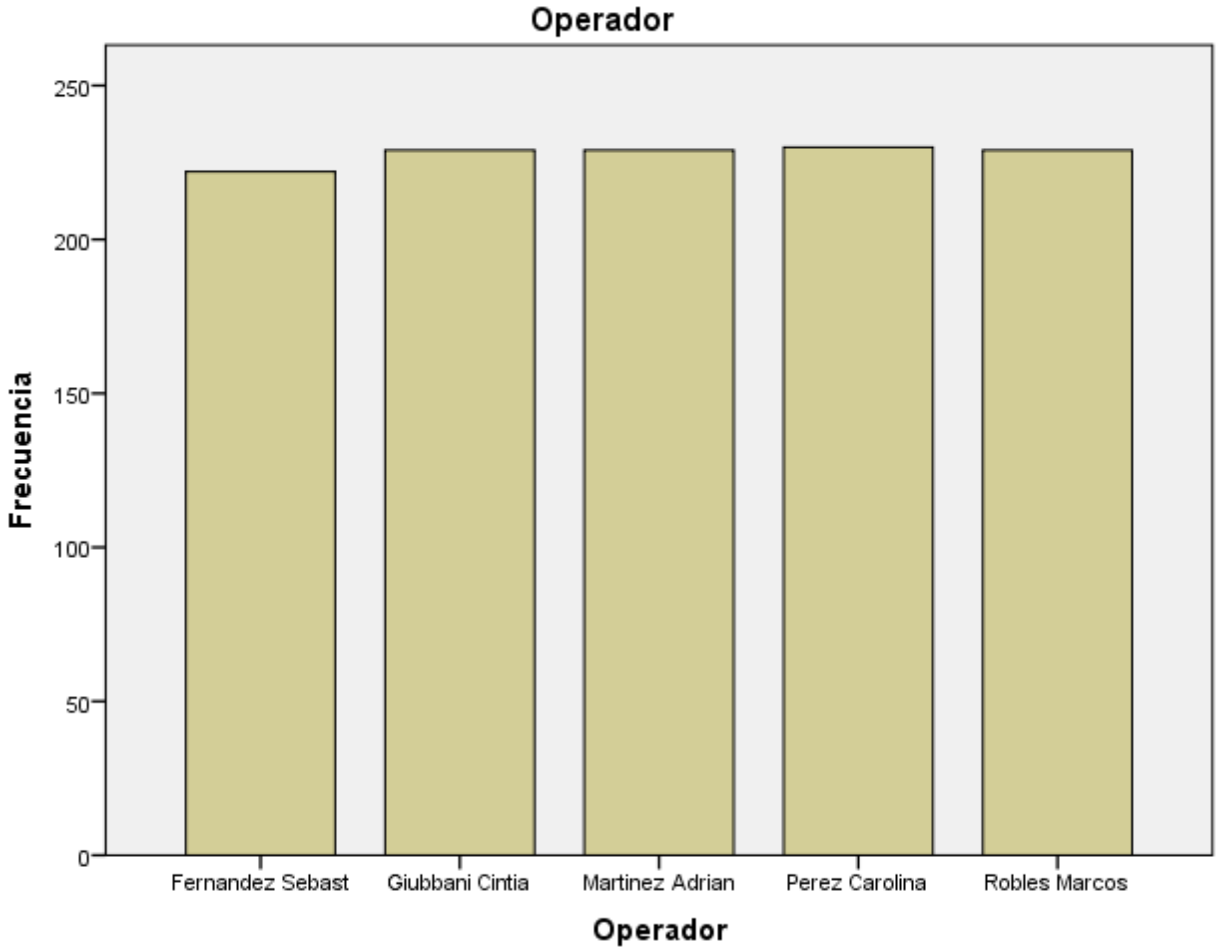


Fig. 4-5: Distribución de frecuencias Operador

Se observa que no hay registros perdidos que no hayan podido ser clasificados.

A continuación se visualizan las distribuciones de frecuencias de las variables Solicitante y Observaciones.

Análisis de la variable Solicitante

Estadísticos

Solicitante

N	Válidos	1139
	Perdidos	0

Tabla de Frecuencia

Solicitante

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		2	,2	,2	,2
	Camara 1 Río IV	1	,1	,1	,3
	Camara 10 Crimen	1	,1	,1	,4
	Camara 4 Crimen	2	,2	,2	,5
	Camara 6 Crimen	1	,1	,1	,6
	Camara 9 Crimen	1	,1	,1	,7
	Camara Criminal y Correccional Vil	2	,2	,2	,9
	CIC	2	,2	,2	1,1
	Comisaria Alejandro Roca	1	,1	,1	1,1
	Comisaria Arias	2	,2	,2	1,3
	Comisaria Bialet Masse	1	,1	,1	1,4
	Comisaría Cabo 1º Cogote	1	,1	,1	1,5
	Comisaría Cabo 1º Cogote Juarez Ce	1	,1	,1	1,6
	Comisaria Capilla del Monte	4	,4	,4	1,9
	Comisaria Colonia Caroya	4	,4	,4	2,3
	Comisaria Colonia Tirolesa	4	,4	,4	2,6
	Comisaria Distrito Rio I	1	,1	,1	2,7
	Comisaria Distrito Unquillo UR 4	1	,1	,1	2,8
	Comisaria Dto.I Villa del Rosario	1	,1	,1	2,9
	Comisaria La Cumbre	2	,2	,2	3,1
	Comisaria malagueño	2	,2	,2	3,2
	Comisaria Malvinas Argentinas	1	,1	,1	3,3
	Comisaría Monte Maiz	4	,4	,4	3,7
	Comisaria Montecristo	1	,1	,1	3,8
	Comisaria Oncativo	1	,1	,1	3,9
	Comisaria Roque Saenz Peña	1	,1	,1	4,0
	Comisaria Saldan	2	,2	,2	4,1

	Comisaria San Antonio Arredondo	1	,1	,1	4,2
	Comisaria San Javier	1	,1	,1	4,3
	Comisaria Santa Maria	6	,5	,5	4,8
	Comisaria Santa Rosa	4	,4	,4	5,2
	Comisaria Santa Rosa Area Mujer	1	,1	,1	5,3
	Comisaria Valle Hermoso	1	,1	,1	5,4
	Comisaria Villa Giardino	2	,2	,2	5,5
	Comisarìa Villa Giardino	1	,1	,1	5,6
	Comsiaria La Cumbre	1	,1	,1	5,7
	Comsiaria Tanti	1	,1	,1	5,8
	Destacamento Falda del Carmen	1	,1	,1	5,9
	Destacamento Villa Rossi	1	,1	,1	6,0
	Division Judicial Salsacate	2	,2	,2	6,1
	Fisc. Inst. 1 Nom Rio IV	6	,5	,5	6,7
	Fisc. Inst. 1 T. Carlos Paz	2	,2	,2	6,8
	Fisc. Inst. 1 T. Rio III	1	,1	,1	6,9
	Fisc. Inst. 1 T. Villa Dolores	3	,3	,3	7,2
	Fisc. Inst. 2 Nom Rio IV	7	,6	,6	7,8
	Fisc. Inst. 2 T. Rio III	1	,1	,1	7,9
	Fisc. Inst. 2 T. Villa Dolores	3	,3	,3	8,2
	Fisc. Inst. 3 Nom Rio IV	4	,4	,4	8,5
	Fisc. Inst. 3 T. Villa María	4	,4	,4	8,9
	Fisc. Inst. 4 Nom Rio IV	1	,1	,1	9,0
	Fisc. Inst. Alta Gracia	3	,3	,3	9,2
	Fisc. Inst. Bell Ville	2	,2	,2	9,4
	Fisc. Inst. Cosquín	2	,2	,2	9,6
	Fisc. Inst. Cruz del Eje	2	,2	,2	9,7
	Fisc. Inst. Cura Brochero	6	,5	,5	10,3
	Fisc. Inst. D1 T1	1	,1	,1	10,4
	Fisc. Inst. D1 T2	1	,1	,1	10,4
	Fisc. Inst. D1 T3	3	,3	,3	10,7
	Fisc. Inst. D1 T4	4	,4	,4	11,1
	Fisc. Inst. D1 T5	1	,1	,1	11,2

	Fisc. Inst. D1 T6	1	,1	,1	11,2
	Fisc. Inst. D2 T1	2	,2	,2	11,4
	Fisc. Inst. D2 T2	9	,8	,8	12,2
	Fisc. Inst. D2 T3	4	,4	,4	12,6
	Fisc. Inst. D2 T5	2	,2	,2	12,7
	Fisc. Inst. D2 T6	1	,1	,1	12,8
	Fisc. Inst. D3 T3	5	,4	,4	13,3
	Fisc. Inst. D3 T4	5	,4	,4	13,7
	Fisc. Inst. D3 T5	5	,4	,4	14,1
	Fisc. Inst. D3 T6	4	,4	,4	14,5
	Fisc. Inst. D3 T7	1	,1	,1	14,6
	Fisc. Inst. D4 T1	9	,8	,8	15,4
	Fisc. Inst. D4 T2	6	,5	,5	15,9
	Fisc. Inst. D4 T3	1	,1	,1	16,0
	Fisc. Inst. D4 T4	4	,4	,4	16,3
	Fisc. Inst. Dean Funes	6	,5	,5	16,9
	Fisc. Inst. Huinca Renanco	1	,1	,1	16,9
	Fisc. Inst. Jesus María	3	,3	,3	17,2
	Fisc. Inst. La Carlota	6	,5	,5	17,7
	Fisc. Inst. Laboulaye	2	,2	,2	17,9
	Fisc. Inst. Men. 1 Turno	1	,1	,1	18,0
	Fisc. Inst. Morteros	4	,4	,4	18,3
	Fisc. Inst. Rio II	4	,4	,4	18,7
	Fisc. Inst. San Francisco	2	,2	,2	18,9
	informatica	3	,3	,3	19,1
	Informatica	15	1,3	1,3	20,5
	Juzg. Inst. Concaran	1	,1	,1	20,5
	Juzg. Men. Y Faltas Cosquin	1	,1	,1	20,6
	Juzg. Penal Juvenil 1º nom. Sec. N	1	,1	,1	20,7
	Juzg.control men.y faltas La Carlo	1	,1	,1	20,8
	Juzg.de Inst. y del Menor Nº 1 Pci	1	,1	,1	20,9
	Juzgado de Niñez Carlos Paz	1	,1	,1	21,0

	Subcomisaria Anisacate	1	,1	,1	21,1
	Subcomisaria Calchin	1	,1	,1	21,2
	Unidad Judicial 1	17	1,5	1,5	22,7
	Unidad Judicial 1 Río IV	2	,2	,2	22,8
	Unidad Judicial 10	14	1,2	1,2	24,1
	Unidad Judicial 11	27	2,4	2,4	26,4
	Unidad Judicial 12	15	1,3	1,3	27,7
	Unidad Judicial 13	11	1,0	1,0	28,7
	Unidad Judicial 14	34	3,0	3,0	31,7
	Unidad Judicial 15	33	2,9	2,9	34,6
	Unidad Judicial 16	11	1,0	1,0	35,6
	Unidad Judicial 17	7	,6	,6	36,2
	Unidad Judicial 18	33	2,9	2,9	39,1
	Unidad Judicial 19	10	,9	,9	39,9
	Unidad Judicial 2	47	4,1	4,1	44,1
	Unidad Judicial 2 Rio IV	28	2,5	2,5	46,5
	Unidad Judicial 20	8	,7	,7	47,2
	Unidad Judicial 21	20	1,8	1,8	49,0
	Unidad Judicial 22	14	1,2	1,2	50,2
	Unidad Judicial 3	13	1,1	1,1	51,4
	Unidad Judicial 4	16	1,4	1,4	52,8
	Unidad Judicial 5	25	2,2	2,2	55,0
	Unidad Judicial 6	26	2,3	2,3	57,2
	Unidad Judicial 7	33	2,9	2,9	60,1
	Unidad Judicial 8	7	,6	,6	60,8
	Unidad Judicial 9	16	1,4	1,4	62,2
	Unidad Judicial Acc. Vial	14	1,2	1,2	63,4
	Unidad Judicial Alta Gracia	14	1,2	1,2	64,6
	Unidad Judicial Carlos Paz	48	4,2	4,2	68,8
	Unidad Judicial Cosquin	10	,9	,9	69,7
	Unidad Judicial Cosquín	9	,8	,8	70,5
	Unidad Judicial Cruz del Eje	4	,4	,4	70,9
	Unidad Judicial de la Mujer y el n	22	1,9	1,9	72,8
	Unidad Judicial Dean Funes	9	,8	,8	73,6

	Unidad Judicial Delitos Economicos	71	6,2	6,2	79,8
	Unidad Judicial Homicidios	24	2,1	2,1	81,9
	Unidad Judicial Jesus Maria	18	1,6	1,6	83,5
	Unidad Judicial La Calera	7	,6	,6	84,1
	Unidad Judicial La Falda	5	,4	,4	84,5
	Unidad Judicial Laboulaye	22	1,9	1,9	86,5
	Unidad Judicial Morteros	20	1,8	1,8	88,2
	Unidad Judicial Rio II	9	,8	,8	89,0
	Unidad Judicial Rio IV	12	1,1	1,1	90,1
	Unidad Judicial Robos y Hurtos	51	4,5	4,5	94,6
	Unidad Judicial San Francisco	2	,2	,2	94,7
	Unidad Judicial Sustracción de Aut	8	,7	,7	95,4
	Unidad Judicial Villa Allende	19	1,7	1,7	97,1
	Unidad Judicial Villa Dolores	14	1,2	1,2	98,3
	Unidad Judicial Violencia Familiar	13	1,1	1,1	99,5
	Unidad Regional Rio I	4	,4	,4	99,8
	Unidad Regional Rio II	1	,1	,1	99,9
	Unidad Regional Tulumba	1	,1	,1	100,0
	Total	1139	100,0	100,0	

Tabla 4.7: Distribución de frecuencias Solicitante

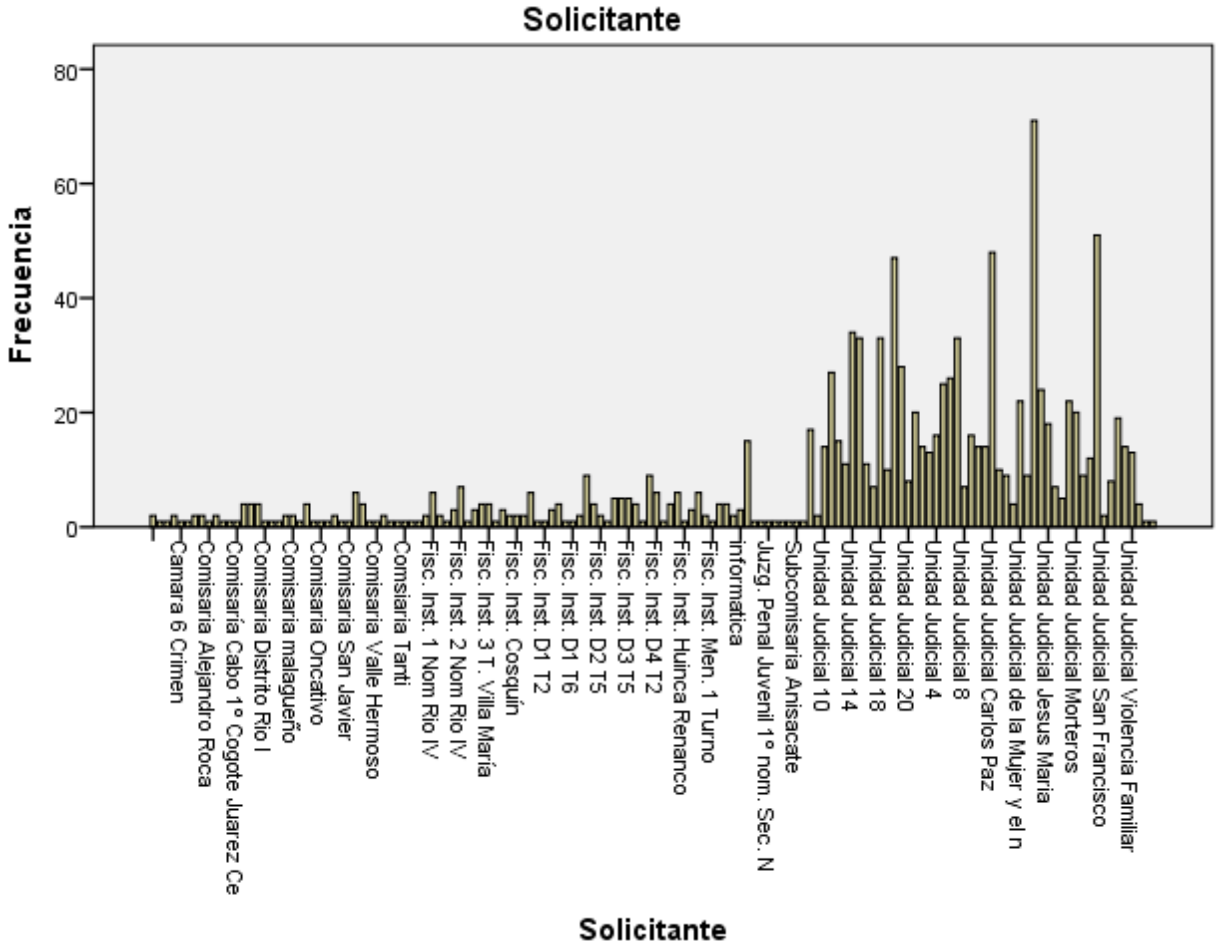


Fig. 4-6: Distribución de frecuencias Solicitante

Se nota nuevamente que no hay registros faltantes, todos han sido clasificados.

Análisis de la variable Observaciones

Estadísticos

Observaciones

N	Válidos	1139
	Perdidos	0

Tabla de Frecuencia

Observaciones

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		1127	98,9	98,9	98,9
	Analisis	5	,4	,4	99,4
	Grafico Antenas	7	,6	,6	100,0
	Total	1139	100,0	100,0	

Tabla 4.8: Distribución de frecuencias Observaciones

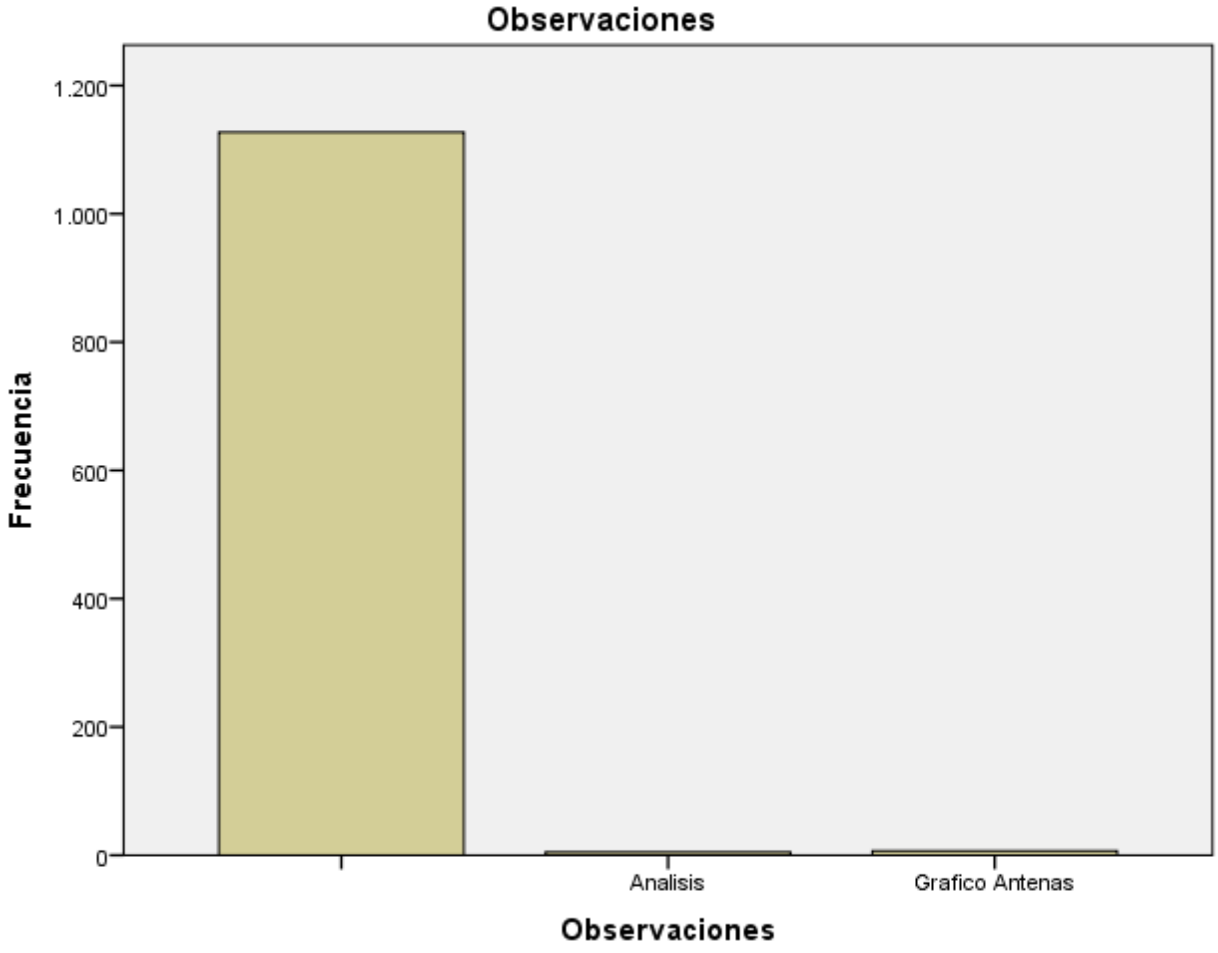


Fig. 4-7: Gráfico Observaciones

Se observa que existen datos cuyos valores son nulos, por lo que se deberá indagar si los mismos son datos faltantes, mal cargados, o cuyo valor está correcto. Vuelve a observarse que no hay pérdida de registros.

Para empezar a comprender la relación entre las variables disponibles, su dependencia o independencia y como una primera observación hacia lo que podría definir luego el tipo de algoritmo a utilizar se realizaron pruebas de Chi-Cuadrado Pearson logrando los siguientes resultados.

Vamos a comenzar por la relación entre las variables CaratulaUJ y Solicitante donde se obtienen las tablas 4.9 y 4.10.

Tabla de contingencia Solicitante * CaratulaUJ

Recuento

		CaratulaUJ												Total
		Abuso Sexual con Acceso	Amenazas	Desaparición de Persona	Estafa	Extorsión	Homicidio	Homicidio Agravado	Null	Robo calificado	Robo Calificado	Secuestro Virtual	Violencia Familiar	
Solicitante	Fisc. Inst. 3 Nom Rio IV	0	0	0	0	1	0	0	1	0	0	0	0	2
	Fisc. Inst. Dean Funes	0	0	0	0	0	0	0	0	0	1	0	0	1
	Unidad Judicial 1 Río IV	0	1	0	0	0	0	0	0	0	0	0	0	1
	Unidad Judicial 10	1	0	0	0	0	0	0	0	0	0	0	0	1
	Unidad Judicial 11	0	0	0	0	1	0	1	1	0	1	1	0	5
	Unidad	0	1	2	0	0	0	0	1	0	2	0	0	6

Judicial 2													
Unidad Judicial 2 Rio IV	0	0	0	0	0	0	0	0	0	1	1	0	2
Unidad Judicial 3	0	1	0	0	0	0	0	0	0	0	0	0	1
Unidad Judicial 4	0	1	0	0	1	0	0	1	0	0	0	0	3
Unidad Judicial 7	0	0	0	0	0	1	0	0	0	1	0	0	2
Unidad Judicial 9	0	0	0	0	0	0	0	0	0	1	0	0	1
Unidad Judicial Cosquin	0	0	1	0	0	0	0	3	0	1	0	0	5
Unidad Judicial de la Mujer y el n	0	0	0	0	0	0	1	0	0	0	0	0	1
Unidad Judicial Delitos Economicos	0	1	0	1	0	0	0	0	1	0	0	0	3
Unidad Judicial Morteros	0	0	0	0	0	0	1	0	0	0	0	0	1
Unidad Judicial Violencia Familiar	0	0	0	0	0	0	0	0	0	0	0	1	1

Total	1	5	3	1	3	1	3	7	1	8	2	1	36

Tabla 4.9: Tabla de frecuencias cruzadas: Solicitante / CaratulaUJ

Pruebas de chi-cuadrado			
	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	2046,832[a]	2288	1,000
Razón de verosimilitudes	1264,390	2288	1,000
N de casos válidos	1139		
a. 2410 casillas (98,4%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es ,00.			

Tabla 4.10: Test de Chi-Cuadrado

Chi cuadrado no debe aplicarse si más del 20% de las frecuencias esperadas es inferior a 5 o alguna inferior a uno, en este caso no se viola la enunciada restricción por lo tanto se considera válido el test.

Realizando las pruebas para las variables Operador y Observaciones se obtienen las tablas 4.11 y 4.12.

Tabla de contingencia Observaciones * Operador							
Recuento							
		Operador					Total
		Fernandez Sebast	Giubbani Cintia	Martinez Adrian	Perez Carolina	Robles Marcos	
Observaciones		219	222	229	230	227	1127
	Analisis	1	4	0	0	0	5
	Grafico Antenas	2	3	0	0	2	7
Total		222	229	229	230	229	1139

Tabla 4.11: Tabla de frecuencias cruzadas: Observaciones / Operador

Pruebas de chi-cuadrado			
	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	17,258[a]	8	,028
Razón de verosimilitudes	18,724	8	,016
N de casos válidos	1139		

a. 10 casillas (66,7%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es ,97.

Tabla 4.12: Test de Chi- Cuadrado

Como resultado de esta prueba, se puede observar que llegamos a una verificación lógica, donde es acertado que haya relación entre las variables, y que las mismas sean dependientes. Aquí también el valor de Sig. Asintótica es menor que 0,05.

Como **conclusión** se verifica que la **variable objetivo está relacionada con las demás**. Esta información sirve a priori para descartar técnicas de minería de datos en las que los supuestos indiquen la necesidad de que las variables que predicen el objetivo sean independientes. Por ejemplo no se podrán aplicar técnicas de regresión lineal.

4.2.4. Verificar la Calidad de los Datos

En cuanto a la consistencia de los datos individuales en el campo CaratulaUJ se detectaron valores nulos.

Teniendo en cuenta el dominio del problema se llega a la conclusión de que cuando el oficio con el requerimiento llega al gabinete, muchas veces, los mismos no poseen caratula del hecho, es decir, la clasificación del mismo no ha sido expuesta, ya sea para mantener en forma confidencial el motivo de la causa, o por omisión del mismo por parte del sumariante en la Unidad Judicial. Por lo que a la hora de cargar dichos datos en el sistema de la oficina, no existe la posibilidad de que el operador encargado, puede determinar la caratula de la causa. En este momento del estudio se decide mantener

estos campos, debido a que al tratarse de causas, las mismas deben ser consideradas importantes a la hora de definir la estrategia de distribución y resolución de los oficios, y se evaluará posteriormente si es conveniente filtrarlas en la construcción del modelo o tratar de obtener información sobre la clasificación de las mismas.

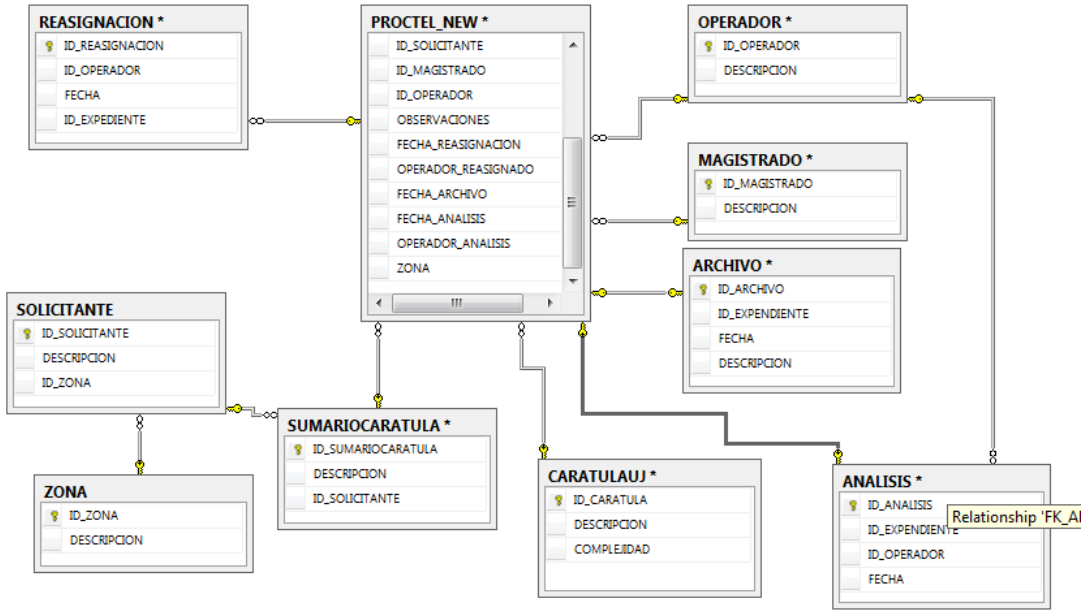
4.3. FASE DE PREPARACIÓN DE LOS DATOS

Esta fase se encuentra estrechamente vinculada a la etapa de modelización, por lo tanto las tareas están relacionadas a las técnicas de minería de datos que se estima serán utilizadas.

4.3.1 Selección de Datos

La selección de datos es una de las tareas más importantes en la fase de preparación de los datos, por lo que se procederá a realizar una selección vertical donde los atributos (campos) seleccionados deberán ser significativos para el conocimiento que se desea obtener y la tarea que se pretende realizar.

A partir de los datos recolectados se construyó la tabla de hechos, cuyas dimensiones se cargarán en el modelo de minería de datos.



4.3.2. Calidad de Datos

Se considera irrelevante el campo Archivo, porque no aporta información en la definición de la estrategia de toma de decisiones con respecto a la distribución de causas. Tampoco se tendrá en cuenta el campo FechaArchivo. Si bien, ambos identifican el estado de un expediente finalizado y pueden considerarse importante para determinar el nivel de resolución de cada empleado, no se tendrán en cuenta por el momento, ya que muchas veces el tiempo de resolución de las causas no depende específicamente de cada operador, sino de los tiempos de respuesta de los requerimientos por parte de las empresas prestadoras de servicios telefónicos. No se vinculan al objetivo propuesto por lo tanto oportunamente se descartarán en el modelo Calidad de datos.

Se mencionaron las causas de los campos con valores nulos con lo cual no existen casos particulares a tener en cuenta. Se considera bueno el dataset disponible.

4.3.3. Estructurar Datos

Examinando los registros disponibles, se puede observar que la cantidad de los mismos es acotada. Esto se debe a que estamos tratando con datos confidenciales. El análisis es específico y orientado hacia un rol, el de los operadores para ayudar a la

toma de decisiones por parte del encargado del área, por lo tanto esto también influye en la cantidad de datos a estudiar.

4.3.4. Integrar datos

La integración de los datos quedará plasmada luego de la ejecución de los procesos de ETL que deberán generarse.

4.3.5. Formateo de los datos

Las transformaciones de los datos se llevaron a cabo creando diferentes paquetes en Integration Services, unificados en un proceso denominado Proceso de Carga. La finalización o salida de este proceso es un data set homogéneo, listo para ser minado.

- Paquete UnirTablas (figura 4-8).
- Tabla Origen: PROCTEL2014.
- Tabla destino: UnionExpedientesExcel.

Aquí se tomó la planilla Excel PROCTEL2014 extraída del actual sistema operacional, se cargaron las hojas correspondientes a la recopilación de datos de cada uno de los registros ingresados durante el año 2014 por el encargado del área. Se generó una nueva base de datos, GABINETE, a través de la conexión con SQL Server 2012 y se generó una tabla unión conteniendo el total de los expedientes ingresados.

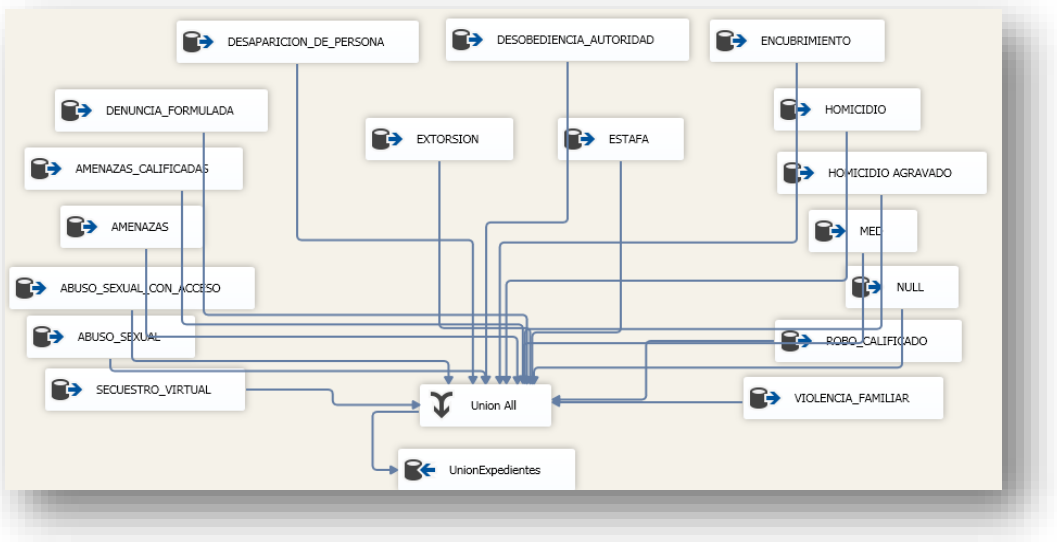
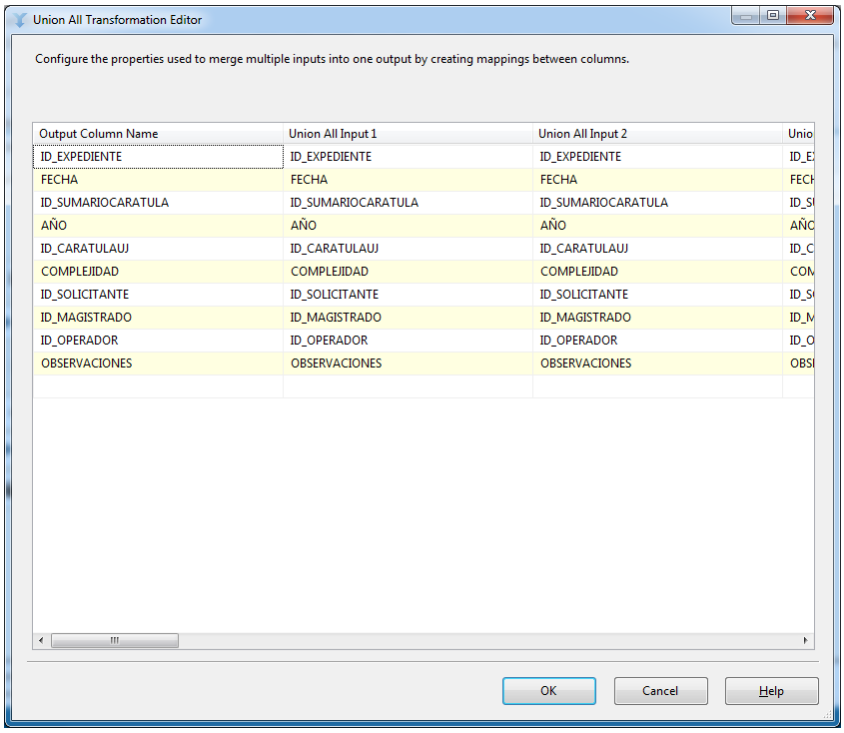
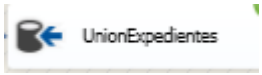
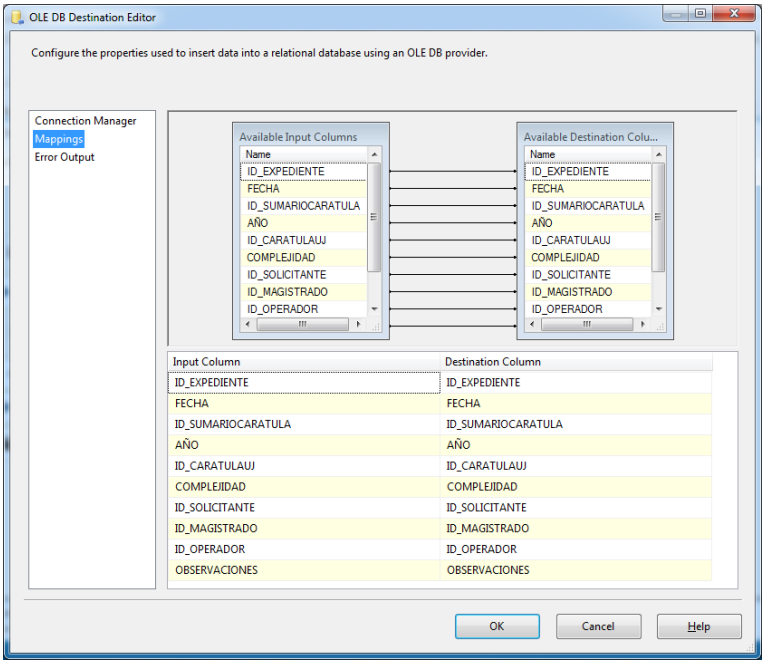


Fig.4-8: Paquete UnirTablas

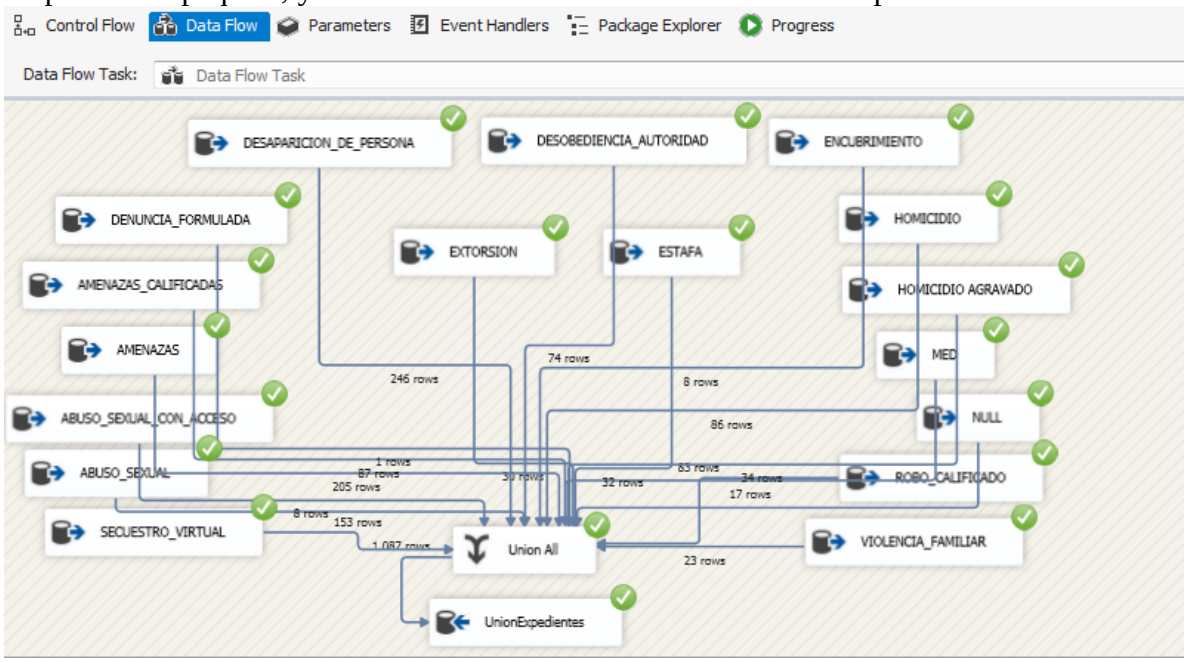
Configuramos el Editor de Transformaciones para la unión de los expedientes en una sola base. 



Luego, configuramos el destino de los datos para poder procesar el paquete y realizar la unión de datos. 



Se procesa el paquete, y se obtiene la nueva tabla de datos UnionExpedientesExcel.



- Paquete OrdenarValores (figura 4-9).
- Tabla origen: UnionExpedientesExcel.
- Tabla destino: ExpedientesOrdenadosSolicitante.

Este paquete es uno de los más importantes en cuanto a que nos permite conocer la cantidad de registros que ingresan por tipo de causa según el solicitante.

La columna CaratulaUJ, considerada como una de las columnas de mayor importancia, fue la que recibió la mayoría de los cambios, ya que se unificaron criterios en cuanto a la categorización de los datos, se redujo la dimensionalidad de la variable, se corrigieron errores detectados.

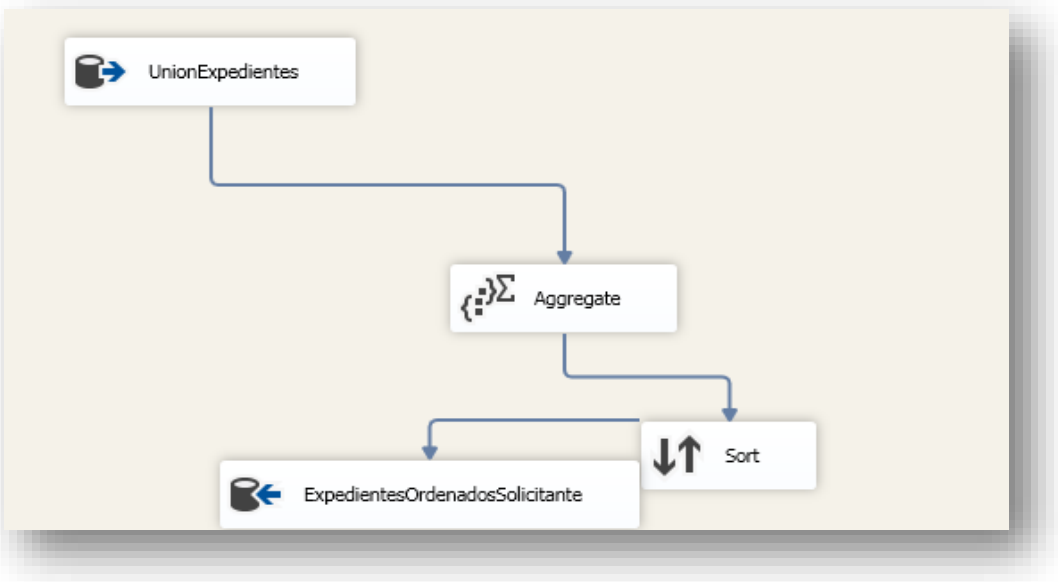
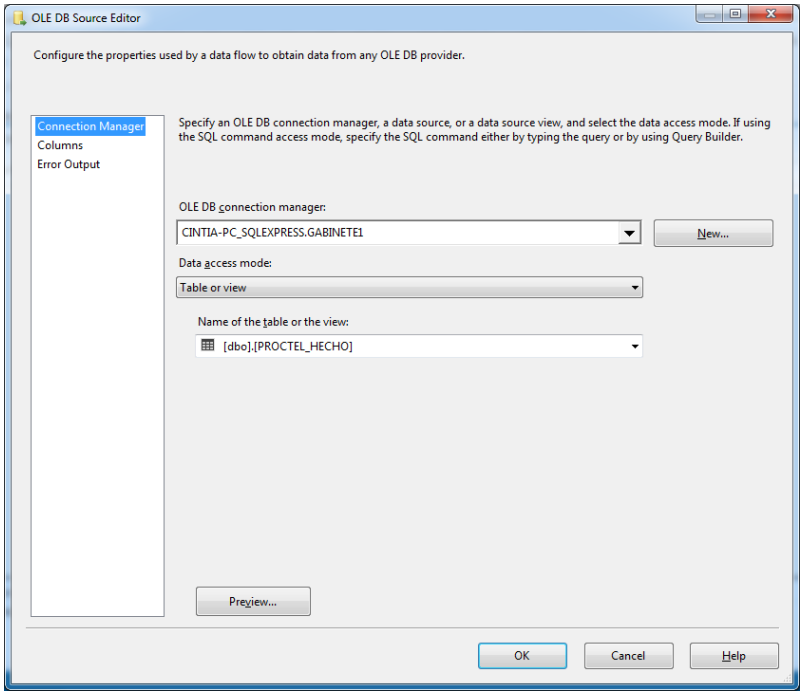
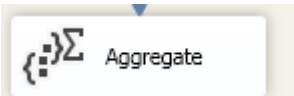


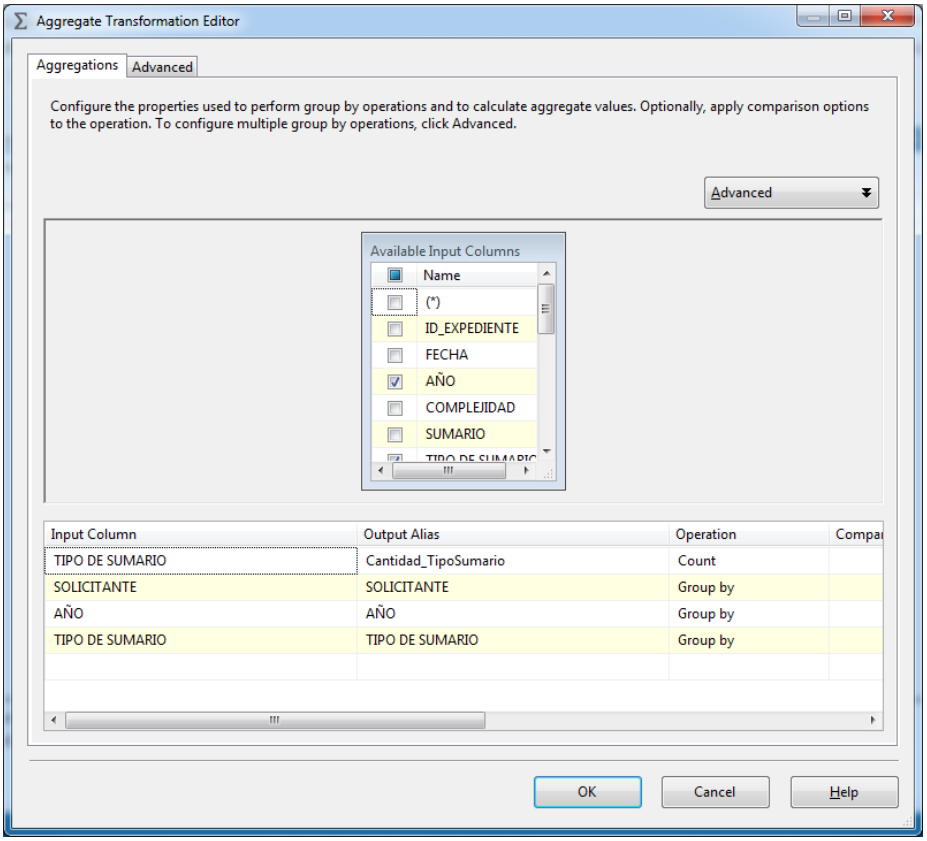
Fig. 4-9: Paquete OrdenarValores

Se configura la fuente de extracción de datos.

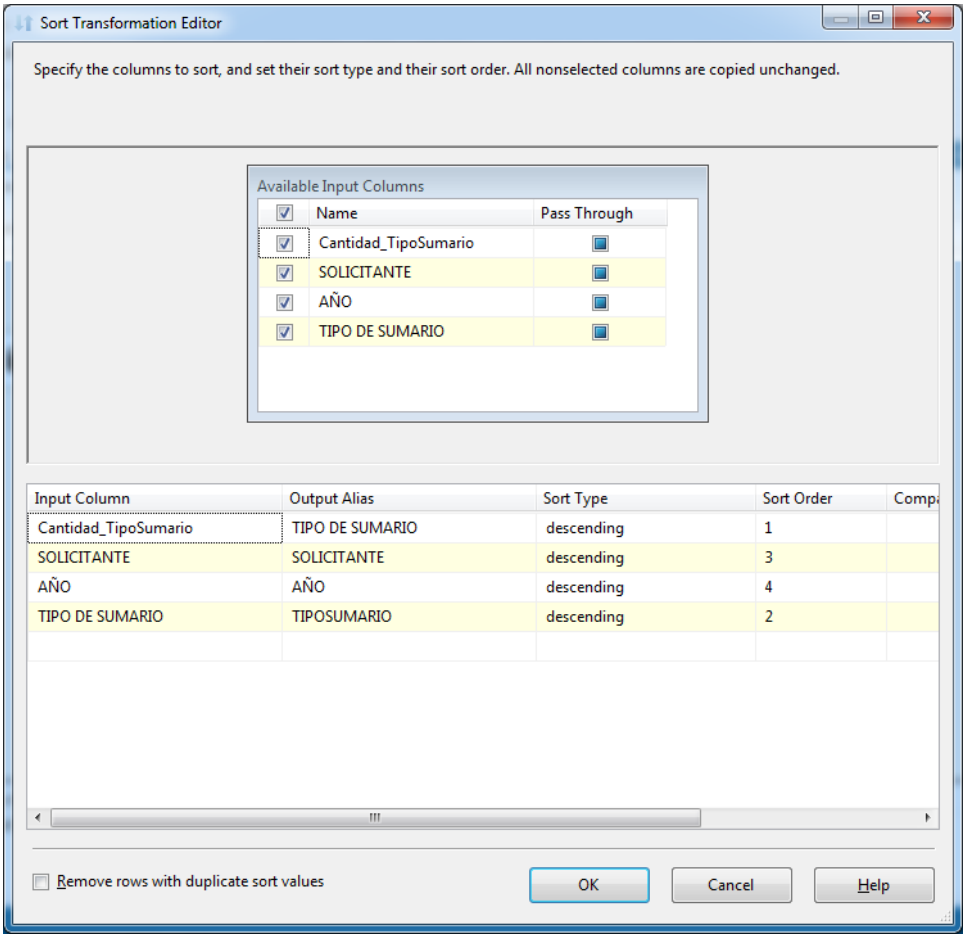


Luego se especifica la operación que deseamos realizar para obtener la cantidad de causas.

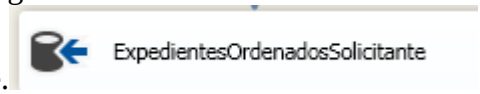


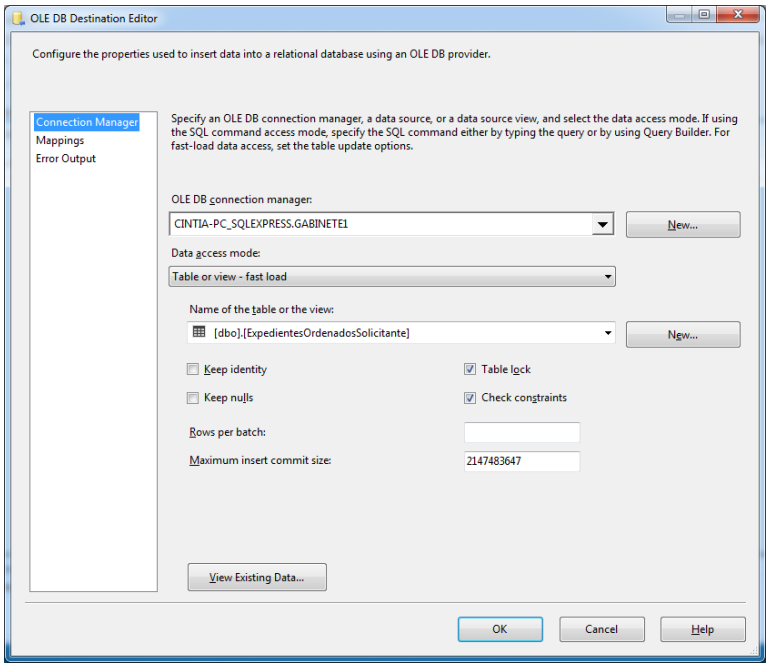


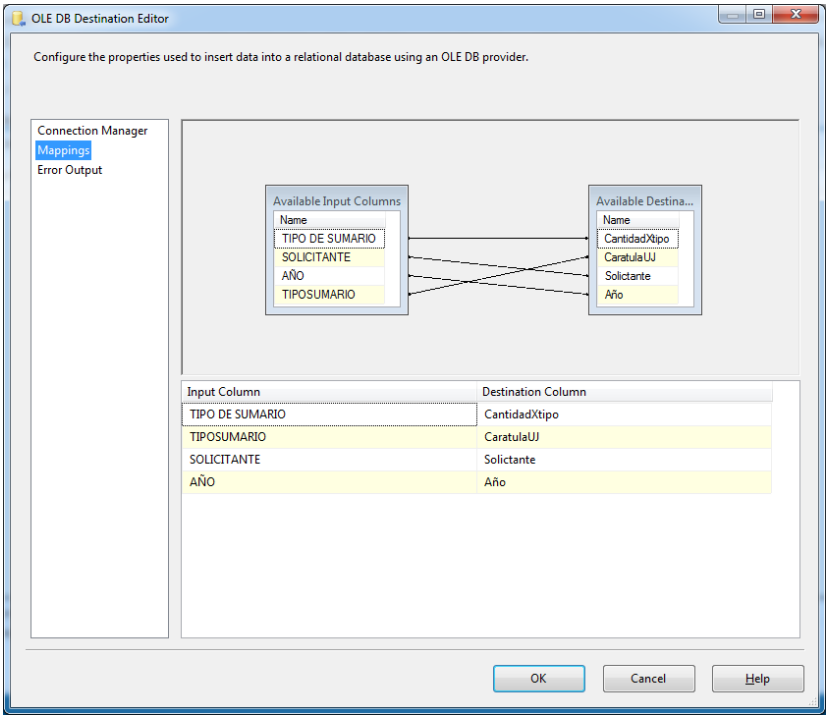
Se ordenan los valores.

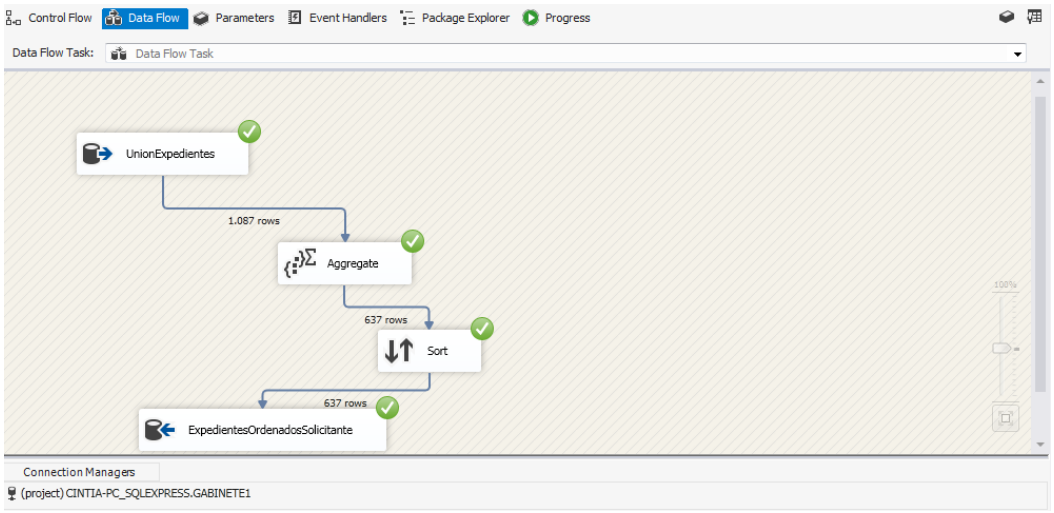


Se configura la tabla destino de la información para el posterior procesamiento del paquete.









A partir de aquí se probaron diferentes modelos y se volvió sobre la limpieza del dataset minable.

Luego de analizar el resultado del paquete anterior, se decidió explorar cuales eran los tipos de causas trabajadas por cada operador; esto nos permitiría predecir el conocimiento y accionar de cada uno de ellos, según los distintos tipos de causas y según la experiencia del mismo.

No serviría mucho identificar cuáles eran la mayor cantidad de causas resueltas por cada uno de ellos, debido a que la resolución de las mismas, como se explicó anteriormente, dependen, en su mayoría, de factores externos.

Dependiendo de la capacidad y conocimiento, es necesario identificar aquellos tipos de causas en las que tienen mayor experiencia, para sus prontas ejecuciones y resoluciones de las mismas. Debido a este análisis se decidió confeccionar un nuevo paquete.

- Paquete OrdenarOperadorCausa (figura 4-10).
- Tabla origen: UnionExpedientesExcel.
- Tabla destino: CausasPorOperador.

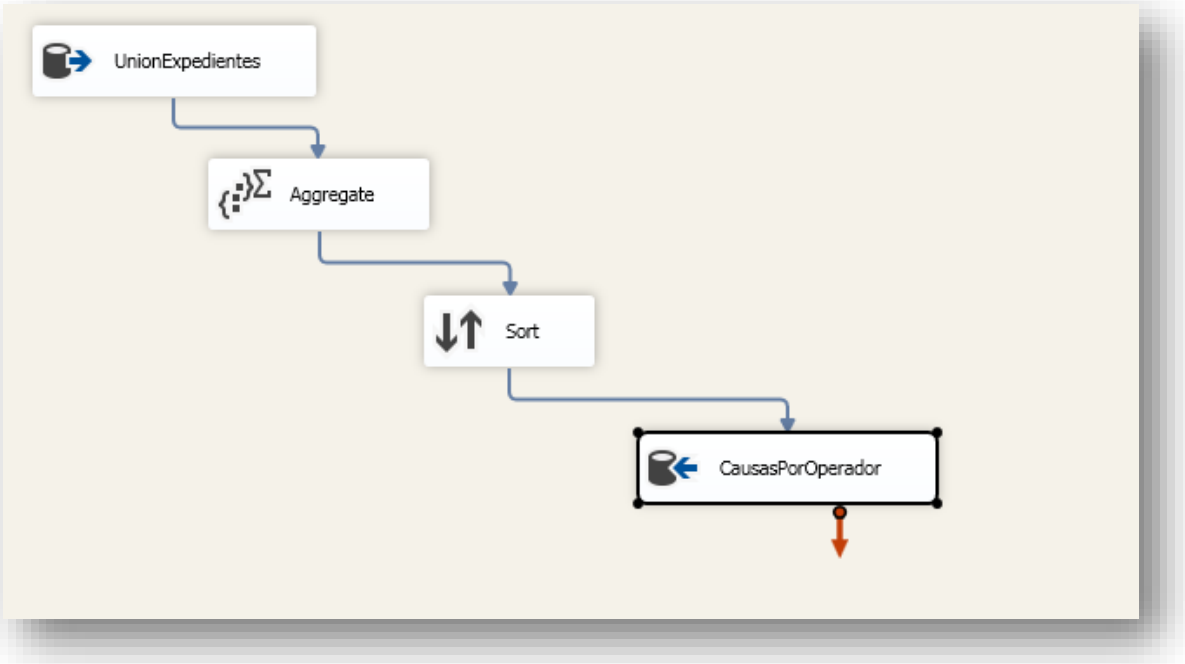
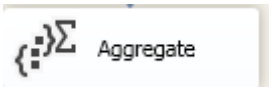
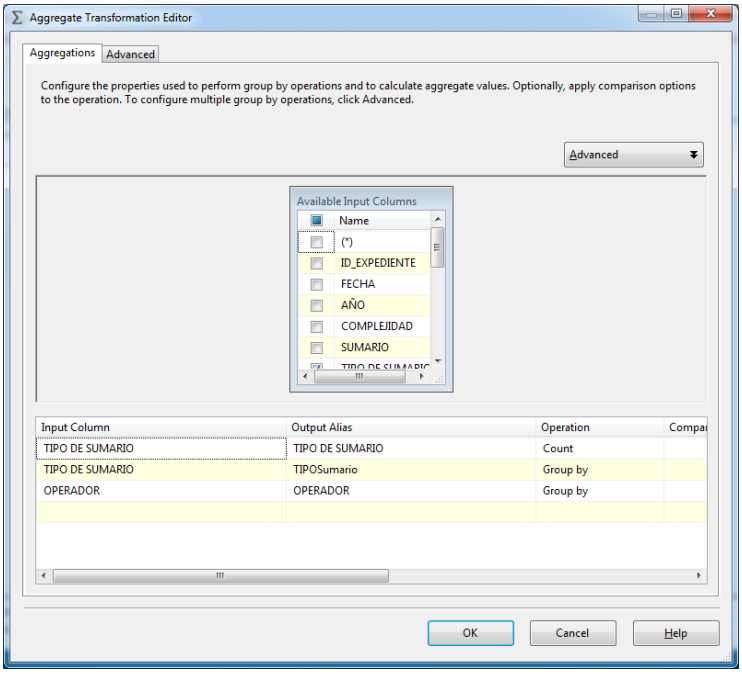


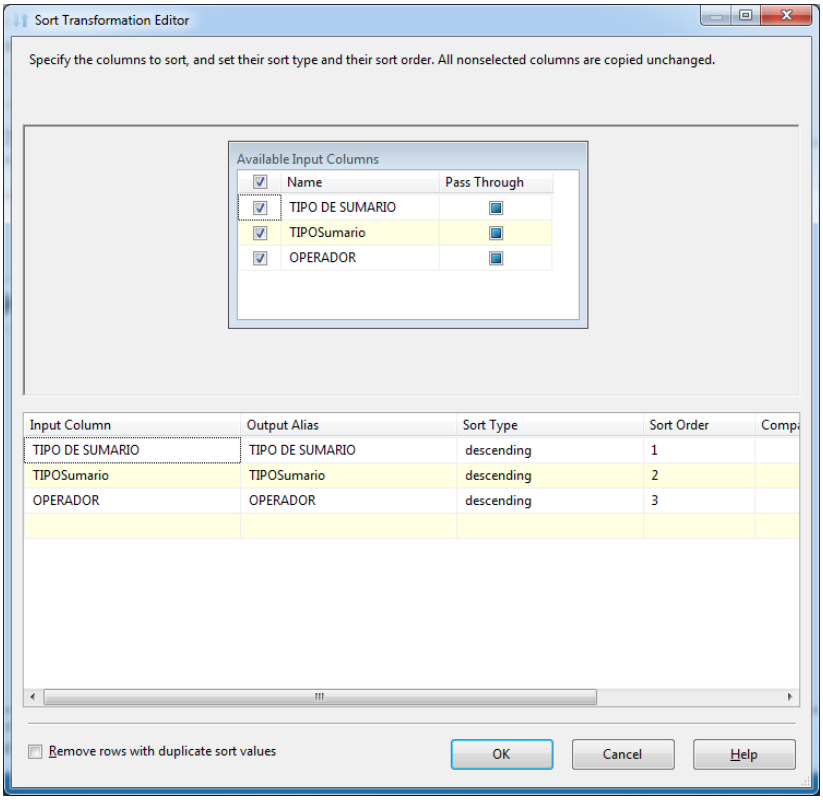
Fig. 4-10: Paquete: OrdenarOperadorCausa

Se configura la operación a realizar para determinar la cantidad de causas por operador.

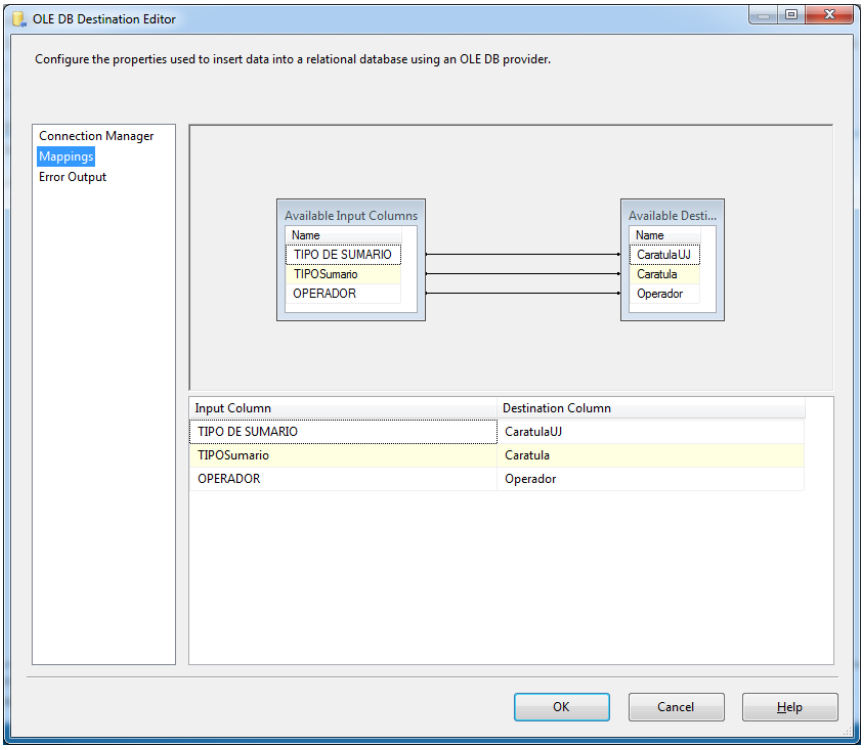


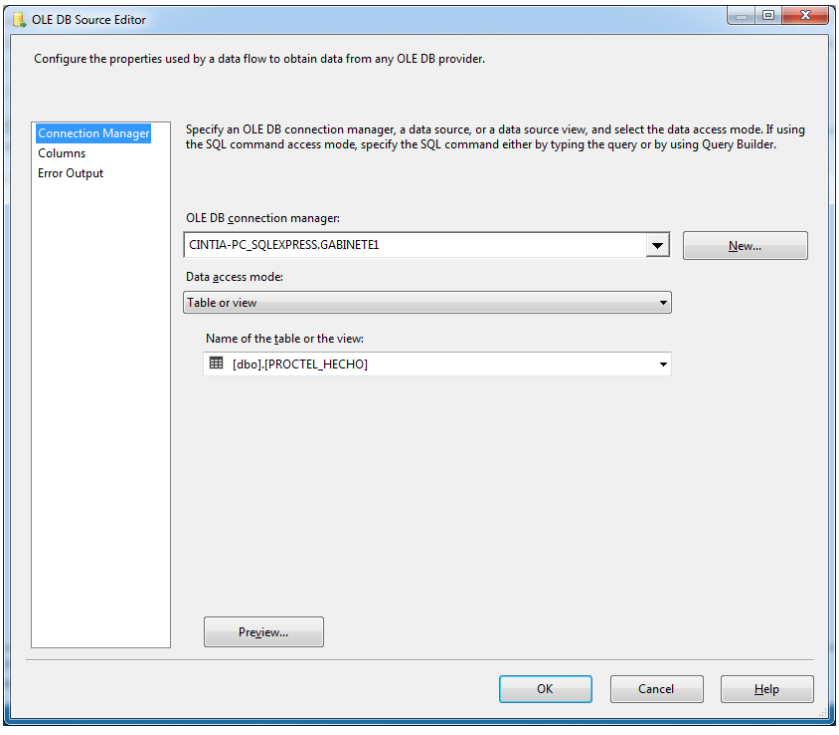


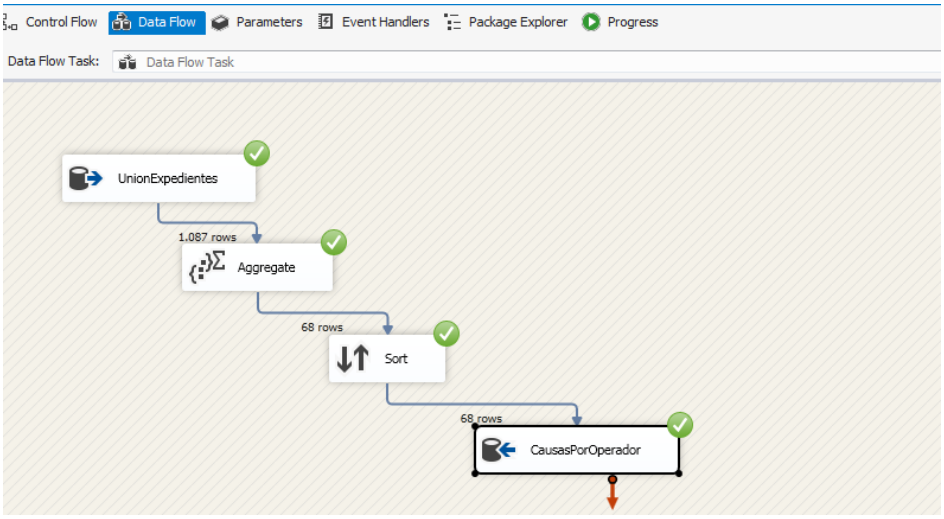
Se ordena los valores.



Se especifica el destino del resultado del paquete.







La figura 4-10 muestra el paquete OrdenarOperadorCausa; en el mismo, se pueden obtener la cantidad de oficios por tipo de causa que tiene asignado cada operador. De esta manera, se puede identificar cuáles son las causas en las que tiene mayor experiencia el operador. Esto es un paso importante en la comprensión de los datos porque este nuevo paquete permite determinar a quién asignarle un determinado oficio según la experiencia y conocimiento del empleado.

- Paquete QuitarNulos (figuras 4-11 y 4-12).

  Tabla Origen: el presenta paquete no es más que una consulta SQL que permite eliminar aquellos registros cuyo campo CaratulaUJ es nulo.

  Tabla Destino: DataSetLimpia (Expedientes1)

  La construcción del modelo se ve afectada por aquellos valores faltantes. Estos valores nulos afectan a la hora de computar las medidas de impurezas como así también el cómo clasificar una instancia con valor faltante. Estos inconvenientes y otros hacen esencial el análisis de estos valores y la toma de decisión sobre lo que se debería hacer respecto a los mismos.

  Se examinó el por qué el campo CaratulaUJ poseía valores nulos, cuáles eran los posibles motivos y si debían ser considerados o no en el estudio propuesto. El campo CaratulaUJ se completa con los datos que proveen los oficios enviados por los magistrados correspondientes. Los valores nulos en dicho campo se deben a la omisión de la clasificación del hecho informado en los oficios que llegan al gabinete. Si bien todos los registros de expedientes son importantes en

este estudio, por el momento aquellos registros con valores nulos en el campo en cuestión no pueden ser clasificados en esta investigación. Cabe mencionar que a raíz de la eliminación de estos campos del dataset preparado para realizar minería de datos, se obtuvo una reducción en la cantidad de registros. Se eliminaron 282 quedando a disposición para el análisis 821. Este resultado es importante para el análisis final como así también para la manera en que se evaluará la eficiencia de los algoritmos.

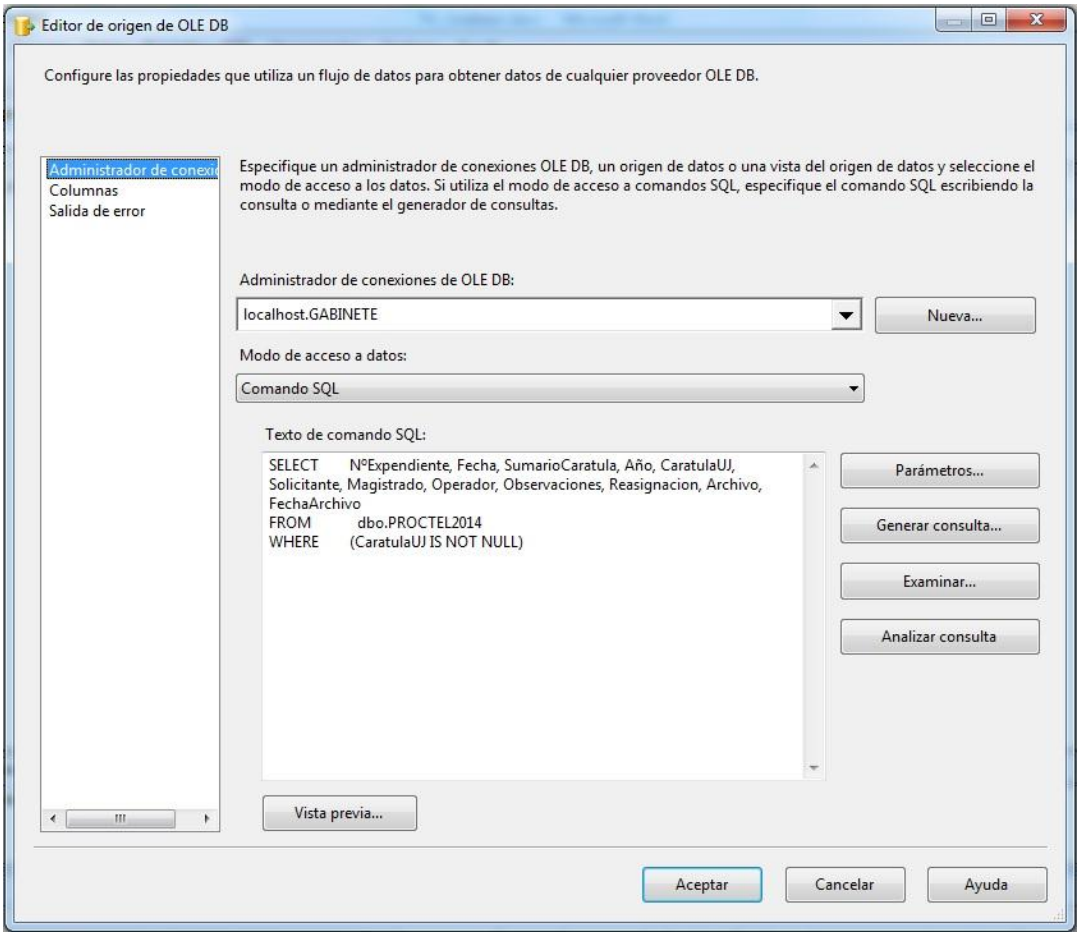


Fig. 4-11: Eliminación valores nulos

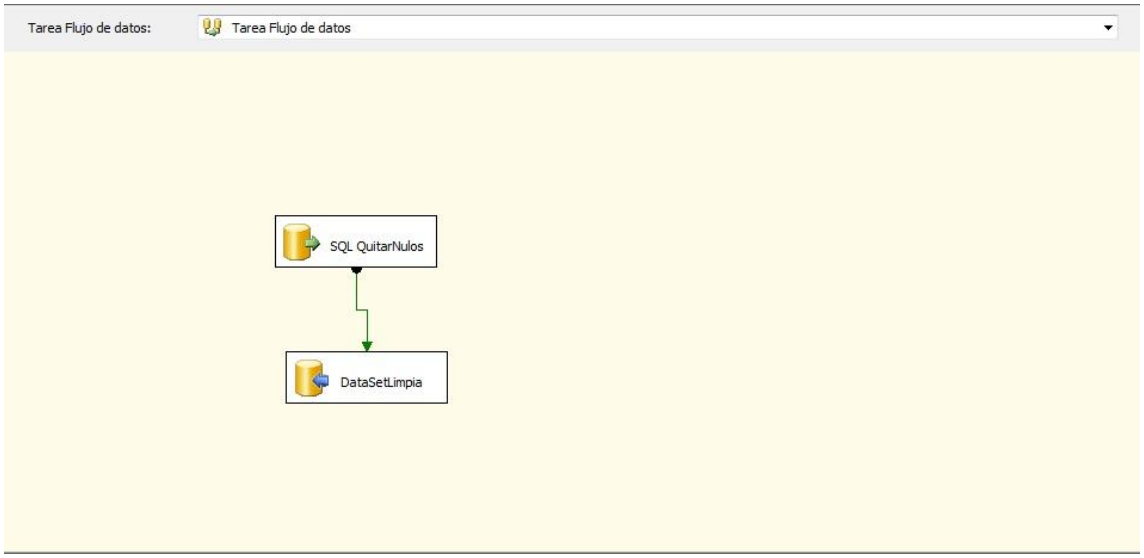


Fig. 4-112: Paquete QuitarNulos

Al finalizar esta fase se obtiene el dataset que será sometido a los algoritmos de minería de datos.

4.4. FASE DE MODELADO

La minería de datos está relacionada con la definición de los modelos. En nuestro estudio se aplicarán métodos predictivos y descriptivos, tales como red neuronal y árbol de decisión entre otros.

Se pretenden encontrar patrones relevantes y significativos entre las diferentes causas que ingresan al área de las diferentes entidades judiciales, al igual que las aptitudes y conocimientos de cada uno de los operadores,

Se deberán definir varios modelos, probando diferentes algoritmos para poder analizarlos y llegar entonces a una conclusión que pueda ser presentada ante el encargado del gabinete y ayude a la toma de decisiones. El análisis de las conclusiones será primordial y para ello deberán presentarse herramientas (a través de modelos) que colaboren en dicha tarea.

4.4.1. Seleccionar técnica de modelado

Nos basaremos en el estudio de diversas técnicas de modelado que nos permita realizar comparaciones entre las mismas, identificar patrones y obtener conclusiones que sean de utilidad para el encargado del área.

4.4.1.1. Modelo de Microsoft Naïve Bayes

Thomas Bayes estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados. El teorema que lleva su nombre se refiere a la probabilidad de un suceso condicionado por la ocurrencia de otro suceso. Más específicamente, con su teorema se resuelve el problema conocido como "de la probabilidad inversa".

Bayes fue pionero en utilizar la probabilidad de forma inductiva y construir una base matemática para la inferencia probabilística. Su principal hallazgo fue calcular la probabilidad de un suceso futuro basándose tanto en eventos previos como en las condiciones actuales y cualquier otro factor relacionado. El Teorema de Bayes permite realizar estimaciones basadas en un conocimiento subjetivo a priori, que puede ser modificado con nueva información adicional.

De alguna forma, sistematiza una práctica casi inconsciente: la intuición humana.

Hoy en día, su teoría es la base de la estadística bayesiana, una filosofía de pensamiento cada vez más utilizada para lograr nuevos conocimientos y aplicaciones.

"Lo sucedido en el pasado va a pasar en el futuro"(Thomas Bayes); La ecuación matemática que enunció Bayes no solo proponía que lo sucedido podría repetirse sino que además era posible calcular lo que pasaría en el futuro a través de sus teorías probabilísticas.

Características del algoritmo

No es robusto pero es veloz, por eso puede usarse para grandes volúmenes de datos. Desde el punto de vista computacional, el algoritmo es menos complejo que otros algoritmos de Microsoft y, por tanto, resulta útil para generar rápidamente modelos de minería de datos que detectan las relaciones entre las columnas de entrada y las columnas de predicción. El algoritmo considera cada par de valores de atributos de entrada y de atributos de salida.

El algoritmo combina la probabilidad condicional e incondicional. Las reglas establecen que si existen evidencias sobre la hipótesis E entonces se puede calcular la probabilidad de la hipótesis H a través de la fórmula propuesta en el algoritmo. Busca correlaciones entre entradas y salidas pudiendo realizar tareas de clasificación con facilidad. Para ello, realiza la selección automática de las características para limitar el número de valores que se consideran al generar el modelo.

El algoritmo está diseñado para reducir al mínimo el tiempo de proceso y seleccionar eficazmente los atributos que tienen la importancia máxima.

Supuestos del modelo

El algoritmo no admite valores nulos en el atributo target y las entradas deben ser sensatamente independientes entre sí. Esto quiere decir que, si bien la teoría exige que sean independientes, en la práctica esto no suele suceder y aun así es baja la tasa de error en clasificación.

En nuestro estudio, esto se cumple debido a la ejecución del proceso de ETL denominado QuitarNulos.

A la vez, será necesario comparar el comportamiento del mencionado algoritmo con otros.

Definición de parámetros

El algoritmo Bayes Naïve de Microsoft admite varios parámetros que influyen en el rendimiento y la precisión del modelo de minería de datos resultante. Estos parámetros se describen a continuación:

- MAXIMUN_IMPUT_ATTRIBUTES

Especifica la cantidad máxima de atributos de entrada que el algoritmo puede procesar antes de invocar la selección de características. El máximo permitido es 255, lo cual no excede la situación presente. La función de selección de atributos de entrada se deshabilita cuando este valor se establece en 0.

- MAXIMUN_OUTPUT_ATTRIBUTES

Especifica la cantidad máxima de atributos que se considerarán para la salida, el máximo permitido es 255, lo cual no excede la situación presente. Se mantiene la definición por default. Si este valor se establece en 0, se deshabilita la selección de características para atributos de salida.

- MAXIMUN_STATES

Especifica el número máximo de estados de atributo que admite el algoritmo. Si el número de estados que tiene un atributo es mayor que el número máximo de estados, el algoritmo utiliza los estados más conocidos del atributo e interpreta que faltan los estados restantes. Se mantiene la definición por default que corresponde a 100 y no excede la situación en estudio.

- MINIMUN_DEPENDENCY_PROBABILITY

Especifica la probabilidad de dependencia mínima entre los atributos de entrada y salida. Este parámetro no impacta en el modelo de predicción pero sí en la cantidad de salidas que se obtienen. El valor de default es 0,5, se realizaron

pruebas para ver si modificando este valor aumentaba la cantidad de correlaciones encontradas pero eso no ha sucedido.

Este valor se utiliza para limitar el tamaño del contenido generado por el algoritmo. El valor de esta propiedad puede establecerse en un valor comprendido entre 0 y 1. Los valores mayores reducen el número de atributos en el contenido del modelo

4.4.1.2. Modelo Árbol de Decisión

Los árboles de decisión son modelos de predicción que se utilizan para organizar gráficamente la información sobre las opciones posibles, las consecuencias y el valor final. Se utilizan para calcular las probabilidades y la extracción de datos.

La tarea primordial que puede realizar un árbol de decisión es la clasificación. Clasificar es el acto de asignar una categoría a cada caso presentado. Cada caso contiene un conjunto de atributos, uno de ellos es el atributo clase (target u objetivo). El trabajo consiste en encontrar un modelo (un árbol de decisión) que describa el atributo clase como función de los atributos de entrada.

La meta es que los registros no vistos previamente puedan ser asignados a una clase tan precisamente como sea posible.

La idea principal de los árboles de decisión es dividir los datos en subconjuntos recursivamente. Cada atributo de entrada es evaluado para ver como éste divide al atributo objetivo en subconjuntos. Cuando finaliza el proceso de recursividad el árbol queda completamente formado.

En conclusión, los árboles de decisión se utilizan para decidir entre diversos cursos de acción. Crean una representación visual de los variados riesgos, las recompensas y los valores potenciales de cada opción.

Características del algoritmo

Es la técnica de minería de datos más usada por su velocidad de ejecución en el entrenamiento de los modelos, el alto grado de precisión y la facilidad con la que se comprenden los patrones hallados.

Se puede enunciar como ventajas frente a otros modelos, su rápida construcción y fácil interpretación. Cada nodo se etiqueta en términos de los atributos de entrada. Cada recorrido del árbol desde la raíz a través de las ramas y hasta las hojas describe una regla acerca del atributo objetivo. Logra una predicción eficiente. En las hojas el valor de predicción está basado en la estadística almacenada en cada nodo.

Supuestos del modelo

El algoritmo de Microsoft Decision Trees permite la elección entre criterios de Entropía, Bayesian Score, Bayesiano con prioridad K2, Equivalente Dirichlet. Las particularidades de cada uno de ellos quedan fuera del alcance de este estudio, limitándose el mismo a la evaluación de la ejecución del algoritmo frente a las distintas posibilidades y la elección del criterio más adecuado según el análisis de los resultados.

Es necesario establecer un límite que determine cuándo es necesario detener el proceso de splitting para evitar sobre entrenamiento (overtraining u overfitting) debido al crecimiento recursivo del árbol cuando el mismo no aporta nueva información al modelo. Los pasos de growning (crecimiento) o pruning (corte) se definen a través de parámetros en base a pruebas debidamente evaluadas.

Definición de parámetros

A continuación se describen los parámetros que pueden utilizarse con el algoritmo de árboles de decisión:

- COMPLEXITY_PENALTY

Controla el crecimiento del árbol de decisión. Busca un equilibrio entre el overtraining (ramas largas) y la pérdida de patrones por el corte de dichas ramas. Un valor bajo aumenta el número de divisiones y un valor alto lo reduce. El valor predeterminado se basa en el número de atributos de un modelo concreto, como se describe en la lista siguiente:

  - De 1 a 9 atributos, el valor predeterminado es 0,5.
  - De 10 a 99 atributos, el valor predeterminado es 0,9.
  - Para 100 o más atributos, el valor predeterminado es 0,99.

- MINIMUN_SUPPORT

Determina el número mínimo de cada nodo del árbol, es decir la cantidad de casos mínimos permitidos. Se verificó después de pruebas que 10 es un número aceptable. Es posible que necesite aumentar este valor si el conjunto de datos es muy grande, para evitar el sobreentrenamiento.

- SCORE_METHOD

Determina el método usado para calcular el resultado de la división; es decir, califica el Split durante el crecimiento del árbol. Se compararon las tres posibilidades: Entropía (definido en el valor 1), Bayesian Dirichlet (valor 4 por default), Bayesian with K2 Prior (valor 3). Se concluyó que el óptimo en este estudio es el método Bayesian Dirichlet.

- SPLIT_METHOD

Determina el método usado para dividir el nodo. Define la manera en que se realizará el Split, si se creará un árbol binario (si el parámetro vale 1), completo (haciendo el Split sobre todas las posibilidades del atributo si el parámetro vale 2) o mixto (si se establece el parámetro en 3). En el caso en estudio se realizaron pruebas entre las tres posibilidades pero la interpretación de los resultados obliga la elección del método completo.

- MAXIMUN_IMPUT_ATTRIBUTES

Define la cantidad máxima de atributos de entrada, es decir, el número de atributos de entrada que el algoritmo puede controlar antes de invocar la selección de características. Si el valor excede el parámetro se ejecutará una rutina interna que seleccionará los atributos a fin de optimizar la performance del algoritmo. No es necesario redefinir un valor en este caso, el valor es inferior a 255 que es el máximo permitido.

- MAXIMUN_OUTPUT_ATTRIBUTES

Define el número de atributos de salida que el algoritmo puede controlar antes de invocar la selección de características. Si el valor excede el parámetro se ejecutaría una rutina interna que seleccionará los atributos a fin de optimizar la performance del algoritmo. No es necesario redefinir un valor en este caso, el valor es inferior a 255 que es el máximo permitido.

- FORCE_REGRESOR

Fuerza al algoritmo a utilizar las columnas indicadas como regresores, independientemente de su importancia según los cálculos del algoritmo. Este parámetro sólo se usa para árboles de decisión que predicen un atributo continuo.

4.4.1.3. Modelo de Clustering

Clustering es el proceso de agrupar datos en clases de tal forma que los objetos de un clúster tengan una similitud alta entre ellos, y baja (sean muy diferentes) con objetos de otros clústers.

La medida de similitud está basada en los atributos que describen a los objetos.

Cuando el problema es multidimensional, se vuelve complejo y allí es necesaria la ayuda del ordenador. Algunas veces a través de este proceso se descubren variables o características ocultas. Los algoritmos de clusterización nos permiten manejar múltiples variables para agrupar los datos de manera óptima.

Características del algoritmo

Es un algoritmo muy flexible porque soporta todo tipo de datos, la manera en que los mismos serán presentados al algoritmo puede contribuir a la solución del problema.

No se tiene que designar una columna de predicción para generar un modelo de agrupación en clústeres. El algoritmo de clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

Puede ser usado para predecir pero más comúnmente se utiliza para detectar categorías, etiquetarlas y poder luego armar modelos dentro de un clúster seleccionado. El objetivo es encontrar grupos donde los elementos pertenecientes a ellos sean lo más similares entre sí y lo más diferentes entre los de los otros grupos.

El algoritmo de Micrsosoft Clustering identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas.

Después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos y, a continuación, intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El

algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los clústeres.

Puede trabajar de dos maneras diferentes, utilizando el método K-means (para cada clúster se elige un punto como centroide, a partir de allí se calculan las distancias Euclideas de los demás elementos y éstos son asignados a un único clúster según dicho valor, la menor distancia determina a qué clúster pertenece cada elemento. Luego se mueve el centro al "centro" de los componentes del clúster y se vuelven a calcular las distancias. Este algoritmo es de tipo "hard clustering" porque cada elemento es asignado a uno y solo un clúster) y a través del método EM cluster-assignment (se calcula una medida probabilística para determinar qué objetos pertenecen a un determinado clúster, se tiene en cuenta la media y el desvío estándar. Los elementos se asignan según una cierta probabilidad. En este caso existe el overlap (solapamiento), es decir un punto puede pertenecer a más de un clúster con una probabilidad asignada en cada uno de ellos).

Uno de los problemas de este tipo de algoritmos es la performance y velocidad de ejecución en las múltiples iteraciones que debe realizar, Microsoft Clustering provee un framework escalable para optimizar ese proceso.

Supuestos del modelo

El algoritmo no tiene en cuenta valores nulos. Si los encuentra los descarta y no los coloca en ningún clúster. La convergencia del modelo estará determinada por el valor de ciertos parámetros, la selección de los mismos deberá ser el resultado de pruebas y análisis hasta obtener el modelo que más se ajuste a las necesidades.

Definición de parámetros

- CLUSTERING_METHOD

Especifica el método de agrupación en clústeres que va a usar el algoritmo, EM o K-means, ambos con la posibilidad de que sean escalables si la cantidad de datos es elevada. Se eligió K-means (opción 4) tras realizar suficientes pruebas.

- CLUSTER_COUNT

Especifica el número aproximado de clústeres que será generado por el algoritmo. Si no se puede generar el número aproximado de clústeres a partir de los datos, el algoritmo genera tantos clústeres como sea posible. Si CLUSTER_COUNT se

establece en 0, el algoritmo usa la heurística para determinar el mejor número de clústeres que debe generar. Cuando el número de atributos es alto se eleva la cantidad de clústeres a encontrar. Después de varias pruebas y análisis respectivos se consideró que 5 era una cantidad óptima para el presente caso.

- MINUMUN_CLUSTER_CASES

Define la cantidad de elementos para el cual el clúster no se considera vacío. Si el número es elevado se podría llegar a resultados incorrectos.

- MODELLING_CARDINALITY

Especifica el número de modelos de ejemplo que se construyen durante el proceso de agrupación en clústeres. Reduciendo el número de modelos candidatos puede mejorar el rendimiento, pero se corre el riesgo de perder algunos modelos buenos.

Permite mejorar la ejecución del modelo controlando los hilos de ejecución disponibles. No es aplicable en este caso debido a la reducida cantidad de registros disponibles.

- STOPPING_TOLERANCE

Especifica el valor que se usa para determinar cuándo se alcanza la convergencia y el algoritmo termina de generar el modelo. La convergencia se alcanza cuando el cambio general de las probabilidades del clúster es inferior al resultado de dividir el parámetro STOPPING_TOLERANCE entre el tamaño del modelo.
Cuando se trabaja con pocos datos el número no debe ser elevado. En este caso se estableció en 5 después de pruebas y análisis de los modelos obtenidos.

- SAMPLE_SIZE

Especifica el número de casos que el algoritmo usa en cada paso si el parámetro CLUSTERING_METHOD está establecido en uno de los métodos de agrupación en clústeres escalables. El método seleccionado no es escalable.

- CLUSTER_SEED

Especifica el número de inicialización usado para generar clústeres aleatoriamente para la fase inicial de generación del modelo. Si al modificar este parámetro el modelo permanece estable, significa que es el modelo correcto. Se han hecho

pruebas modificando este valor y se obtuvieron probabilidades similares en los clústeres obtenidos.

- MAXIMUN_IMPUT_ATTRIBUTES

Especifica el número máximo de atributos de entrada que el algoritmo puede procesar antes de invocar la selección de características. Si se establece este valor en 0, se especifica que no existe un número máximo de atributos.

Si la cantidad máxima de atributos excede este número el algoritmo invoca el proceso de selección automática donde se elegirán los atributos que más se repiten. Los no elegidos quedarán fuera del proceso de clusterización. No sucede esta particularidad en el caso en estudio.

- MAXIMUN_STATES

Especifica el número máximo de estados de atributo admitido por el algoritmo. Si un atributo tiene más estados que el máximo permitido, el algoritmo usa los estados más conocidos y pasa por alto los estados restantes. Si se excede este límite se agrega la categoría otros que agrupa los estados excedidos. Queda inalterado el valor por default, que corresponde al número 100.

4.4.1.4. Modelo de Reglas de Asociación

Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Se describió el modelo como el análisis y la presentación de reglas fuertes descubiertas en bases de datos utilizando diferentes medidas de interés. Se basa en obtener reglas de asociación que descubran relaciones entre los datos recopilados a gran escala. Por ejemplo, las acciones que realiza un individuo cuando va al supermercado pueden seguir una secuencia de patrones que describan su comportamiento. Se puede observar que las personas que compran el artículo A también compran siempre el artículo B y a su vez las que compran A y B compran o no siempre C.

Encontrar reglas de asociación puede definirse entonces de la siguiente manera: “Dado un conjunto de transacciones, hallar reglas que permitan predecir la ocurrencia de un ítem basado en la ocurrencia de otros ítems en la transacción”.

Características del algoritmo

El algoritmo genera una red entre los estados de los atributos, esta particularidad lo hace diferente a otros algoritmos que encuentran relaciones entre atributos sin importar el estado de los mismos.

La ejecución del algoritmo se realiza en dos pasos:

El primero es una fase intensa de cálculo donde el objetivo es encontrar "itemset" (conjunto de ítems, cada uno formado por el valor de un atributo).

Cada itemset posee un tamaño según la cantidad de ítems que lo componen y se caracteriza por las siguientes medidas:

- Soporte: Cuenta la cantidad de ocurrencias que el conjunto de ítems se encuentra en el total de los datos.
- Confianza o Probabilidad: es la probabilidad de ocurrencia de una regla calculada usando el soporte del itemset. Se refiere al número de registros de la base de datos que cubre la regla entre el número de registros que cubre el antecedente de la misma. Por ejemplo en un itemset formado por los ítems A y B el soporte de (A, B) dividido el soporte de A me da la confianza para el itemset (A, B).
- Lift o importancia: indica si los itemset son dependientes o independientes. Se refiere a cuántas veces el antecedente y el consecuente aparecen juntos más a menudo de lo esperado en el conjunto de datos suponiendo independencia estadística entre antecedente y consecuente. Mide el grado de dependencia entre el entre el antecedente y el consecuente. Un valor superior a 1 indica dependencia positiva, mientras que un valor inferior a 1 indica dependencia negativa. El cálculo matemático se realiza utilizando el logaritmo de la probabilidad de (B/A) dividido la probabilidad de (B/not A).

El segundo es un paso que genera reglas de asociación basadas en la frecuencia de los itemsets. Consume mucho menos tiempo de ejecución respecto de la primera fase. Si se define un atributo de predicción el mismo aparecerá siempre del lado derecho de la regla, mientras que los atributos de entrada conformarán el lado izquierdo.

El algoritmo de Microsoft Association Rules pertenece a la familia de algoritmos de asociación a priori porque integra en base a la longitud de los itemsets encontrados, iniciando por los de longitud 1 e incrementando la misma hasta no encontrar más conjuntos que se repitan.

Supuestos del modelo

No acepta atributos continuos porque el motor cuenta la cantidad de correlaciones entre atributos discretos.

Definición de parámetros

Particularmente este algoritmo es muy sensible a la definición del valor de los parámetros.

- MINUMUN_SUPPPORT

Es el soporte mínimo (cantidad de ocurrencias) a partir del cual se considerarán útiles las reglas encontradas por el algoritmo.

Si desea generar menos conjuntos de elementos, o limitar su tamaño, se debe usar dicho parámetro. En el presente estudio y luego de pruebas y comparaciones se estableció el valor en 10. Este número indica que para que una regla sea considerada por el algoritmo deberá ocurrir al menos 10 veces en el total de datos.

- MAXIMUN_SUPPORT

Indica la cantidad máxima de soporte permitido. Puede expresarse como un número entero o como un porcentaje del total en un rango de 0 a 1. Se definió el valor 1 como valor apropiado.

- MINIMUN_PROBABILITY

Define la probabilidad para una regla de asociación hallada. El valor de default es 0,4. Las pruebas realizadas modificando este valor no fueron satisfactorias.

- MINUMUN_IMPORTANCE

Filtra las reglas cuya importancia mínima sea inferior a este valor. No se declaró un valor con el fin de poder observar el total de reglas detectadas.

- MAXIMUN_ITEMSET_SIZE

Define el tamaño máximo para tener en cuenta un itemset. Afecta el tiempo de ejecución del algoritmo. La cantidad reducida de datos hace que este parámetro no sea relevante.

- MINIMUN_ITEMSET_SIZE

Se usa cuando es necesario considerar solamente aquellos itemset que contengan una cantidad mínima de ítems. Teniendo en cuenta el acotado conjunto de datos con el que se trabaja no fue necesario redefinirlo.

- MAXIMUN_ITEMSET_COUNT

Detiene el algoritmo una vez hallada la cantidad máxima de itemsets definidas por el parámetro. Teniendo en cuenta el acotado conjunto de datos con el que se trabaja no fue necesario redefinirlo.

- OPTIMIZED_PREDICTION_COUNT

Optimiza la cantidad máxima de predicciones que al menos debería encontrar el algoritmo en una query. Un valor igual a cero reportará todas las predicciones posibles. No fue oportuno modificarlo.

- AUTODETECT_MINIMUN_SUPPORT

Cuando el valor es 1.0 el algoritmo detectará el valor mínimo apropiado para el soporte; cuando es 0.0 se utilizará el soporte mínimo definido como parámetro. Se definieron los valores manualmente, no se utilizó esta característica.

4.4.1.5. Modelo de Red Neuronal

Es un modelo matemático óptimo para mejorar o entender las relaciones complejas entre entradas y salidas. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida. El algoritmo combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente estas probabilidades pueden usarse para la regresión, la clasificación o para predecir el resultado de un atributo de predicción basándose en los atributos de entrada.

Es decir, que parte de un conjunto de datos de entrada suficientemente significativo y el objetivo es conseguir que la red aprenda automáticamente las propiedades deseadas.

Un algoritmo de red neuronal puede realizar trabajos de clasificación y regresión. Encuentra relaciones no lineales. Puede ser considerado un instrumento de aprendizaje más sofisticado que los algoritmos de Árboles de Decisión y de Bayes Naïve por su

mayor complejidad pero tiene la desventaja del alto tiempo de ejecución que puede requerir y de la dificultad en la interpretación de los resultados.

Se usa cuando el número de datos que se dispone es elevado. Existen distintos tipos de redes neuronales, Microsoft en su algoritmo utiliza el tipo feed forward, es decir inicia con una cantidad de entradas similar a la cantidad de neuronas de la primera capa y la transmisión es hacia adelante.

Características del algoritmo

Nodos de entrada: forman la primera capa de la red, cada nodo mapea un atributo (un único estado de un atributo cuando se trata de valores discretos). Los valores deben ser normalizados en una misma escala para que las comparaciones sean posibles.

Nodos ocultos: se encuentran en la capa intermedia. Reciben entradas de las neuronas de entrada y proporcionan salidas a las neuronas de salida. En esta capa oculta se asignan pesos a las distintas probabilidades de las entradas. Un peso describe la importancia de una entrada determinada para la neurona oculta. Cuanto mayor sea el peso asignado a una entrada, más importante será el valor de dicha entrada. Los pesos pueden ser negativos, lo que significa que la entrada puede desactivar, en lugar de activar, un resultado concreto.

Nodos de salida: representan valores del atributo de predicción para el modelo de minería de datos. Para atributos discretos una neurona de salida representa un único estado del atributo target.

Combinación y activación: cada neurona de la red es una unidad de procesamiento básico. Para combinar las entradas existen diferentes métodos, el algoritmo de Microsoft Neural Network utiliza una aproximación de los pesos ponderados. *Backpropagation:* Al iniciar el algoritmo se asignan los pesos de manera aleatoria a los diferentes nodos, luego el algoritmo calcula las salidas. A continuación se calcula el error para cada salida y neurona de la capa oculta. De esta manera se actualizan los pesos de la red. Este proceso se repite hasta que exista la condición de finalización.

Topología de la red: la topología debe fijarse antes de iniciar el proceso. La cantidad de capas puede generar overtraining y está relacionada a la performance del

algoritmo. Debido a estudios que garantizan un óptimo funcionamiento el algoritmo, Microsoft Neural Network no permite más de una capa oculta.

Supuestos del modelo

Todas las entradas pueden estar relacionadas a alguna o a todas las salidas y la red considera estas relaciones en el entrenamiento.

El algoritmo puede predecir tanto atributos discretos como continuos.

Debe contener por lo menos una columna de entrada y una columna de salida.

Las relaciones detectadas por el algoritmo de Microsoft Neural Network pueden tratarse de dos maneras. En un nivel simple, en cuyo caso se trata de una Regresión Logística o en dos niveles cuando se define una capa oculta y las entradas no pasan directamente a la salida como en el caso anterior sino que existe esa capa intermedia.

Definición de parámetros

- MAXIMUN_IMPUT_ATTRIBUTES

Determina el número máximo de atributos de entrada que se pueden proporcionar al algoritmo antes de emplear la función de selección de características (que elige los más significativos). Esta función se deshabilita cuando el valor se establece en 0. El valor predeterminado es 255. Para el presente estudio no es necesario modificarlo.

- MAXIMUN_OUTPUT_ATTRIBUTES

Establece la cantidad máxima de atributos de salida que se pueden proporcionar al algoritmo antes de emplear la función de selección de características (que elige los más significativos). Esta función se deshabilita cuando el valor se establece en 0. El valor predeterminado es 255. Para el presente estudio no es necesario modificarlo.

- MAXIMUM_STATES

Determina el número máximo de estados discretos por atributo que admite el algoritmo. Si para un determinado atributo dicho número es mayor que el número especificado para este parámetro, el algoritmo utiliza los estados más frecuentes de este atributo y trata al resto como estados que faltan. No debe modificarse en este caso el valor 100 que es el que se usa por default.

- HOLDOUT_PERCENTAGE

Define el porcentaje de escenarios de los datos de entrenamiento utilizados para calcular el error de exclusión, que se utiliza como parte de los criterios de detención

durante el entrenamiento del modelo de minería de datos. Se deja el valor por default que corresponde al 30%.

- HOLDOUT_SEED

Establece el número que se utiliza para inicializar el generador pseudoaleatorio cuando el algoritmo determina aleatoriamente los datos de exclusión. Si este parámetro se establece en 0, el algoritmo genera la inicialización basada en el nombre del modelo de minería de datos, para garantizar que el contenido del modelo permanece intacto al volver a realizar el proceso. Se realizaron pruebas modificando este valor pero no se obtuvieron resultados interesantes, se decide dejarlo en cero.

- HIDDEN_NODE_RATIO

Especifica la proporción entre neuronas ocultas y neuronas de entrada y de salida. La fórmula *Total neuronas de entrada * total neuronas de salida* determina el número inicial de la capa oculta. El valor predeterminado es 4. Este parámetro no está disponible en el algoritmo de Regresión Logística. Si se construye un modelo con este valor definido en 0 se obtendrá exactamente el mismo resultado que en la ejecución del algoritmo de Regresión Logística. Para este estudio se deja el valor por default.

- SAMPLE_SIZE

Determina el número de escenarios que se van a utilizar para realizar el entrenamiento del modelo. El algoritmo utiliza el valor menor entre este número o el porcentaje del total de escenarios que no están incluidos en los datos de exclusión, según especifica el parámetro HOLDOUT_PERCENTAGE. Se mantiene el valor por default.

4.4.2. Generar Plan de Pruebas

Un plan de pruebas permite especificar lo que desea probar y cómo ejecutar dichas pruebas.

Para evaluar los modelos expuestos es importante tener en cuenta la *precisión*, que muestra hasta donde el modelo pone en correspondencia un resultado con los atributos que se han proporcionado, la *utilidad* (determina si el modelo proporciona

información útil, y la *confiabilidad*, que es donde calcula la manera en que se comporta un modelo en conjuntos de datos diferentes.

En Analysis Services, dentro de las técnicas disponibles se realizarán las siguientes pruebas:

a) Medir la mejora del modelo respecto al modelo predictivo: para ellos se representa gráficamente la mejora que proporciona un modelo de minería de datos en comparación con una estimación aleatoria, y mide el cambio en términos de puntuación de la *mejora respecto al modelo predictivo*. Al comparar las puntuaciones de mejora respecto al modelo predictivo para las distintas partes del conjunto de datos y para los distintos modelos, puede determinar cuál es el mejor modelo y qué porcentaje de casos del conjunto de datos se beneficiaría de aplicar las predicciones del modelo El mismo muestra los resultados del modelo predictivo encontrado, los resultados de una previsión aleatoria y los que generaría un modelo ideal. Al comparar las puntuaciones de varios modelos se podrá determinar cuál de todos es el mejor. En este caso particular se podrá verificar qué tanto mejora una predicción respecto a un modelo aleatorio.

b) Matriz de clasificación (Classification Matrix): Una *matriz de clasificación* ordena todos los casos del modelo en categorías, determinando si el valor de predicción coincide con el valor real. Es una manera de examinar la precisión del modelo. Es una herramienta valiosa porque no solo muestra la frecuencia con que el modelo predice un valor correctamente, sino que también muestra qué valores predice incorrectamente.

c) Validación cruzada (Cross Validation): *La validación cruzada* es una herramienta estándar de análisis que resulta muy útil al desarrollar y ajustar los modelos de minería de datos. Se usa después de crear una estructura de minería de datos y los modelos de minería de datos relacionados para determinar la validez del modelo sirve para comparar la eficiencia de los algoritmos y la exactitud con la cual pueden interpretarse los datos. Es fundamentalmente valiosa cuando no se dispone de un alto volumen de datos porque la validación se realiza sobre el mismo conjunto de datos sobre el que se generó el modelo. La técnica consiste en construir un modelo para cada una de las particiones, entonces se validan los mismos contra la partición seleccionada. Se obtienen tantos resultados como particiones se hayan establecido y se comparan los resultados. Es posible seleccionar el conjunto de datos de la

estructura de minería de datos sobre la cual se ejecutará el modelo. De esta manera esta técnica sirve para comparar cuál es el modelo que mejor está trabajando sin tener que entrenar el mismo con el consecuente consumo de tiempo y de recursos.

4.4.3. Construir el modelo

Los modelos propuestos, se han construido sobre la misma estructura de datos configurada a partir del dataset obtenido en el proceso de ETL en la tabla denominada DatasetLimpia (Expedientes1). Todo se almacena y ejecuta utilizando el motor de Analysis Services.

La figura 4-13 muestra la definición de la estructura de datos y sus campos. La selección de los campos puede variar en función del tipo de algoritmo a utilizar. La figura 4-14 muestra los tipos de datos de las columnas de la estructura.

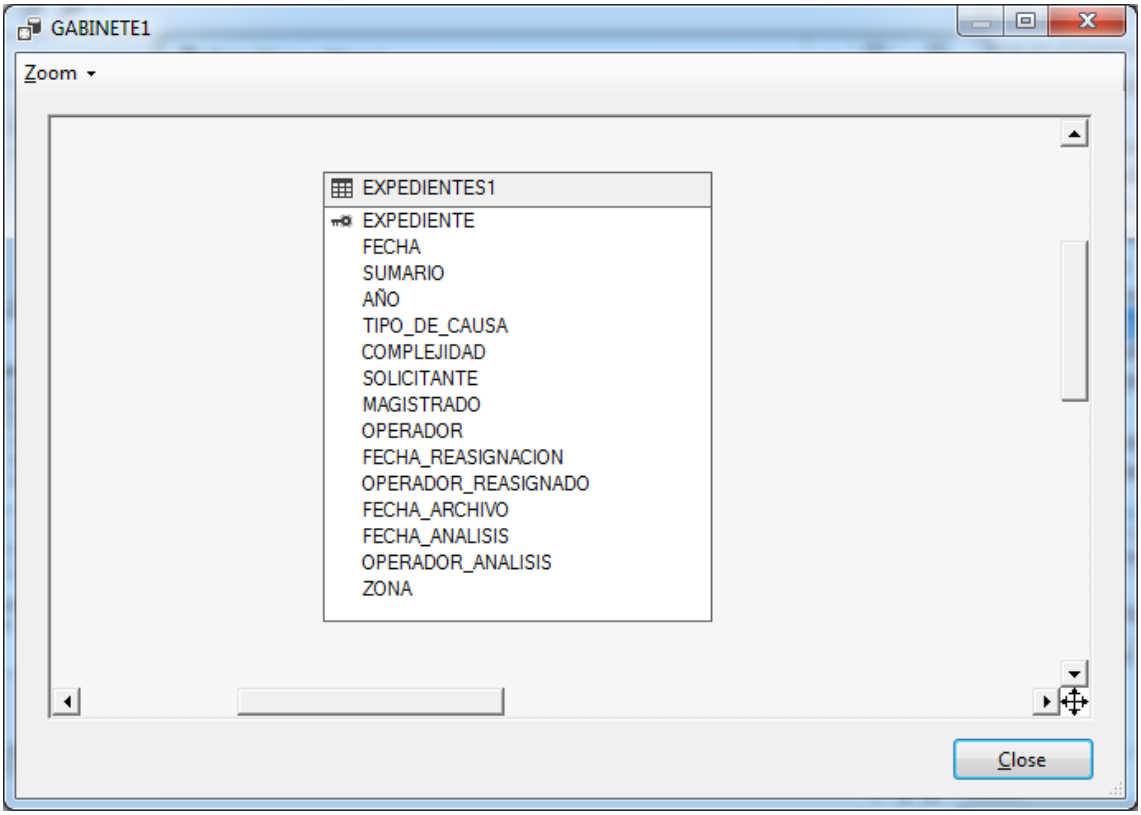


Fig.4.4-13: Estructura PROCTEL

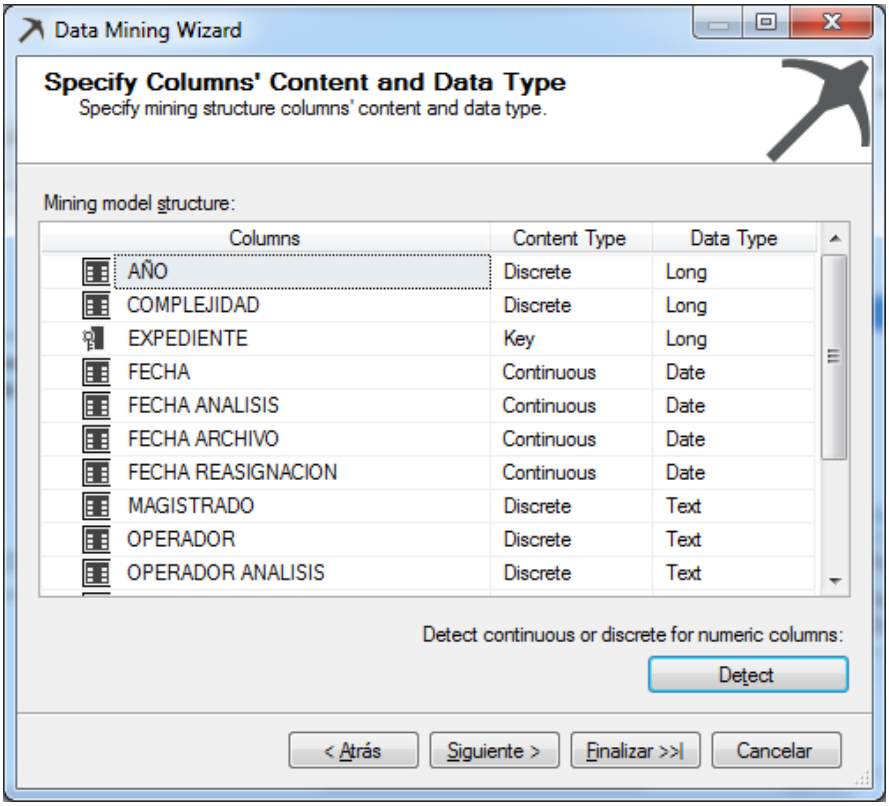
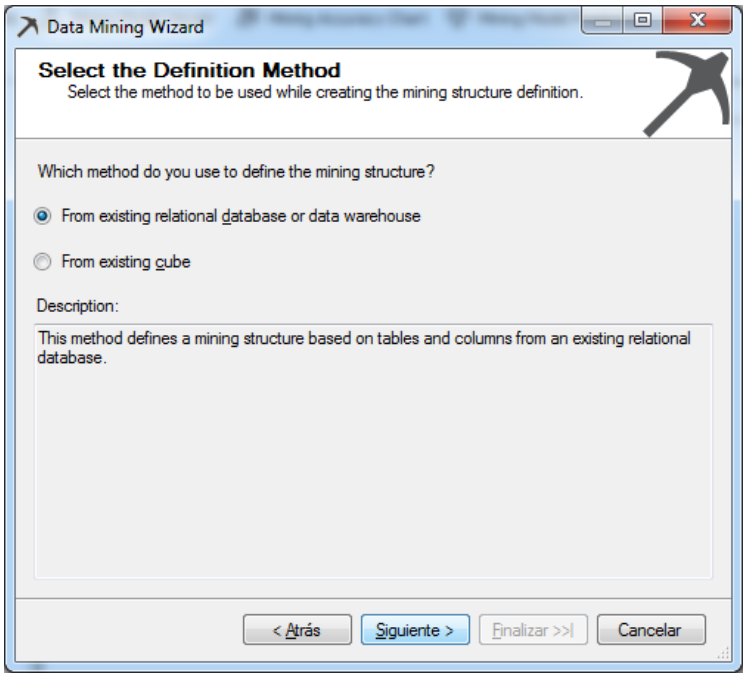


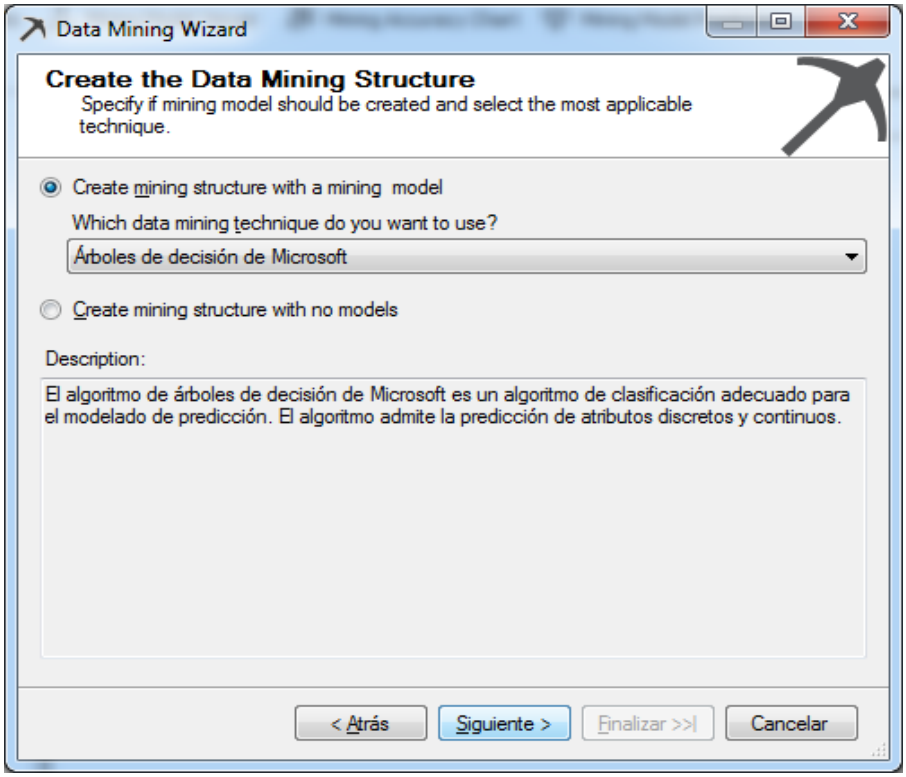
Fig. 4.4-14: Tipos de Datos “Estructura PROCTEL”

Comenzamos a crear una nueva estructura de Minería de Datos

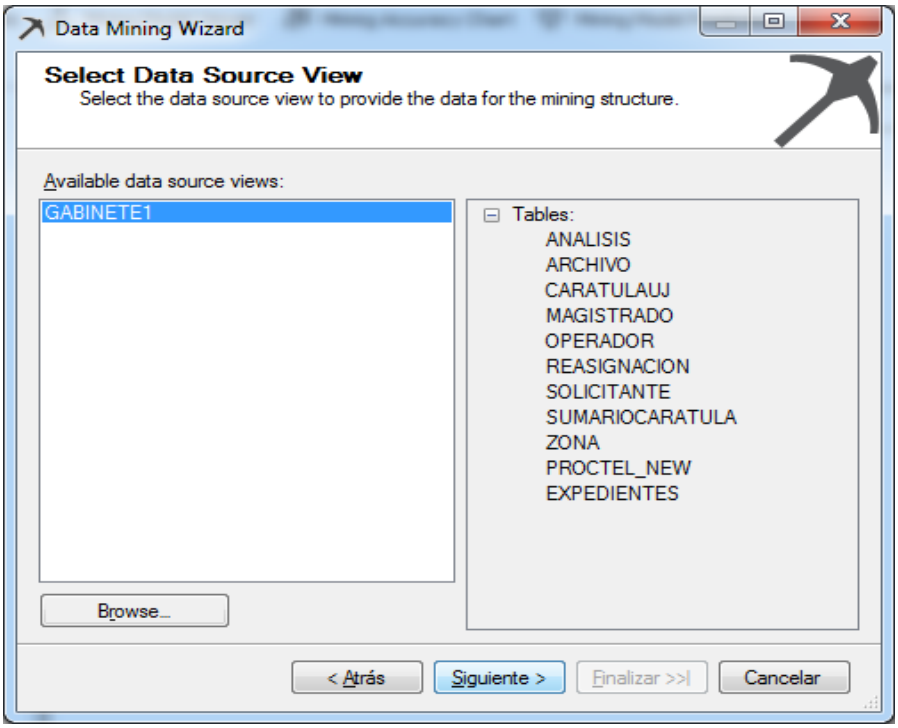
Seleccionamos el método de Estructura de Minería a utilizar



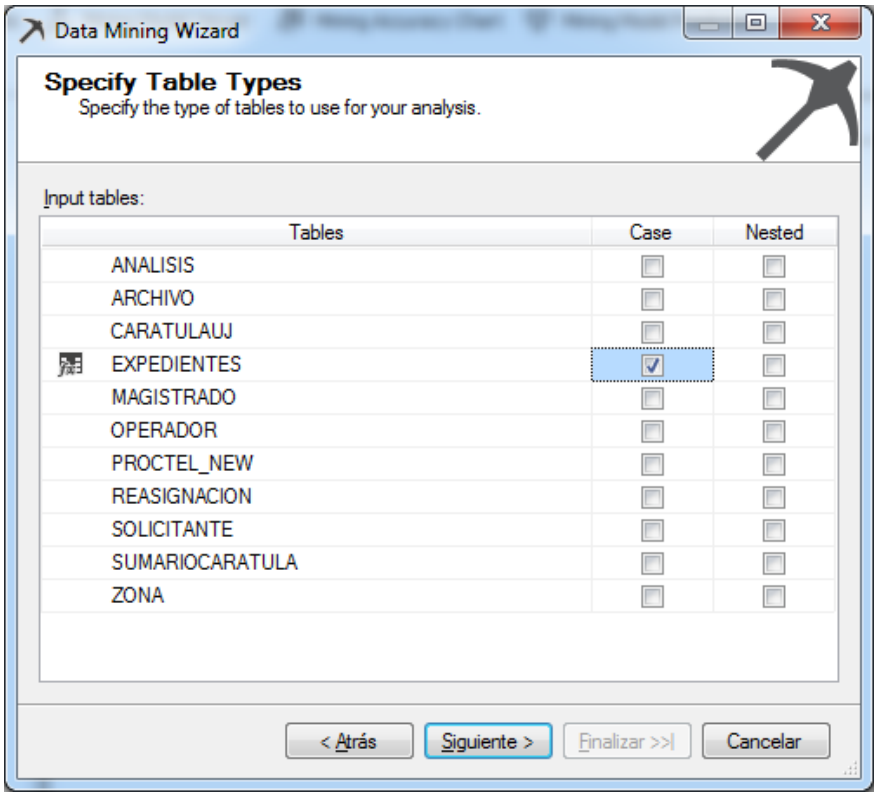
Seleccionamos la técnica, en este caso, se comenzó por el algoritmo de Árbol de Decisión.



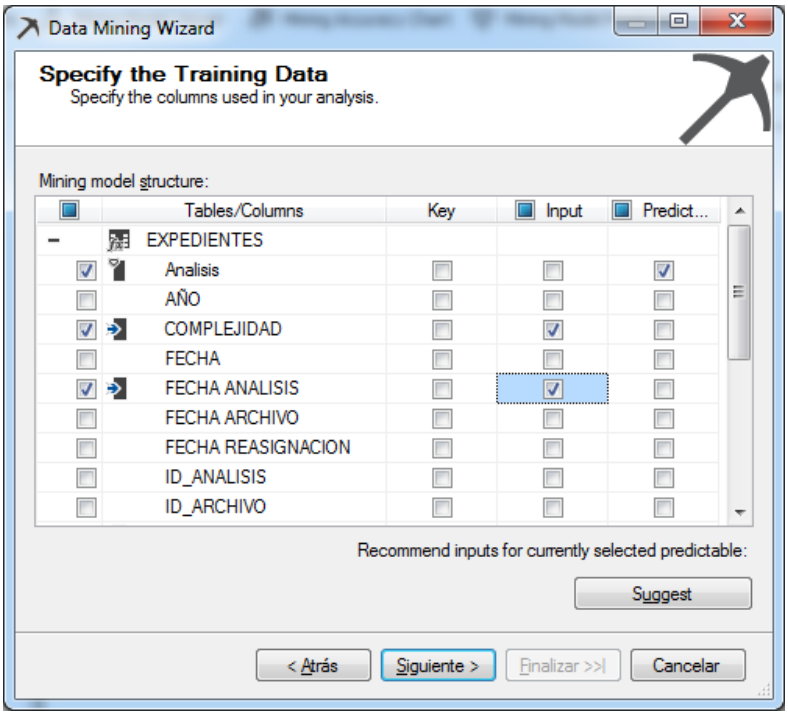
Seleccionamos el Data Source



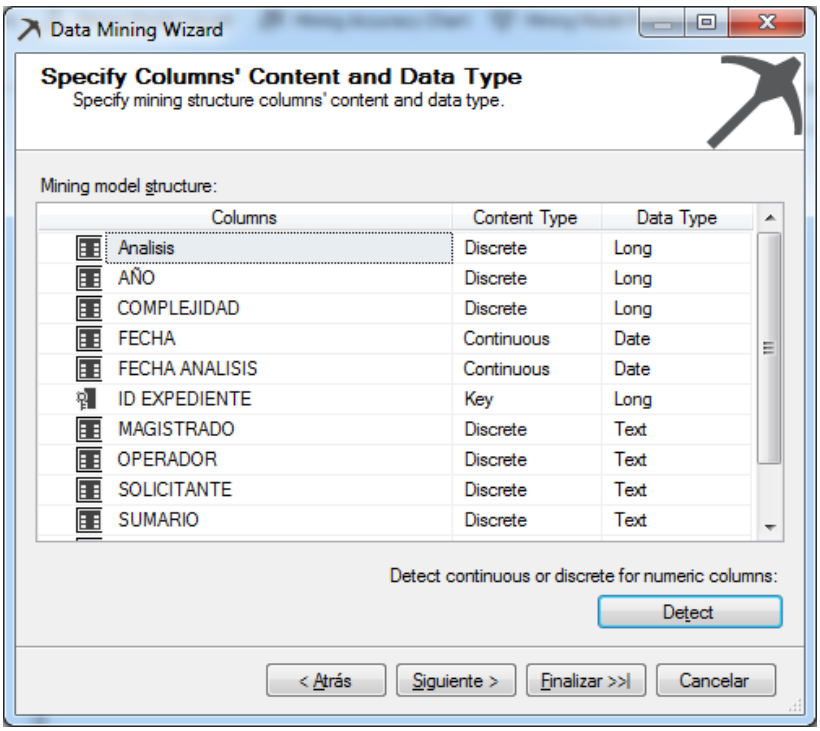
Seleccionamos la Tabla a Utilizar para el modelo



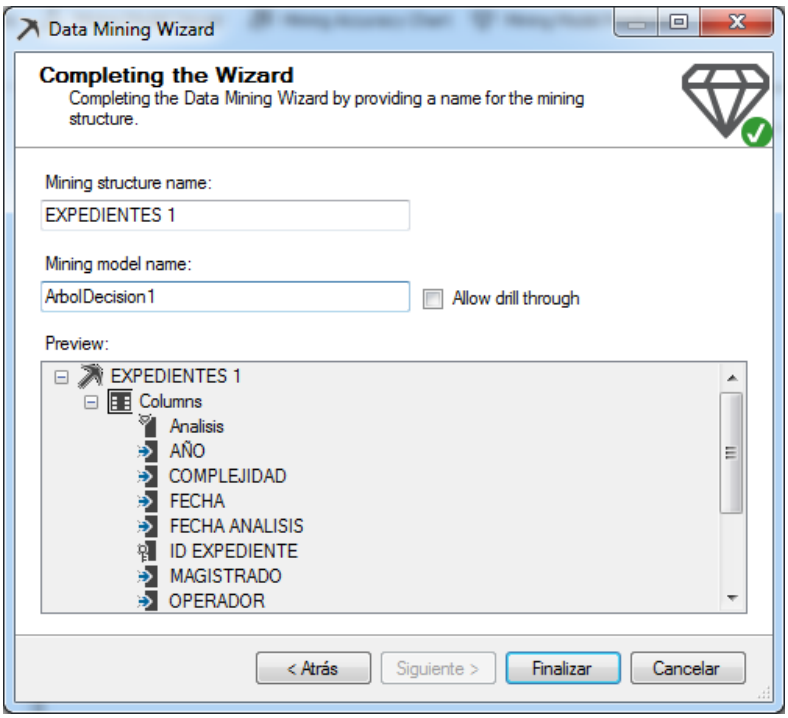
Escogemos la clave principal, los datos de entrada, y el dato a predecir.



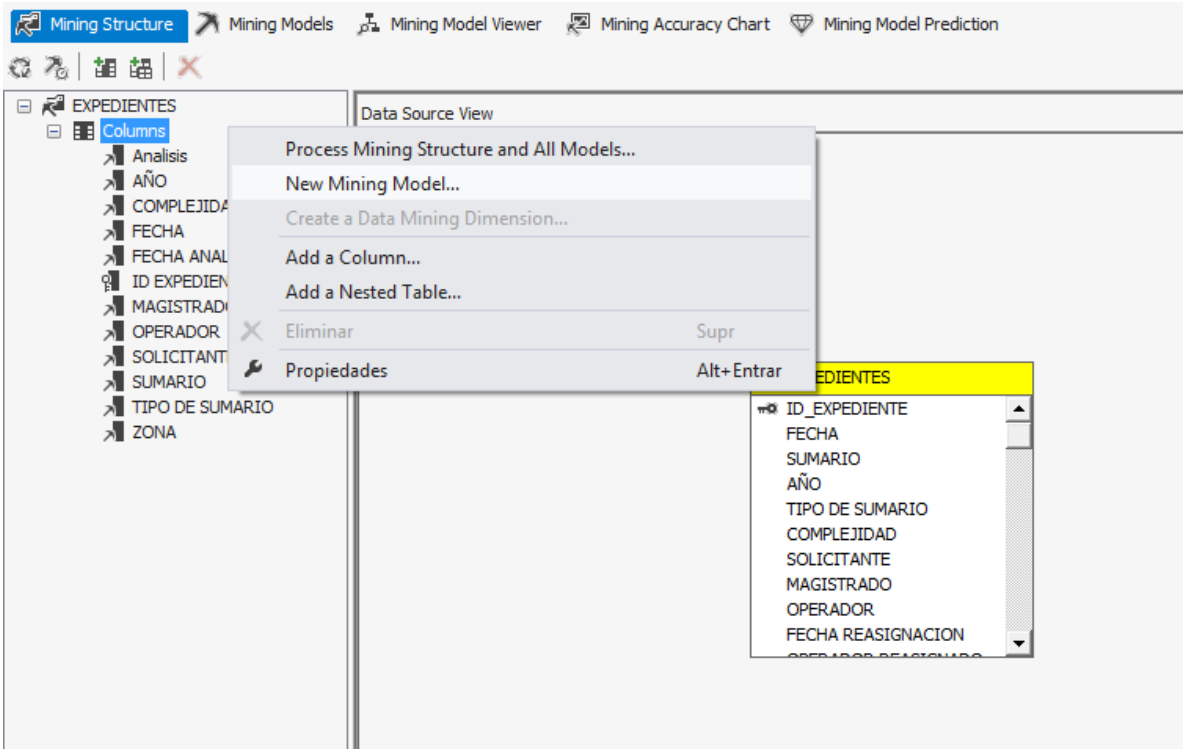
Especificamos el tipo de dato de cada columna



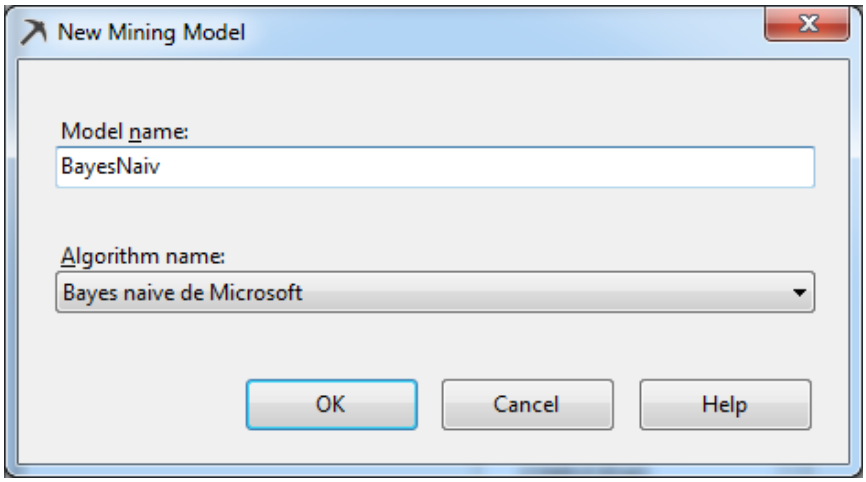
Se define el nombre de la Estructura y el nombre del modelo generado.



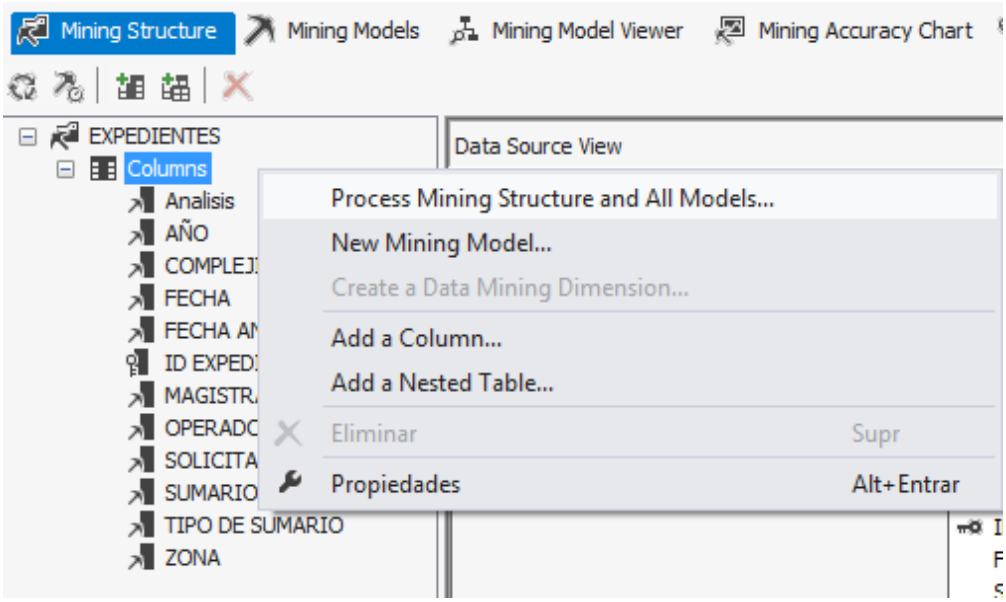
Posteriormente, una vez generado el primer modelo, se procede a generar los restantes. Para ello, se selecciona, Nuevos Modelos.

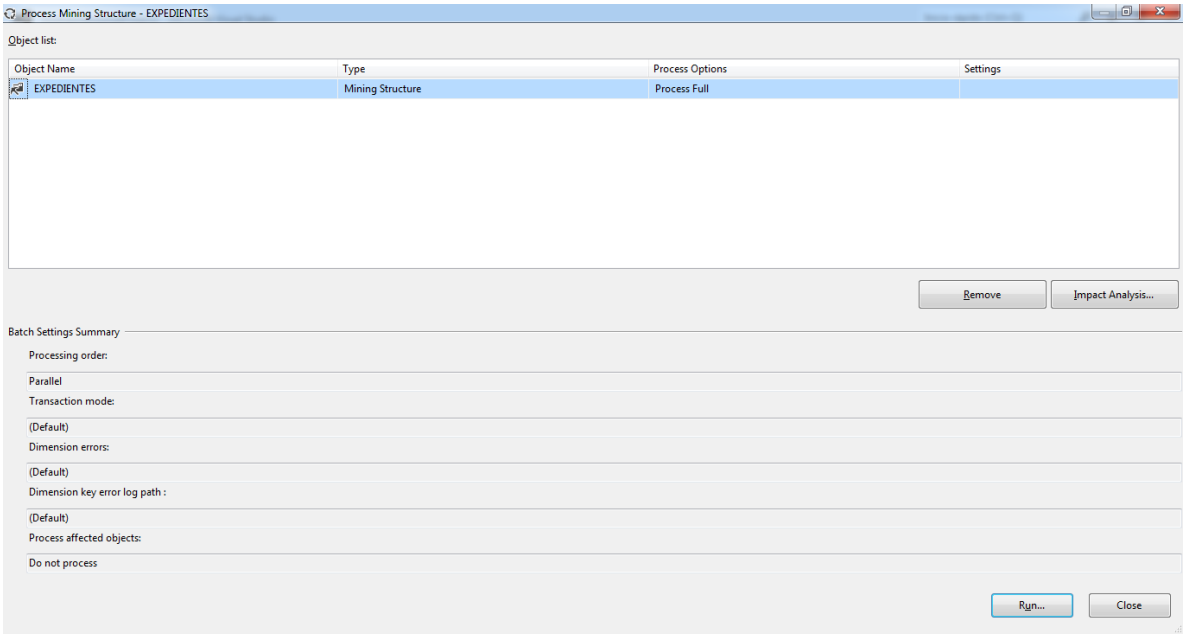


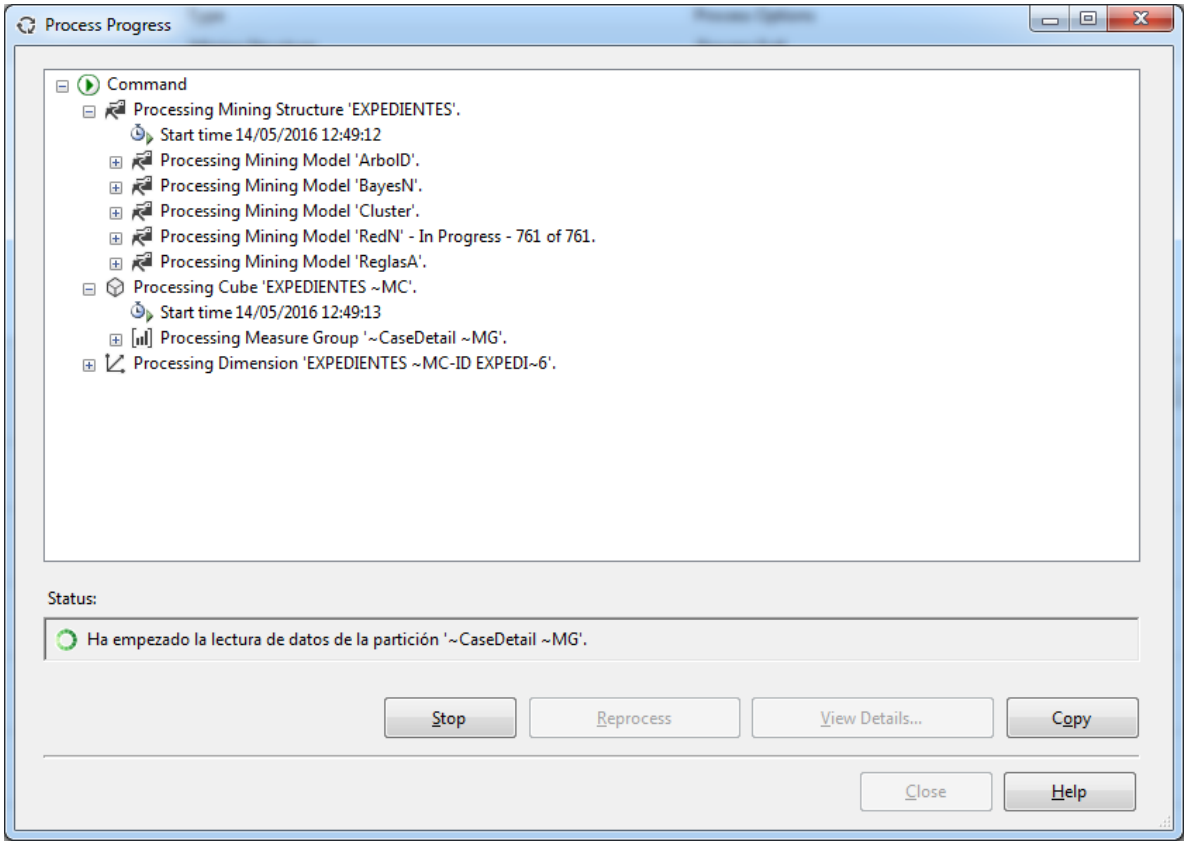
Se procede a colocar el nombre que llevará el modelo y el tipo de algoritmo a utilizar. Así se debe hacer por cada algoritmo que se desee generar.

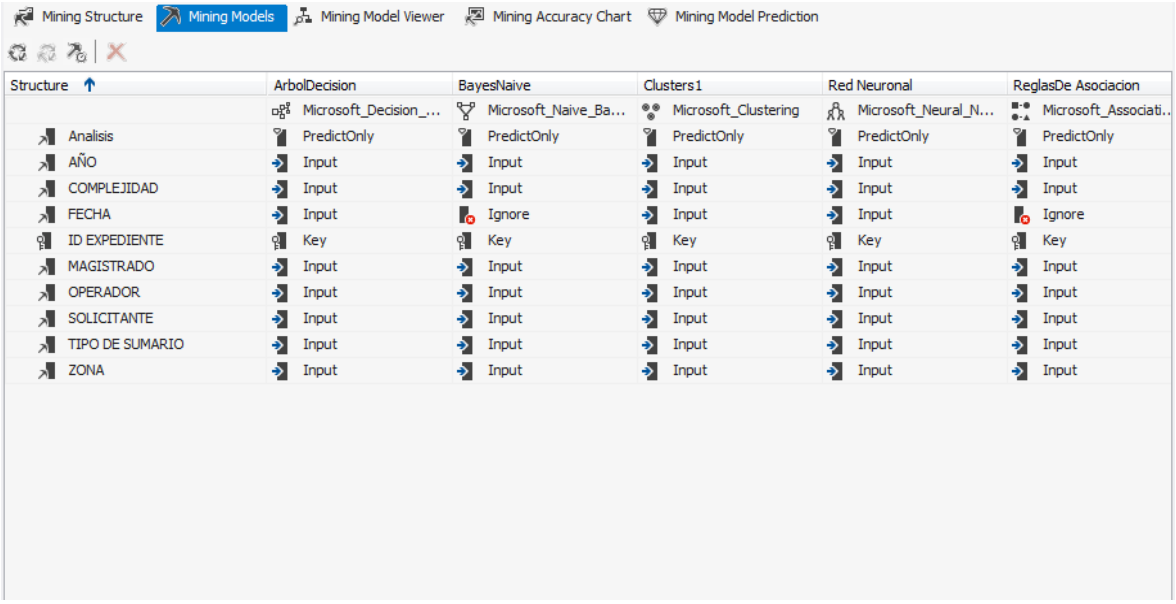


Una vez que contamos con todos los algoritmos, se procede a Procesar las estructuras de Minería de Datos y todos los modelos generados.


Mining Structure
Mining Models
Mining Model Viewer
Mining Accuracy Chart
EXPEDIENTES
Columns
Analisis
AÑO
FECHA
TIPO DE SUMARIO
ZONA
Data Source View
Process Mining Structure and All Models...
New Mining Model...
Create a Data Mining Dimension...
Add a Column...
Add a Nested Table...
Eliminar
Supr
Propiedades
Alt+Entrar


Process Mining Structure - EXPEDIENTES
Object list:
Object Name
Type
Process Options
Settings
EXPEDIENTES
Mining Structure
Process Full
Remove
Impact Analysis...
Batch Settings Summary
Processing order:
Parallel
Transaction mode:
(Default)
Dimension errors:
(Default)
Dimension key error log path :
(Default)
Process affected objects:
Do not process
Run...
Close





Structure	ArbolDecision	BayesNaive	Clusters1	Red Neuronal	ReglasDe Asociacion
	Microsoft_Decision_...	Microsoft_Naive_Ba...	Microsoft_Clustering	Microsoft_Neural_N...	Microsoft_Associati...
Analisis	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly
AÑO	Input	Input	Input	Input	Input
COMPLEJIDAD	Input	Input	Input	Input	Input
FECHA	Input	Ignore	Input	Input	Ignore
ID EXPEDIENTE	Key	Key	Key	Key	Key
MAGISTRADO	Input	Input	Input	Input	Input
OPERADOR	Input	Input	Input	Input	Input
SOLICITANTE	Input	Input	Input	Input	Input
TIPO DE SUMARIO	Input	Input	Input	Input	Input
ZONA	Input	Input	Input	Input	Input

4.4.3.1. Construcción de Bayes Naïve

Parámetros

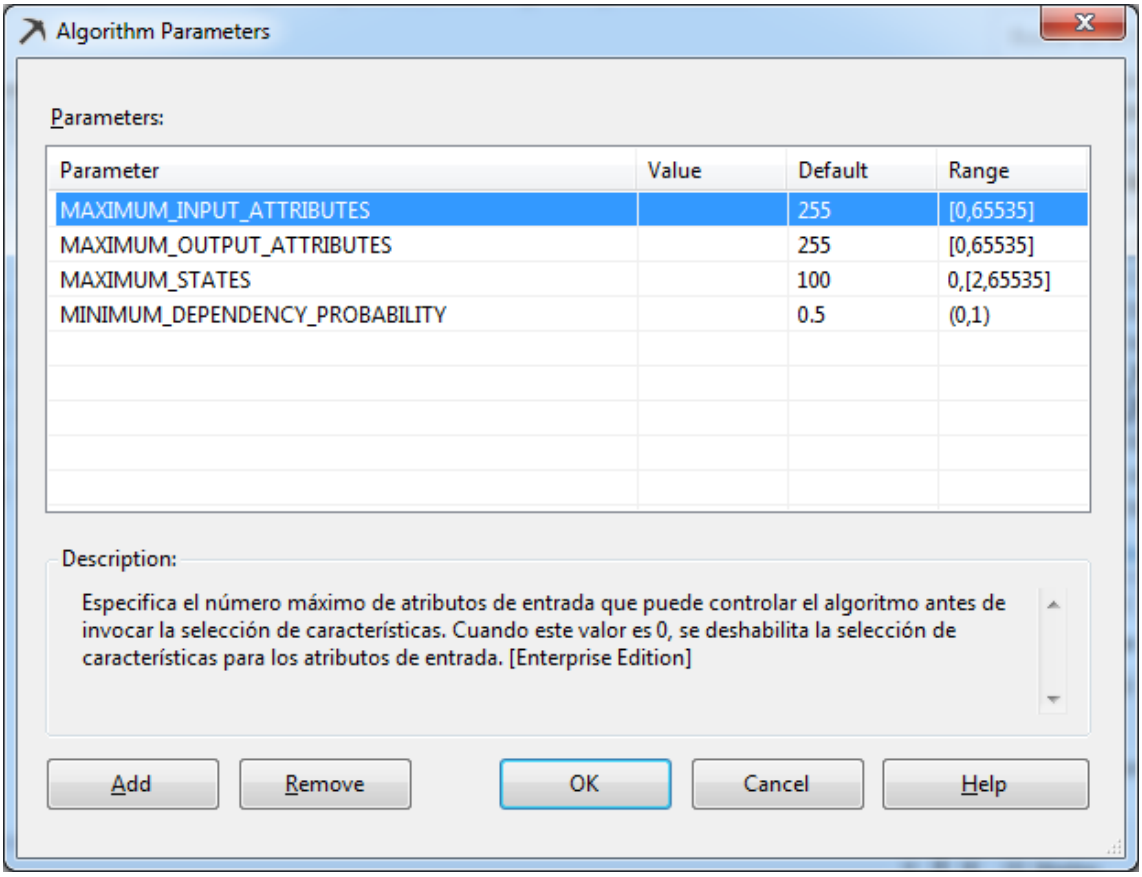


Fig. 4.4-15: Parámetros Bayes Naïve

Exploración del Modelo

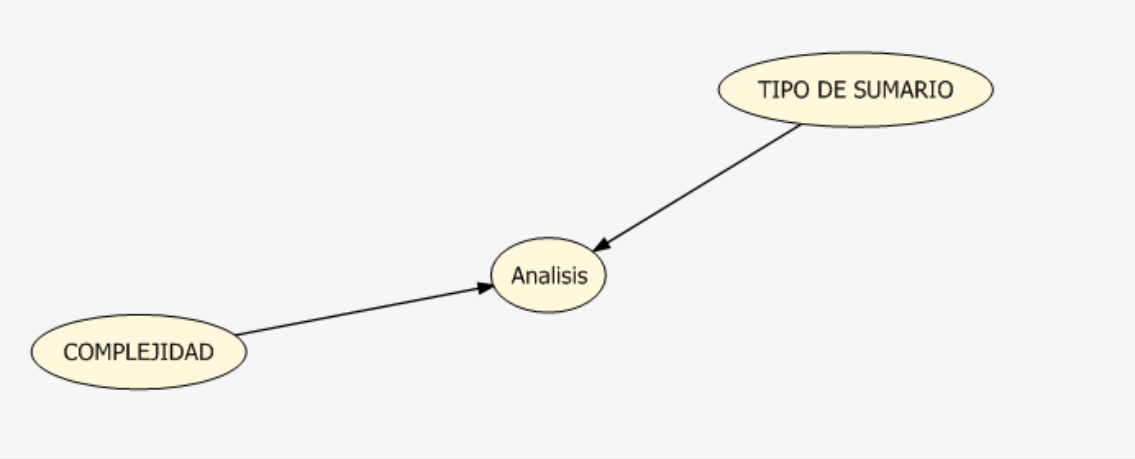


Fig. 4.4-16: Red de dependencias Bayes Naïve

Se observa que el nodo seleccionado Análisis es predecido por Complejidad y Tipo de Causa.

Fig. 4.4-1: Perfiles del Atributo Bayes Naïve

En la figura 4-17 se observa que para cada valor del atributo Análisis (columnas) se obtiene una cantidad de cada uno de los valores de los atributos de entrada. La tabla 4.13 muestra la caracterización para el valor del atributo Setter en 2T-2.

Atributos	Valores	Probabilidad
Complejidad	5	35,498%
Complejidad	1	30,514%
Tipo de Sumario	Amenazas	22,054%

Tipo de Sumario	Desaparición de Persona	21,903%
Complejidad	4	18,429%
Complejidad	2	10,725%
Tipo de Sumario	Amenazas Calificadas	9,970%
Tipo de Sumario	Secuestro Virtual	9,668%
Tipo de Sumario	Homicidio	9366%
Tipo de Sumario	Desobediencia a la Autoridad	8,308%
Tipo de Sumario	Robo Calificado	5,740%
Complejidad	3	4,834%
Tipo de Sumario	Violencia Familia	3,021%
Tipo de Sumario	Estafa	2,417%
Tipo de Sumario	Extorsión	2,417%
Tipo de Sumario	Homicidio Agravado	2,417%
Tipo de Sumario	MED	1,360%
Tipo de Sumario	Encubrimiento	0,755%

Tabla 4.4.1: Características del atributo 2T-2 Bayes Naïve

Esta tabla muestra para el valor del atributo Análisis = 0 la probabilidad con que cada valor de los atributos de entrada caracteriza al objetivo. Los valores están ordenados de manera decreciente según el valor del campo probabilidad. Es de notar que si bien la lectura de estos valores podrá ayudar a identificar patrones, los valores que no se encuentran aquí también servirán a extraer conclusiones.

4.4.3.2. Construcción del Árbol de Decisión

Parámetros

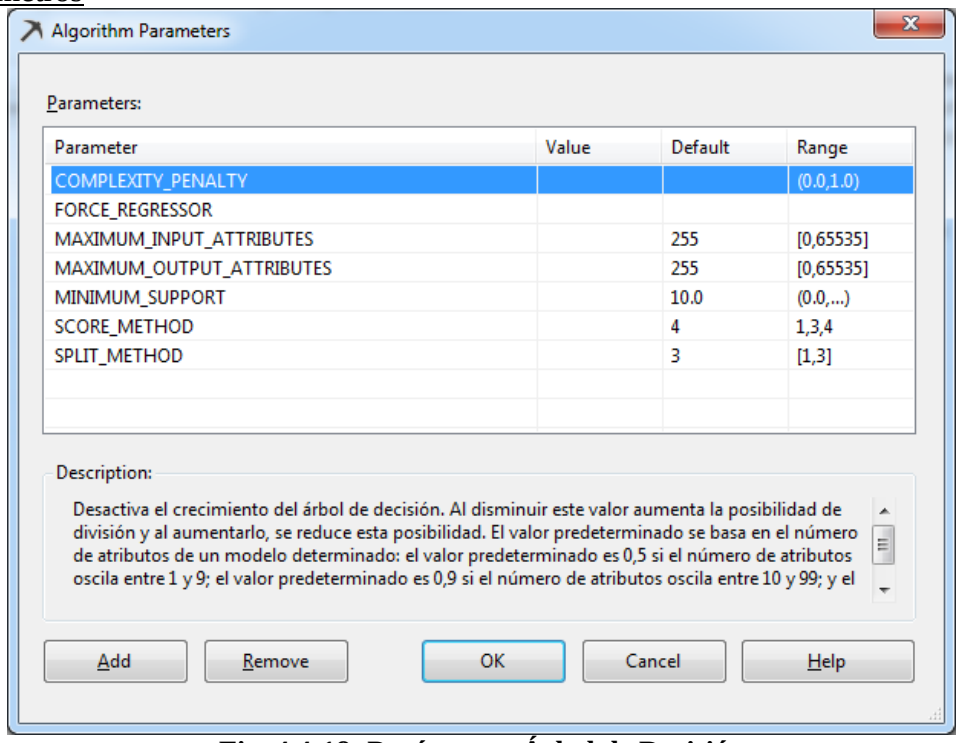


Fig. 4.4-18: Parámetros Árbol de Decisión

Exploración del modelo

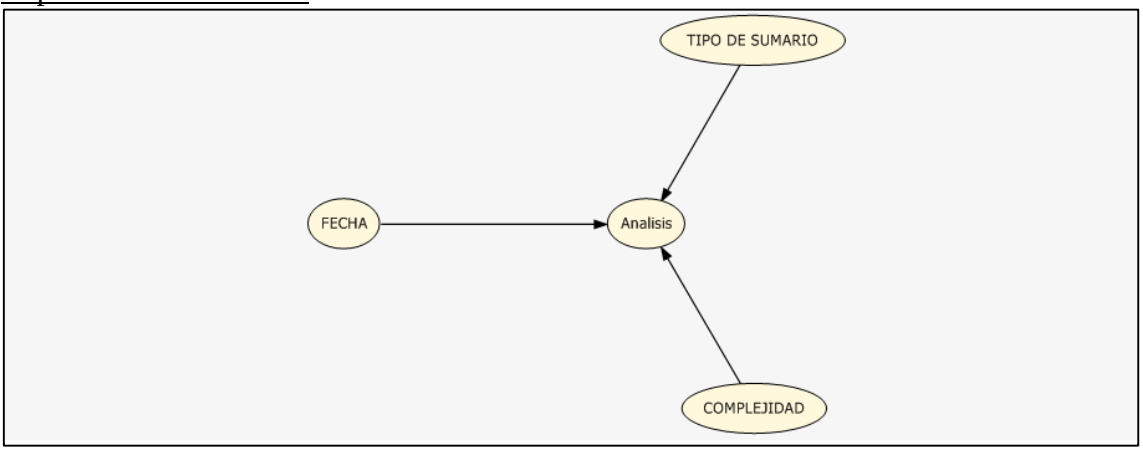


Fig. 4.4-19: Red de dependencias Árbol de Decisión

La figura 4-19 muestra, al igual que en el modelo Bayes Naïve que el nodo Análisis es predicho por los atributos Complejidad y Tipo de Sumario, pero también se le agrega el atributo Fecha, el cual no es omitido en el modelo Bayes por tratarse de un dato continuo.

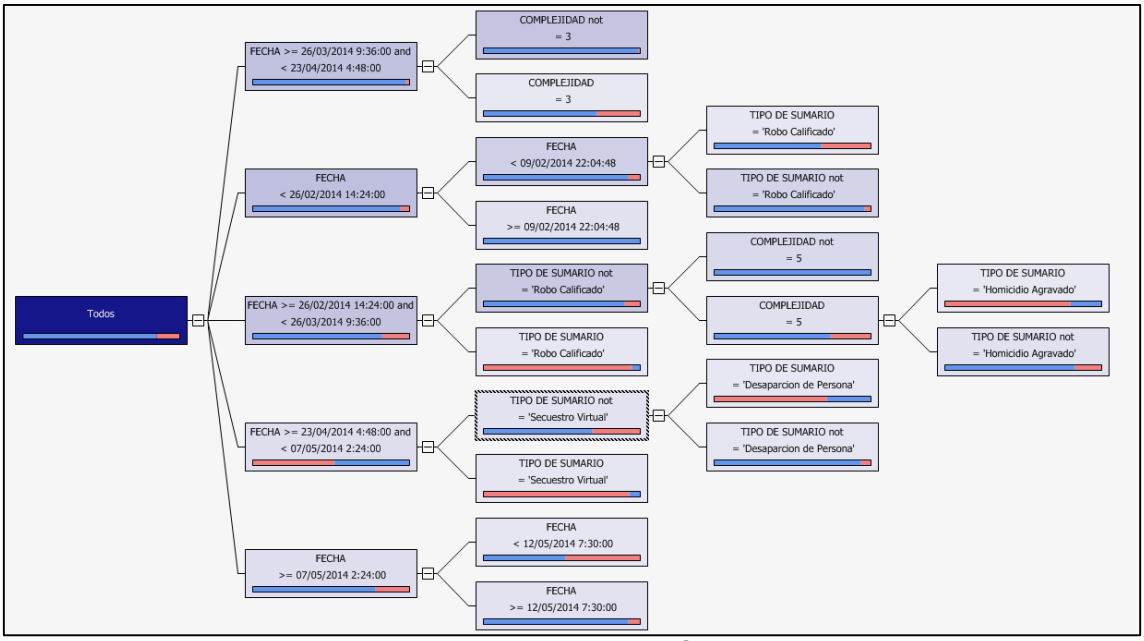


Fig. 4.4-20: Esquema de visualización Árbol de decisión

La figura 4-20 muestra que se han logrado ramas cuya pureza es importante, es decir las hojas de algunos recorridos del árbol logran predecir aparentemente bien un valor del atributo target.

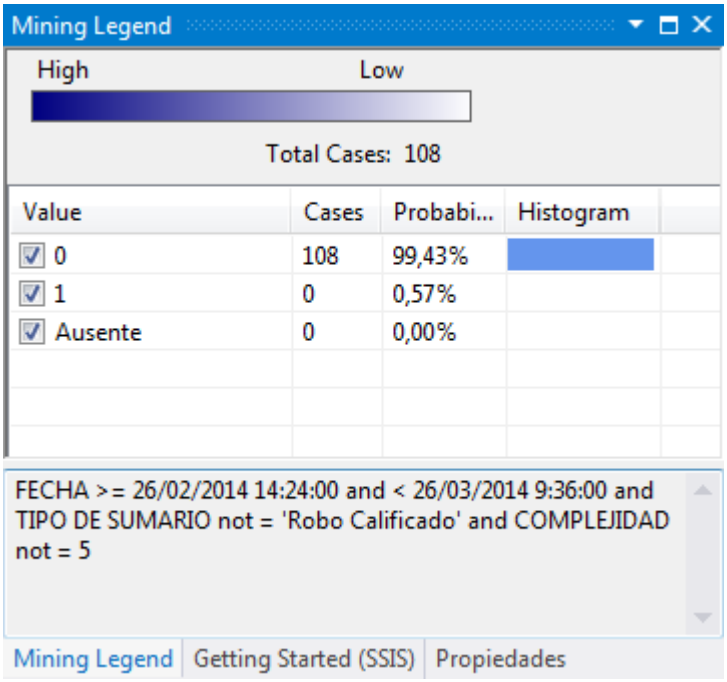


Fig. 4.4-21: Leyenda visualización Árbol de Decisión

La figura 4-21 describe una regla de asociación que concluye con un 99,43% de probabilidad de que en el caso de que suceda esa secuencia de eventos, habría Oficios a los cuales realizarle Análisis en 108 casos aproximadamente.

4.4.3.3. Armado de Clusteres

Parámetros

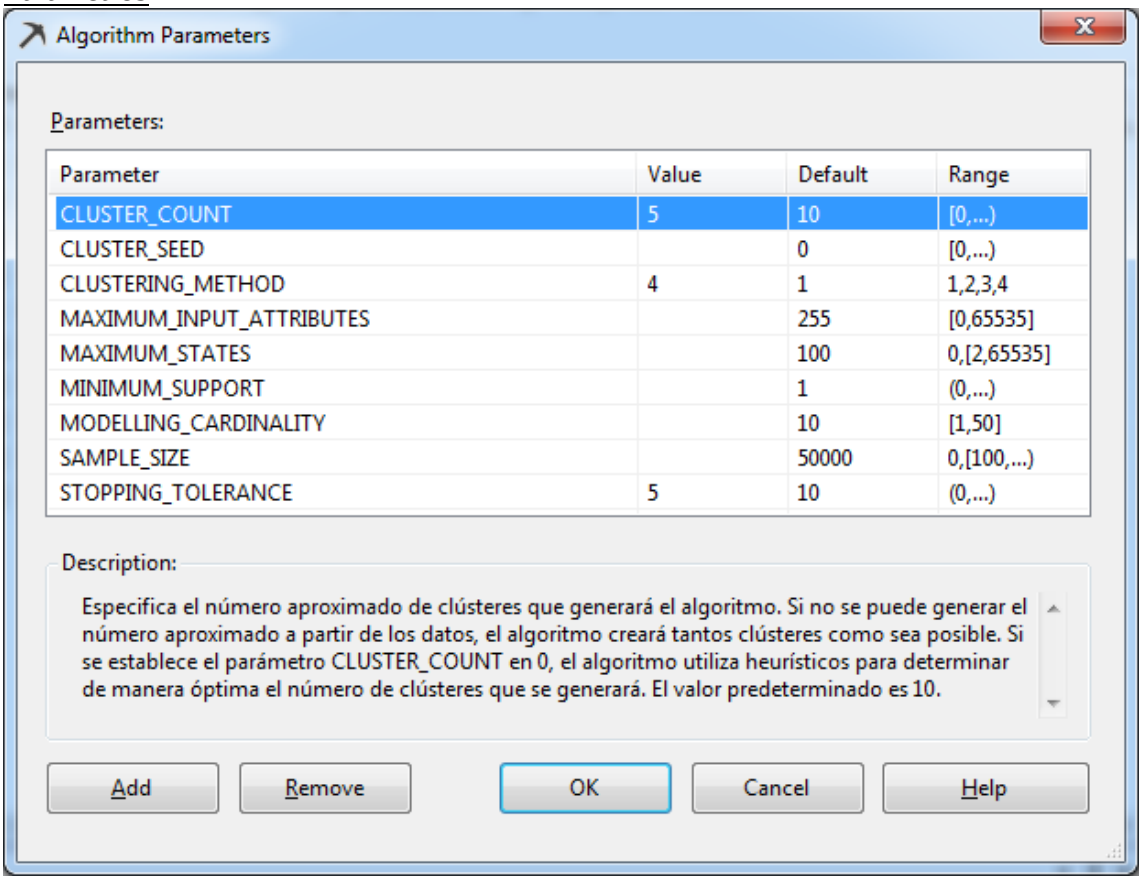


Fig. 4.4-22: Parámetros Clusterización

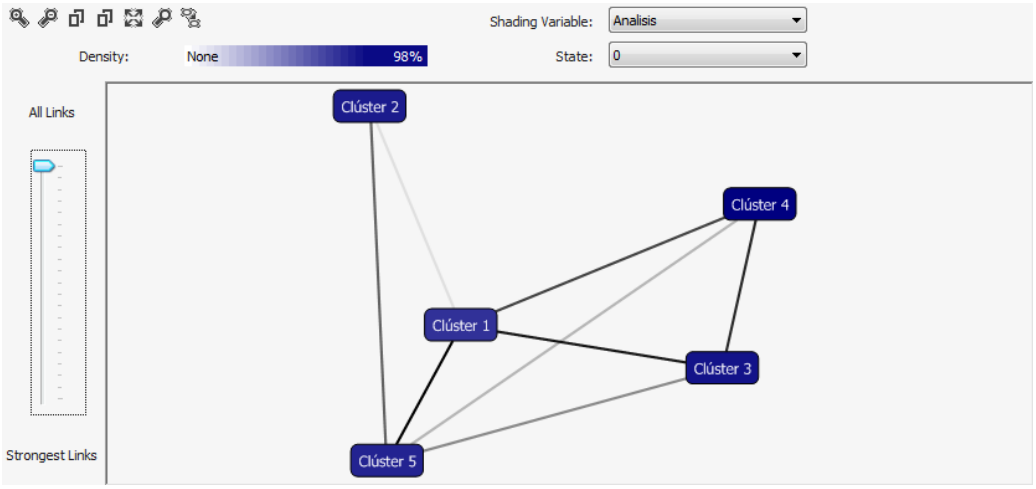


Fig. 4.4-23: Diagrama de Clústeres

En la figura 4-23 se observa que las relaciones directas de los clústers.

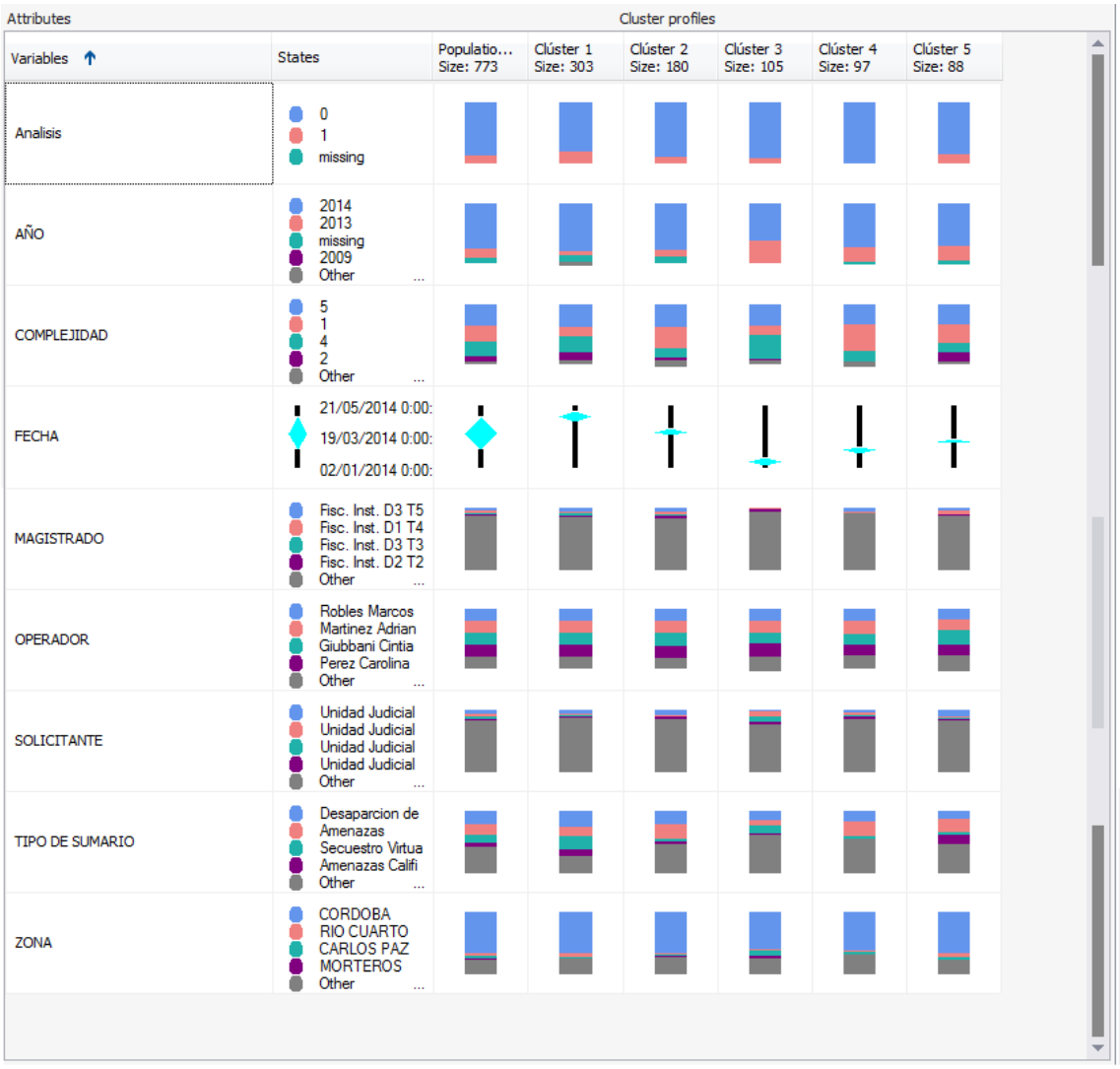


Fig. 4.4-24: Visualización Perfiles del Clúster

En la figura 4-24 aparecen los 5 clústeres obtenidos con las características de cada uno de ellos de manera gráfica. Esto permite observar por ejemplo que el clúster 1 puede definirse por un mayor porcentaje de causas cuya complejidad es alta. Estas particularidades, se evidenciarán y analizarán aún más en el momento de sacar conclusiones.

A modo de ejemplo la tabla 4.15 muestra las características de los casos que pertenecen al Clúster 1 mostrando los atributos de manera decreciente según los valores de probabilidad de ocurrencia.

Variables	Values	Probability
AÑO	2014	80,841%
FECHA	22/02/2014 15:06:51 - 19/03/2014 16:54:55	80,135%
ZONA	CORDOBA	72,430%
Analisis	0	71,028%
COMPLEJIDAD	1	37,850%
COMPLEJIDAD	5	32,243%
Analisis	1	28,972%
TIPO DE SUMARIO	Amenazas	24,299%
OPERADOR	Giubbani Cintia	22,430%
OPERADOR	Martinez Adrian	20,093%
OPERADOR	Robles Marcos	19,626%
OPERADOR	Perez Carolina	19,159%
AÑO	2013	19,159%
OPERADOR	Fernandez Sebast	18,692%
COMPLEJIDAD	4	16,355%
TIPO DE SUMARIO	Desaparcion de Persona	15,421%
FECHA	19/03/2014 16:54:55 - 13/04/2014 18:43:00	14,922%
TIPO DE SUMARIO	Desobediencia a la Autoridad	13,551%
SOLICITANTE	Unidad Judicial Delitos Economicos	10,748%
TIPO DE SUMARIO	Homicidio	9,813%
TIPO DE SUMARIO	Amenazas Calificadas	9,346%
COMPLEJIDAD	2	9,346%
TIPO DE SUMARIO	Robo Calificado	8,879%
MAGISTRADO	Fisc. Inst. D3 T5	7,009%
ZONA	RIO CUARTO	5,607%
SOLICITANTE	Unidad Judicial 18	5,140%
MAGISTRADO	Fisc. Inst. D1 T4	5,140%
FECHA	02/01/2014 0:00:00 - 22/02/2014 15:06:51	4,934%
SOLICITANTE	Unidad Judicial 7	4,673%
TIPO DE SUMARIO	Homicidio Agravado	4,206%
SOLICITANTE	Unidad Judicial 2	4,206%
SOLICITANTE	missing	4,206%
MAGISTRADO	Fisc. Inst. D2 T2	4,206%
MAGISTRADO	Fisc. Inst. 1 Nom Rio IV	4,206%
COMPLEJIDAD	3	4,206%
TIPO DE SUMARIO	Secuestro Virtual	4,206%
MAGISTRADO	Fisc. Inst. D4 T5	3,738%
ZONA	CARLOS PAZ	3,738%
MAGISTRADO	Fisc. Inst. D3 T3	3,738%
MAGISTRADO	Fisc. Inst. D2 T4	3,738%
MAGISTRADO	Fisc. Inst. D2 T6	3,738%
SOLICITANTE	Unidad Judicial Homicidios	3,738%
SOLICITANTE	Unidad Judicial 15	3,271%
SOLICITANTE	Unidad Judicial 5	3,271%
SOLICITANTE	Unidad Judicial 6	3,271%
SOLICITANTE	Unidad Judicial 14	3,271%
TIPO DE SUMARIO	Violencia Familiar	3,271%
SOLICITANTE	Unidad Judicial Carlos Paz	3,271%

MAGISTRADO	Fisc. Inst. D1 T2	2,804%
MAGISTRADO	Fisc. Inst. D4 T3	2,804%
TIPO DE SUMARIO	Estafa	2,804%
SOLICITANTE	Unidad Judicial Rio IV	2,804%
SOLICITANTE	Unidad Judicial 2 Rio IV	2,804%
MAGISTRADO	Fisc. Inst. D4 T4	2,804%
MAGISTRADO	Fisc. Inst. 1 T. Carlos Paz	2,804%
MAGISTRADO	Fisc. Inst. D1 T3	2,804%
MAGISTRADO	Fisc. Inst. D2 T1	2,804%
MAGISTRADO	Fisc. Inst. D2 T3	2,804%
MAGISTRADO	Fisc. Inst. D1 T5	2,804%
MAGISTRADO	Fisc. Inst. D3 T7	2,336%
MAGISTRADO	Fisc. Inst. Jesus María	2,336%
MAGISTRADO	Fisc. Inst. D4 T6	2,336%
ZONA	VILLA DOLORES	2,336%
MAGISTRADO	Fisc. Inst. D1 T6	2,336%
MAGISTRADO	Fisc. Inst. D3 T2	2,336%
MAGISTRADO	Fisc. Inst. D3 T4	2,336%
MAGISTRADO	Fisc. Inst. D3 T1	2,336%
MAGISTRADO	Fisc. Inst. D2 T5	2,336%
TIPO DE SUMARIO	MED	2,336%
SOLICITANTE	Informatica	2,336%
MAGISTRADO	missing	1,869%
ZONA	DEAN FUNES	1,869%
SOLICITANTE	Unidad Judicial Robos y Hurtos	1,869%
ZONA	VILLA ALLENDE	1,869%
SOLICITANTE	Unidad Judicial 12	1,869%
SOLICITANTE	Unidad Judicial Jesus Maria	1,869%
SOLICITANTE	Unidad Judicial 10	1,869%
MAGISTRADO	Fisc. Inst. Cosquin	1,869%
MAGISTRADO	Fisc. Inst. D4 T1	1,869%
MAGISTRADO	Fisc. Inst. Dean Funes	1,869%
SOLICITANTE	Unidad Judicial Villa Allende	1,869%
ZONA	JESUS MARIA	1,869%
SOLICITANTE	Unidad Judicial 13	1,402%
SOLICITANTE	Unidad Judicial Violencia Familiar	1,402%
ZONA	COSQUIN	1,402%
SOLICITANTE	Unidad Judicial 3	1,402%
SOLICITANTE	Unidad Judicial 21	1,402%
SOLICITANTE	Unidad Judicial Dean Funes	1,402%
TIPO DE SUMARIO	Extorsión	1,402%
SOLICITANTE	Fisc. Inst. D2 T2	1,402%
SOLICITANTE	Unidad Judicial 11	1,402%
SOLICITANTE	Unidad Judicial de la Mujer y el niño	1,402%
MAGISTRADO	Fisc. Inst. D4 T2	1,402%
MAGISTRADO	Fisc. Inst. Alta Gracia	1,402%
MAGISTRADO	Fisc. Inst. 2 T. Carlos Paz	1,402%
SOLICITANTE	Unidad Judicial 22	1,402%
SOLICITANTE	Unidad Judicial Laboulaye	0,935%

MAGISTRADO	Fisc. Inst. Morteros	0,935%
SOLICITANTE	Fisc. Inst. D3 T5	0,935%
MAGISTRADO	Fisc. Inst. Laboulaye	0,935%
SOLICITANTE	Comisaria Santa Maria	0,935%
ZONA	ALTA GRACIA	0,935%
SOLICITANTE	Unidad Judicial Morteros	0,935%
SOLICITANTE	Unidad Judicial Sustracción de Automotor	0,935%
SOLICITANTE	Unidad Judicial 19	0,935%
MAGISTRADO	Fisc. Inst. 3 Nom Rio IV	0,935%
SOLICITANTE	Unidad Judicial 8	0,935%
MAGISTRADO	Fisc. Inst. 2 T. Villa Dolores	0,935%
SOLICITANTE	Unidad Judicial Alta Gracia	0,935%
SOLICITANTE	Unidad Judicial Villa Dolores	0,935%
SOLICITANTE	Unidad Judicial Rio II	0,935%
MAGISTRADO	Fisc. Inst. 4 Nom Rio IV	0,935%
MAGISTRADO	Fisc. Inst. Rio II	0,935%
ZONA	MORTEROS	0,935%
MAGISTRADO	Camara Criminal y Correccional Villa Dolores	0,935%
SOLICITANTE	Camara Criminal y Correccional Villa Dolores	0,935%
ZONA	RIO SEGUNDO	0,935%
ZONA	LABOULAYE	0,935%
SOLICITANTE	Unidad Judicial 20	0,935%

Tabla 4.4.2: Características de Clúster 1

4.4.3.4. Construcción reglas de Asociación

Parámetros

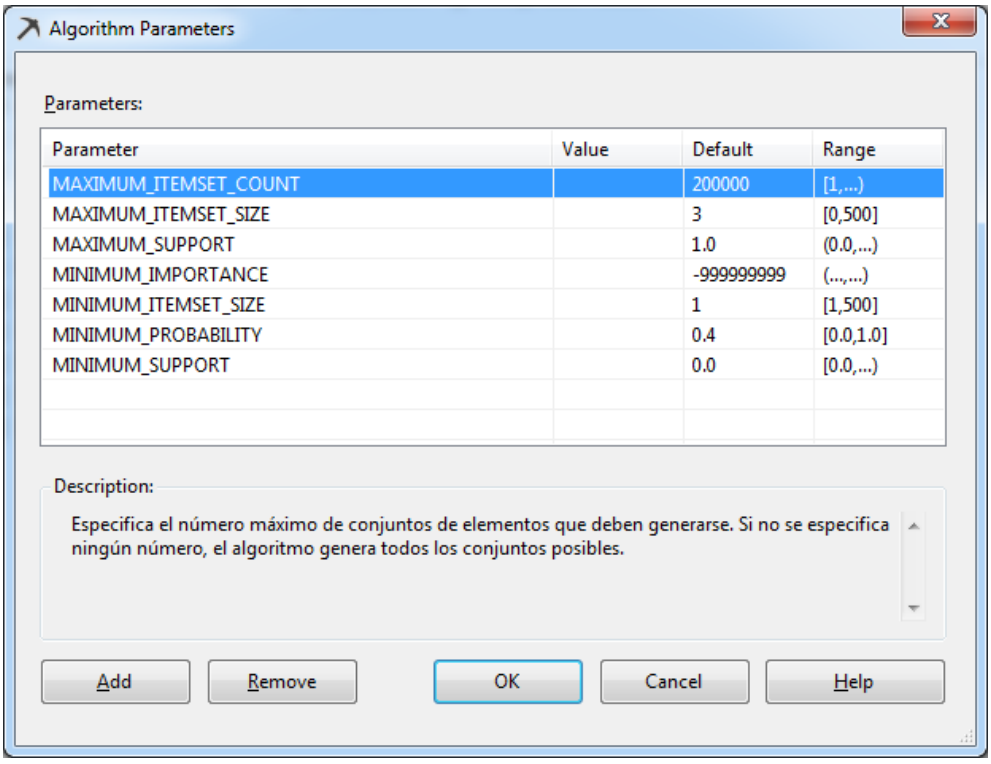


Fig. 4.4-3: Parámetros Reglas de Asociación

Interpretación del modelo

Support	S.	Itemset
298	3	ZONA = CORDOBA, Analisis = 0, AÑO = 2014
175	3	COMPLEJIDAD = 1, Analisis = 0, AÑO = 2014
155	3	COMPLEJIDAD = 5, ZONA = CORDOBA, AÑO = 2014
148	3	TIPO DE SUMARIO = Desaparcion de Persona, COMPLEJIDAD = 5, AÑO = 2014
146	3	TIPO DE SUMARIO = Amenazas, COMPLEJIDAD = 1, Analisis = 0
145	3	TIPO DE SUMARIO = Desaparcion de Persona, COMPLEJIDAD = 5, Analisis = 0
137	3	Analisis = 1, ZONA = CORDOBA, AÑO = 2014
135	3	COMPLEJIDAD = 5, Analisis = 0, AÑO = 2014
135	3	COMPLEJIDAD = 1, ZONA = CORDOBA, Analisis = 0
129	3	TIPO DE SUMARIO = Desaparcion de Persona, Analisis = 0, AÑO = 2014
124	3	TIPO DE SUMARIO = Amenazas, Analisis = 0, AÑO = 2014
124	3	TIPO DE SUMARIO = Amenazas, COMPLEJIDAD = 1, AÑO = 2014
119	3	COMPLEJIDAD = 1, ZONA = CORDOBA, AÑO = 2014
109	3	TIPO DE SUMARIO = Desaparcion de Persona, COMPLEJIDAD = 5, ZONA = CORDOBA
100	3	TIPO DE SUMARIO = Amenazas, ZONA = CORDOBA, Analisis = 0
100	3	TIPO DE SUMARIO = Amenazas, COMPLEJIDAD = 1, ZONA = CORDOBA
98	3	COMPLEJIDAD = 4, ZONA = CORDOBA, AÑO = 2014

Itemsets: 1216

Fig. 4.4-4: Itemsets Reglas de Asociación

La figura 4-27 muestra el resultado de la ejecución de la primera fase del algoritmo, define los itemsets hallados. En este caso en particular no fue necesario filtrar por una cantidad específica de casos debido al objetivo del problema a resolver y a la cantidad de registros disponibles.

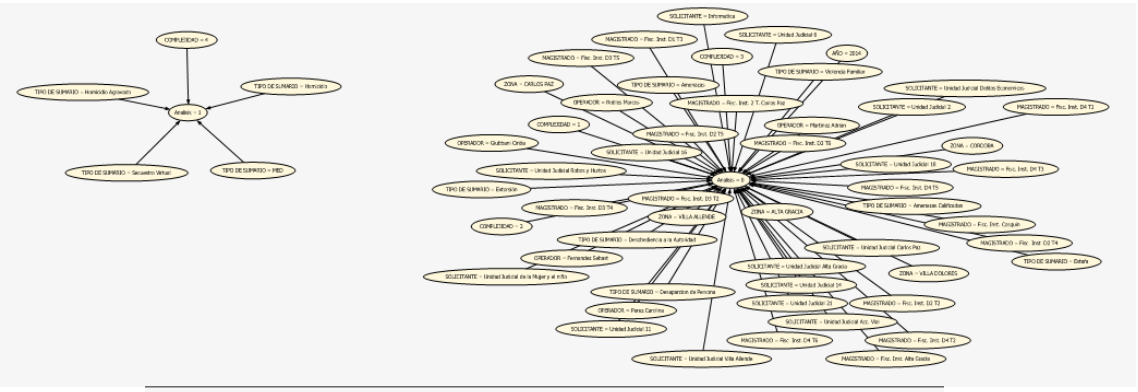
Probability	Importance	Rule
0,912	0,597	TIPO DE SUMARIO = Secuestro Virtual -> Analisis = 1
0,912	0,597	TIPO DE SUMARIO = Secuestro Virtual, COMPLEJIDAD = 4 -> Analisis = 1
0,895	0,555	TIPO DE SUMARIO = Secuestro Virtual, AÑO = 2014 -> Analisis = 1
0,943	0,531	TIPO DE SUMARIO = Homicidio, AÑO = 2014 -> Analisis = 1
0,908	0,529	TIPO DE SUMARIO = Homicidio -> Analisis = 1
0,908	0,529	TIPO DE SUMARIO = Homicidio, COMPLEJIDAD = 5 -> Analisis = 1
0,891	0,517	TIPO DE SUMARIO = Secuestro Virtual, ZONA = CORDOBA -> Analisis = 1
0,935	0,515	TIPO DE SUMARIO = Homicidio, ZONA = CORDOBA -> Analisis = 1
1,000	0,510	TIPO DE SUMARIO = Homicidio Agravado -> Analisis = 1
1,000	0,510	TIPO DE SUMARIO = Homicidio Agravado, COMPLEJIDAD = 5 -> Analisis = 1
1,000	0,499	TIPO DE SUMARIO = Secuestro Virtual, OPERADOR = Martinez Adrian -> Analisis = 1
1,000	0,492	TIPO DE SUMARIO = Secuestro Virtual, AÑO = 2013 -> Analisis = 1
1,000	0,492	TIPO DE SUMARIO = Homicidio Agravado, AÑO = 2014 -> Analisis = 1
1,000	0,486	TIPO DE SUMARIO = Homicidio, OPERADOR = Giubbani Cintia -> Analisis = 1
1,000	0,486	TIPO DE SUMARIO = Homicidio Agravado, ZONA = CORDOBA -> Analisis = 1
0,950	0,479	TIPO DE SUMARIO = Secuestro Virtual, OPERADOR = Perez Carolina -> Analisis = 1
1,000	0,471	TIPO DE SUMARIO = Homicidio Agravado, OPERADOR = Perez Carolina -> Analisis = 1

Rules: 196

Fig. 4.4-5: Reglas del algoritmo Reglas de Asociación

En la figura 4-28 se examinan las reglas encontradas teniendo en cuenta la puntuación conseguida en la probabilidad e importancia de la misma. La importancia podría ser considerada como una medida de la utilidad de la regla; a mayor importancia mayor calidad de la misma. La máxima probabilidad de

ocurrencia no garantiza utilidad, para ello es necesario tener en cuenta ambos parámetros para obtener el verdadero valor de la regla.



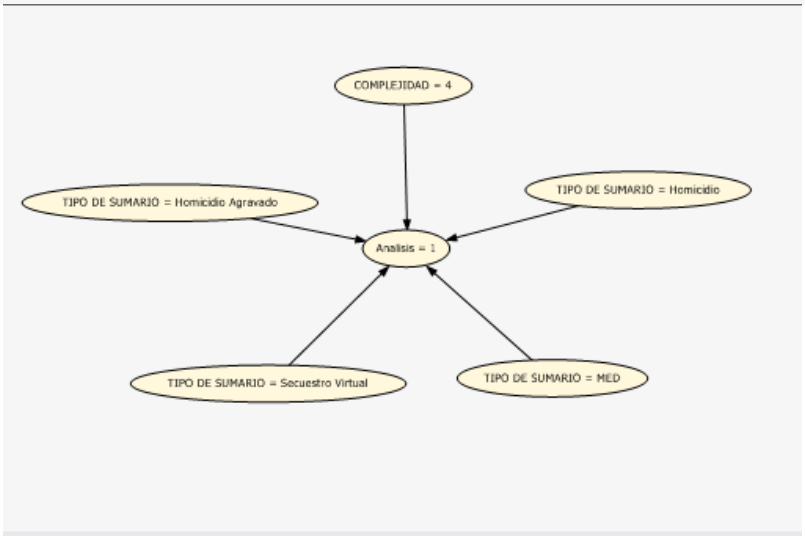


Fig. 4.4-6: Red de Dependencias Reglas de Asociación

En la figura 4-29, se puede observar que el modelo encuentra una relación ya verificada previamente por el algoritmo Bayes Naïve, y es que en las causas que llevan análisis son aquellas cuya tipo de causa son homicidios, muertes de etiología dudosa, secuestros virtuales, las cuales tiene en común complejidad entre 4 y 5. Estas peculiaridades darán validez a las conclusiones extraídas dado que se repiten más allá del algoritmo seleccionado.

4.4.3.5. Construcción Red Neuronal

Parámetros

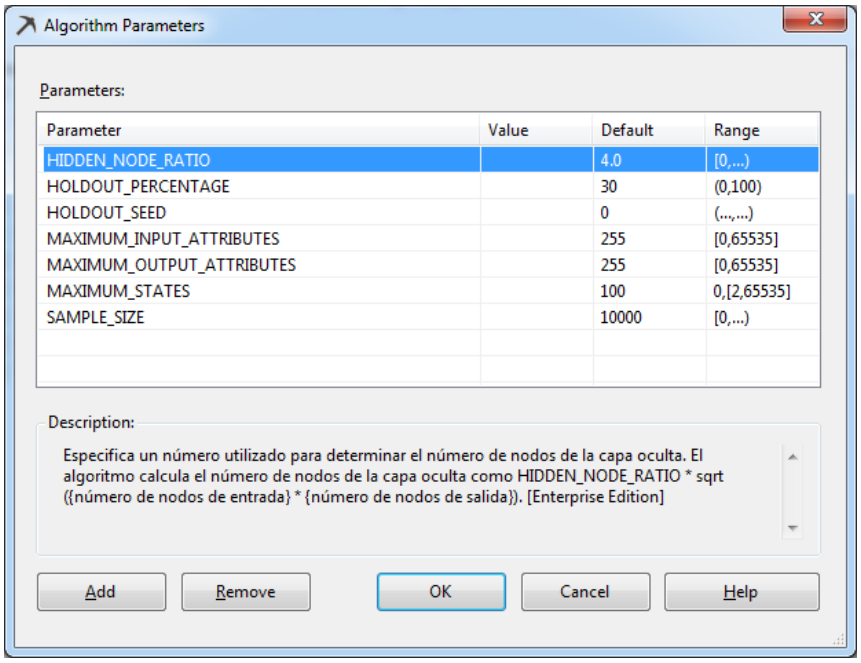


Fig. 4.4-7: Parámetros Red Neuronal

Interpretación del modelo

Al observar el modelo de Red Neuronal, podemos darnos cuenta de que difiere de los algoritmos anteriormente expuestos. En dicho modelo, se visualiza el impacto de los atributos de entrada respecto el atributo objetivo, es vez de mostrar la distribución de la red de dependencias. La tabla se ordena según la puntuación calculada de manera probabilística y se refleja en la figura 4-13.

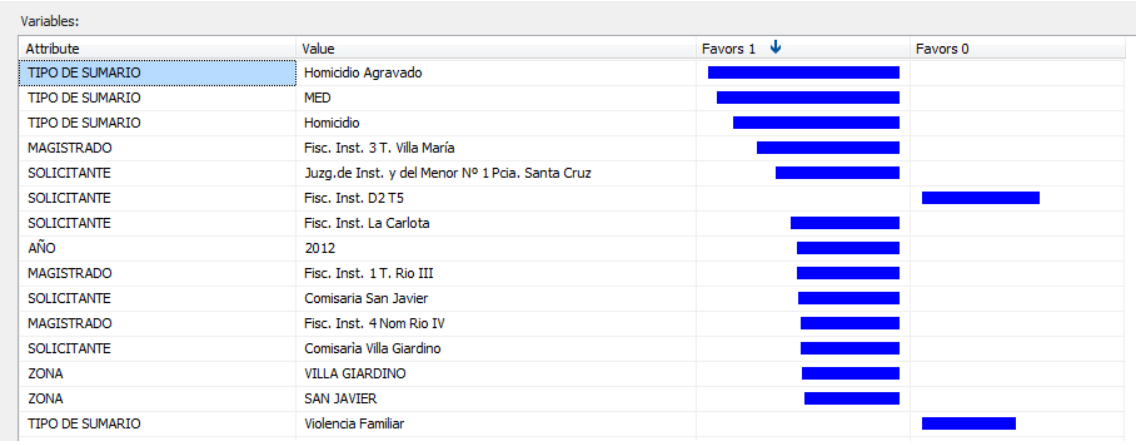


Fig. 4.4-8: Visor del modelo Red Neuronal

En la tabla 4.17 se muestran las puntuaciones obtenidas al comparar las causas que tienen solicitud de análisis (1), con aquellas que aún no lo tienen (0). Se verifica

visiblemente que cuando el tipo de Sumario es de complejidad alta (homicidio agravado), este atributo predice con una puntuación de 100 que las causas sean de ese tipo sean analizada. Si bien estas particularidades podrían parecer obvias confirman el buen funcionamiento del modelo.

Attribute	Value	Favors 1	Favors 0
TIPO DE SUMARIO	Homicidio Agravado	100	
TIPO DE SUMARIO	MED	95,5	
TIPO DE SUMARIO	Homicidio	87,14	
MAGISTRADO	Fisc. Inst. 3 T. Villa María	74,86	
SOLICITANTE	Juzg.de Inst. y del Menor Nº 1 Pcia. Santa Cruz	65,19	
SOLICITANTE	Fisc. Inst. D2 T5		61,55
SOLICITANTE	Fisc. Inst. La Carlota	57,45	
AÑO	2012	53,9	
MAGISTRADO	Fisc. Inst. 1 T. Rio III	53,87	
SOLICITANTE	Comisaria San Javier	53,04	
MAGISTRADO	Fisc. Inst. 4 Nom Rio IV	51,94	
SOLICITANTE	Comisarìa Villa Giardino	51,92	
ZONA	VILLA GIARDINO	51,43	
ZONA	SAN JAVIER	50	
TIPO DE SUMARIO	Violencia Familiar		49,37
ZONA	LA FALDA	49,13	
MAGISTRADO	Juzg.de Inst. y del Menor Nº 1 Pcia. Santa Cruz	48,58	
SOLICITANTE	Fisc. Inst. D2 T3	48,31	
SOLICITANTE	Fisc. Inst. D1 T3		48,03
TIPO DE SUMARIO	Secuestro Virtual	47,3	
MAGISTRADO	Juzg. Men. 4 Nom. Corr.3	47,12	
MAGISTRADO	Fisc. Inst. D4 T3		46,89
MAGISTRADO	Camara Criminal y Correccional Villa Dolores		45,69
MAGISTRADO	Fisc. Inst. D3 T2		43,18
SOLICITANTE	Unidad Judicial 20		42,71
SOLICITANTE	Unidad Judicial 5	42,3	
SOLICITANTE	Fisc. Inst. Dean Funes	41,51	
MAGISTRADO	Fisc. Inst. Oliva		41,06
MAGISTRADO	Camara 4 Crimen	40,99	
SOLICITANTE	Juzgado de Niñez Carlos Paz	40,86	
ZONA	VALLE HERMOSO	39,66	
COMPLEJIDAD	1		39,35
SOLICITANTE	Unidad Judicial 13	39,05	
SOLICITANTE	Unidad Judicial 4	38,32	
ZONA	RIO SEGUNDO		37,81
SOLICITANTE	Unidad Judicial La Calera	36,67	
MAGISTRADO	Juzg. Men. Y Faltas Cosquin	36,64	
SOLICITANTE	Unidad Judicial Villa Allende		36,46
TIPO DE SUMARIO	Desaparcion de Persona		36,29
SOLICITANTE	Camara 4 Crimen	35,91	
SOLICITANTE	Fisc. Inst. 3 T. Villa María	35,87	
MAGISTRADO	Fisc. Inst. D3 T7		35,18

TIPO DE SUMARIO	Desobediencia a la Autoridad		34,64
MAGISTRADO	Fisc. Inst. D2 T6	34,23	
SOLICITANTE	Unidad Judicial Rio II		33,44
SOLICITANTE	Comisaria Distrito Rio I		32,83
ZONA	SALDAN	32,01	
MAGISTRADO	Juzg. de Control 3	31,85	
TIPO DE SUMARIO	Amenazas		31,21
MAGISTRADO	Fisc.Inst. La Falda		31,15
SOLICITANTE	Fisc. Inst. D1 T5	30,9	
SOLICITANTE	Fisc. Inst. D3 T4	30,83	
ZONA	RIO TERCERO	30,72	
MAGISTRADO	Fisc. Inst. Cruz del Eje		30,49
MAGISTRADO	Fisc. Inst. 2 T. Villa Dolores		30,09
SOLICITANTE	Unidad Judicial Dean Funes		29,3
TIPO DE SUMARIO	Extorsión		29,07
COMPLEJIDAD	2		27,99
ZONA	VILLA ROSSI	27,84	
MAGISTRADO	Camara 9 Crimen		27,43
SOLICITANTE	Fisc. Inst. Jesus María		27,42
ZONA	MONTE MAIZ		26,26
MAGISTRADO	Fisc. Inst. Laboulaye	26,22	
SOLICITANTE	Unidad Judicial Morteros	26,12	
SOLICITANTE	Comisaria Valle Hermoso	25,71	
SOLICITANTE	Unidad Judicial 16		25,67
MAGISTRADO	Fisc. Inst. Jesus María	25,32	
MAGISTRADO	Fisc. Inst. La Carlota	25,22	
SOLICITANTE	Fisc. Inst. D4 T4	25,19	
MAGISTRADO	Fisc. Inst. D2 T1	24,95	
SOLICITANTE	Camara 1 Río IV	24,57	
ZONA	LA CALERA	24,32	
MAGISTRADO	Juzg. Men. 6 Nom. Corr. 5		24,28
COMPLEJIDAD	4	24,24	
ZONA	RIO PRIMERO		24,1
MAGISTRADO	Fisc. Inst. D3 T1	23,88	
SOLICITANTE	Subcomisaria Anisacate		23,42
MAGISTRADO	Fisc. Inst. Cura Brochero		23,41
MAGISTRADO	Juzg. De control en lo Penal y Ec.		23,33
MAGISTRADO	Fisc. Inst. 1 T. Carlos Paz	23,2	
MAGISTRADO	Fisc. Inst. D3 T3	23,02	
COMPLEJIDAD	5	22,9	
SOLICITANTE	Unidad Judicial Cosquin	22,7	
ZONA	ANISACATE	22,17	
SOLICITANTE	Unidad Judicial Acc. Vial		21,92
MAGISTRADO	Unidad Judicial Violencia Familiar	21,83	
TIPO DE SUMARIO	Abuso Sexual con Acceso Carnal	21,76	
SOLICITANTE	Unidad Judicial 11		21,69
SOLICITANTE	Fisc. Inst. D3 T5		21,69
MAGISTRADO	Fisc. Inst. D2 T3	21,53	
ZONA	ALEJANDRO ROCA		21,44

SOLICITANTE	Unidad Judicial 2 Rio IV		21,19
SOLICITANTE	Unidad Regional Rio I	20,9	
MAGISTRADO	Fisc. Inst. Morteros	20,81	
SOLICITANTE	Unidad Judicial 3		20,74
ZONA	FALDA DEL CARMEN	20,64	
SOLICITANTE	Unidad Judicial 22	20,53	
MAGISTRADO	Fisc. Inst. D4 T4	20,23	
TIPO DE SUMARIO	Amenazas Calificadas		20,2
MAGISTRADO	Fisc. Inst. Bell Ville		19,92
ZONA	VILLA ALLENDE		19,81
TIPO DE SUMARIO	Robo Calificado	19,7	
SOLICITANTE	Fisc. Inst. Rio II		19,59
SOLICITANTE	Unidad Judicial Violencia Familiar	19,58	
MAGISTRADO	Fisc. Inst. Men. 1 Turno		19,13
ZONA	MALAGUEÑO		18,95
MAGISTRADO	Fisc. Inst. D3 T4	18,86	
SOLICITANTE	Unidad Judicial La Falda	18,74	
SOLICITANTE	Comsiaria La Cumbre	18,55	
SOLICITANTE	Comisaria Capilla del Monte		17,92
MAGISTRADO	Fisc. Inst. 2 Nom Rio IV	17,62	
SOLICITANTE	Unidad Judicial 10		17,55
SOLICITANTE	Unidad Judicial Cruz del Eje	17,49	
SOLICITANTE	Unidad Judicial Laboulaye	16,88	
TIPO DE SUMARIO	Estafa	16,88	
SOLICITANTE	Unidad Judicial 8	16,88	
SOLICITANTE	Fisc. Inst. Bell Ville	16,84	
SOLICITANTE	Unidad Judicial 18	16,22	
ZONA	CAPILLA DEL MONTE		16,13
MAGISTRADO	Juzg. Inst. Concaran	15,97	
ZONA	LA CARLOTA	15,7	
ZONA	MALVINAS ARGENTINAS		15,69
SOLICITANTE	Unidad Judicial Delitos Economicos	15,56	
MAGISTRADO	Fisc. Inst. 2 T. Rio III	15,21	
MAGISTRADO	Juzg. Juvenil 7 Correcc. 8	15,2	
MAGISTRADO	Fisc. Inst. D3 T5		15,1
MAGISTRADO	Fisc. Inst. Corral de Bustos		14,59
MAGISTRADO	Camara 1 Río IV		14,59
ZONA	CARLOS PAZ		14,5
SOLICITANTE	Fisc. Inst. 2 T. Villa Dolores	14,42	
MAGISTRADO	Fisc. Inst. D1 T1		14,24
SOLICITANTE	Informatica		14,22
MAGISTRADO	Fisc. Inst. D4 T6		14,14
SOLICITANTE	Unidad Judicial San Francisco	14,1	
ZONA	COLONIA TIROLESA	14,01	
SOLICITANTE	Unidad Judicial Sustracción de Automotor	13,99	
SOLICITANTE	Comisaria Colonia Tirolesa		13,47
SOLICITANTE	Unidad Judicial 9	12,92	
MAGISTRADO	Fisc. Inst. 1 Nom Rio IV	12,77	
SOLICITANTE	Unidad Judicial Robos y Hurtos	12,56	

SOLICITANTE	Fisc. Inst. 1 Nom Rio IV	12,55	
SOLICITANTE	Unidad Judicial 2		12,12
SOLICITANTE	Unidad Judicial 12	11,88	
ZONA	SAN FRANCISCO	11,3	
SOLICITANTE	Fisc. Inst. D4 T2		11,25
SOLICITANTE	Unidad Judicial 21		11,2
MAGISTRADO	Fisc. Inst. D4 T2		10,91
MAGISTRADO	Fisc. Inst. Las Varillas	10,87	
SOLICITANTE	Unidad Judicial 19	10,81	
ZONA	BELL VILLE	10,78	
ZONA	COLONIA CAROYA		10,77
MAGISTRADO	Fisc. Inst. D1 T4	10,63	
SOLICITANTE	Comisaria Santa Maria	10,52	
MAGISTRADO	Fisc. Inst. D2 T5	10,45	
SOLICITANTE	Unidad Judicial Homicidios	10,28	
SOLICITANTE	Fisc. Inst. D3 T6		10,24
MAGISTRADO	Fisc. Inst. D1 T3		10,17
SOLICITANTE	Unidad Judicial Jesus Maria	10,08	
MAGISTRADO	Fisc. Inst. D2 T4		10,03
TIPO DE SUMARIO	Encubrimiento	9,99	
MAGISTRADO	Fisc. Inst. D4 T5		9,95
SOLICITANTE	Comisaria Alejandro Roca	9,42	
SOLICITANTE	Unidad Judicial 7		9,15
ZONA	ALTA GRACIA	8,83	
AÑO	2013	8,69	
SOLICITANTE	Unidad Judicial Alta Gracia	8,62	
MAGISTRADO	Fisc. Inst. D1 T6		8,5
ZONA	BIALET MASSE		8,36
MAGISTRADO	Juzg.control men.y faltas La Carlota		7,87
SOLICITANTE	Unidad Judicial Villa Dolores		7,78
ZONA	DEAN FUNES	7,72	
SOLICITANTE	Unidad Judicial Rio IV		7,71
MAGISTRADO	Fisc. Inst. D1 T2	7,49	
MAGISTRADO	Fisc. Inst. Cosquin		7,47
ZONA	CONCARAN		7,45
MAGISTRADO	Fisc. Inst. D4 T1	7,17	
MAGISTRADO	Fisc. Inst. San Francisco	7,13	
MAGISTRADO	Fisc. Inst. Rio II	7,08	
MAGISTRADO	Fisc. Inst. 2 T. Carlos Paz	6,98	
SOLICITANTE	Fisc. Inst. Morteros		6,79
SOLICITANTE	Unidad Judicial Carlos Paz		6,73
SOLICITANTE	Fisc. Inst. D1 T1	6,72	
SOLICITANTE	Unidad Judicial 17		6,69
SOLICITANTE	Fisc. Inst. Cura Brochero		6,43
SOLICITANTE	Fisc. Inst. San Francisco	6,23	
ZONA	LABOULAYE	6,06	
SOLICITANTE	Unidad Judicial 14		5,94
OPERADOR	Fernandez Sebast	5,93	
ZONA	COSQUIN		5,66

SOLICITANTE	Fisc. Inst. D3 T3	5,55	
MAGISTRADO	Fisc. Inst. 1 T. Villa Dolores		5,42
SOLICITANTE	Unidad Judicial 6	5,37	
MAGISTRADO	Unidad Judicial Alta Gracia	5,21	
MAGISTRADO	Fisc. Inst. D1 T5		5,19
SOLICITANTE	Fisc. Inst. D2 T2		5
MAGISTRADO	Fisc. Inst. 3 Nom Rio IV		4,98
SOLICITANTE	Comisaria Santa Rosa		4,79
ZONA	VILLA MARÍA		4,76
SOLICITANTE	Fisc. Inst. D1 T2	4,56	
SOLICITANTE	Fisc. Inst. D2 T1	4,35	
ZONA	VILLA DOLORES	4,34	
SOLICITANTE	Fisc. Inst. 1 T. Villa Dolores		4,23
ZONA	SALSACATE		4,08
MAGISTRADO	Fisc. Inst. Alta Gracia	4,06	
SOLICITANTE	Fisc. Inst. 2 Nom Rio IV	4,04	
FECHA	02/01/2014 0:00:00 - 22/02/2014 20:28:35		3,87
MAGISTRADO	Fisc. Inst. Dean Funes		3,69
SOLICITANTE	CIC	3,62	
COMPLEJIDAD	3	3,6	
FECHA	14/04/2014 4:16:06 - 21/05/2014 0:00:00	3,49	
ZONA	CORDOBA	3,36	
SOLICITANTE	Fisc. Inst. D1 T6	3,21	
SOLICITANTE	Unidad Judicial 15		3,17
SOLICITANTE	Fisc. Inst. D4 T1	3,15	
SOLICITANTE	Comisaria Saldan	3,15	
SOLICITANTE	Unidad Judicial 1	2,94	
MAGISTRADO	Juzgado de Niñez Carlos Paz	2,68	
TIPO DE SUMARIO	Denuncia Formulada		2,57
OPERADOR	Martinez Adrian		2,55
SOLICITANTE	Fisc. Inst. D1 T4	2,52	
MAGISTRADO	Fisc. Inst. D2 T2		2,46
ZONA	ARIAS		2,35
AÑO	2014		1,64
SOLICITANTE	Fisc. Inst. Cruz del Eje		1,58
OPERADOR	Perez Carolina		1,45
SOLICITANTE	Comisaría Monte Maiz		1,32
ZONA	MORTEROS		1,15
ZONA	LA CUMBRE	1,05	
FECHA	20/03/2014 0:22:21 - 14/04/2014 4:16:06	1	
FECHA	22/02/2014 20:28:35 - 20/03/2014 0:22:21		0,98
SOLICITANTE	Fisc. Inst. Alta Gracia	0,89	
ZONA	CRUZ DEL EJE	0,8	
ZONA	CURA BROCHERO		0,78
OPERADOR	Robles Marcos		0,68
SOLICITANTE	Camara Criminal y Correccional Villa Dolores	0,56	
SOLICITANTE	Comisaria Colonia Caroya		0,45
SOLICITANTE	Unidad Judicial de la Mujer y el niño	0,38	
ZONA	RIO CUARTO	0,36	

ZONA	JESUS MARIA	0,36
OPERADOR	Giubbani Cintia	0,31
ZONA	SANTA ROSA	0,17

Tabla 4.4.3: Comparación Puntuación 1(Solicitud de Análisis) y 0 (No solicitado)

4.4.4. Valoración de los modelos

Para la valoración de los modelos, se tiene en cuenta el punto de vista de los expertos en el dominio del problema, los beneficios y las limitaciones de las técnicas de minería de datos.

Se ha podido llevar a cabo pruebas sobre cada uno de los modelos a través de las iteraciones realizadas. Además, gracias a las pruebas mencionadas, se ha logrado someter el conjunto de datos disponibles a distintos algoritmos, donde se pudo extraer información útil, tanto en conocimiento como en demostración cruzada de resultados de otros modelos. Durante la construcción y exploración, se ha podido descubrir características comunes en algunos modelos.

4.5. FASE DE EVALUACIÓN

Para comenzar con la evaluación, se debe tener en cuenta los criterios de éxito del proyecto de minería de datos. En general, se utiliza la exactitud de la clasificación o la tasa de error para medir el desempeño de un modelo de clasificación en el conjunto de pruebas.

Se han observado relaciones importantes descubiertas a través de la ejecución de los modelos. Además, será de suma importancia evaluar la puntuación de los modelos para de esta manera poder determinar la eficiencia y la eficacia de los mismos. Se puede concluir que el proyecto resulta positivo.

4.5.1. Resultados

Se procede a utilizar los métodos enumerados en el plan de pruebas a cada uno de los modelos generados para poder calificarlos y a su vez compararlos.

4.5.1.1. Gráfico de Elevación o Lift Chart

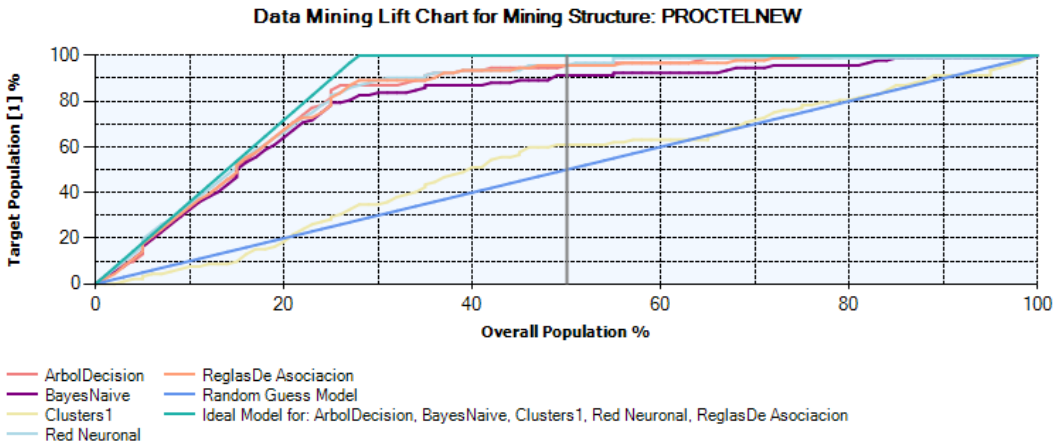


Fig. 4.5-1: Gráfico de Elevación para 1 – Solicitud de Análisis

Mining Legend

Population percentage: 50,00%

Series, Model	Score	Target population	Predict probability
ArbolDecision	0,96	95,65%	6,58%
BayesNaive	0,92	91,30%	4,02%
Clusters1	0,61	60,87%	27,68%
Red Neuronal	0,96	95,65%	5,05%
ReglasDe Asociacion	0,96	95,65%	55,56%
Random Guess Model		50,00%	
Ideal Model for: ArbolDecisio...		100,00%	

Fig. 4.5-2: Leyenda gráfico de Elevación

La figura 4-32 nos permite observar que todos los modelos se encuentran por encima de la línea de un modelo aleatorio y a si mismo, están por debajo de un modelo ideal. Esto se podría considerar positivo, ya que nos permite suponer que para un planteo estratégico de las causas ante un posible pedido de análisis podemos utilizar los modelos hallados y seguramente la elección será superior a que si dejamos la decisión librada al azar.

El eje X del gráfico representa el porcentaje del conjunto de datos de prueba que se usa para comparar las predicciones. El eje Y del gráfico representa el porcentaje de valores de predicción, en este caso se ha seleccionado la opción “Solicitud de Análisis=1”.

La línea recta diagonal, mostrada aquí en azul, aparece en cada gráfico. Representa los resultados de la estimación aleatoria y es la línea base con la que evaluar la elevación. Las demás líneas de colores (coral, violeta, amarilla, celeste y naranja) representan cada una un modelo hallado y la línea verde el modelo ideal para cada uno de los modelos entrenados.

La línea gris está posicionada en el punto en el cuál se describe la leyenda.

La puntuación es una fracción de la performance respecto al modelo ideal. Ayuda a comparar la efectividad de los modelos utilizando una población normalizada.

El modelo ideal puede captar el 100% del objetivo utilizando aproximadamente el 14% del total de la población.

El modelo de Árbol de Decisión en cambio puede captar el 95,65% del objetivo a partir del 50% del total de la población. EL umbral de probabilidad necesario para

incluir un caso entre los casos con probabilidad de tener "Solicitud de Análisis (ST1)" es 6,58 %.

El modelo ideal, como dijimos anteriormente, utiliza el 14% de la población, por lo que suponiendo que disponemos de 1000 casos sabemos que 140 (14%) responderán positivamente a la predicción. Ordenando los casos disponibles según el modelo ideal los primeros 140 serán los correctos, aunque también podemos decir que suelen existir predicciones incorrectas, como por ejemplo, ordenando los casos según el modelo del árbol de decisión encontraré 134 casos exactos (el 95,65% de los 140 que me da el modelo ideal) entre los primeros 260 casos (26% del total de la población).

Interpretando la leyenda vemos que la línea gris dice que en el 50% de los casos presentados, el modelo ideal puede predecir correctamente el 100% de ellos, es decir que para ese punto en el eje X predecirá 500 casos perfectos. El Árbol de Decisión puede en cambio predecir correctamente 479 casos, el 95,65% del total de la población en ese punto, Bayes Naïve puede predecir el 91,30% es decir 456 casos, Clúster 60,87% lo que significa 304 casos predecibles, el algoritmo de Reglas de Asociación el 95,65%, 479 casos y la Red Neuronal 95,65%, es decir 479 casos.

Es importante comprender que para el dominio del que se trata este estudio no es de utilidad conocer la cantidad de casos que se pueden predecir pero sí qué tan bien trabaja el modelo respecto al modelo aleatorio y si se acerca o no a un modelo ideal.

Hasta el momento, podemos decir que es válido avanzar con la verificación de la precisión de los resultados debido a que se considera positiva la primera prueba, todos los modelos superan la línea aleatoria y podrían estar en condiciones de ayudar a un encargado en la identificación de patrones y toma de decisiones. Es decir que, para el objetivo planteada, el modelo es de utilidad.

4.5.1.2. Matriz de Clasificación o Matriz de Confusión

Una matriz de clasificación es un método para ordenar las estimaciones buenas y malas en una tabla, para analizar con qué precisión predice el modelo el valor de destino.

A continuación se presenta la matriz hallada para cada uno de los modelos realizados.

Counts for BayesNaive on Analisis			
	Predicted	0 (Actual)	1 (Actual)
	0	209	15
	1	29	77

Tabla 4.5.1: Recuentos para Bayes Naïve en Setter

Counts for Arbol Decision on Analisis			
	Predicted	0 (Actual)	1 (Actual)
	0	218	12
	1	20	80

Tabla 4.5.2: Recuentos para Árbol de Decisión en Setter

Counts for Clusters1 on Analisis			
	Predicted	0 (Actual)	1 (Actual)
	0	199	83
	1	39	9

Tabla 4.5.3: Recuentos para Clustering en Setter

Counts for ReglasDe Asociacion on Analisis			
	Predicted	0 (Actual)	1 (Actual)
	0	226	16
	1	12	76

Tabla 4.5.4: Recuentos para Reglas Asociación en Setter

Counts for Red Neuronal on Analisis			
	Predicted	0 (Actual)	1 (Actual)
	0	220	10
	1	18	82

Tabla 4.5.5: Recuentos para Red Neuronal en Setter

Para generar una matriz de clasificación se cuenta el número de predicciones buenas y erróneas, utilizando los valores reales existentes en el conjunto de datos de prueba. La matriz es una herramienta valiosa porque no solo muestra la frecuencia con que el modelo predice un valor correctamente, sino que también muestra qué valores predice incorrectamente. Una matriz de clasificación muestra el recuento real de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada atributo de predicción.

Teniendo en cuenta los resultados de la matriz de cada modelo se puede saber en cuántas ocasiones ha sido exacta la predicción del mismo.

Las filas de cada matriz representan los valores de predicción del modelo, mientras que las columnas representan los valores reales.

Las matrices se elaboraron con el 25% del total de datos disponibles, según lo especificado en la estructura de minería de datos y en los modelos, no con los que se han utilizado para generar los mismos. Esta particularidad otorga confiabilidad a la prueba realizada.

Las matrices han dado confiabilidad y exactitud a los modelos, por lo tanto se puede considerar satisfactoria la prueba realizada.

4.5.1.3. Validación Cruzada o Cross Validation

La validación cruzada permite particionar un conjunto de datos en muchas secciones transversales de menor tamaño y crear varios modelos en dichas secciones para probar la validez del conjunto de datos completo. Los datos se dividen en particiones, cada una se utiliza a su vez como datos de pruebas, mientras que los datos restantes se utilizan para entrenar un nuevo modelo.

Esta prueba es sin dudas la que más información puede aportar. Las pruebas anteriores pudieron asegurar exactitud en la predicción, la validación cruzada deberá además asegurar si el modelo se ajusta al trabajo para el cuál ha sido creado. Deberá otorgar solidez y permitirá comparar los modelos entre sí desde el punto de vista estadístico.

Se llevará a cabo un análisis de todos los modelos, excepto el modelo de Clustering, dado que éste no puede ser sometido a las mismas pruebas por el tipo de resultado que arroja.

La prueba se basa en diez particiones. Los resultados obtenidos para los modelos de Bayes Naïve, Árbol de Decisión, Red Neuronal y Reglas de Asociación se muestran a continuación en la tabla 4.23.

BayesNaive				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	76	Classification	Pass	64
2	77	Classification	Pass	65
3	78	Classification	Pass	64

4	78	Classification	Pass	71
5	78	Classification	Pass	63
6	78	Classification	Pass	66
7	78	Classification	Pass	67
8	78	Classification	Pass	67
9	76	Classification	Pass	69
10	76	Classification	Pass	68
			Average	66,3972
			Standard Deviation	2,3769
1	76	Classification	Fail	12
2	77	Classification	Fail	12
3	78	Classification	Fail	14
4	78	Classification	Fail	7
5	78	Classification	Fail	15
6	78	Classification	Fail	12
7	78	Classification	Fail	11
8	78	Classification	Fail	11
9	76	Classification	Fail	7
10	76	Classification	Fail	8
			Average	10,9133
			Standard Deviation	2,6242
1	76	Likelihood	Log Score	-0,7738
2	77	Likelihood	Log Score	-0,6603
3	78	Likelihood	Log Score	-0,6322
4	78	Likelihood	Log Score	-0,4102
5	78	Likelihood	Log Score	-0,7724
6	78	Likelihood	Log Score	-0,5016
7	78	Likelihood	Log Score	-0,4821
8	78	Likelihood	Log Score	-0,5428
9	76	Likelihood	Log Score	-0,4472
10	76	Likelihood	Log Score	-0,5061
			Average	-0,5727
			Standard Deviation	0,1233
1	76	Likelihood	Lift	-0,1501
2	77	Likelihood	Lift	-0,0398
3	78	Likelihood	Lift	-0,0049
4	78	Likelihood	Lift	0,217
5	78	Likelihood	Lift	-0,1452
6	78	Likelihood	Lift	0,1257
7	78	Likelihood	Lift	0,1451
8	78	Likelihood	Lift	0,0844
9	76	Likelihood	Lift	0,1764
10	76	Likelihood	Lift	0,1175
			Average	0,0528
			Standard Deviation	0,1238
1	76	Likelihood	Root Mean Square Error	0,2302
2	77	Likelihood	Root Mean Square Error	0,1741
3	78	Likelihood	Root Mean Square Error	0,1755
4	78	Likelihood	Root Mean Square Error	0,1508
5	78	Likelihood	Root Mean Square Error	0,1328

6	78	Likelihood	Root Mean Square Error	0,193
7	78	Likelihood	Root Mean Square Error	0,0853
8	78	Likelihood	Root Mean Square Error	0,1781
9	76	Likelihood	Root Mean Square Error	0,1477
10	76	Likelihood	Root Mean Square Error	0,1492
			Average	0,1615
			Standard Deviation	0,0366

ArbolDecision

Partition Index	Partition Size	Test	Measure	Value
1	76	Classification	Pass	68
2	77	Classification	Pass	68
3	78	Classification	Pass	69
4	78	Classification	Pass	67
5	78	Classification	Pass	69
6	78	Classification	Pass	72
7	78	Classification	Pass	72
8	78	Classification	Pass	67
9	76	Classification	Pass	72
10	76	Classification	Pass	69
			Average	69,2988
			Standard Deviation	1,9019
1	76	Classification	Fail	8
2	77	Classification	Fail	9
3	78	Classification	Fail	9
4	78	Classification	Fail	11
5	78	Classification	Fail	9
6	78	Classification	Fail	6
7	78	Classification	Fail	6
8	78	Classification	Fail	11
9	76	Classification	Fail	4
10	76	Classification	Fail	7
			Average	8,0116
			Standard Deviation	2,1439
1	76	Likelihood	Log Score	-0,2448
2	77	Likelihood	Log Score	-0,3024
3	78	Likelihood	Log Score	-0,2957
4	78	Likelihood	Log Score	-0,3416
5	78	Likelihood	Log Score	-0,2867
6	78	Likelihood	Log Score	-0,176
7	78	Likelihood	Log Score	-0,234
8	78	Likelihood	Log Score	-0,3585
9	76	Likelihood	Log Score	-0,1614
10	76	Likelihood	Log Score	-0,2002
			Average	-0,2605
			Standard Deviation	0,0644
1	76	Likelihood	Lift	0,3788
2	77	Likelihood	Lift	0,318
3	78	Likelihood	Lift	0,3316

4	78	Likelihood	Lift	0,2856
5	78	Likelihood	Lift	0,3405
6	78	Likelihood	Lift	0,4513
7	78	Likelihood	Lift	0,3933
8	78	Likelihood	Lift	0,2688
9	76	Likelihood	Lift	0,4623
10	76	Likelihood	Lift	0,4234
			Average	0,365
			Standard Deviation	0,0639
1	76	Likelihood	Root Mean Square Error	0,1712
2	77	Likelihood	Root Mean Square Error	0,1439
3	78	Likelihood	Root Mean Square Error	0,181
4	78	Likelihood	Root Mean Square Error	0,1258
5	78	Likelihood	Root Mean Square Error	0,1496
6	78	Likelihood	Root Mean Square Error	0,1487
7	78	Likelihood	Root Mean Square Error	0,1728
8	78	Likelihood	Root Mean Square Error	0,1798
9	76	Likelihood	Root Mean Square Error	0,1626
10	76	Likelihood	Root Mean Square Error	0,1555
			Average	0,1591
			Standard Deviation	0,0168

Red Neuronal				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	76	Classification	Pass	63
2	77	Classification	Pass	66
3	78	Classification	Pass	65
4	78	Classification	Pass	68
5	78	Classification	Pass	67
6	78	Classification	Pass	66
7	78	Classification	Pass	70
8	78	Classification	Pass	70
9	76	Classification	Pass	68
10	76	Classification	Pass	70
			Average	67,304
			Standard Deviation	2,2328
1	76	Classification	Fail	13
2	77	Classification	Fail	11
3	78	Classification	Fail	13
4	78	Classification	Fail	10
5	78	Classification	Fail	11
6	78	Classification	Fail	12
7	78	Classification	Fail	8
8	78	Classification	Fail	8
9	76	Classification	Fail	8
10	76	Classification	Fail	6
			Average	10,0065
			Standard Deviation	2,2739
1	76	Likelihood	Log Score	-0,5019

2	77	Likelihood	Log Score	-0,3926
3	78	Likelihood	Log Score	-0,4362
4	78	Likelihood	Log Score	-0,3266
5	78	Likelihood	Log Score	-0,431
6	78	Likelihood	Log Score	-0,431
7	78	Likelihood	Log Score	-0,3263
8	78	Likelihood	Log Score	-0,3245
9	76	Likelihood	Log Score	-0,2685
10	76	Likelihood	Log Score	-0,1799
			Average	-0,3622
			Standard Deviation	0,0897
1	76	Likelihood	Lift	0,1217
2	77	Likelihood	Lift	0,2279
3	78	Likelihood	Lift	0,1911
4	78	Likelihood	Lift	0,3007
5	78	Likelihood	Lift	0,1962
6	78	Likelihood	Lift	0,1962
7	78	Likelihood	Lift	0,3009
8	78	Likelihood	Lift	0,3027
9	76	Likelihood	Lift	0,3552
10	76	Likelihood	Lift	0,4438
			Average	0,2634
			Standard Deviation	0,0895
1	76	Likelihood	Root Mean Square Error	0,1363
2	77	Likelihood	Root Mean Square Error	0,1407
3	78	Likelihood	Root Mean Square Error	0,1403
4	78	Likelihood	Root Mean Square Error	0,1995
5	78	Likelihood	Root Mean Square Error	0,176
6	78	Likelihood	Root Mean Square Error	0,1418
7	78	Likelihood	Root Mean Square Error	0,1185
8	78	Likelihood	Root Mean Square Error	0,1491
9	76	Likelihood	Root Mean Square Error	0,1408
10	76	Likelihood	Root Mean Square Error	0,1355
			Average	0,1479
			Standard Deviation	0,022

ReglasDe Asociacion				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	76	Classification	Pass	64
2	77	Classification	Pass	65
3	78	Classification	Pass	66
4	78	Classification	Pass	69
5	78	Classification	Pass	68
6	78	Classification	Pass	70
7	78	Classification	Pass	68
8	78	Classification	Pass	66
9	76	Classification	Pass	70
10	76	Classification	Pass	70
			Average	67,6003

			Standard Deviation	2,0996
1	76	Classification	Fail	12
2	77	Classification	Fail	12
3	78	Classification	Fail	12
4	78	Classification	Fail	9
5	78	Classification	Fail	10
6	78	Classification	Fail	8
7	78	Classification	Fail	10
8	78	Classification	Fail	12
9	76	Classification	Fail	6
10	76	Classification	Fail	6
			Average	9,7102
			Standard Deviation	2,2728
1	76	Likelihood	Log Score	-0,0801
2	77	Likelihood	Log Score	-0,0813
3	78	Likelihood	Log Score	-0,0929
4	78	Likelihood	Log Score	-0,0607
5	78	Likelihood	Log Score	-0,0809
6	78	Likelihood	Log Score	-0,0682
7	78	Likelihood	Log Score	-0,0823
8	78	Likelihood	Log Score	-0,0747
9	76	Likelihood	Log Score	-0,0462
10	76	Likelihood	Log Score	-0,0452
			Average	-0,0713
			Standard Deviation	0,0151
1	76	Likelihood	Lift	0,5435
2	77	Likelihood	Lift	0,5392
3	78	Likelihood	Lift	0,5344
4	78	Likelihood	Lift	0,5666
5	78	Likelihood	Lift	0,5463
6	78	Likelihood	Lift	0,559
7	78	Likelihood	Lift	0,545
8	78	Likelihood	Lift	0,5525
9	76	Likelihood	Lift	0,5774
10	76	Likelihood	Lift	0,5785
			Average	0,5542
			Standard Deviation	0,0147
1	76	Likelihood	Root Mean Square Error	0,0773
2	77	Likelihood	Root Mean Square Error	0,0741
3	78	Likelihood	Root Mean Square Error	0,0599
4	78	Likelihood	Root Mean Square Error	0,0594
5	78	Likelihood	Root Mean Square Error	0,0809
6	78	Likelihood	Root Mean Square Error	0,0845
7	78	Likelihood	Root Mean Square Error	0,078
8	78	Likelihood	Root Mean Square Error	0,0745
9	76	Likelihood	Root Mean Square Error	0,0578
10	76	Likelihood	Root Mean Square Error	0,0598
			Average	0,0707
			Standard Deviation	0,0097

Tabla 4.5.6: Validación Cruzada Bayes Naïve, Árbol de Decisión,

El análisis de la exactitud de los resultados obtenidos debe leerse teniendo en cuenta:

1. Similitud de los resultados entre las particiones

Se observa que existe homogeneidad de resultados entre las distintas particiones. Esto se cumple prácticamente en la totalidad de las pruebas y de las particiones. Por lo tanto el primer análisis dice que el conjunto de datos sobre el cual se ha realizado la validación es bueno para ejecutar la tarea.

2. Calidad de los resultados en base a métricas estadísticas

Prueba de clasificación

Esta métrica representa el recuento de casos clasificados correctamente.

Se divide en dos opciones "PASS" y "FAIL". La primera muestra la cantidad de clasificaciones correctas para el atributo target (sin determinar un valor específico) en cada una de las particiones.

De acuerdo a los resultados obtenidos, se indica que el modelo de árbol de decisión, presenta en promedio una mejor probabilidad de predicción con un 90% y presenta una desviación estándar (1.90) más baja que el resto de los modelos.

El modelo de Red Neuronal y Reglas de decisión, presentan probabilidades de predicción del 87%, pero con una desviación estándar de 2,23 y 2.09 respectivamente.

En cuanto al modelo de Bayes Naïve, la probabilidad de predicción es de 86%; sin embargo la desviación estándar (2,37) es más alta que la obtenida en el resto.

La clasificación "FAIL" se comporta de la misma manera que "PASS" en los 4 algoritmos pero con porcentajes ligeramente superiores. De todas maneras lo importante es observar que siempre, todos los algoritmos clasifican más cantidad de valores como "PASS" que como "FAIL", es decir lo que clasifican bien es superior a lo que clasifican mal.

Respecto a la clasificación errónea, el modelo árbol de decisión presenta una mejor desviación estándar (2,14) que el modelo de Red Neuronal (2,27), Reglas de Asociación (2,27) y Bayes Naïve (2,62).

Prueba estadística de logaritmo (Likelihood Log Score)

Denominada también "puntuación del registro", esta métrica es el logaritmo de la probabilidad real de cada caso, sumada y después dividida por el número de filas del conjunto de datos de entrada. Como la probabilidad se representa como una fracción decimal, las puntuaciones del registro son siempre números negativos. Si bien se detectan valores dispares en algunos algoritmos, los promedios de esta métrica son todos valores pequeños, nunca mayores a uno, lo cual es un indicador de que el modelo supera esta prueba.

Prueba estadística de mejora respecto al modelo predictivo (Likelihood Lift)

Este indicador representa la proporción entre la probabilidad de predicción real y la probabilidad marginal en los casos de prueba y muestra hasta qué punto mejora la probabilidad cuando se utiliza el modelo.

Es un número calculado utilizando la media del logaritmo para todas las filas con valores para el atributo de destino y las probabilidades actuales y marginales. Puede obtenerse un valor positivo o negativo, pero un valor positivo significa un modelo efectivo que supera la estimación aleatoria. Los resultados para todas las particiones de todos los algoritmos, exceptuando el modelo Bayes Naïve, son positivos, por lo tanto los demás modelos superan desde este punto de vista la estimación aleatoria.

Prueba estadística raíz cuadrada del error promedio (Root Mean Square Error)

Este indicador corresponde a la raíz cuadrada del error promedio para todos los casos de partición, dividido por el número de casos en la partición. Denominada también RMSE, es un estimador para los modelos predictivos. Cuanto menos sea la variación, más acertado será el modelo. El RMSE promedio mayor encontrado es 0,1615 y corresponde al modelo Bayes Naïve, con un error de 0,036. Estos valores y los menores a ellos que corresponden a los demás algoritmos indican que los modelos pueden ser acertados.

La prueba de Validación Cruzada para modelo de Clustering arroja la tabla 4.24.

Clusters1				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	78	Clustering	Case Likelihood	0,4802
2	78	Clustering	Case Likelihood	0,4935
3	78	Clustering	Case Likelihood	0,4542
4	77	Clustering	Case Likelihood	0,475

5	77	Clustering	Case Likelihood	0,4941
6	77	Clustering	Case Likelihood	0,4832
7	77	Clustering	Case Likelihood	0,4996
8	77	Clustering	Case Likelihood	0,4806
9	77	Clustering	Case Likelihood	0,5163
10	77	Clustering	Case Likelihood	0,4114
			Average	0,4788
			Standard Deviation	0,0273

Tabla 4.5.7: Validación Cruzada Clustering

Prueba de probabilidad de los casos (Case Likelihood)

Es la suma de las puntuaciones de probabilidad de clúster para todos los casos, divididas por el número de casos de la partición, exceptuando las filas con valores ausentes para el atributo de destino.

Indica la probabilidad de que un caso pertenezca a un clúster determinado. Las puntuaciones se suman y luego se dividen entre el número total de casos de manera que la puntuación es una media de la probabilidad de los casos. De lo anteriormente expuesto, se puede deducir la probabilidad media como 0,4788 con un desvío estándar de 0,0273, por lo tanto es una probabilidad bastante cercana a uno lo cual indica que el modelo es aceptable.

4.5.2 Valoración de los resultados

En esta etapa los resultados obtenidos se evaluaron, así como los modelos aprobados en la etapa anterior. Los resultados fueron evaluados por la comprensión e interpretación de los obtenidos en cada modelo, así como el impacto de los resultados de minería de datos para los objetivos del negocio.

Una vez probados y evaluados los modelos de minería de datos, se pudo comprobar los factores principales.

Las variables que se consideran en la utilización de la estructura de los modelos de minería de datos fueron consideradas y evaluadas para el comportamiento adecuado de cada modelo.

Los resultados obtenidos a lo largo de todas las pruebas superan los criterios de éxito planteados y cumplen con los requisitos mínimos para la solidez, exactitud y confianza.

Todos los modelos han superado, según las particularidades, las distintas pruebas en forma diferente. El modelo de Árbol de Decisión fue uno de los más óptimo, seguido por el modelo de Bayes Naïve, ambos son los que en general logran un estándar superior pero vale destacar que el modelo Red Neuronal si bien no obtuvo buenos resultados, en alguna de las pruebas es óptima en una parte del análisis de la performance.

4.5.2. Próximos Pasos

Se continuará con el proyecto planteado, debido a la satisfacción de los resultados encontrados. Para continuar, se crearán informes que permitan mostrar los resultados de los modelos entrenados y sirvan de apoyo al jefe del área en la toma de decisiones. La herramienta de trabajo seleccionada para llevar adelante dichos informes será Reporting Services. Esta herramienta es una plataforma de informes basada en un servidor que proporciona la funcionalidad completa de generación de informes para una gran variedad de orígenes de datos. Incluye una gran cantidad de herramientas para crear, administrar y entregar informes. Funciona en el entorno de Visual Studio y está totalmente integrado con las herramientas de SQL Server. Los informes pueden ser requeridos en distintos formatos, recibidos vía mail o accedidos a través de una URL. Para nuestro proyecto, el usuario se deberá conectar al servidor de informes y a través de un navegador podrá acceder a los mismos para su posterior análisis. Es importante destacar que se dejará la posibilidad de que el usuario pueda exportar los informes en diferentes formatos según su preferencia. Se confeccionará una carpeta para cada uno de los modelos realizados y allí se guardarán el o los informes necesarios para describir el contenido de cada uno de ellos.

4.6. IMPLANTACIÓN

4.6.1. Plan de Implantación

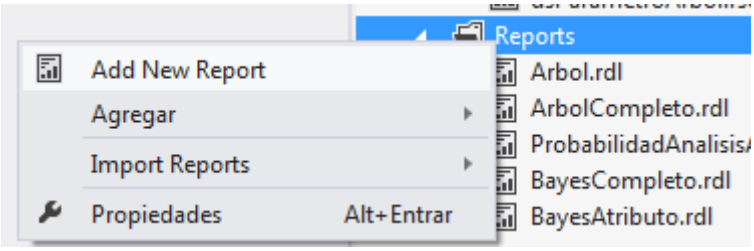
El plan de implantación estará constituido por cuatro tareas que le permitirán al usuario final leer los resultados hallados:

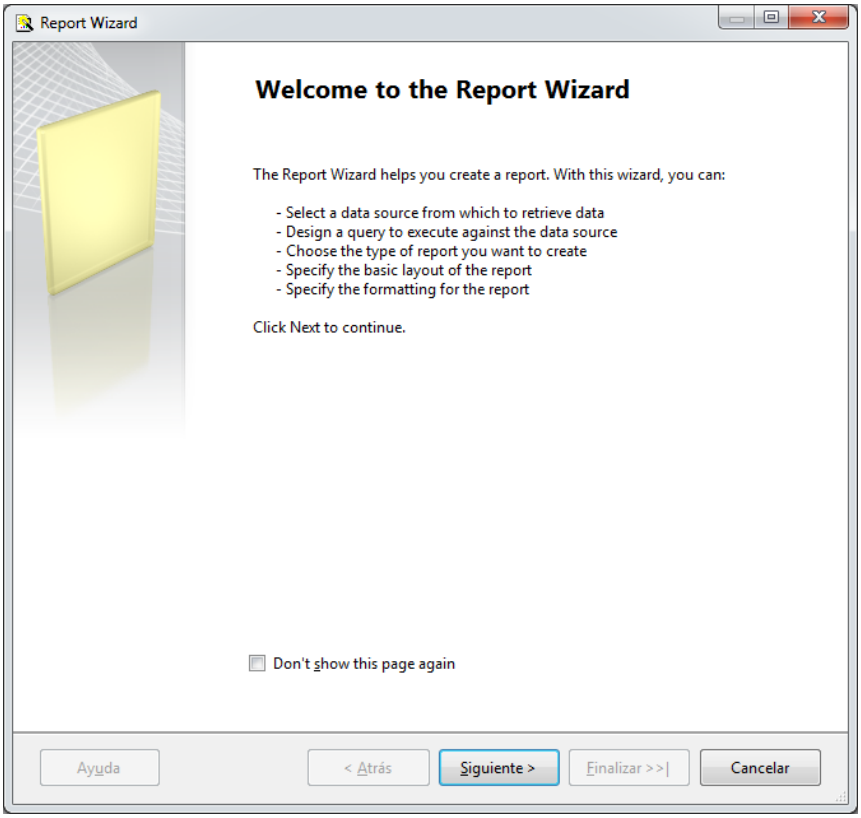
1. Crear para cada modelo una carpeta en el Servidor de Informes.
2. Crear consultas que muestren resultados útiles al usuario final.
3. Plasmar las consultas de contenido de cada modelo en tablas. Esta presentación será de fácil exportación en el formato preferido por el usuario (Word, Excel, Adobe Acrobat Reader, xml, html).
4. Relacionar variables de entrada con el atributo de predicción o variable objetivo. Identificar la mayor cantidad de ocurrencias de un atributo de predicción o la probabilidad de alcanzar el mismo a través de una condición en los atributos de entrada.

Creación de Reportes

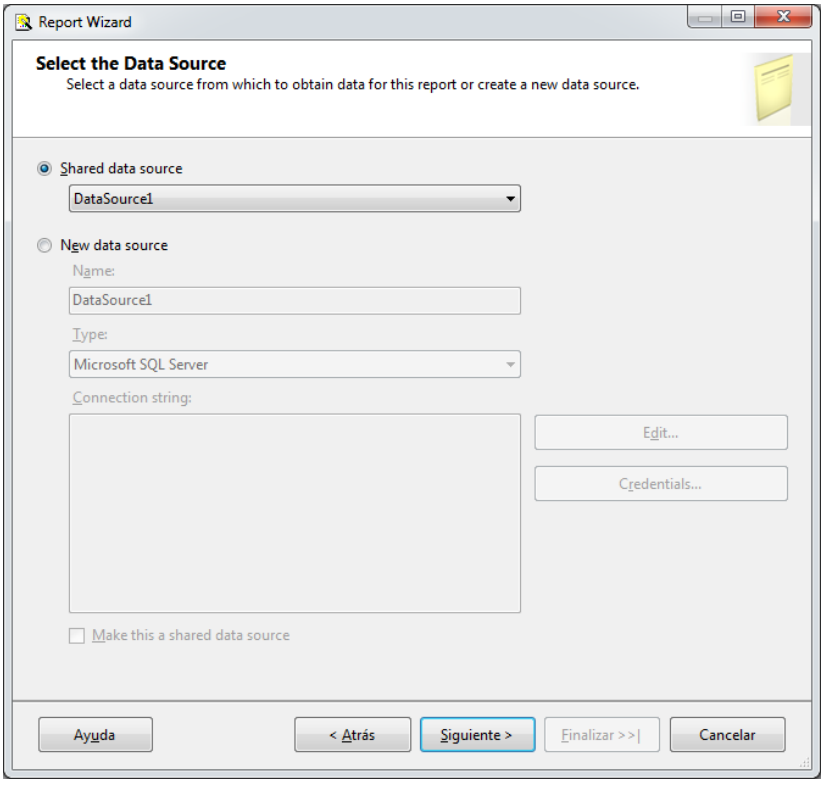
Para poder dar a conocer al usuario final los resultados de los algoritmos creados, se procedió a la elaboración de reportes.

Para ello, en el Visual Studio, creamos una carpeta Reports, Agregamos archivos de Reporting Services.

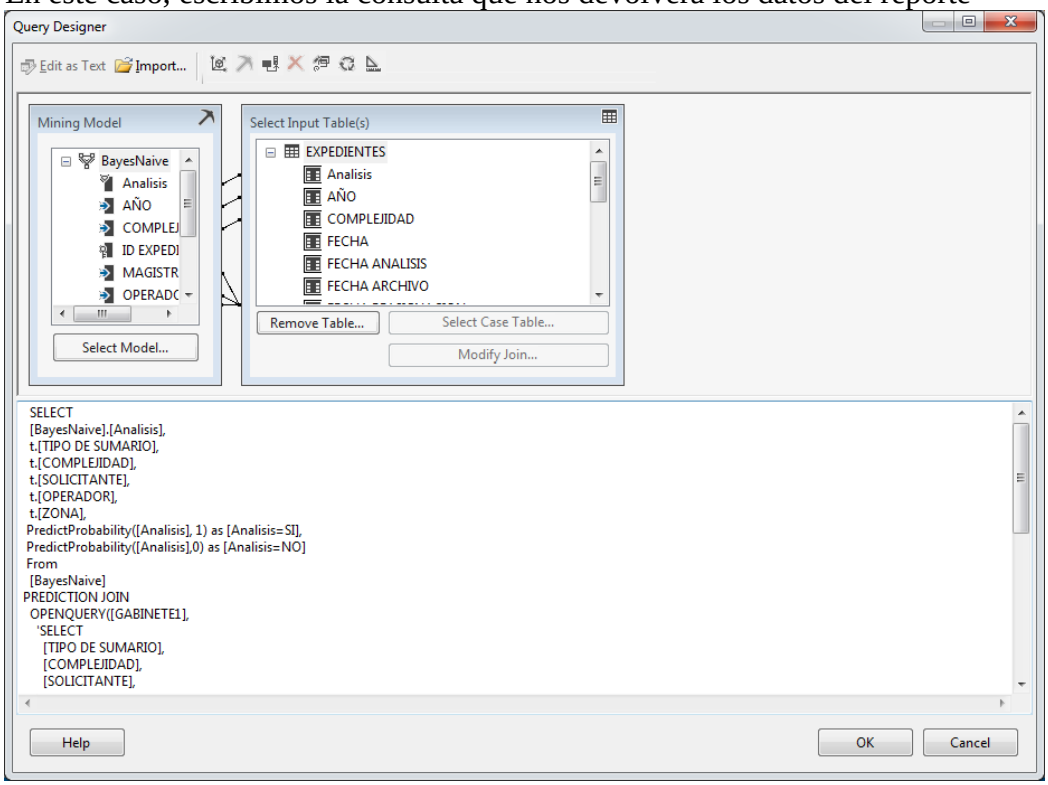


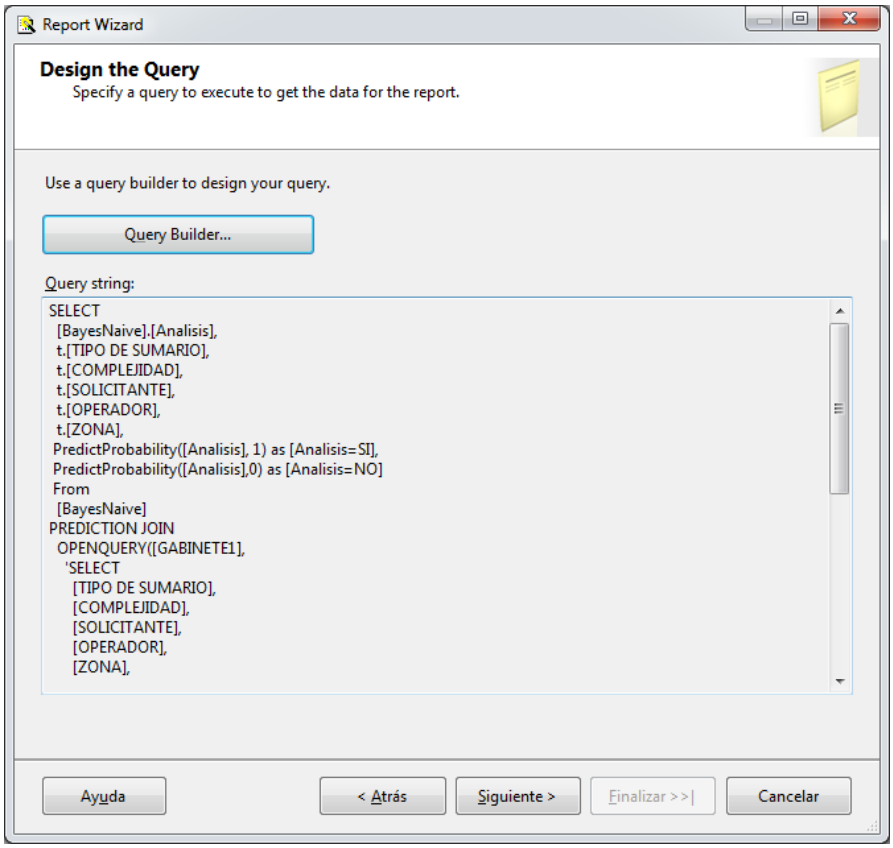


Seleccionamos el Data Source a utilizar:

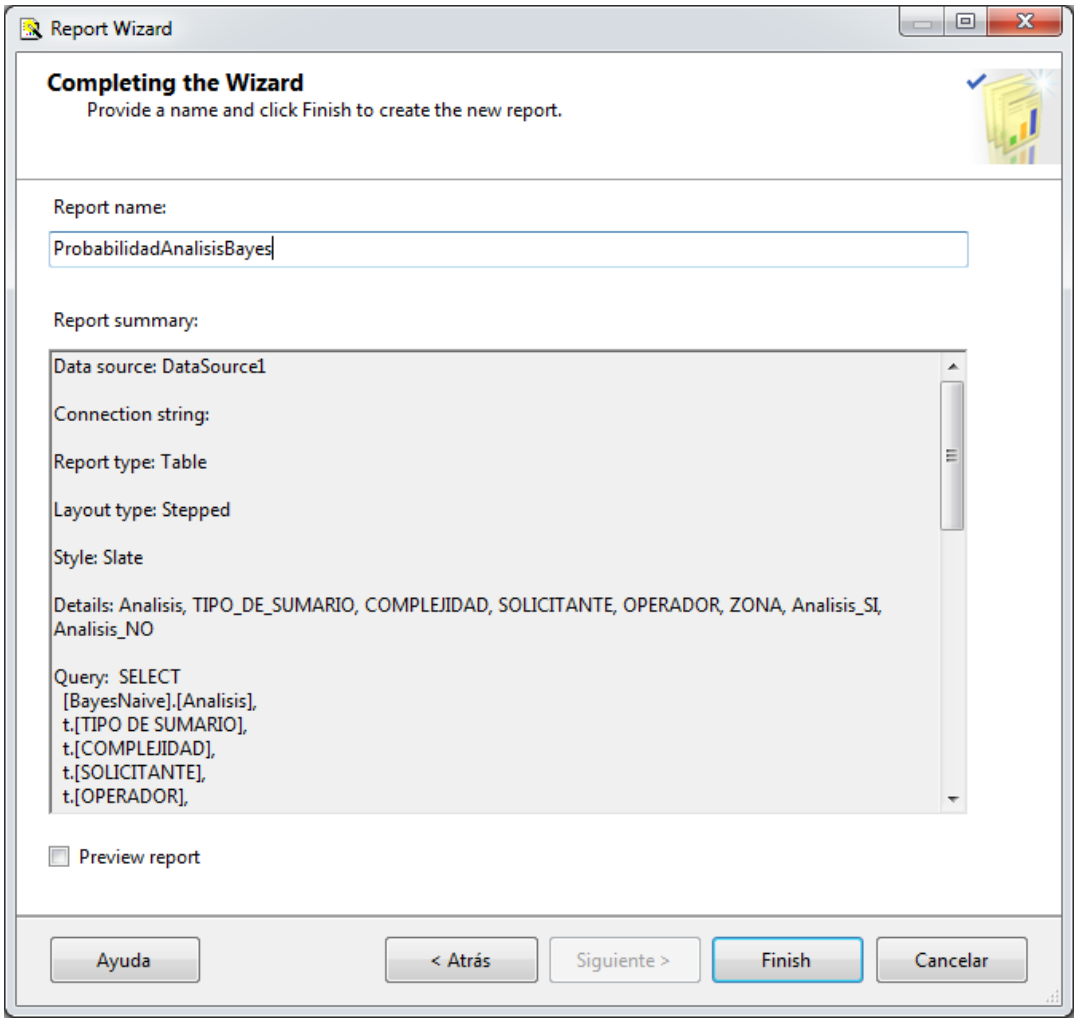


En este caso, escribimos la consulta que nos devolverá los datos del reporte

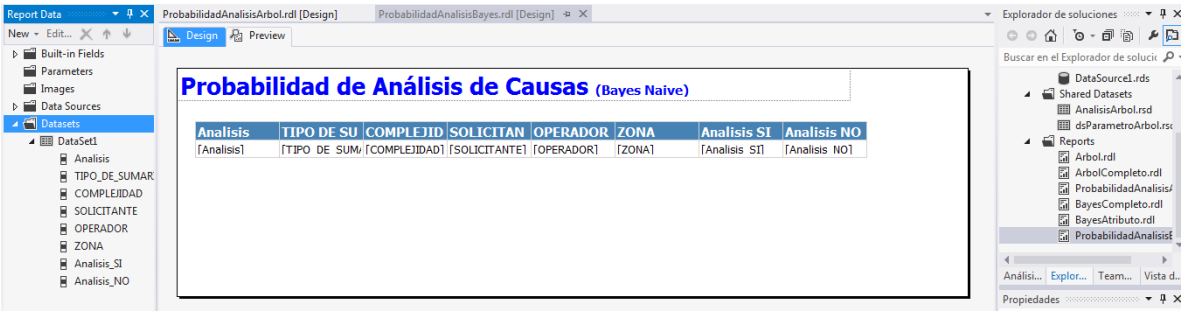




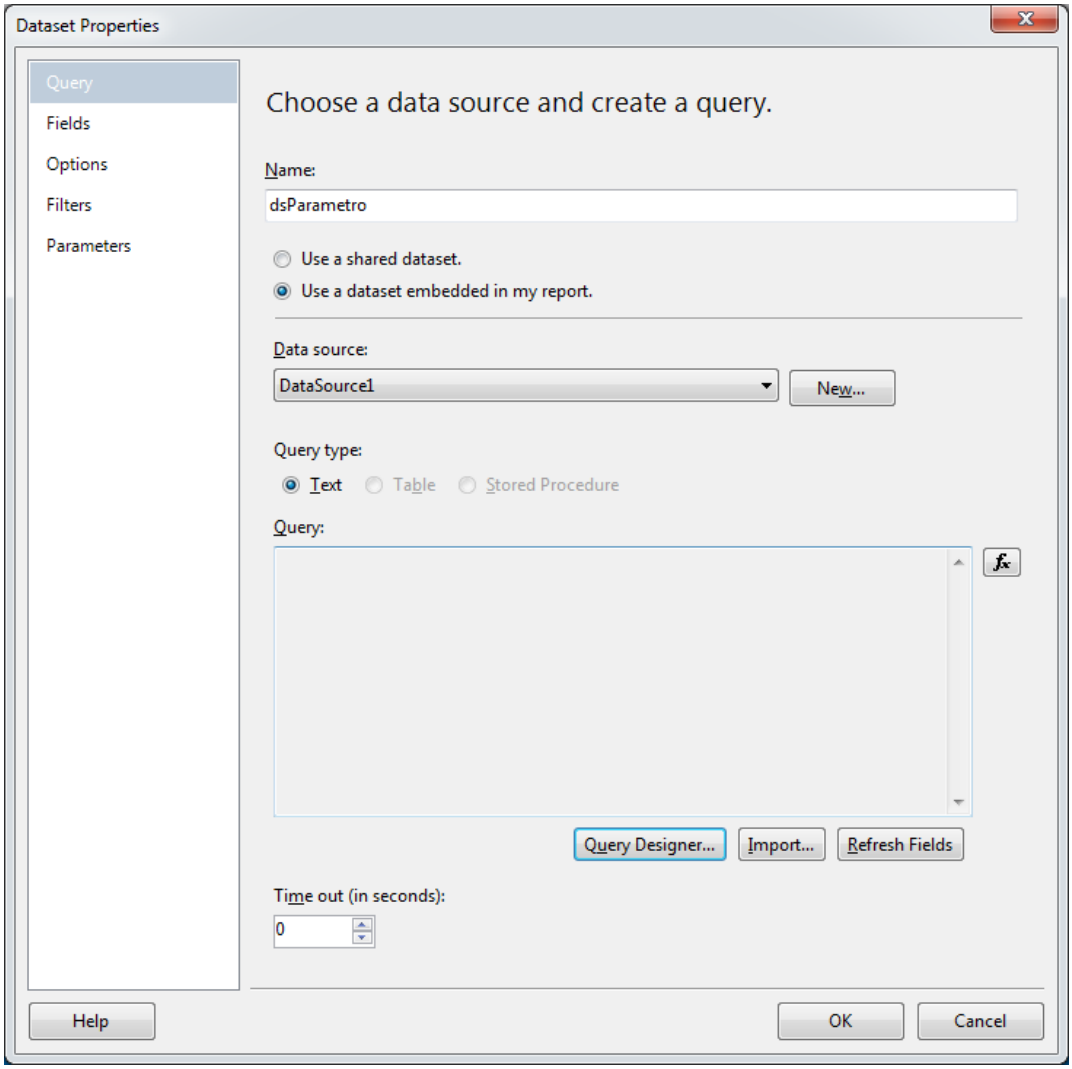
Posteriormente, colocamos el nombre del archivo y finalizamos.

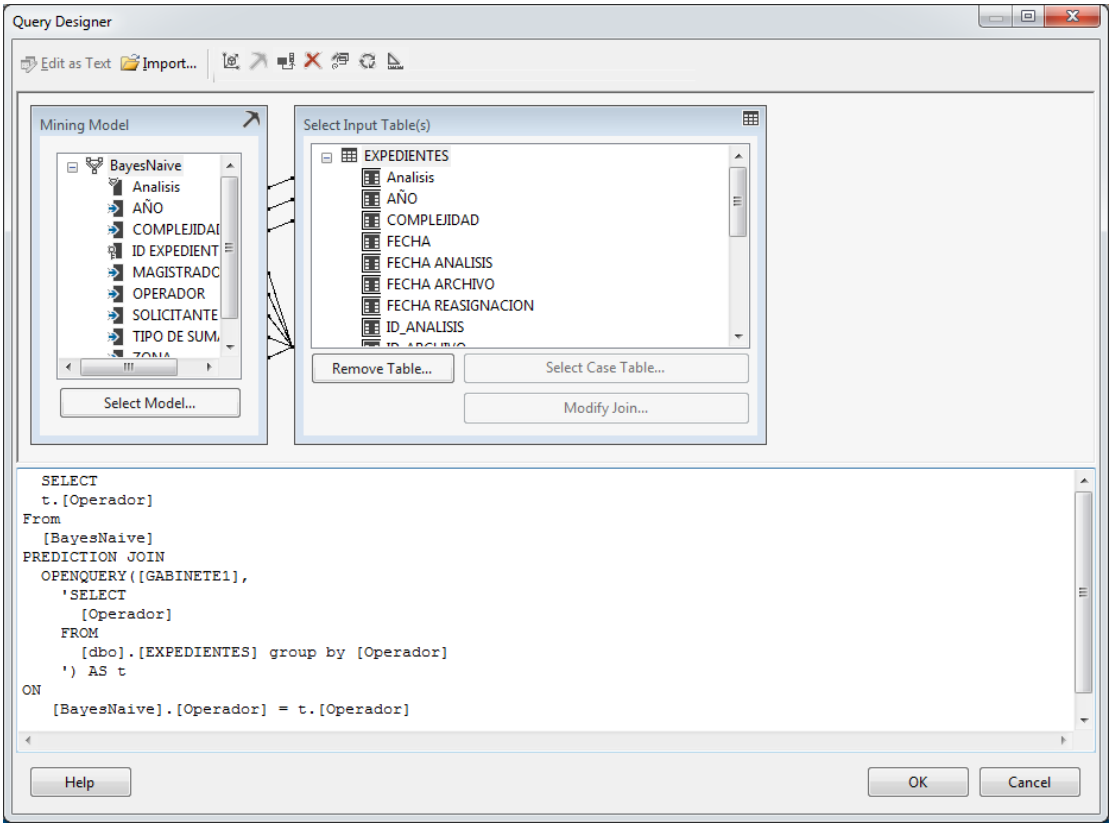


Una vez creado el reporte, se procede a editar los datos que queremos que muestre.

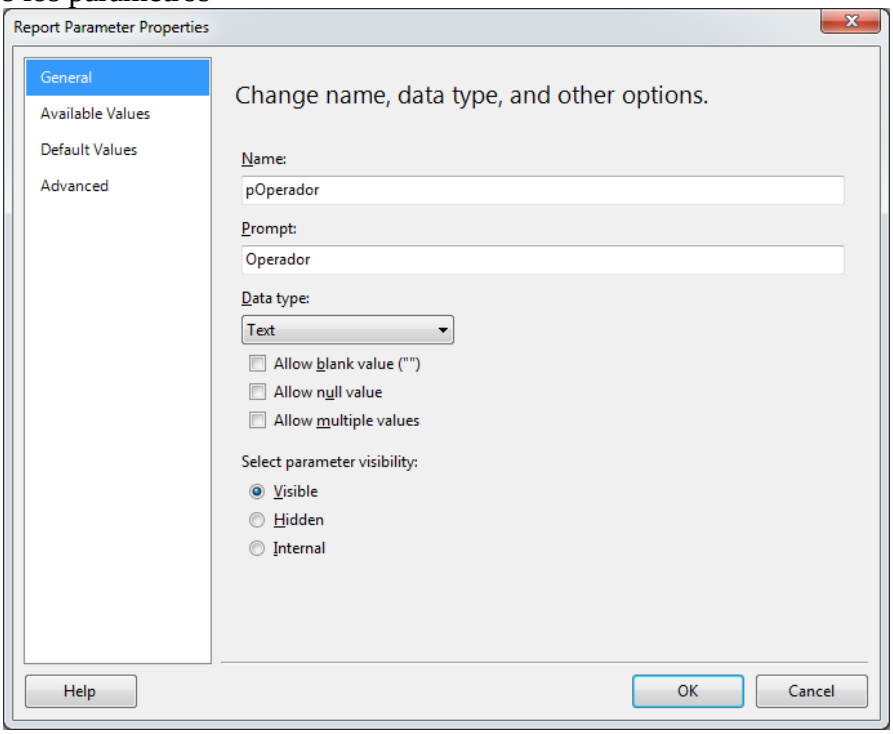


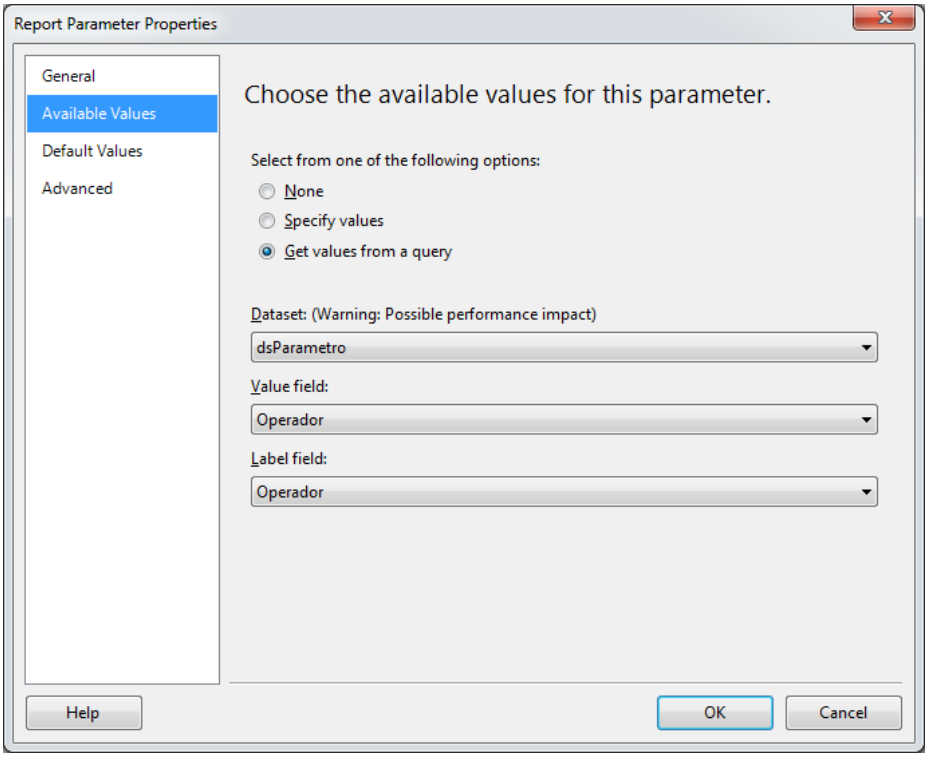
Creamos DataSet para los parámetro a usar.





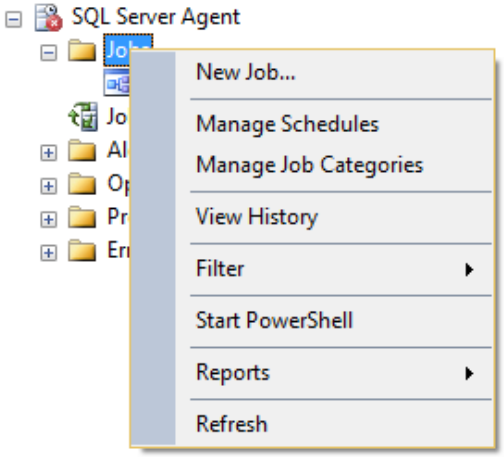
Definimos los parámetros



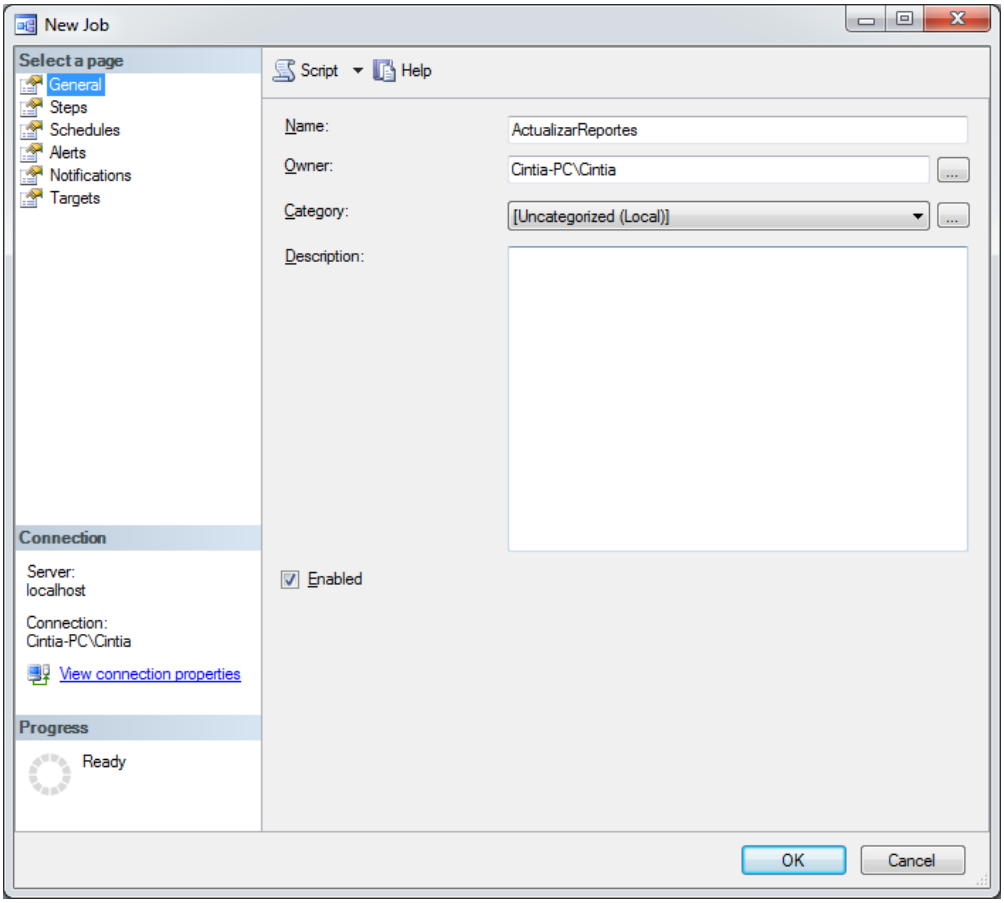


Actualización de datos de las Estructuras generadas.

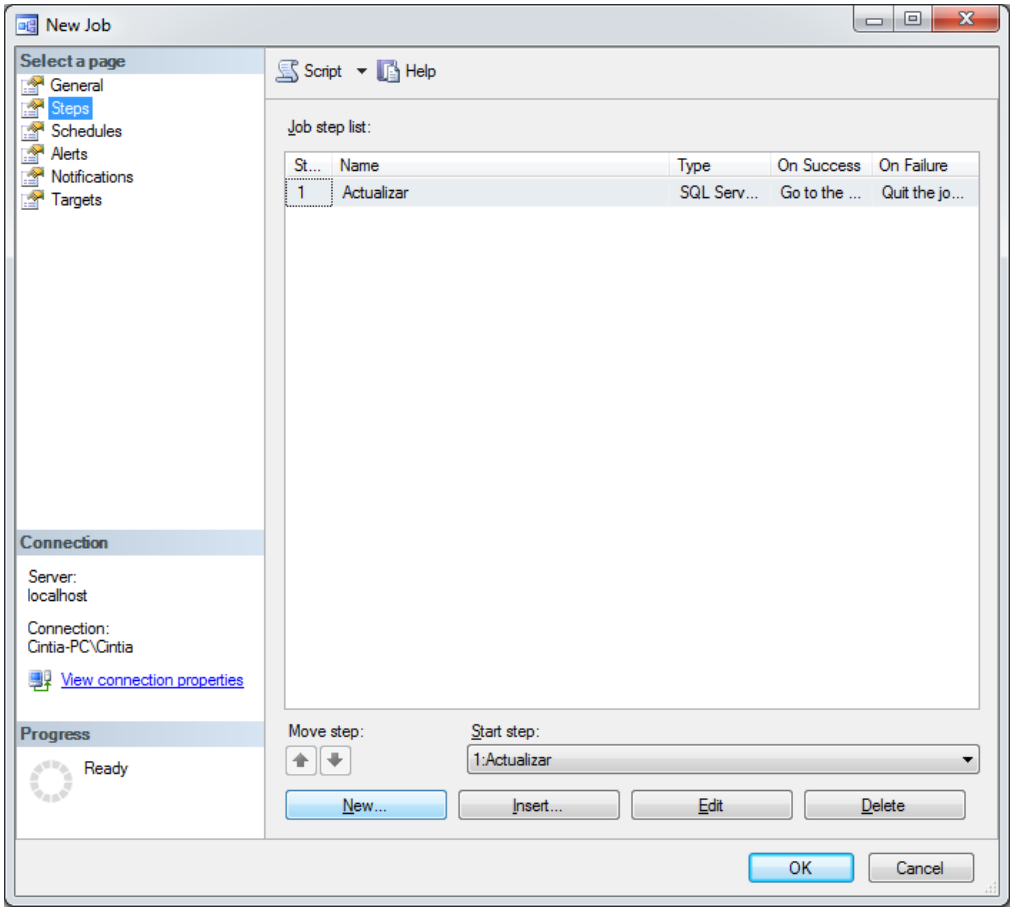
Los reportes serán actualizados mediante una tarea programada utilizando SQL Agent. Para ello se creó una tarea que ejecuta un Script XMLA, que permite la actualización de las consultas de cada uno de los modelos utilizados.



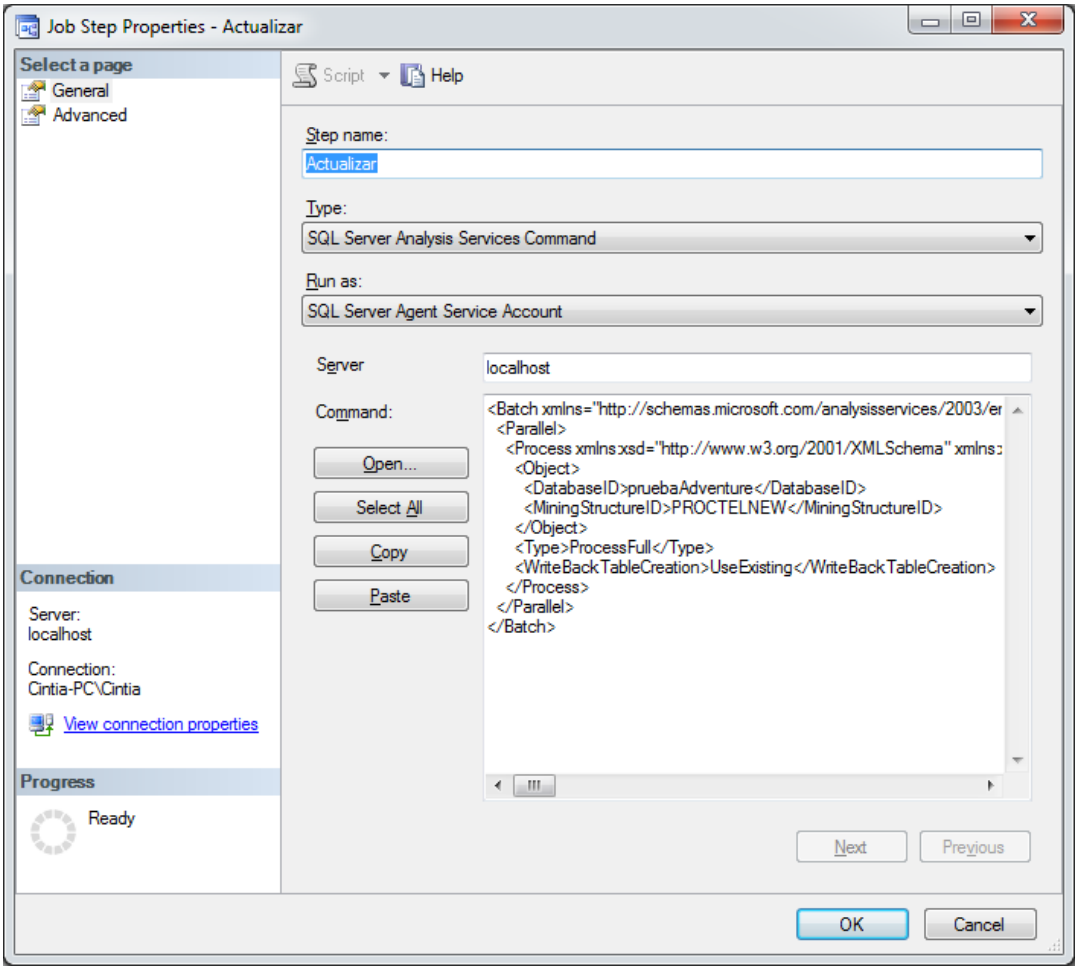
Ingresamos el nombre de la tarea.



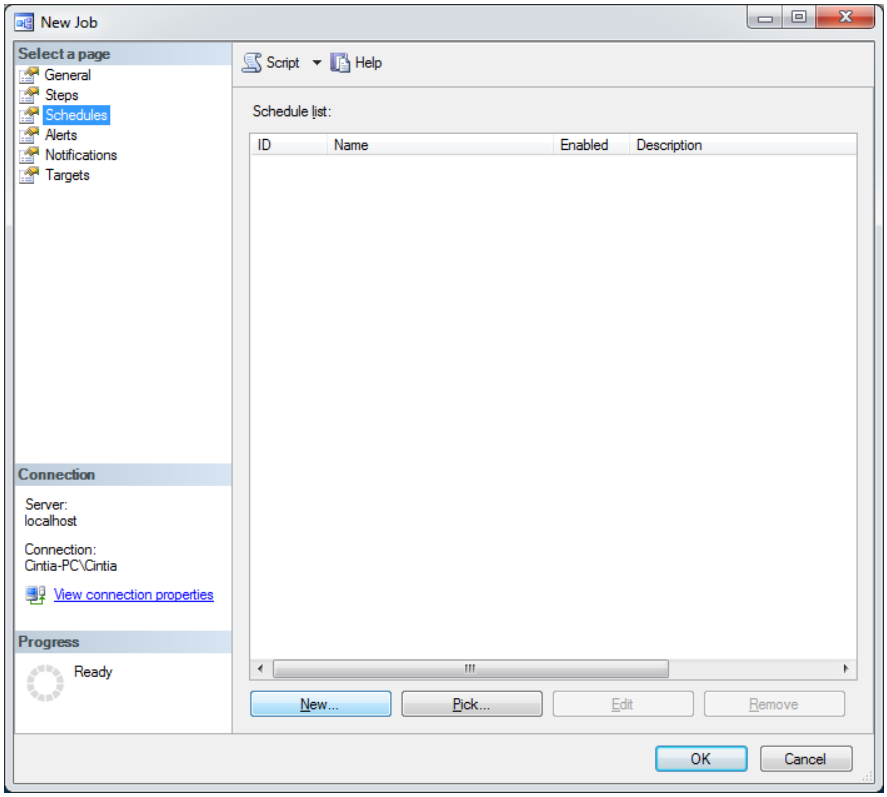
En la opción step creamos un nuevo paso.



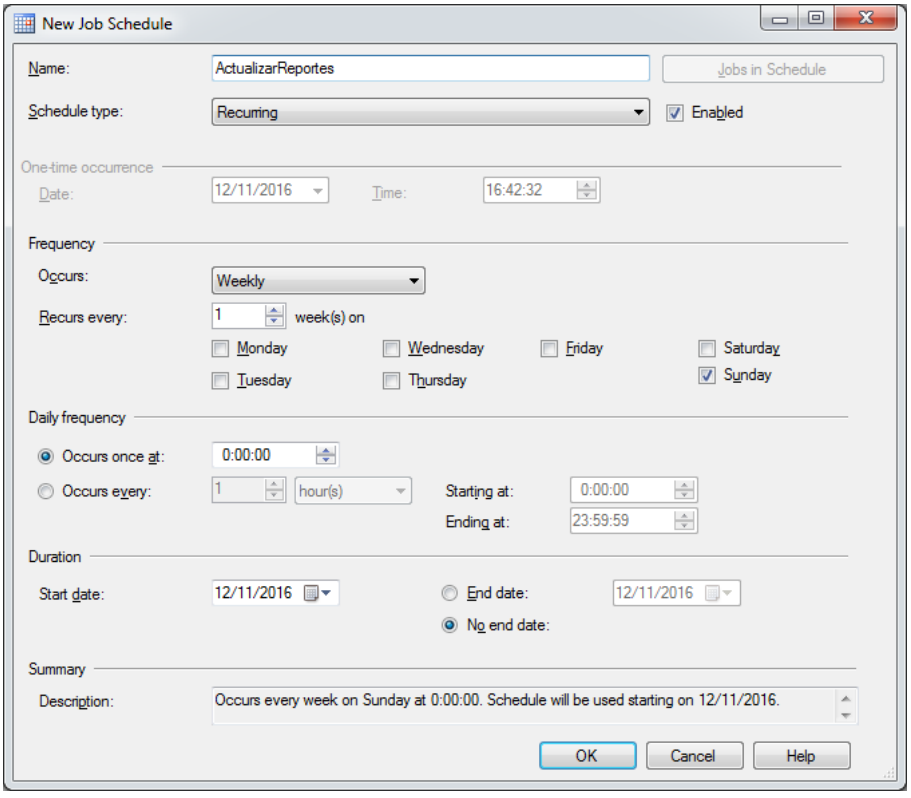
En Tipo, seleccione Comando de SQL Server Analysis Services.
En Servidor, escribimos localhost para una instancia predeterminada de Analysis Services.
En comando, colocamos el script XMLA que generamos para actualizar los reportes.



Posteriormente, en la opción programación, seleccionamos la opción nuevo.



La tarea fue programada para ejecutarse una vez al día. Por lo que los reportes no podrán ser actualizados en línea.



4.6.2. Informe final

Se pueden obtener reportes en cada carpeta para cada uno de los modelos estudiados, utilizando una conexión al servidor de informes. Estos ayudarán al encargado en las decisiones tácticas para una mejor y más eficiente resolución de expedientes. El encargado podrá interpretar los reportes presentados desde diferentes perspectivas. Si su objetivo es encontrar relaciones que detecten patrones de comportamiento en los oficios judiciales que ingresen, dependiendo de los tipos de hechos, solicitantes, etc., deberá leer los mismos observando esos atributos. Si en cambio su objetivo fuera identificar la cantidad de oficios que ingresan, se podrán analizar los datos buscando estas similitudes. Esta variabilidad muestra una vez más que la comprensión de los resultados finales es válida cuando existe un acabado y completo conocimiento del dominio del problema.

Para acceder a los reportes, se creó un pequeño Sitio Web, donde están contenidos los mismos y algunas opciones básicas que pueden ser de ayuda al Operador, como acceso online de planillas de cálculos, Documentos, Calendario, Calculadora, y acceso a las aplicaciones de escritorio de Microsoft Word y Microsoft Excel.
Cabe aclarar que las aplicaciones en línea contienen funciones reducidas, por lo que se procedió a colocar acceso a las de escritorio.
A modo de ejemplo se muestra la figura 4-34.

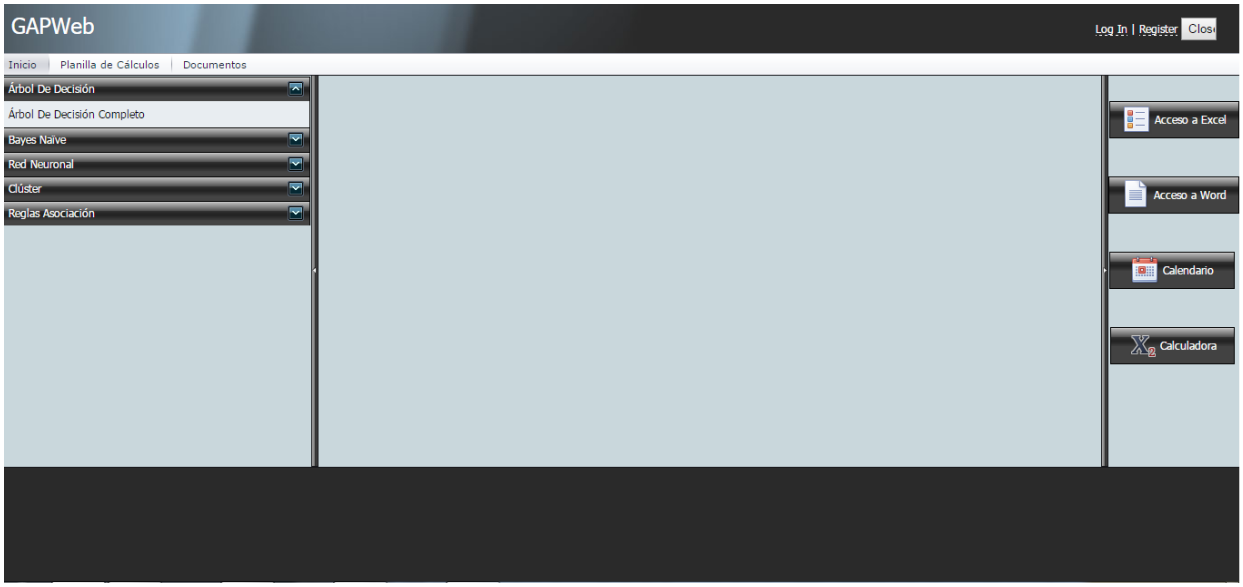
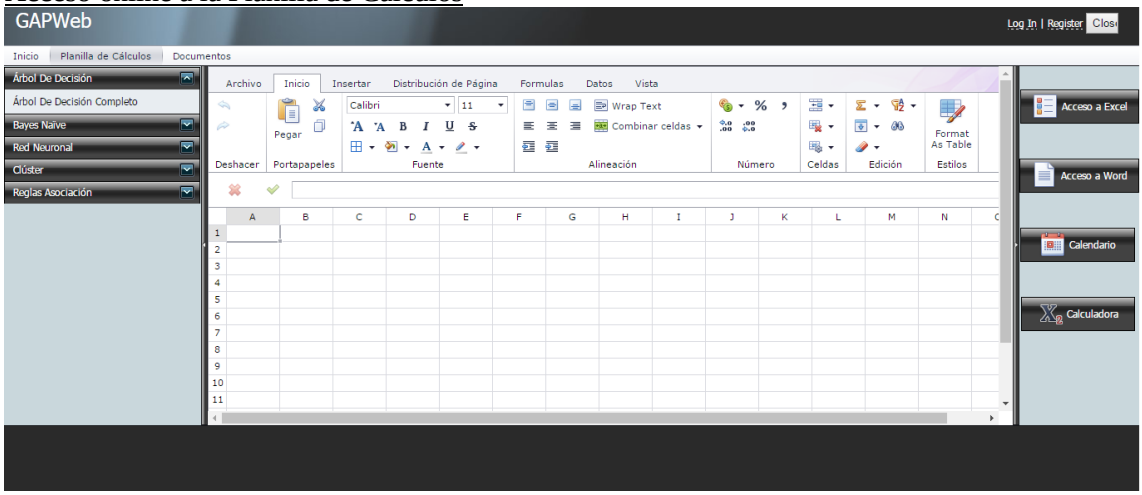
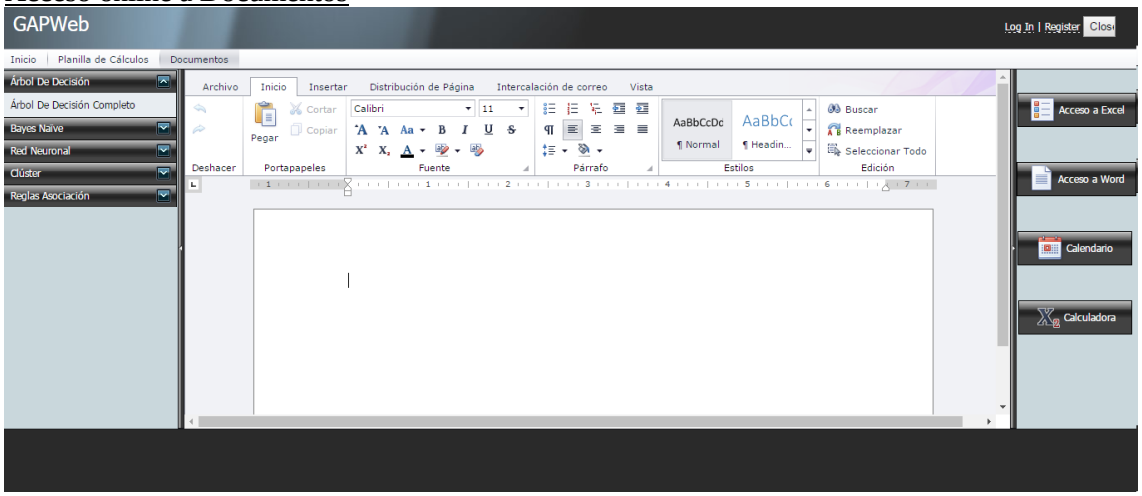


Fig. 4.6-1: Sitio Web Servidor de Informes

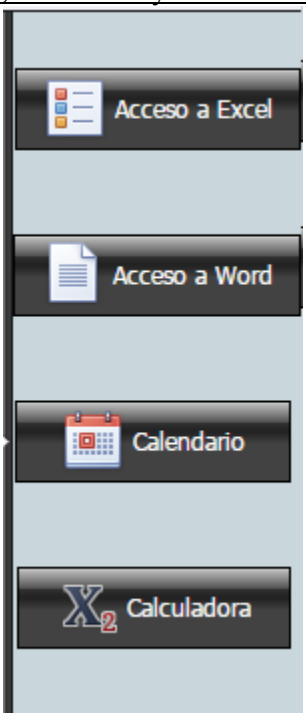
Acceso online a la Planilla de Cálculos



Acceso online a Documentos



Opciones de Acceso a Word, Excel, Calendario y Calculadora

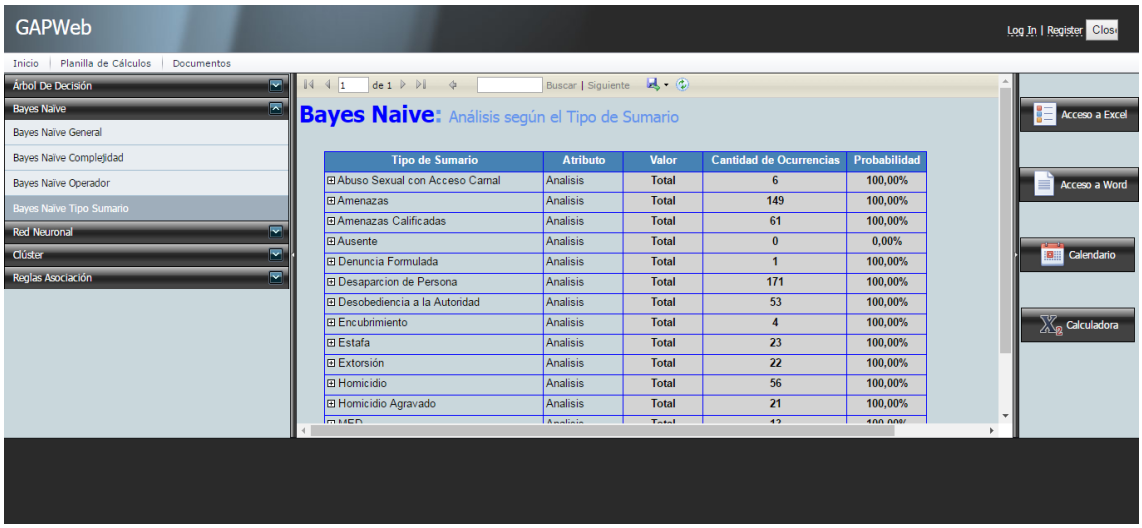


4.6.2.1. Carpeta de informes Proyecto Bayes Naïve

Se elaboró un informe general que brinda información sobre los datos de entrada. Se puede observar la cantidad de ocurrencia y probabilidad de cada uno de los posibles valores de entrada y salida del modelo. Se llama Bayes Naïve General.

Se realizaron, tres informes diferentes para cada uno de los atributos de entrada. Al igual que el anterior, se observan cantidad de ocurrencia y probabilidad de cada uno de los posibles valores en relación a un valor del atributo de salida o variable objetivo. Los nombres de los informes son: Bayes Naïve Complejidad, Bayes Naïve Operador y Bayes Naïve Tipo_Sumario.

La figura 4-35 muestra como ejemplo el informe Bayes Naïve Tipo_Sumario.



Tipo de Sumario	Atributo	Valor	Cantidad de Ocurrencias	Probabilidad
Abuso Sexual con Acceso Carnal	Analisis	Total	6	100,00%
Amenazas	Analisis	Total	149	100,00%
Amenazas Calificadas	Analisis	Total	61	100,00%
Ausente	Analisis	Total	0	0,00%
Denuncia Formulada	Analisis	Total	1	100,00%
Desaparcion de Persona	Analisis	Total	171	100,00%
Desobediencia a la Autoridad	Analisis	Total	53	100,00%
Encubrimiento	Analisis	Total	4	100,00%
Estafa	Analisis	Total	23	100,00%
Extorsión	Analisis	Total	22	100,00%
Homicidio	Analisis	Total	56	100,00%
Homicidio Agravado	Analisis	Total	21	100,00%

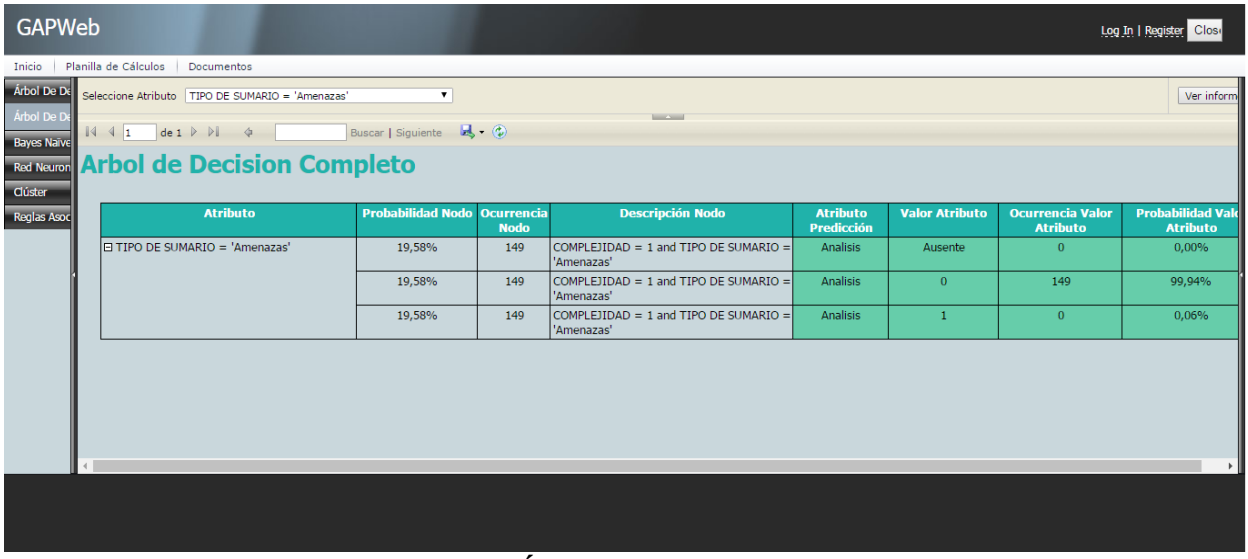
Fig.4.6-2: Informe Bayes Naïve Tipo_Sumario

4.6.2.2. Carpeta de informes Proyecto Árbol Decisión

Para el modelo Árbol de Decisión se elaboró un único informe llamado Árbol de Decisión Completo. En él, se agrupan los atributos de entrada con su respectivo valor, identificando la regla asociada a ese caso y determinando la cantidad de ocurrencia y probabilidad de los distintos valores del atributo de salida.

En la figura 4-36 se muestran las características del atributo Tipo de Sumario en la hoja del árbol de decisión identificada por el valor atributo "Amenazas"

Se aclara que este reporte tiene la opción de seleccionar el atributo a consultar.



Atributo	Probabilidad Nodo	Ocurrencia Nodo	Descripción Nodo	Atributo Predicción	Valor Atributo	Ocurrencia Valor Atributo	Probabilidad Valor Atributo
TIPO DE SUMARIO = 'Amenazas'	19,58%	149	COMPLEJIDAD = 1 and TIPO DE SUMARIO = 'Amenazas'	Analisis	Ausente	0	0,00%
	19,58%	149	COMPLEJIDAD = 1 and TIPO DE SUMARIO = 'Amenazas'	Analisis	0	149	99,94%
	19,58%	149	COMPLEJIDAD = 1 and TIPO DE SUMARIO = 'Amenazas'	Analisis	1	0	0,06%

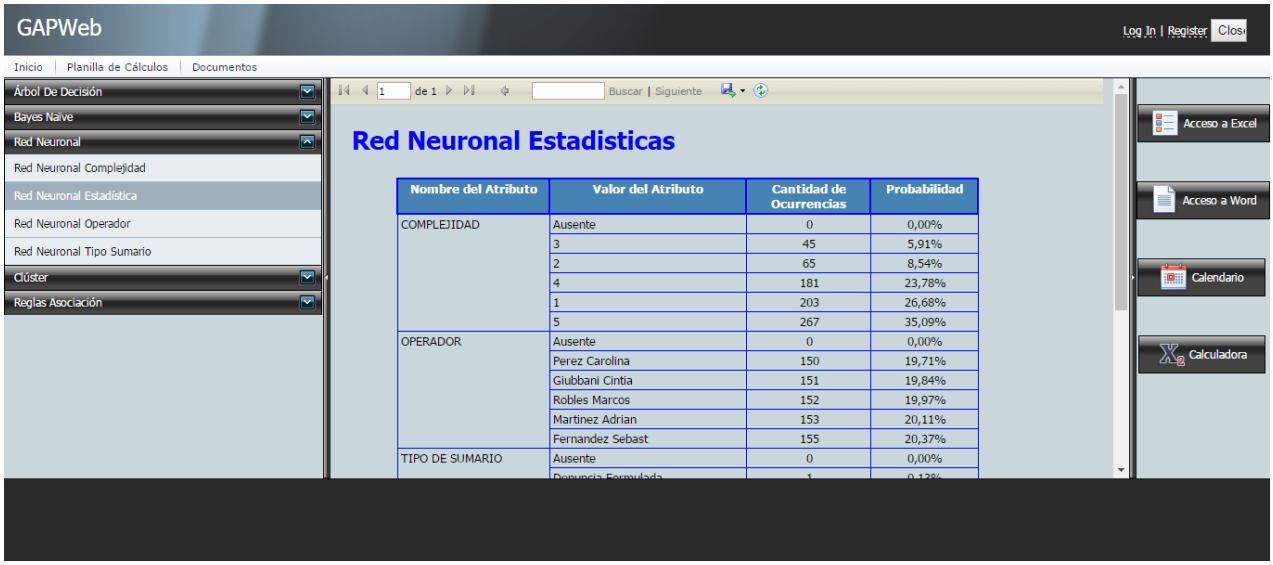
Fig. 4.6-3: Informe Árbol de Decisión Completo

4.6.2.3. Carpeta de informes Proyecto Red Neuronal

Para el modelo Red Neuronal se confeccionó un informe denominado Red Neuronal Estadísticas, el cuál brinda información general sobre los datos de entrada. Además de exponer al usuario la posibilidad de conocer los datos a partir de los cuáles se realizó el estudio sirve para dar consistencia en la comparación de los modelos utilizados.

Además, se confeccionaron cuatro informes que muestran la cantidad de ocurrencia y probabilidad para cada atributo de entrada y su relación con los valores del atributo de salida. Los mismos se denominan: Red Neuronal Complejidad, Red Neuronal Operador y Red Neuronal Tipo de Sumario.

La figura 4-37 muestra el informe general sobre las estadísticas marginales del modelo.



Nombre del Atributo	Valor del Atributo	Cantidad de Ocurrencias	Probabilidad
COMPLEJIDAD	Ausente	0	0,00%
	3	45	5,91%
	2	65	8,54%
	4	181	23,78%
	1	203	26,68%
	5	267	35,09%
OPERADOR	Ausente	0	0,00%
	Perez Carolina	150	19,71%
	Giubbani Cintia	151	19,84%
	Robles Marcos	152	19,97%
	Martinez Adrian	153	20,11%
	Fernandez Sebast	155	20,37%
TIPO DE SUMARIO	Ausente	0	0,00%

Fig.4.6-4: Informe Red Neuronal Estadísticas

4.6.2.4. Carpeta de informes Proyecto Reglas de Asociación

Para el modelo Reglas de Asociación, se realizó un único informe que expone de manera ordenada las reglas de asociación detectadas durante la fase de entrenamiento del modelo. La probabilidad de ocurrencia de las mismas es lo que da soporte al usuario para su aceptación como patrones de comportamiento. El informe se denomina Reglas de Asociación y se muestra en la figura 4-38.

Fig.4.6-5: Informe Reglas de Asociación

4.6.2.5. Carpeta de informes Proyecto Clústeres

Se elaboraron cinco informes, uno para cada clúster: Clúster 1, Clúster 2 ,Clúster 3, Clúster 4, Clúster 5. En cada grupo se aglutinan características especiales que identifican la asociación entre distintos valores de los atributos. La figura 4-39 muestra el Clúster 5.

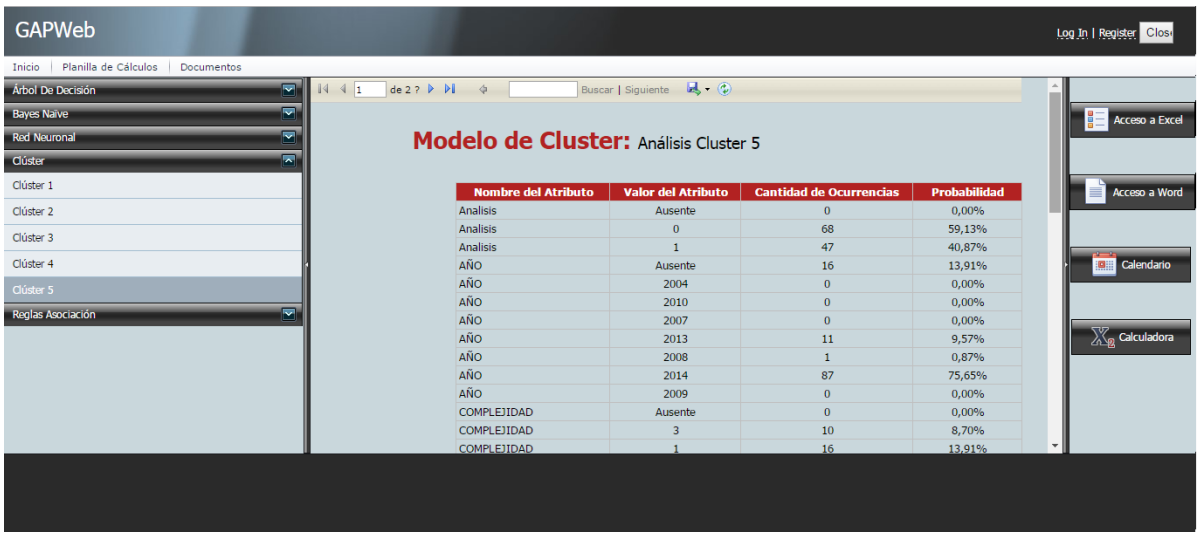


Fig.4.6-6: Informe Clúster 5

4.6.2.6. Patrones detectados

- *El variable Magistrado no es de gran importancia en los resultados observados.*

Los modelos de Bayes Naïve y Árbol de Decisión no lo incluyen en las redes de dependencia halladas. En el modelo Red Neuronal, no lo destaca con alta puntuación en la comparación de atributos. El modelo de Reglas de Asociación y Clustering lo incluyen en las agrupaciones, con muy baja probabilidad, pero esto es debido a la manera de trabajar de los algoritmo. No es suficiente para poder ser significativo de una tendencia.

- *En Complejidad, la preferencia son las que tienen una complejidad tipo 4 y 5.*
  El modelo de Bayes Naïve no arroja muestras de análisis en causas cuya complejidad es de tipo 1 o 2, lo mismo sucede en las Reglas de Asociación. El modelo clúster posee en su mayoría, una agrupación importante en los tipos 4 y 5. En cuanto a los modelos Red Neuronal y el Árbol de Decisión, el porcentaje hallado en Complejidad de Tipo 5, es superior al resto. Esto es suficiente para afirmar que son las causas de mayor complejidad las que se requiere análisis.
- *En Operador, la preferencia para causas sin Análisis es "Fernández".*
  En todos los modelos utilizados se identifica esta tendencia, teniendo en cuenta la relación Cantidad de Ocurrencias - Probabilidad. Por lo que es suficiente para definir un patrón de comportamiento.
- *En Operador, la preferencia para causas con Análisis es "Giubbani".*
  Tanto los modelos de Bayes Naïve, como Reglas de Asociación y Árbol de Decisión, identifican esta tendencia. En cuanto al modelo de Red Neuronal no logran enunciar esta tendencia, pero es oportuno mencionar que de 108 ocurrencias, el 13% no son causas con análisis, por lo que eso no es suficiente para definir o contradecir un patrón de comportamiento. En lo que respecta al modelo Clúster, los porcentajes obtenidos son con valores de ocurrencias menores a 20 del total de casos analizados. Se podría considerar un patrón de comportamiento, teniendo en cuenta otros aspectos.
- *En Solicitante, datos relevantes.*
  Si bien este atributo no es considerado un patrón importante de comportamiento, parece aporta algo interesante, permitirá ayudar al encargado a tomar decisiones de análisis. El modelo Árbol de Decisión, permite identificar que el 99% de las causas solicitadas por la Unidad Judicial de Homicidios, suele pedirse análisis

de la información. En cuanto al resto de oficios solicitados por las demás Unidades Judiciales sólo el 80% se requiere análisis. El modelo Reglas de Asociación, abala dicho patrón. En cuanto al modelo Bayes Naïve, no arroja muestras en cuanto a Solicitante. Los restantes modelos no muestran resultados relevantes.

- *El variable Zona no adquiere importancia en los resultados observados*.

  Los modelos de Bayes Naïve y Árbol de Decisión no la incluyen en las redes de dependencia halladas. Red Neuronal y Clúster no la destacan con alta puntuación en la comparación de atributos. Solamente el modelo de Reglas de Asociación lo incluye en las agrupaciones pero esto es debido a la manera de trabajar del algoritmo. No es suficiente para poder ser significativo de una tendencia, aunque, cabe aclarar, que dichos resultados, suman a la hora de observar cuales son las zonas de las que se reciben más pedidos.

- *En Tipo de Sumario, se observa que existe una tendencia hacia análisis en Desapariciones de persona, Secuestros Virtuales y Homicidios.*

  Esta es una característica importante que puede aportar un valor agregado al encargado. Esto podría indicar al encargado, la anticipación de solicitudes de análisis, cuando ingresan este tipo de causas. Cabe mencionar que si bien los porcentajes que apoyan la identificación de esta característica son elevados en Bayes Naïve y Reglas de Asociación, en los modelos Red Neuronal y Clúster, deben considerarse suficientes porque en los otros tipos de causas, los mismos son nulos o no superan el 10%.

- *El Operador "Robles Marcos", tiene preferencia por los análisis cuyas causas son de Tipo de Sumario "Secuestro Virtual".*

  Esta afirmación es importante para un planteo táctico, la misma se obtiene del análisis de la resolución de informes realizados por el operador mencionado. Bayes Naïve, Red Neuronal, Árbol de Decisión y Reglas de Asociación marcan esta tendencia.

- *Las causas de Tipo "Desobediencia a la autoridad" no se solicitan análisis.*

  Si bien no aporta muchos datos al encargado sirve para verificar el buen funcionamiento de los algoritmos, debido a que todos coinciden en dicho patrón.

Esta información permitirá al encargado evaluar a quién debe otorgar este tipo de causas.

4.6.3. Revisión del proyecto

No se considera necesario modificar o agregar algún detalle a los objetivos originales del proyecto. Esto es gracias a la retroalimentación permanente por parte del usuario final. Los modelos han sido probados, evaluados y modificados tratando de encontrar aquellos que en su definición de parámetros, estructura y funcionamiento tenían un comportamiento ideal para el caso en estudio. Se da por concluido el proyecto en su especificación actual y se encuentran totalmente satisfactorios los resultados hallados.

5. Proyecto a Futuro

Se intentará integrar las demás áreas con el fin de compartir información criminal, integrando distintas fuentes de información. Esto tendrá por objetivo detectar tipos de asociaciones entre las bases de datos mediante técnicas de minería de datos, lo que nos permitirá identificar: integrantes de bandas criminales, como se relacionan entre sí, patrones de comportamiento.
Se buscará extraer relaciones entre los sumarios y construir una posible red de sospechosos y causas para identificar patrones de interacción entre los mismos.

Se tendrá como objetivo identificar los recursos críticos para establecer estrategias de prevención y detección más eficientes; proveer de fundamentos empíricos para el desarrollo de planes, entre los diferentes departamentos, orientados a la reducción del delito e identificar la información relevante a ser recolectada en el lugar del hecho para poder desbaratar bandas criminales.

6. Conclusión

La minería de datos está orientada al desarrollo de métodos para explorar datos; se utilizaron técnicas que permitieron transformar dichos datos en información que permitiera entender ciertos patrones de comportamiento en el trabajo con expedientes.

En este proyecto, se ha desarrollado una herramienta que permite detectar patrones de comportamiento en las distintas investigaciones judiciales que ingresan día a día, y de esta manera usar dichos patrones para diseñar estrategias proactivas. La herramienta se ha probado exitosamente en el campo.

Durante el desarrollo del trabajo se reconoce que fueron alcanzados los objetivos propuestos. Se estudiaron diferentes técnicas para desarrollar modelos de predicción, basados en sistemas de soporte de decisiones, utilizando minería de datos, tales como Bayes Naïve, Árbol de Decisión, Redes Neuronales, Reglas de Asociación y Clusters.

Se obtuvo similitudes en los patrones obtenidos, lo que permite considerar a los resultados hallados como coherentes, lo que nos permite concluir que se ha alcanzado el objetivo propuesto. Se logró relacionar atributos de entrada que permitieron identificar tendencias en un atributo de salida. Además, fue posible verificar que los patrones obtenidos son de utilidad para el encargado del área.

Se puede concluir que para el objetivo propuesto es suficiente la creación de tres modelos: Bayes Naïve, Árbol de Decisión y Reglas de Asociación. Dichos modelos se destacan del resto de los estudiados por su eficiencia y facilidad de uso. Se decidió elegir tanto Bayes Naïve como Árbol de Decisión, debido a que ambos aportan la mayoría de los patrones y las pruebas de validación de los mismos son ampliamente superadas. Los dos se acercan al comportamiento de un modelo ideal. Con respecto al modelo Reglas de Asociación, se escogió ya que es simple de entender y sobre todo práctico en su uso debido a la claridad con que arroja las conclusiones.

El estudio debe seguir, deben analizarse más patrones de comportamiento, analizar cómo se relacionan los mismos y recién allí se estará en presencia de una emulación del conocimiento que permita definir un planteo táctico íntegro y completo a la hora de llevar adelante una investigación eficaz y eficiente de las causas judiciales.

7. BIBLIOGRAFIA

7.6. Referencias Bibliográficas

[1] - Defense Intelligence Journal; (2005) - Colleen McCue, Ph.D

[2] - https://www.jusmisiones.gov.ar/files/MaterialCursoUnidos/POLICIA%20JUDICIAL%20historia%20y%20presente.doc

[3] - http://aulavirtual.derecho.proed.unc.edu.ar/file.php/110/Carpeta_Dr._Vivas_Ussher/Capitulo_7/policiajudicial.ppt

[4] - https://www.jusmisiones.gov.ar/files/MaterialCursoUnidos/Polica%20Judicial%202014.pdf

[5] - M. de la Puente, *Minería de Datos* [online] Recuperado: 12 de abril 2012 disponible en

http://profesionalesdecienciasdelainformacion.wordpress.com/2010/05/04/mineria-de-datos/

[6] Blackwelder, J.K., L.L. Jonson, 1984. *Estadística Criminal y Acción Policial en Buenos Aires, 1887-1914*. Desarrollo Económico, 93, Vol. 24, 1984, pp. 109-122.

[7] Rubial B.C., 1993. *Ideología del Control Social,1880-1920*. Centro Editor de América Latina, Buenos Aires, Argentina.