UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y DE SISTEMAS

TRABAJO DE TESIS PARA OBTENER EL TITULO PROFESIONAL DE INGENIERO DE COMPUTACION Y SISTEMAS

"APLICACIÓN DE UN MODELO DE MINERÍA DE DATOS PARA IDENTIFICACIÓN DE PATRONES QUE INFLUYEN EN LA DESERCIÓN ACADEMICA EN EL INSTITUTO SUPERIOR LEONARDO DAVINCI USANDO IBM SPSS MODELER Y LA METODOLOGÍA CRISP-DM"

Línea de Investigación:

Gestión de Datos y de Información.

Autores:

Br. PANDO CUEVA, AUREA DAJANA MAKARENA

Br. ZARATE OBESO, WINNY DEL ROSARIO

Asesor:

ING. AGUSTIN EDUARDO ULLON RAMIREZ

2020

Fecha de Sustentación: 10/12/20

“APLICACIÓN DE UN MODELO DE MINERÍA DE DATOS PARA IDENTIFICACIÓN DE PATRONES QUE INFLUYEN EN LA DESERCIÓN ACADEMICA EN EL INSTITUTO SUPERIOR LEONARDO DAVINCI USANDO IBM SPSS MODELER Y LA METODOLOGÍA CRISP-DM”

Elaborado por:

Br. Pando Cueva, Aurea Dajana

Br. Zarate Obeso, Winny Del Rosario

Aprobada por:

Ing. Heber Gerson Abanto Cabrera
Presidente
CIP: 106421

Ing. Edward Fernando Castillo Robles
Secretario
CIP: 192352

Ing. Karla Vanessa Meléndez Revilla
Vocal
CIP: 120097

Ing. Agustín Eduardo Ullón Ramírez
Asesor
CIP: 137602

PRESENTACIÓN

De acuerdo a los requerimientos establecidos por el reglamento de grados y Títulos de la Universidad y al reglamento interno de la escuela profesional de Ing. de Computación y Sistemas, presentamos nuestro Trabajo de Tesis: "**APLICACIÓN DE UN MODELO DE MINERÍA DE DATOS PARA IDENTIFICACIÓN DE PATRONES QUE INFLUYEN EN LA DESERCIÓN ACADEMICA EN EL INSTITUTO SUPERIOR LEONARDO DAVINCI USANDO IBM SPSS MODELER Y LA METODOLOGÍA CRISP-DM"** para obtener el Título Profesional de Ingeniero de Computación y Sistemas.

El trabajo fue desarrollado de acuerdo al marco de referencia los lineamientos establecidos por la Facultad de Ingeniería y procedimientos de la Escuela Profesional de Ing. de Computación y Sistemas, así como de los conocimientos alcanzados durante nuestra formación en la universidad.

DEDICATORIA

A Dios por guiarnos en la vida y a lo largo de nuestra carrera. A las personas que han estado con nosotros siempre, especialmente a mis padres, quienes han sido los grandes pilares para este objetivo de mi vida.

Br. Pando Cueva, Aurea Dajana

Agradezco en primer lugar a Dios por haberme permitido lograr este objetivo y por darme lo necesario para seguir adelante día a día. Agradezco infinitamente a mi familia por darme las bases necesarias para culminar con éxito este gran proyecto.

Br. Zarate Obeso, Winny Del Rosario

AGRADECIMIENTO

Se agradece de una manera especial a todas las personas de la institución educativa que nos han brindado su apoyo en la obtención de la información, dando todas las facilidades para el término del trabajo.

También un agradecimiento a nuestro asesor por la paciencia y apoyo en cada fase del proyecto. También se agradece a los ingenieros y docentes que con sus enseñanzas dentro de la universidad nos proporcionaron el conocimiento necesario para abordar el tema de la tesis.

Muchas Gracias.

Las autoras.

RESUMEN

"APLICACIÓN DE UN MODELO DE MINERÍA DE DATOS PARA IDENTIFICACIÓN DE PATRONES QUE INFLUYEN EN LA DESERCIÓN ACADEMICA EN EL INSTITUTO SUPERIOR LEONARDO DAVINCI USANDO IBM SPSS MODELER Y LA METODOLOGÍA CRISP-DM"

Por:

Br. Pando Cueva, Aurea Dajana
Br. Zarate Obeso, Winny Del Rosario

En la actualidad para las instituciones educativa se ha convertido en un problema conocer los patrones que influyen en la deserción académica y de esta manera tratar de reducir este número, convirtiéndose en un dolor de cabeza para los tomadores de decisiones de las Instituciones educativas. Es por ello que es importante conocer porque los estudiantes deciden abandonar sus estudios y cuáles son las circunstancias que lleva a ello.

Las herramientas que permiten crear un modelo de minería de datos y más el análisis de la información de los datos de los estudiantes que fueron proporcionados por los sistemas informáticos del Instituto Superior Leonardo Davinci, nos ha llevado ha crear un modelo de minería de datos que nos lleva a obtener patrones que influyen en un estudiante desertor. La presente modelo se implementó a través del análisis de la información: personal, académica y de la interacción de los estudiantes.

Para contribuir con la solución al problema de la deserción estudiantil se plantea desarrollar un "Modelo de Minería de datos para identificación de patrones que influyen en la Deserción Académica en el Instituto Superior Leonardo Davinci" con el objetivo de conocer cuáles son las posibles causas o patrones que llevan a un alumno a abandonar sus estudios, basado en del análisis de las características de los datos de los estudiantes.

ABSTRACT

"APPLICATION OF A DATA MINING MODEL FOR IDENTIFICATION OF PATTERNS THAT INFLUENCE ACADEMIC DROPOUT AT THE LEONARDO DAVINCI SUPERIOR INSTITUTE USING IBM SPSS MODELER AND THE CRISP-DM METHODOLOGY"

By:

Br. Pando Cueva, Aurea Dajana
Br. Zarate Obeso, Winny Del Rosario

At present for educational institutions it has become a problem to know the patterns that influence academic dropout and thus try to reduce this number, becoming a headache for decision makers in educational institutions. That is why it is important to know why students decide to abandon their studies and what are the circumstances that lead to it.

The tools that allow creating a data mining model and more the analysis of the information of the student data that were provided by the computer systems of the Leonardo Davinci Higher Institute, has led us to create a data mining model that leads to obtain patterns that influence a dropout student. This model was implemented through the analysis of information: personal, academic, and student interaction.

To contribute to the solution to the problem of student dropout, it is proposed to develop a "Data Mining Model to identify patterns that influence Academic Dropout at the Leonardo Davinci Higher Institute" with the aim of knowing what the possible causes or patterns are that lead a student to abandon their studies, based on the analysis of the characteristics of student data..

ÍNDICE DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1. INTRODUCCION

1.1. PLANTEAMIENTO DEL PROBLEMA

Las empresas hoy en día generan gran cantidad de información, pero mucha de esta información almacenada no suele proporcionar algún beneficio directo para las empresas, el valor real de esta información se encuentra oculto y debemos de utilizar diferentes técnica para extraerla, es decir es posible encontrar mayor información que ayude a tomar decisiones o a mejorar la comprensión de lo que sucede dentro de las empresas.

En este contexto es que la Minería de Datos (MD) permite extraer información sensible que reside de manera implícita en los datos. La Minería de Datos está orientada a la exploración de los datos que una empresa posee y que puede ser un material muy importante en la búsqueda de conocimiento sobre cómo está evolucionando su negocio.

En la actualidad la deserción de los estudiantes se vuelve un problema muy grande que repercute en las finanzas de muchas instituciones educativas y tratar de reducir este número es un dolor de cabeza para los tomadores de decisiones de las Instituciones educativas. Es por ello que es importante conocer porque los estudiantes deciden abandonar sus estudios y cuáles son las diferentes circunstancias.

El rendimiento académico y la deserción estudiantil es una preocupación constante y para ello uno de los objetivos principales es determinar los patrones o factores que repercuten a ello. Particularmente en las instituciones educativas cada vez que los alumnos se inscriben a cursar o rendir una materia, cuando aprueban o reprueban la misma, se está generando gran cantidad de datos.

Lo que se busca con la Mineria de Datos es revelar conocimiento oculto útil y no evidente a partir de grandes BD. Desde la década pasada la MD se ha

ido incorporando a las organizaciones para constituirse en un apoyo esencial en el proceso de toma de decisiones.

Para contribuir con la solución al problema de la deserción estudiantil se plantea desarrollar un "Modelo de Minería de datos para identificación de patrones que influyen en la Deserción Académica en el Instituto Superior Leonardo Davinci" con el objetivo de entender cuáles son las causas en que un alumno decide desertar en sus estudios, basándose en un análisis de las características y datos de los estudiantes.

La toma de decisiones implementada a través de herramientas de minería de datos, contribuirá de gran manera a una mejor planeación en el área administrativa, docente y psicopedagógica, para tratar de evitar la deserción estudiantil y apoyar en todo momento al alumnado.

1.2. DELIMITACIÓN DEL PROBLEMA

El siguiente proyecto se realizará analizando la realidad en que se encuentra el área académica del Instituto Leonardo Davinci y las dificultades que este presenta para analizar los datos respecto a los factores que influyen en el desempeño académico que lleva a que los alumnos deserten de la institución.

1.3. FORMULACIÓN DEL PROBLEMA

¿Cómo identificar los patrones que influye en la deserción académica en el Instituto Superior Leonardo Davinci?

1.4. FORMULACIÓN DEL HIPÓTESIS

El desarrollo de un Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler permite conocer los patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci.

1.5. OBJETIVOS DEL ESTUDIO

El Objetivo general es:

Desarrollar un modelo de Minería de Datos para la Identificación de Patrones que influye en la deserción académica en el Instituto Superior Leonardo Davinci usando IBM SPSS Modeler y la Metodología Crisp-DM.

.

Los objetivos específicos son los siguientes:

- Identificar los problemas y determinar los requerimientos y necesidades del área mediante un análisis de modelo de negocio basada en la información obtenida de la institución educativa.
- Desarrollar el análisis y preparación de datos de los estudiantes de la institución obtenidos de los sistemas transaccionales
- Diseñar y Construir el modelo de predicción y búsqueda de patrones basados en las técnicas de modelado de minería de Datos utilizando IBM SPSS Modeler.
- Evaluar los resultados de los informes que muestran el modelo de la Minería de Datos implementada.

1.6. JUSTIFICACIÓN DEL ESTUDIO

1.6.1. Importancia de la investigación

- El proyecto basado en la minería de datos nos dará información relevante extraída de los datos de la institución, con el objetivo de descubrir patrones y tendencias estructurando la información obtenida de un modo comprensible para su posterior utilización.
- La presente investigación se justifica en la base a que la empresa mediante la aplicación de minería de datos mejoraría su proceso de toma de decisiones respecto a los alumnos que probablemente desertarían, teniendo un panorama veraz de la información basada en patrones de comportamientos e información de sus estudiantes.

1.6.2. Aportes

El desarrollo de este trabajo proporcionará considerables beneficios a la institución, como por ejemplo:

- Celeridad en la obtención de información de los perfiles de los estudiantes en la institución.
- Implementación de un modelo ajustado a la identificación de patrones que influyen en la deserción académica.
- La minería de datos va permitir descubrir patrones escondidos en la bases de datos, conociendo de antemano comportamientos futuros de los estudiantes.

1.6.3. Viabilidad de la investigación

- Es viable ya que se cuenta con herramientas básicas para poder desarrollar la investigación teniendo en cuenta su nivel de dificultad y aprendizaje por parte de los autores.
- Es viable ya que se ha planteado un cronograma para el desarrollo paso a paso de la investigación.

1.6.4. Limitaciones del estudio

Existen algunas limitaciones en el estudio:

- Como la información que maneja la institución es confidencial algunos datos fueron modificados.
- La institución solo brindará un rango de datos para la evaluación del modelo de minería de datos propuesto.
- No se podrá medir el impacto real del proyecto hasta que sea implantado en la institución educativa.
- La limitación más importante será la fidelidad y veracidad de los datos que nos proporcione la empresa.
- Otra limitación es el tiempo que se tiene para el desarrollo del proyecto (un año).

2. MARCO TEÓRICO

2.1. Antecedentes

- **Autor:** Saldaña Valqui, Edwin John

  Título de la investigación: "Modelo predictivo de Minería de datos de apoyo a la gestión hospitalaria sobre morbilidad de pacientes hospitalizados".

  Repositorio Académico Universidad Privada Antenor Orrego, 2015

  Descripción y análisis del trabajo:

  El presente trabajo de investigación, "propone aplicar un marco estándar de actividades de minería datos, creando un modelo predictivo, que sirva de apoyo a la Gestión Hospitalaria sobre la morbilidad con pacientes hospitalizados, basado en el algoritmo de análisis de serie de tiempo, Modelo ARIMA, con información histórica de los últimos 7 años de los pacientes del Hospital Víctor Ramos Guardia. En la investigación, se tomó como referencia la metodología CRISP-DM". Tambien en sus **objetivos específicos** "desarrollaron un analisis el estado del arte en modelos predictivos minería de datos de apoyo a la gestión hospitalaria, para luego Diseñar el proceso de preparación los datos y Aplicar la metodología Diseñando un Modelo de Minería de Datos". Por lo tanto se obtuvo como **resultado** "la extracción de los datos, transformación de los datos, carga de datos, limpieza de datos, diseño del datamart HEALTHMINING, la selección y creación de variables que sirvieron como datos de entrada para mi modelo, para posteriormente crear un modelo de pronósticos, que permitió conocer los casos de morbilidad en pacientes hospitalizados del hospital VRG para los próximos tres años".

- **Autores:** Mendoza Castillo, Sandra Jaqueline y Zavaleta Henrriquez, Fernando

  Título de la investigación: "Desarrollo de un Modelo de Minería de Datos para la Toma de Decisiones en la Gestión de Inventarios en Empresas Comerciales".

  Repositorio Académico Universidad Nacional de Trujillo, 2015

  Descripción y análisis del trabajo:

  El trabajo tiene como **objetivo** "desarrollar un modelo de minería de datos para mejorar el proceso de toma de decisiones en la gestión de inventarios en las empresas comerciales". Para ello se basó en los siguientes **objetivos específicos**: "Analizar las diferentes técnicas de la minería de datos, Diseñar un modelo de minería de datos, Implementar el modelo de minería de datos diseñado y luego Analizar la mejora en el proceso de la toma de decisiones en la gestión de inventarios al aplicar el modelo de minería de datos implementado". El trabajo obtuvo como **resultado** "la implementación de arboles de decisión apoyados de regresión lineal múltiple, proponiendo una arquitectura por niveles de manera que haya una separación entre los distintos procesos que involucra la minería de datos y de forma que la mayoría de la interacción ocurra únicamente entre niveles vecinos".

- **Autor:** Azabache Santa María Araceli y Figueroa Gutiérrez Joshua

  Título de la investigación: "Diseño de un algoritmo basado en técnicas de minería de datos para el eficiente abastecimiento del inventario en las MYPES".

  Repositorio Académico Universidad Nacional de Trujillo, 2015

  Descripción y análisis del trabajo:

  El presente trabajo tiene como **objetivo general** "diseñar un algoritmo basado en técnicas de minería de datos" y, para su cumplimiento se estableció el estudio en **objetivos específicos** que "comprenden el proceso de minería, el análisis de la técnica de minería de datos regresión lineal aplicada a los pronósticos de ventas para planificar el

abastecimiento del inventario en MYPES, el análisis y determinación de una técnica de diseño de algoritmo, que permita obtener un algoritmo de pronóstico soportado en minería de datos". El trabajo obtuvo como **resultados** que "la mejor técnica de minería de datos que se adapta para este pronóstico es la regresión lineal múltiple y logrando realizar el diseño de un algoritmo basado en técnicas de minería de datos con el fin de abastecer de manera eficiente el inventario en Mype comercializadores de artículos de primera necesidad".

- **Autor:** Lázaro Rodríguez, Stefhanny y Moreno Chávez, Irvin
  Título de la investigación: "Aplicación de técnicas de minería de datos para la predicción de Clientes rentables en las pymes de Trujillo". Repositorio Académico Universidad Nacional de Trujillo, 2014
  Descripción y análisis del trabajo:
  En el presente trabajo tiene como **objetivo "**Predecir clientes rentables para las pymes de Trujillo mediante la aplicación de técnicas de minería de datos", en donde definió sus **objetivos específicos** donde determino que "primero debe de Analizar las técnicas de minería de datos para predicción de datos, Comprender el negocio y los datos, Construir el modelo de minería de datos para predecir clientes rentables y luego Evaluar el modelo de minería de datos". Finalmente el trabajo obtuvo como **resultados** que al "aplicar Minería de datos ayuda a la economía de la empresa, ya que al identificar clientes rentables para la empresa y la captación de estos generaran aumentos en los ingresos y en definitiva los beneficios, permitiendo así el crecimiento progresivo de la empresa".

- **Autor:** Zoraida Emperatriz Mamani Rodríguez
  Título de la investigación: "Aplicación de la Minería de Datos Distribuida usando Algoritmo de Clustering K-Means para mejorar la calidad de servicios de las organizaciones modernas". Repositorio Académico UNMSM, Lima 2015

Descripción y análisis del trabajo:

El presente trabajo tiene como **objetivo** principal "implementar una aplicación de la Minería de Datos Distribuida usando Algoritmo de Clustering K-Means para mejorar la calidad de servicios de las organizaciones modernas", teniendo como **objetivos específicos** "realizar una revisión bibliográfica de las técnicas clustering k-means, elabora una propuesta concreta, desarrolla un prototipo de aplicación. La propuesta se centra básicamente en la detección de patrones de comportamiento basado en un proceso de negocio particular correspondiente a una organización del sector judicial para lo cual aplica la minería de datos distribuida debido a su naturaleza física". El trabajo obtuvo como **resultados** "beneficios con su implementación, fortaleciendo mecanismos para la reducción del volumen procesal, plazos procesales y nivel de litigiosidad, con lo cual se lograría una mejora en la calidad de los servicios que esta brinda a los ciudadanos".

2.2. DEFINICIONES

2.2.1. TOMA DE DECISIONES:

Toma de decisiones. Es el proceso por medio del cual se obtiene como resultado una o más decisiones con el propósito de dar solución a una situación. Pueden participar uno o más actores y se elige entre varias alternativas. (Ecured, 2018)

La toma de decisiones es una capacidad netamente humana, deriva del poder de la razón y el poder de la voluntad, es decir, pensamiento y querer unidos en la misma dirección. (Webyempresas, 2018)

Decidir significa hacer que las cosas sucedan en vez de simplemente dejar que ocurran como consecuencia del azar u otros factores externos.

Esta habilidad ofrece a las personas herramientas para evaluar las diferentes posibilidades, teniendo en cuenta, necesidades, valores, motivaciones, influencias y posibles consecuencias presentes y futuras. Esta competencia se relaciona con la capacidad de tomar riesgos pero difiere en que no siempre las decisiones implican necesariamente un riesgo o probabilidad de fracaso, sino dos vías diferenciales y alternativas de acción para resolver un problema. (Universidad de Cadiz, 2017)

2.2.2. PROCESO DE TOMA DE DECISIONES:

"El proceso de la toma de decisiones en una organización comienza con la detección de una situación que rodea algún problema. Seguidamente viene el análisis y la definición del problema. Para ello se requiere contar con un sistema de información confiable, oportuno, y actualizado, que permitan comprender claramente la naturaleza del problema a resolver". (ConexionEsan, 2016)

También es necesario "conocer los factores internos formales e informales de la organización, como son la cultura, organizaciones, manuales, políticas, estructura, recursos disponibles, etc. y los informales como las políticas implícitas, los hábitos, la experiencia, etc. A ello se añade el conocimiento de los factores externos de la organización: clientes, proveedores, economía, competencia, entre otros. Es preciso también elegir las técnicas o herramientas a utilizar. A cada problema específico le corresponde una combinación de metodologías para abordarlo, comprenderlo y resolverlo". (ConexionEsan, 2016)

Otro factor clave es "la evaluación y establecimiento del costo-beneficio que tendría la decisión a tomar. Se debe especificar los rendimientos esperados que justifiquen la decisión a tomar. En ese mismo sentido se debe evaluar las posibles consecuencias". (ConexionEsan, 2016)

Igualmente “es importante especificar los objetivos y las metas esperadas. Tomar una decisión "por tomarla" no es adecuado. Todo debe tener un fin. Luego viene la búsqueda de las opciones más adecuadas para alcanzar los objetivos. Esas opciones deberán ser evaluadas y comparadas entre sí, con el fin de escoger la que mejor se ajuste a las necesidades de la organización en términos de costo-beneficio y de cumplimiento de las metas y objetivos trazados”. (ConexionEsan, 2016)

“Finalmente, tenemos la implementación de la opción elegida y su ulterior evaluación”. (ConexionEsan, 2016)

2.2.3. MINERÍA DE DATOS

El datamining (minería de datos), (Sinexus, 2017) “es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto”.

Básicamente, “el datamining surge para intentar ayudar 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”. (Sinexus, 2017)

De forma general, “los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.Los datos que vemos son sólo la punta del iceberg. Aunque en datamining cada caso concreto puede ser

radicalmente distinto al anterior, el proceso común a todos ellos se suele componer de cuatro etapas principales" (ConexionEsan, 2016):

- ✓ "Determinación de los objetivos. Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data Mining". (ConexionEsan, 2016)
- ✓ "Preprocesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data Mining". (ConexionEsan, 2016)
- ✓ "Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial". (ConexionEsan, 2016)
- ✓ "Análisis de los resultados. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones". (ConexionEsan, 2016)
- ✓ "Esfuerzo en cada etapa del datamining y Carga de trabajo en las fases de un proyecto de datamining" (ConexionEsan, 2016)

En resumen, "el datamining se presenta como una tecnología emergente, con varias ventajas: por un lado, resulta un buen punto de encuentro entre los investigadores y las personas de negocios; por otro, ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. Además, no hay duda de que trabajar con esta tecnología implica cuidar un sinnúmero de detalles debido a que

el producto final involucra toma de decisiones". (Sinexus, 2017)

(Pérez López, 2007) "Data Mining (minería de datos) es el proceso de extracción de información significativa de grandes bases de datos, información que revela inteligencia del negocio, a través de factores ocultos, tendencias y correlaciones para permitir al usuario realizar predicciones que resuelven problemas del negocio proporcionando una ventaja competitiva. Las herramientas de Data Mining predicen las nuevas perspectivas y pronostican la situación futura de la empresa, esto ayuda a los mismos a tomar decisiones de negocios proactivamente".

La minería de datos, Data Mining, "es un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining. Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos". (Sinnexus, 2016)

La minería de datos "está incluida en un proceso mayor denominado Descubrimiento de Conocimientos en Base de Datos, Knowledge Discovery in Database (KDD). Rigurosamente el Data Mining se restringe a la obtención de modelos, restando las etapas anteriores y el propio Data Mining como instancias del KDD". (Sinnexus, 2016)

2.2.4. SISTEMA DE APOYO A LA TOMAS DE DECISIONES:

Un sistema de apoyo a la toma de decisiones o de soporte a la decisión (DSS por sus siglas en inglés) “es un sistema basado en ordenadores destinado a ser utilizado por un gerente particular o por un grupo de gerentes a cualquier nivel organizacional para tomar una decisión en el proceso de resolver una problemática semiestructurada. Los sistemas de apoyo a la toma de decisiones son un tipo de sistema computarizado de información organizacional que ayuda al gerente en la toma de decisiones cuando necesita modelar, formular, calcular, comparar, seleccionar la mejor opción o predecir los escenarios”. (Transformacion Digital, 2017)

Los sistemas de apoyo a la toma de decisiones “están específicamente diseñados para ayudar al equipo directivo a tomar decisiones en situaciones en las que existe incertidumbre sobre los posibles resultados o consecuencias. Ayuda a los gerentes a tomar decisiones complejas”. (Transformacion Digital, 2017)

Un DSS, en términos muy generales, es “un sistema basado en computador que ayuda en el proceso de toma de decisiones” (Finlay, 1994).

En términos bastante más específicos, un DSS es “un sistema de información basado en un computador interactivo, flexible y adaptable, especialmente desarrollado para apoyar la solución de un problema de gestión no estructurado para mejorar la toma de decisiones. Utiliza datos, proporciona una interfaz amigable y permite la toma de decisiones en el propio análisis de la situación” (Turban, 2005)

Otras definiciones intermedias entre las dos anteriores serían:

- Un DSS es un "conjunto de procedimientos basados en modelos para procesar datos y juicios para asistir a un gerente en si toma de decisiones" (Sinexus, 2016)
- Un DSS "combina recursos intelectuales individuales con las capacidades de un ordenador para mejorar la calidad de las decisiones (son un apoyo informático para los encargados de tomar decisiones sobre problemas semiestructurados)" (Microsoft, 2016)

2.2.5. BUSINESS INTELLIGENCE (BI):

"Es una estrategia empresarial que persigue incrementar el rendimiento de la empresa o la competitividad del negocio, a través de la organización inteligente de sus datos históricos (transacciones u operaciones diarias), usualmente residiendo en Data Warehouse corporativos o Data Marts departamentales". (Sinnexus, 2016)

De forma general, el BI suele definirse "como la transformación de datos de la compañia en conocimiento para obtener una ventaja competitiva. Si lo asociamos directamente a las tecnologías de la información, podemos definir Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañia) en información estructurada para su explotación directa (reporting, análisis OLAP, mineria de datos, etc.) o para su análisis y conversión en conocimiento como soporte a la toma de decisiones sobre el negocio". (Dataprix, 2010)

2.2.6. SISTEMAS DE INFORMACIÓN

Un sistema de información "se puede definir técnicamente como un conjunto de componentes relacionados que recolectan (o recuperan), procesan, almacenan y distribuyen información para apoyar la toma de decisiones y el control en una organización". (Itson, 2018)

Actividades de un Sistema de Información

Hay tres actividades en un sistema de información que "producen la información que esas organizaciones necesitan para tomar decisiones, controlar operaciones, analizar problemas y crear nuevos productos o servicios. Estas actividades son": (Itson, 2018)

- ✓ Entrada: captura o recolecta datos en bruto tanto del interior de la organización como de su entorno externo.
- ✓ Procesamiento: convierte esa entrada de datos en una forma más significativa.
- ✓ Salida: transfiere la información procesada a la gente que la usará o a las actividades para las que se utilizará.

El objetivo de un sistema de información "es ayudar al desempeño de las actividades que desarrolla la empresa, suministrando la información adecuada, con la calidad requerida, a la persona o departamento que lo solicita, en el momento y lugar especificados con el formato más útil para el receptor". (Kendall & Kendall, 2005)

2.2.7. BENCHMARKING DE HERRAMIENTAS DE MINERIA DE DATOS

Se realizó un benchmarking de herramientas de minería de datos las cuales lo presentaremos en una tabla, en donde las comparamos y verificamos que la herramienta de IBM SPSS Modeler es una buena alternativa para estas soluciones. Para esta comparación se tomaron en cuenta las siguientes variables:

- ✓ Tipo de Software
- ✓ Conexión a la nube
- ✓ Text Analytics
- ✓ Social Network Analytics
- ✓ Big Data
- ✓ Integración lenguaje R
- ✓ Rendimiento
- ✓ Reportes Mobile
- ✓ Tipo de Aplicación
- ✓ Procesamiento de Datos
- ✓ Ejecución de modelos

Herramienta / Característica	Tipo de Software	Conexión a la nube	Text Analytics	Social Network Analytics	Big Data	Integración lenguaje R	Rendimiento	Reportes Mobile	Tipo de Aplicación	Procesamiento de Datos	Ejecución de Modelos
SAP Predictive Analytics	Propietario	Por medio de SAP HANA	x	x	x	x	Capaz de ejecutar modelos en la nube	x	Desktop Cliente-Servidor	▪ Hadoop ▪ In-Database ▪ In-Memory	▪ In-Database ▪ In-Memory ▪ Real-Time
IBM SPSS Predictive	Propietario	Por medio de SPSS Analytics Server (Paquete Gold)	x	x	x		Capaz de ejecutar modelos en la nube	x	Desktop Cliente-Servidor	▪ Hadoop Distribution ▪ In-Database	▪ Batch ▪ In-Memory ▪ Real-Time
SAS Predictive Analytics	Propietario		x		x		De acuerdo a la RAM	x	Online Cliente-Servidor	▪ Hadoop ▪ In-Database	▪ In-Database ▪ In-Memory
RapidMiner	Propietario		x	x		x	De acuerdo a la RAM		Desktop Cliente-Servidor	▪ In-Database ▪ In-Memory	▪ In-Memory ▪ Real-Time
Angoss Predictive Analytics	Propietario	Por medio de FunGUARD (ventas) y ClaimGUARD (fraudes)			x	x	De acuerdo a la RAM	x	Online Cliente-Servidor	▪ Hadoop ▪ In-Database ▪ In-Memory	▪ In-Database ▪ In-Memory
GraphLab Create	Libre		x				De acuerdo a la RAM		Desktop	▪ In-Database ▪ In-Memory	▪ In-Memory ▪ Real-Time
TIBCO - Spotfire Predictive Analytics	Propietario	Por medio de Spotfire Cloud	x	x		x	De acuerdo a la RAM		Online Cliente-Servidor	▪ Hadoop ▪ In-Database	▪ In-Database ▪ In-Memory
Weka Data Mining	Libre		x		x		De acuerdo a la RAM		Desktop	▪ In-Database ▪ In-Memory	▪ In-Memory ▪ Real-Time

Tabla 01. Benchmarking de Herramientas de Minería de Datos

2.2.8. IBM SPSS MODELER

(IBM, 2017) “IBM SPSS Modeler es una plataforma de Análisis Predictivo diseñada para aportar inteligencia predictiva a las decisiones de negocio. Utiliza una amplia gama de técnicas predictivas y descriptivas para mostrar los patrones y tendencias de sus datos. Esta información ayuda a mejorar los procesos actuales y tomar decisiones que influyan de forma positiva en su negocio”.

IBM SPSS Modeler “descubre tendencias y patrones ocultos en los datos. Sólo con IBM SPSS Modeler puede acceder directamente y de forma sencilla a datos de texto, datos web y datos de encuesta”. (IBM, 2017)

IBM SPSS Modeler “se integra en la infraestructura tecnológica existente de su organización y ofrece opciones de distribución flexibles para garantizar que tendrá a su disposición información predictiva precisa dónde y cuándo se necesite”. (IBM, 2017)

IBM SPSS Modeler “es compatible con una amplia gama de bases de datos, hojas de cálculo y archivos planos, entre los que se incluyen archivos SPSS Statistics Base, SAS y Microsoft Excel y una amplia gama de plataformas, de modo que puede aprovechar todos los datos y obtener mejores resultados”. (IBM, 2017)

La arquitectura abierta de IBM SPSS Modeler le “permite acceder a datos y distribuir modelos, predicciones e información a los responsables de la toma de decisiones y sistemas operativos automatizados”. (IBM, 2017)

IBM SPSS Modeler está diseñado para:

Acelerar las tareas de Data Mining

La pionera interfaz gráfica de IBM SPSS Modeler "facilita que los analistas se centren en los problemas empresariales sin perder tiempo en tareas de programación más rutinarias". (IBM, 2017)

Al liberar a los analistas de tareas técnicas no productivas, IBM SPSS Modeler permite que se concentren en la búsqueda de respuestas a sus problemas de negocio, de modo que puedan obtener y distribuir resultados con mayor rapidez y repercusión.

Mejorar las decisiones y los resultados

- "Construir modelos predictivos con una amplia gama de algoritmos avanzados". (IBM, 2017)
- "Combinar modelos predictivos, reglas de negocio y técnicas de optimización para mejorar la toma de decisiones". (IBM, 2017)
- "Ofrecer recomendaciones a personas y sistemas, que redundan en una mejora de las decisiones y las acciones". (IBM, 2017)
- "Integrar resultados analíticos en procesos empresariales existentes y aplicaciones operativas". (IBM, 2017)

Extraer valor de los datos

- "Descubrir información y modelos atrapados en datos con algoritmos estadísticos y análisis de texto" (IBM, 2017).
- "Analizar no sólo los datos estructurados, como la edad, el precio, el producto, la ubicación, etc., sino también los datos no estructurados, como texto, correos electrónicos, datos de medio de comunicación social, etc" (IBM, 2017)

Integrarse de forma más sencilla en los sistemas existentes

- "Utilizar con bases de datos de IBM o bases de datos de otros proveedores para desarrollar e implementar modelos con una mayor velocidad y eficiencia". (IBM, 2017)
- "Habilitar un flujo de trabajo dinámico a partir de la integración con IBM SPSS Statistics, Cognos Business Intelligence, Cognos TM1 e InfoSphere Streams". (IBM, 2017)
- "Minimizar el movimiento de datos y mejorar el rendimiento con las versiones del servidor que permiten la funcionalidad en IBM Pure Data Systems, InfoSphere Warehouse, IBM DB2 y Linux en IBM System z" (IBM, 2017).

2.3.METODOLOGIA PARA EL DESARROLLO DEL PROYECTO

2.3.1. Metodología Crisp-DM

La metodología CRISP, está dividida en 4 niveles de forma jerárquica, en tareas que van desde el nivel más general, hasta los casos más específicos como se muestran a continuación (Smartbase Group, 2016):

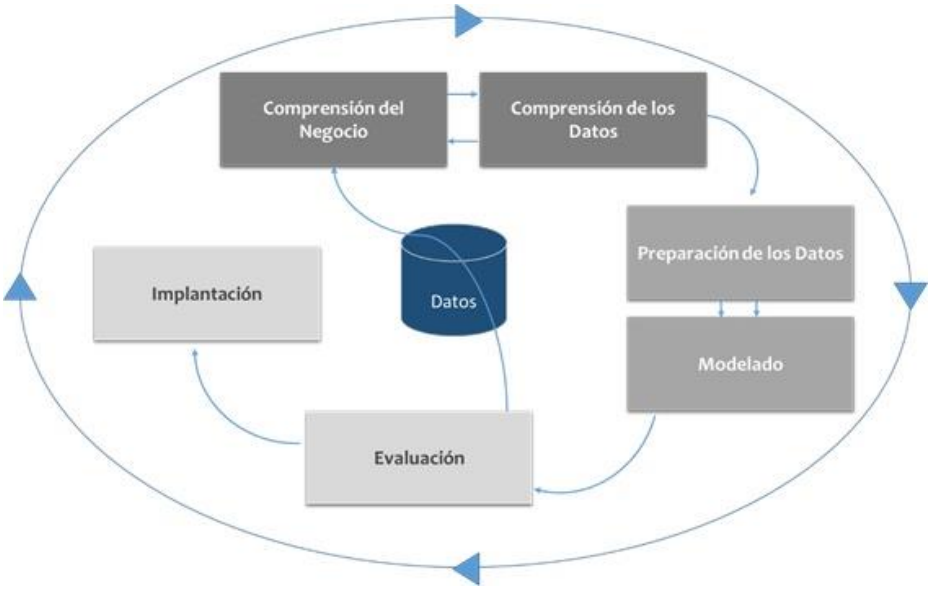


Figura 01: Modelo de proceso de la Metodología CRISP–DM

Fuente: (Smartbase Group, 2016)

La sucesión de fases no es necesariamente rígida.

Fase 1: Comprensión del negocio o problema

"La primera fase de la guía de referencia de la Metodología CRISP-DM , denominada fase de comprensión del negocio o problema.

Se realizan las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, debido a que convierte en objetivos técnicos y en un plan de proyecto" (Smartbase Group, 2016).

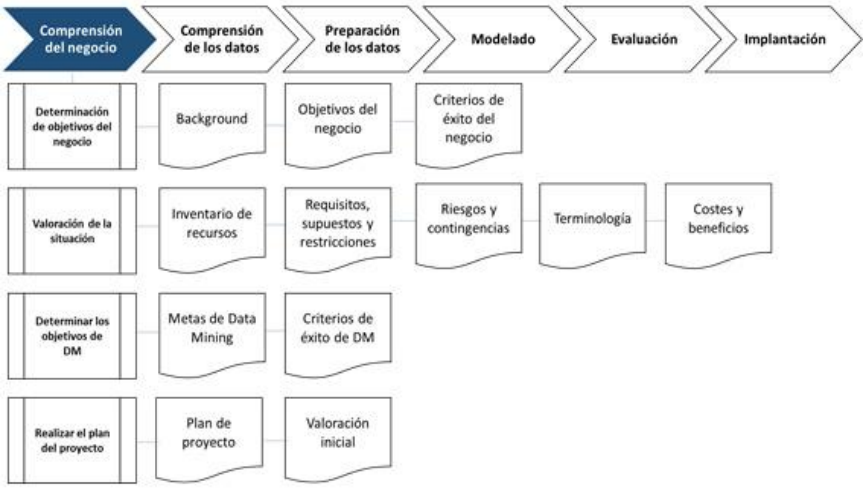


Figura 02: Fase de compresión del negocio o problema

Fuente: (Smartbase Group, 2016)

Fase 2 Comprensión de los datos

"La fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema" (Smartbase Group, 2016).

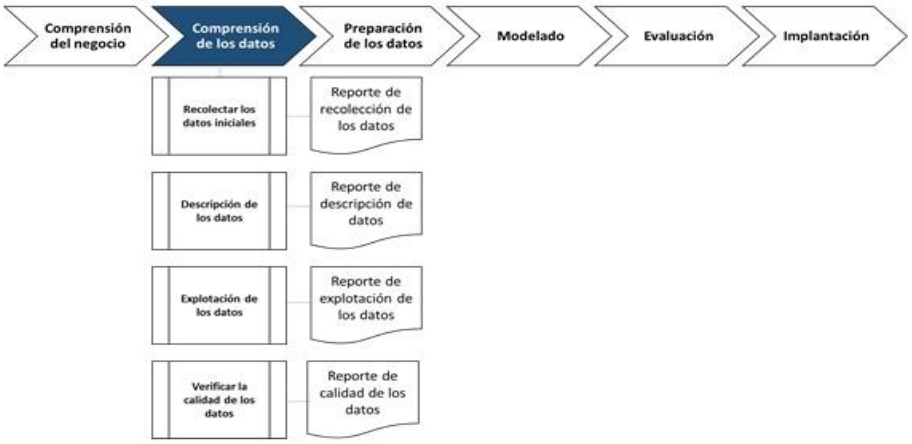


Figura 03: Fase de comprensión de los datos.

Fuente: (Smartbase Group, 2016)

Fase 3: Preparación de los datos

"En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente" (Smartbase Group, 2016).

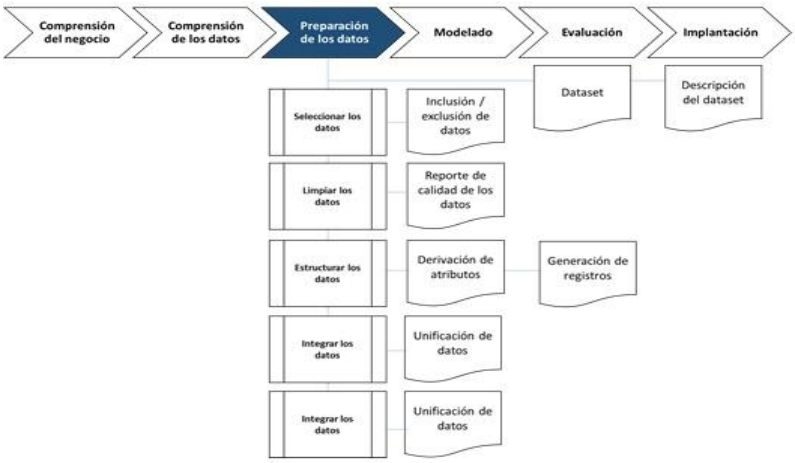


Figura 04: Fase de preparación de los datos.

Fuente: (Smartbase Group, 2016)

Fase 4: Modelado

"En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico" (Smartbase Group, 2016).

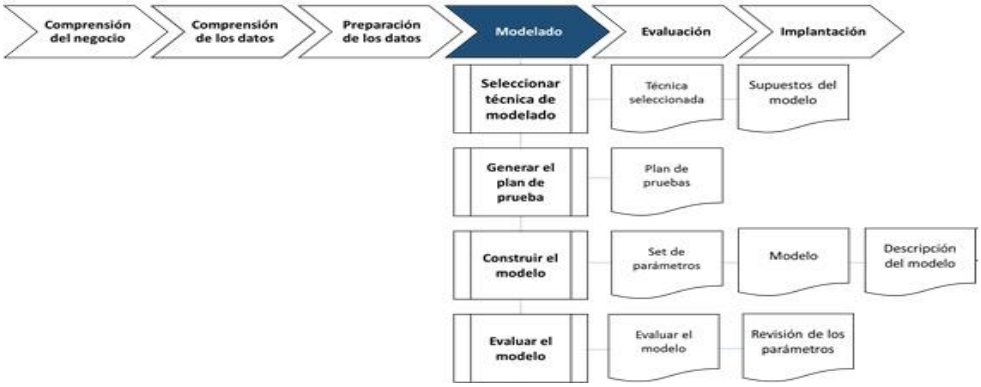


Figura 05: Fase del modelado.

Fuente: (Smartbase Group, 2016)

Fase 5: de evaluación

"En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema" (Smartbase Group, 2016).

Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados.

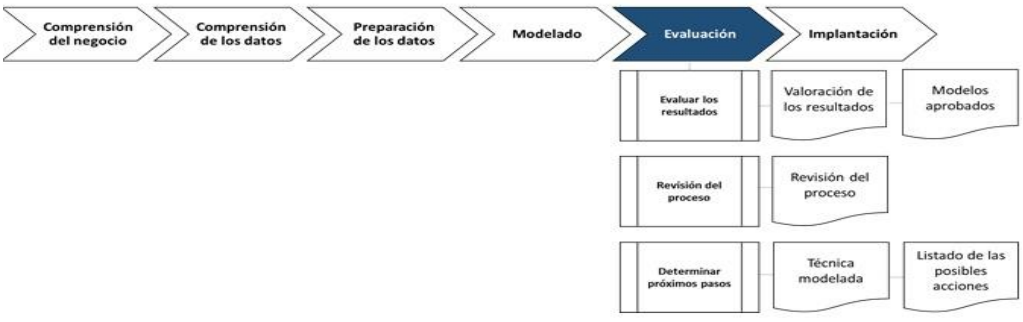


Figura 06: Fase de evaluación

Fuente: (Smartbase Group, 2016)

Fase 6: Implementación

"En esta fase , y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso" (Smartbase Group, 2016).

Las tareas que se ejecutan en esta fase son las siguientes:

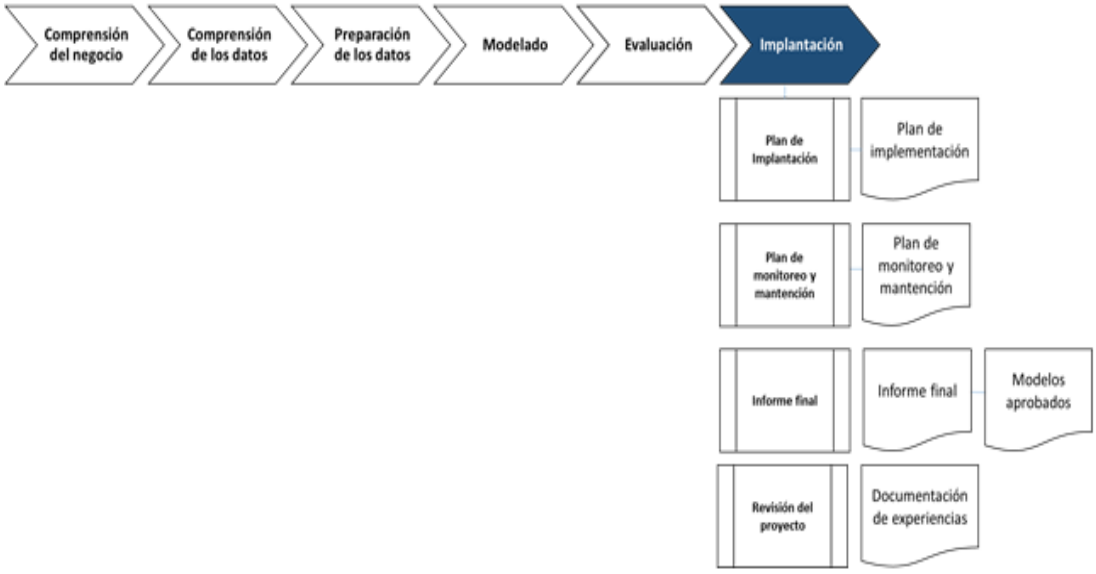


Figura 07: Fase de implementación

Fuente: (Smartbase Group, 2016)

3. MATERIALES Y METODOS

3.1.Material

3.1.1. Población

Todos los estudiantes del Instituto Superior Leonardo Davinci.

3.1.2. Muestra

Registros de estudiantes de los años 2018 -2019 almacenados en la base de datos del Instituto Superior Leonardo Davinci.

3.1.3. Unidad de Análisis

Registros de los Datos de los Estudiantes

3.2.Método

3.2.1. Método de la Investigación:

Se utilizará el **método hipotético - deductivo**

Nivel de Investigación:

Deductivo – Inductivo.

Diseño Investigación:

Diseño Pre-experimental con pre-prueba y post-prueba

Este diseño se presenta de la siguiente manera:

Diseño del modelo pre-experimental	**G -> O_1 -> X -> O_2**
G (Grupo a investigar)	Datos de estudiantes del instituto.
X (Tratamiento)	Aplicación del Modelo
O (Observación)	O_1 : Observación pre-test
	O_2: Observación post-test

3.2.2. Variables de estudio y Operacionalización

- Independiente (VI): Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler.
- Dependiente (VD): Identificación de patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci.

3.2.3. Operacionalización de las variables

Variable	Dimensión	Indicador	Unidad de medida	Instrumento de Investigación
VI	Precisión	Porcentaje de precisión del modelo	% Precisión	Hoja de captura de datos
VD	Satisfacción	Grado de Satisfacción	Rango de satisfacción	Tabla de satisfacción

Tabla 01.variables de estudio y Operacionalización

3.2.4. Instrumentos de recolección de Datos

- Búsqueda de información:
  - Ficha bibliográfica.
  - Hoja resumen.
- Entrevista:
  - Guion de entrevista.
  - Hoja resumen.
- Observación:
  - Guion de Observación.

3.2.5. Procedimientos y análisis de datos

3.2.5.1.Procesamiento de datos

El recojo de datos se realizará a través de los cuestionarios, siendo una fase esencial para toda la investigación, referida a la clasificación o agrupación de los datos referentes a cada variable objetivo de estudio y su presentación conjunta.

Los resultados se presentan mediante ecuaciones, gráficos y tablas para su interpretación.

3.2.5.2.Análisis de datos

El análisis de los datos se llevará a cabo por medio de cuadros estadísticos descriptivos (Pruebas hipótesis nula y alternativa y las Pruebas Distribución normal (Prueba Z), ya que la información obtenida será analizada y mostrada por medio de cuadros y gráficos.

4. RESULTADOS : APLICACIÓN DE LA METODOLOGIA

4.1.ANALISIS DEL PROBLEMA

4.1.1. Objetivos Institucionales:

El Grupo Educativo Leonardo Da Vinci fue fundado en 1992 por el Dr. Alberto Escudero, con el objetivo de brindar a miles de jóvenes peruanos las herramientas de formación superior necesarias para trabajar y progresar en corto tiempo, potenciando al máximo el poder de sus alumnos para alcanzar el éxito

Otro de los objetivos institucionales es disminuir la problemática descrita "deserción estudiantil", por lo que, los objetivos esenciales perseguidos para este trabajo se presentan a manera de un modelo de minería de datos basado en la metodología CRISP. Esto permitirá:

- Mejorar el proceso de toma de Decisiones del área académica.
- Conocer los patrones que influyen en la deserción académica del Instituto Leonardo Davinci.

4.1.2. Evaluación de la Situación:

- En la actualidad en la institución educativa "Leonardo Davinci", se están preparando informes para el soporte de tomas de decisiones por parte de la dirección y del jefe del área de sistemas de los cuales no brindan la información necesaria o importante para dar este soporte.
- Al no contar con información oportuna o suficiente sobre los estudiantes no es posible mejorar algunos factores dentro del área académica y de esa manera brindar un mejor servicio, además estos reportes deberían contener más datos exactos y con una mejor

presentación gráfica en Excel, pero en muchos casos estos se reprocesan con la ayuda del personal de sistemas o por el propio jefe del área.

- El tomador de decisión en la institución necesita toda la información que se pueda obtener desde la base de datos y más aún una clasificación de sus estudiantes y conocer que factores influyen en él y que lleva a una deserción académica.

4.1.3. Recursos Computacionales:

Hardware y Software:	
Resumen Software:	
Sistema operativo:	Microsoft Windows 10 (x64)
Ofimática:	Microsoft Office 2016
Herramienta de Minería:	IBM SPSS Modeler
Resumen Hardware:	
Procesador:	Intel Core i7-960
Memoria:	16 GBytes
Unidad de disco duro:	Seagate 1 Tb
Unidad óptica:	HL-DT-ST BD-RE BH10LS30
Adaptador de pantalla:	GIGABYTE GeForce GT 220
Monitor:	LG E2240

4.2. ANALISIS DE DATOS

4.2.1. Recolección de Datos Iniciales:

- Los datos de los estudiantes fueron obtenidos desde la base de datos de la institución educativa que está en un servidor en MS SQL Server.

- Por motivos de confidencialidad de información de sus estudiantes, que se proporcionó fue una parte de la información del año 2018-2019 y en algunos datos se realizaron modificaciones.

- Datos seleccionados para esta investigación:

TABLA ESTUDIANTE
✓ Nivel_Socioeconomico
✓ Ocupacion_Padre
✓ Ocupacion_Madre
✓ Tipo_residencia
✓ Sexo
✓ Estado_Civil
✓ Tipo_colegio
✓ Edad_Ingreso
✓ Vive_con_familia
✓ Tiene_Hermanos
✓ Ingresos_familiares
✓ Horas_dedicas_Estudiar

TABLA NOTAS
✓ Promedio_curso_anterior
✓ Cursos_desaprobados

TABLA ESPECIALIDAD
✓ Especialidad de estudio

TABLA MATRICULA
✓ Turno
✓ Ciclo

TABLA ASISTENCIAS
✓ N°Inasistencias
✓ Acceso_Plataforma_Virtual

TABLA ENCUESTA
✓ Grado_satisfaccion

4.2.2. Descripción de los Datos:

Datos seleccionados para el análisis:

Tabla 02: Descripción de los Datos a utilizar.

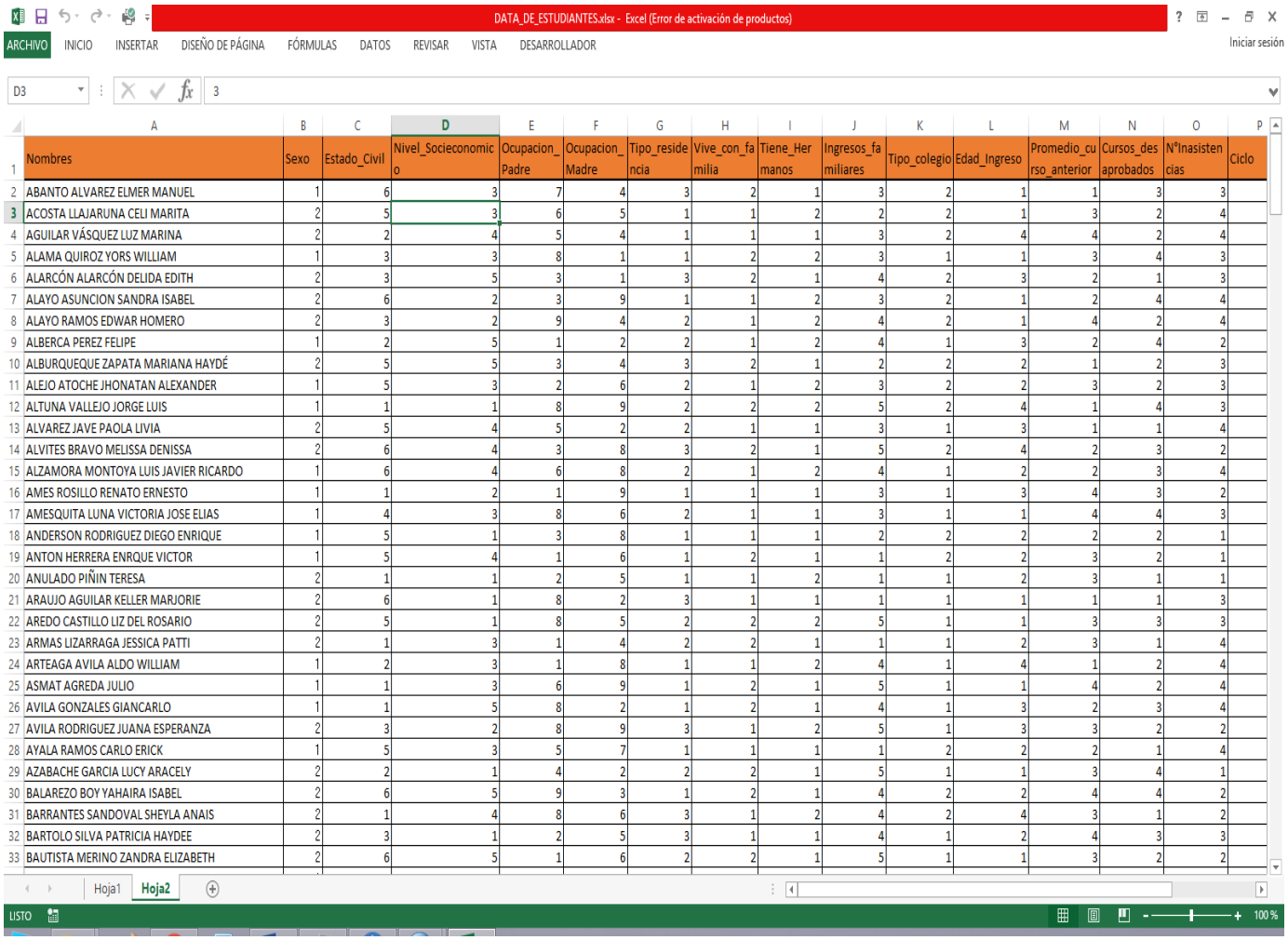
N°	Tabla	Campo
1	Estudiante	Nivel_Socioeconomico
2	Estudiante	Ocupacion_Padre
3	Estudiante	Ocupacion_Madre

4	Estudiante	Tipo_residencia
5	Estudiante	Sexo
6	Estudiante	Estado_Civil
7	Estudiante	Tipo_colegio
8	Estudiante	Edad_Ingreso
9	Estudiante	Vive_con_familia
10	Estudiante	Tiene_Hermanos
11	Estudiante	Ingresos_familiares
12	Estudiante	Horas_dedicas_Estudiar
13	Notas	Promedio_curso_anterior
14	Notas	Cursos_desaprobados
15	Especialidad	Especialidad
16	Asistencias	N°Inasistencias
17	Asistencias	Acceso_plataforma_virtual
18	Matricula	Turno
19	Matricula	Ciclo
20	Encuesta	Grado_satisfaccion

4.2.3. Exploración y Validación de los datos:

Se listan las circunstancias encontradas al examinar y validar los datos.

4.2.3.1.Datos a Analizar: Registros de los campos a utilizar (20 campos, 599 registros)



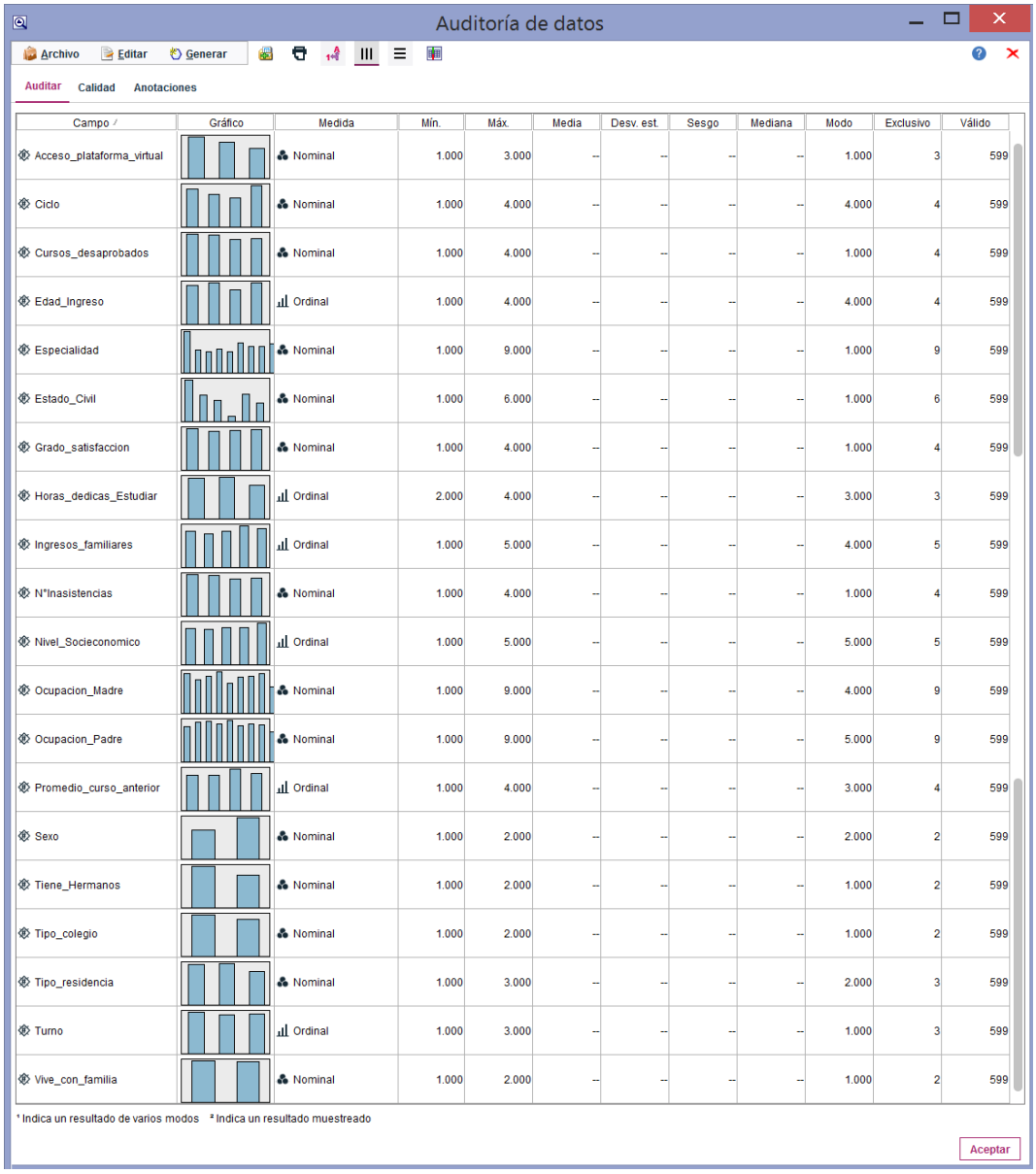
DATA_DE_ESTUDIANTES.xlsx - Excel (Error de activación de productos)

	Nombres	Sexo	Estado_Civil	Nivel_Socieconomico	Ocupacion_Padre	Ocupacion_Madre	Tipo_residencia	Vive_con_familia	Tiene_Hermanos	Ingresos_familiares	Tipo_colegio	Edad_Ingreso	Promedio_curso_anterior	Cursos_desaprobados	N°Inasistencias	Ciclo
2	ABANTO ALVAREZ ELMER MANUEL	1	6	3	7	4	3	2	1	3	2	1	1	3	3	
3	ACOSTA LLAJARUNA CELI MARITA	2	5	3	6	5	1	1	2	2	2	1	3	2	4	
4	AGUILAR VÁSQUEZ LUZ MARINA	2	2	4	5	4	1	1	1	3	2	4	4	2	4	
5	ALAMA QUIROZ YORS WILLIAM	1	3	3	8	1	1	2	2	3	1	1	3	4	3	
6	ALARCÓN ALARCÓN DELIDA EDITH	2	3	5	3	1	3	2	1	4	2	3	2	1	3	
7	ALAYO ASUNCION SANDRA ISABEL	2	6	2	3	9	1	1	2	3	2	1	2	4	4	
8	ALAYO RAMOS EDWAR HOMERO	2	3	2	9	4	2	1	2	4	2	1	4	2	4	
9	ALBERCA PEREZ FELIPE	1	2	5	1	2	2	1	2	4	1	3	2	4	2	
10	ALBURQUEQUE ZAPATA MARIANA HAYDÉ	2	5	5	3	4	3	2	1	2	2	2	1	2	3	
11	ALEJO ATOCHE JHONATAN ALEXANDER	1	5	3	2	6	2	1	2	3	2	2	3	2	3	
12	ALTUNA VALLEJO JORGE LUIS	1	1	1	8	9	2	2	2	5	2	4	1	4	3	
13	ALVAREZ JAVE PAOLA LIVIA	2	5	4	5	2	2	1	1	3	1	3	1	1	4	
14	ALVITES BRAVO MELISSA DENISSA	2	6	4	3	8	3	2	1	5	2	4	2	3	2	
15	ALZAMORA MONTOYA LUIS JAVIER RICARDO	1	6	4	6	8	2	1	2	4	1	2	2	3	4	
16	AMES ROSILLO RENATO ERNESTO	1	1	2	1	9	1	1	1	3	1	3	4	3	2	
17	AMESQUITA LUNA VICTORIA JOSE ELIAS	1	4	3	8	6	2	1	1	3	1	1	4	4	3	
18	ANDERSON RODRIGUEZ DIEGO ENRIQUE	1	5	1	3	8	1	1	1	2	2	2	2	2	1	
19	ANTON HERRERA ENRQUE VICTOR	1	5	4	1	6	1	2	1	1	2	2	3	2	1	
20	ANULADO PIÑIN TERESA	2	1	1	2	5	1	1	2	1	1	2	3	1	1	
21	ARAUJO AGUILAR KELLER MARJORIE	2	6	1	8	2	3	1	1	1	1	1	1	1	3	
22	AREDO CASTILLO LIZ DEL ROSARIO	2	5	1	8	5	2	2	2	5	1	1	3	3	3	
23	ARMAS LIZARRAGA JESSICA PATTI	2	1	3	1	4	2	2	1	1	1	2	3	1	4	
24	ARTEAGA AVILA ALDO WILLIAM	1	2	3	1	8	1	1	2	4	1	4	1	2	4	
25	ASMAT AGREDA JULIO	1	1	3	6	9	1	2	1	5	1	1	4	2	4	
26	AVILA GONZALES GIANCARLO	1	1	5	8	2	1	2	1	4	1	3	2	3	4	
27	AVILA RODRIGUEZ JUANA ESPERANZA	2	3	2	8	9	3	1	2	5	1	3	3	2	2	
28	AYALA RAMOS CARLO ERICK	1	5	3	5	7	1	1	1	1	2	2	2	1	4	
29	AZABACHE GARCIA LUCY ARACELY	2	2	1	4	2	2	2	1	5	1	1	3	4	1	
30	BALAREZO BOY YAHAIRA ISABEL	2	6	5	9	3	1	2	1	4	2	2	4	4	2	
31	BARRANTES SANDOVAL SHEYLA ANAIS	2	1	4	8	6	3	1	2	4	2	4	3	1	2	
32	BARTOLO SILVA PATRICIA HAYDEE	2	3	1	2	5	3	1	1	4	1	2	4	3	3	
33	BAUTISTA MERINO ZANDRA ELIZABETH	2	6	5	1	6	2	2	1	5	1	1	3	2	2	

La Validación de datos con la herramienta IBM SPSS Modeler se muestra en tablas de informe y gráficos que proporciona estadísticos de resumen, como los histogramas y gráficos de distribución que pueden ser útiles para obtener una primera información de los datos y su distribución.

Para lo cual se desarrolla primero una auditoria de datos, como se demuestra en la siguiente imagen.

4.2.3.2. Auditoria de los campos:



Campo	Gráfico	Medida	Mín.	Máx.	Media	Desv. est.	Sesgo	Mediana	Modo	Exclusivo	Válido
Acceso_plataforma_virtual		Nominal	1.000	3.000	--	--	--	--	1.000	3	599
Ciclo		Nominal	1.000	4.000	--	--	--	--	4.000	4	599
Cursos_desaprobados		Nominal	1.000	4.000	--	--	--	--	1.000	4	599
Edad_Ingreso		Ordinal	1.000	4.000	--	--	--	--	4.000	4	599
Especialidad		Nominal	1.000	9.000	--	--	--	--	1.000	9	599
Estado_Civil		Nominal	1.000	6.000	--	--	--	--	1.000	6	599
Grado_satisfaccion		Nominal	1.000	4.000	--	--	--	--	1.000	4	599
Horas_dedicas_Estudiar		Ordinal	2.000	4.000	--	--	--	--	3.000	3	599
Ingresos_familiares		Ordinal	1.000	5.000	--	--	--	--	4.000	5	599
N°Inasistencias		Nominal	1.000	4.000	--	--	--	--	1.000	4	599
Nivel_Socieconomico		Ordinal	1.000	5.000	--	--	--	--	5.000	5	599
Ocupacion_Madre		Nominal	1.000	9.000	--	--	--	--	4.000	9	599
Ocupacion_Padre		Nominal	1.000	9.000	--	--	--	--	5.000	9	599
Promedio_curso_anterior		Ordinal	1.000	4.000	--	--	--	--	3.000	4	599
Sexo		Nominal	1.000	2.000	--	--	--	--	2.000	2	599
Tiene_Hermanos		Nominal	1.000	2.000	--	--	--	--	1.000	2	599
Tipo_colegio		Nominal	1.000	2.000	--	--	--	--	1.000	2	599
Tipo_residencia		Nominal	1.000	3.000	--	--	--	--	2.000	3	599
Turno		Ordinal	1.000	3.000	--	--	--	--	1.000	3	599
Vive_con_familia		Nominal	1.000	2.000	--	--	--	--	1.000	2	599

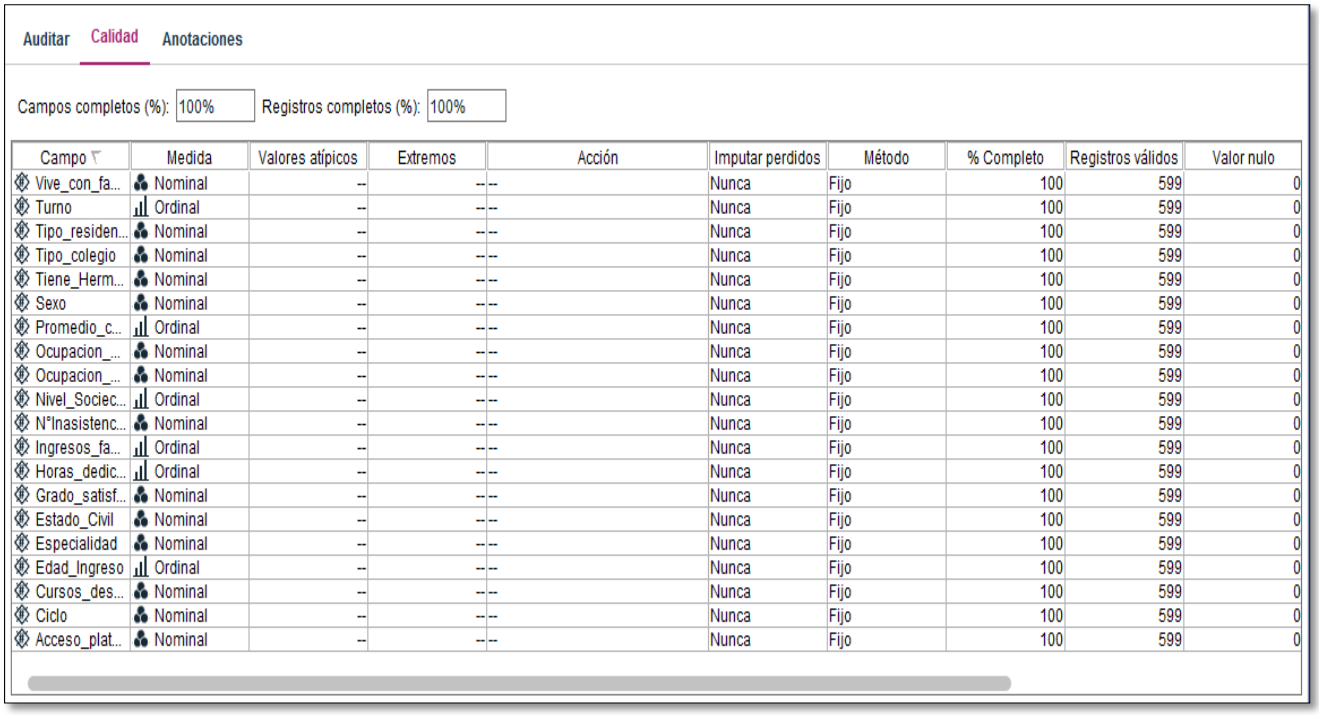
* Indica un resultado de varios modos # Indica un resultado muestreado

Figura 08: Auditoria de datos

Fuente Propia

Además en la pestaña de Calidad del informe de auditoría muestra información sobre valores extremos, atípicos y perdidos, y proporciona herramientas para tratar dichos valores

4.2.3.3. Calidad de los datos:



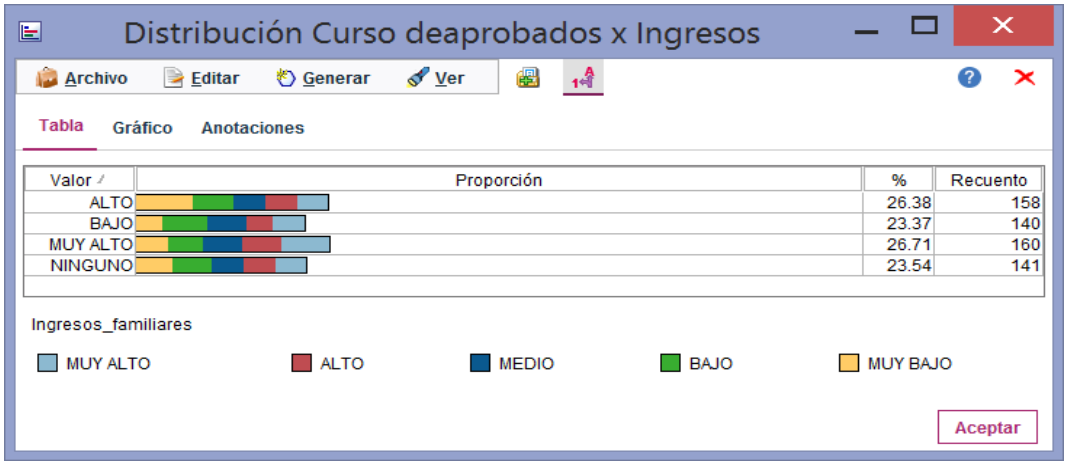
Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo	Registros válidos	Valor nulo
Vive_con_fa...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Turno	Ordinal	--	--	--	Nunca	Fijo	100	599	0
Tipo_residen...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Tipo_colegio	Nominal	--	--	--	Nunca	Fijo	100	599	0
Tiene_Herm...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Sexo	Nominal	--	--	--	Nunca	Fijo	100	599	0
Promedio_c...	Ordinal	--	--	--	Nunca	Fijo	100	599	0
Ocupacion_...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Ocupacion_...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Nivel_Sociec...	Ordinal	--	--	--	Nunca	Fijo	100	599	0
N°Inasistenc...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Ingresos_fa...	Ordinal	--	--	--	Nunca	Fijo	100	599	0
Horas_dedic...	Ordinal	--	--	--	Nunca	Fijo	100	599	0
Grado_satisf...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Estado_Civil	Nominal	--	--	--	Nunca	Fijo	100	599	0
Especialidad	Nominal	--	--	--	Nunca	Fijo	100	599	0
Edad_Ingreso	Ordinal	--	--	--	Nunca	Fijo	100	599	0
Cursos_des...	Nominal	--	--	--	Nunca	Fijo	100	599	0
Ciclo	Nominal	--	--	--	Nunca	Fijo	100	599	0
Acceso_plat...	Nominal	--	--	--	Nunca	Fijo	100	599	0

En este resultado de Calidad de Datos se aprecia que no se han encontrado datos atípicos, ni extremos, así como ningún dato vacío o nulo

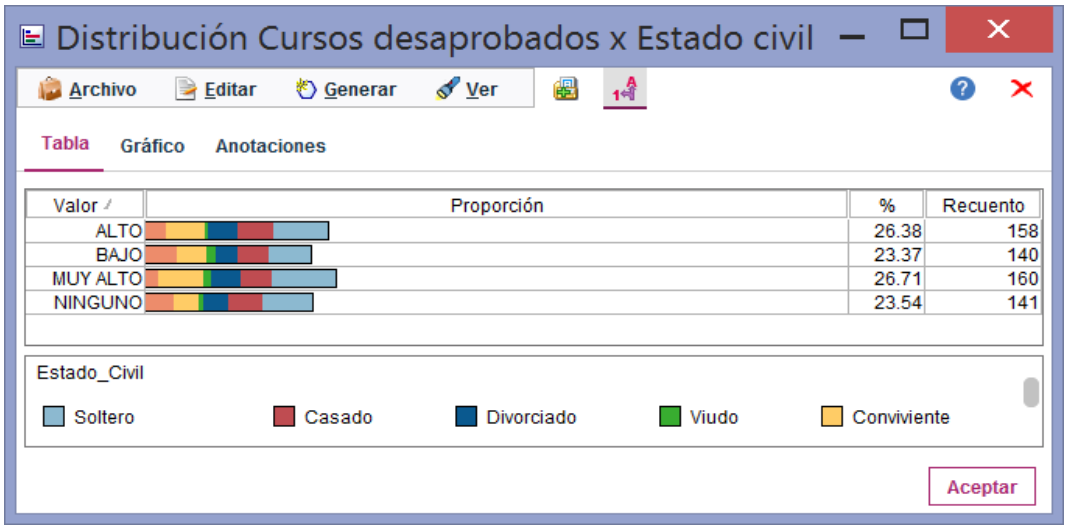
4.2.3.4. Exploración de los datos:

Los gráficos utilizados para la exploración de datos fueron los Histogramas y de distribución. Estos gráficos de distribución se utilizaron para mostrar valores simbólicos en un conjunto de datos. Se suelen utilizar para corregir cualquier desequilibrio. Por ejemplo, si las instancias de los estudiantes con estado civil “Soltero” suceden con mucha más frecuencia que el otro tipo de estado civil, tal vez desee reducir estas instancias para que se pueda generar una regla más útil en posteriores operaciones de minería de datos.

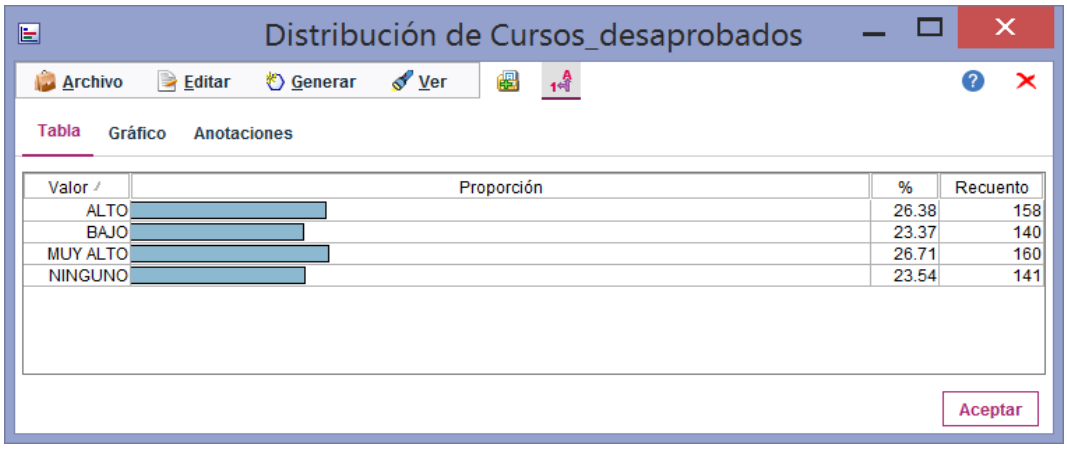
✓ **Cursos desaprobados por ingresos+**



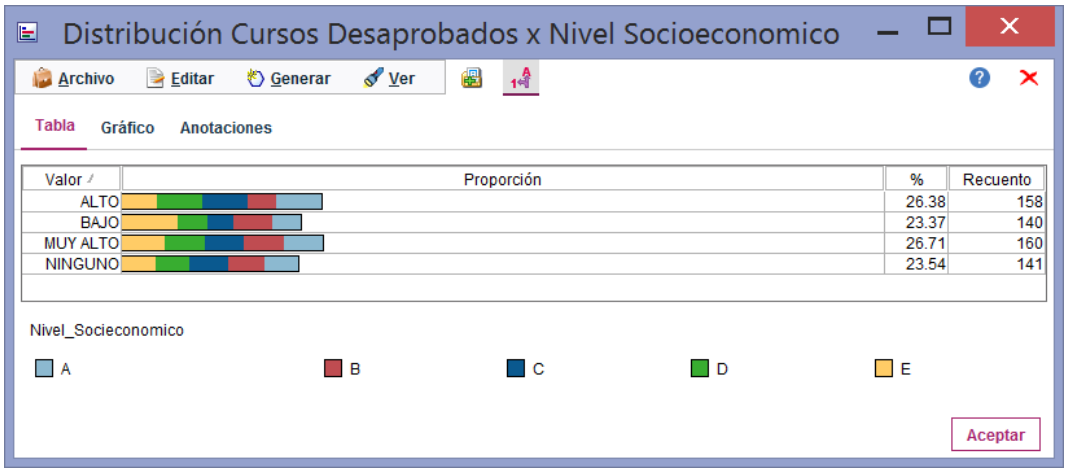
✓ **Cursos desaprobados por Estado Civil**



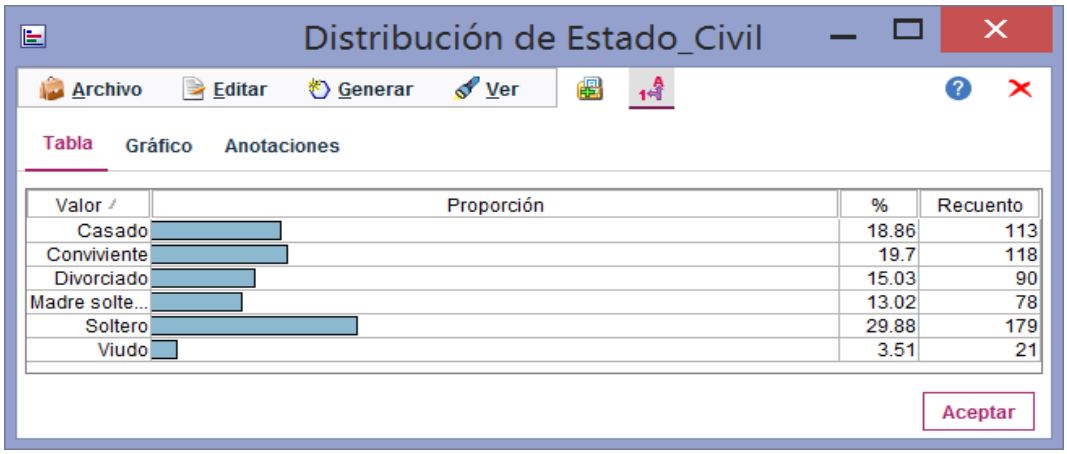
✓ **Distribución de Cursos desaprobados**



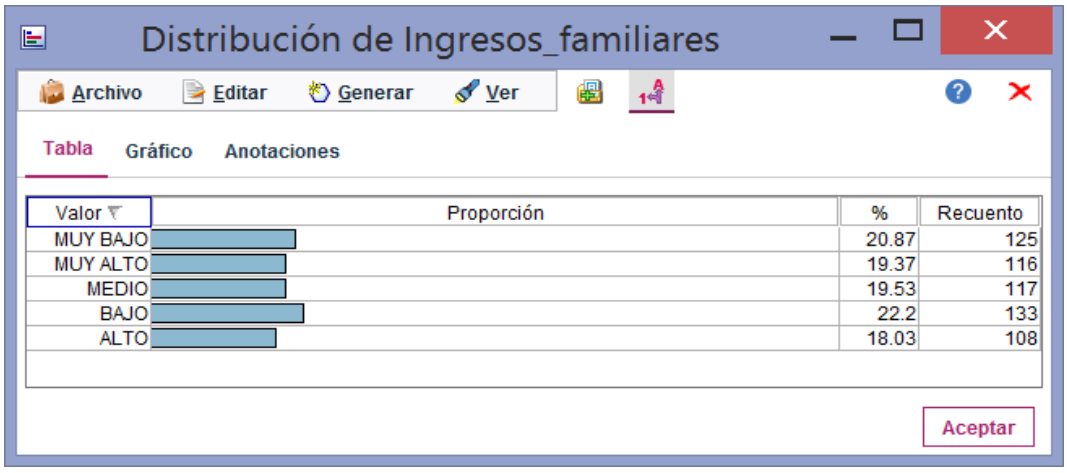
✓ **Cursos desaprobados por Nivel Socioeconómico**



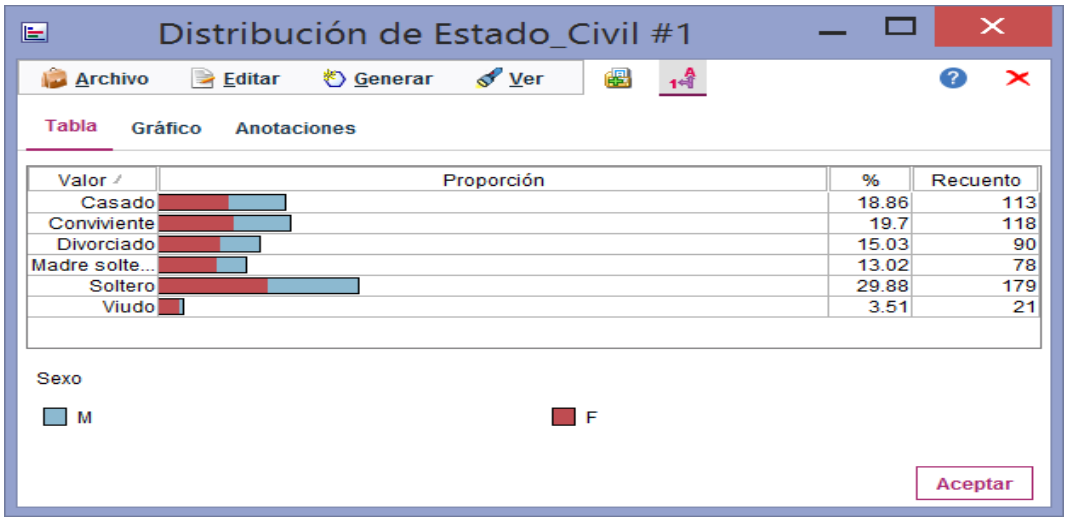
✓ **Distribución de Estado Civil**



✓ **Distribución de Ingresos Familiares**



✓ **Distribución de Datos x Estado Civil x Sexo de los estudiantes**



4.2.4. Selección y Limpieza de datos:

No es necesario aplicar ajustes a los datos ya que estos están en condición correcta para la elaboración del modelo.

4.2.5. Construcción de datos e Integración de Datos:

Esta etapa tiene como objetivo construir un conjunto de datos o vista minable que tiene la siguiente estructura por estudiante.

Tabla 03: Descripción de los campos

N°	Tabla	Campo	Descripción	Valores
1	Estudiante	Nivel_Socieconomico	Nivel socioeconómico del Estudiante	A, B, C, D, E
2	Estudiante	Ocupacion_Padre	Ocupación del padre del Estudiante	construcción, minería, hogar, profesionales, Operarios, artesanos, industria

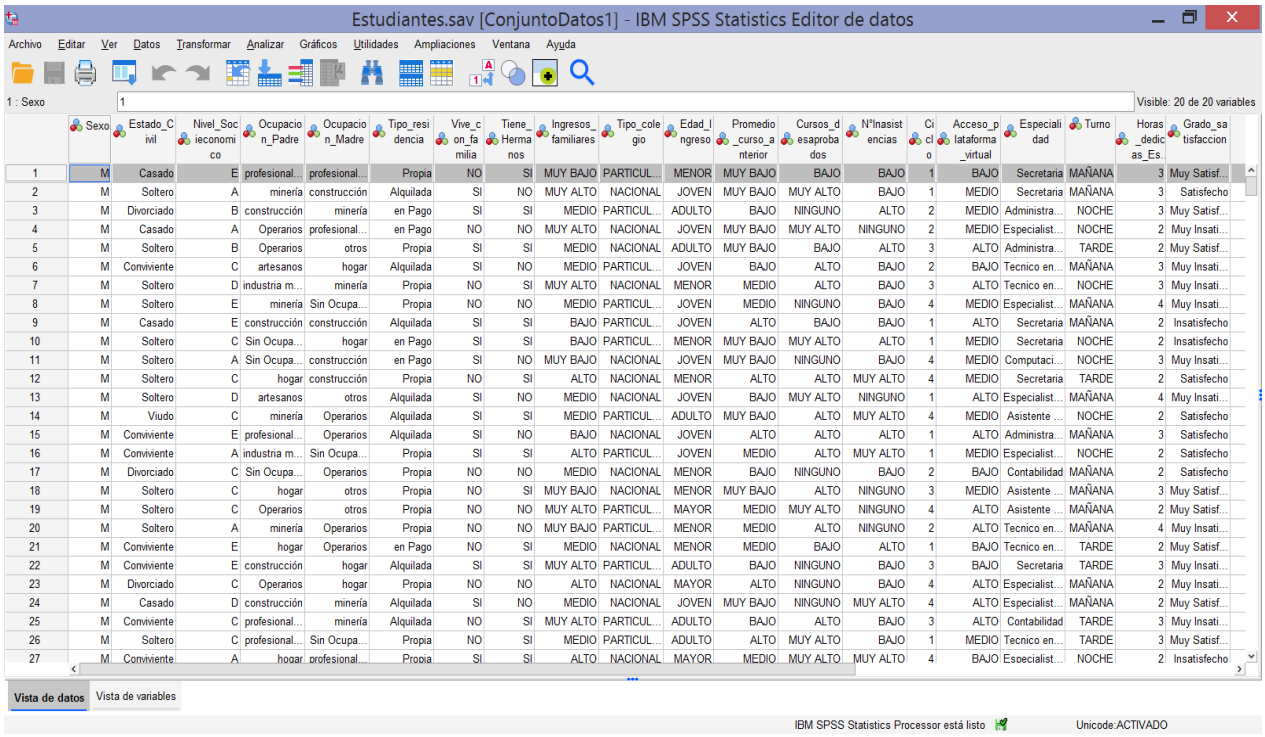
				manufacturera, universitarios, Sin Ocupación, otros
3	Estudiante	Ocupacion_Madre	Ocupación de la madre del Estudiante	construcción, minería, hogar, profesionales, universitarios Operarios, artesanos, industria manufacturera, , Sin Ocupación, otros
4	Estudiante	Sexo	Sexo del Estudiante	M , F
5	Estudiante	Estado_Civil	Estado civil del Estudiante	Soltero, Casado, Divorciado, Viudo, Madre soltera, Solo pareja
6	Estudiante	Tipo_residencia	Tipo de residencia del estudiante	Propia, Alquilada, Hipoteca
7	Estudiante	Vive_con_familia	Si vive con más integrantes familiares	Si / No
8	Estudiante	Tiene_Hermanos	Si tiene hermanos	Si / No
9	Estudiante	Ingresos_familiares	Ingresos familiares	Muy Alto: de 10000 a más Alto: de 6000 a 9999 Medio: de 3000 a 5999 Bajo: de 1000 a 2999 Muy bajo: de 999 a menos
10	Estudiante	Tipo_colegio	Colegio de donde proviene	Nacional, Particular
11	Estudiante	Edad_Ingreso	Edad del estudiante	Menor : menor a 18 Joven: de 18 a 25 Adulto: de 26 a 35 Mayor: de 35 a más
12	Estudiante	Horas_dedicas_Estudiar	Horas de dedicación a los estudios	2, 3, 4
13	Especialidad	Especialidad	Especialidad que	Administrador de

			estudia	empresa, Contabilidad, Computación e Informática, Marketing, Secretaria, Asistente de Negocios, Especialista en Ofimática, Técnico en computación, Técnico en diseño grafico
14	Asistencias	N°Inasistencias	N° inasistencias hasta el momento	Muy Alto: de 6 a más Alto: de 4 a 6 Bajo: de 1 a 3 Ninguno: 0
15	Asistencias	Acceso_plataforma_virtual	Frecuencia de acceso a la plataforma de apoyo	Alto, Medio, Bajo
16	Notas	Promedio_curso_anterior	Promedio del estudiante del curso anterior	Alto: de 15 a 20 Medio: de 11 a 14 Bajo: de 5 a 9 Muy bajo: de 0 a 4
17	Notas	Cursos_desaprobados	N° cursos desabrobados	Muy Alto: de 5 a 6 Alto: de 3 a 4 Bajo: de 1 a 2 Ninguno: 0
18	Matricula	Turno	Turno de estudios	Mañana, Tarde, Noche
19	Matricula	Ciclo	Ciclo al que pertenece	1, 2, 3, 4
20	Encuesta	Grado_satisfaccion	Grado de satisfacción respecto al servicio ofrecido por la institución en la última encuesta	Muy Satisfecho, Satisfecho, Insatisfecho, Muy Insatisfecho

La construcción de la siguiente manera:

1. Los datos fueron almacenados en una archivo .sav del programa de IBM SPSS Statistics para un mejor análisis.

Datos a utilizar: Estudiantes.sav



	Sexo	Estado_Civil	Nivel_Socieconomico	Ocupacion_Padre	Ocupacion_Madre	Tipo_residencia	Vive_con_familia	Tiene_Hermanos	Ingresos_familiares	Tipo_colegio	Edad_Ingreso	Promedio_curso_anterior	Cursos_desaprobados	NºInasistencias	Ciclo	Acceso_plataforma_virtual	Especialidad	Turno	Horas_dedicadas_Es.	Grado_satisfaccion
1	M	Casado	E	profesional...	profesional...	Propia	NO	SI	MUY BAJO	PARTICUL...	MENOR	MUY BAJO	BAJO	BAJO	1	BAJO	Secretaria	MAÑANA	3	Muy Satisf...
2	M	Soltero	A	minería	construcción	Alquilada	SI	NO	MUY ALTO	NACIONAL	JOVEN	MUY BAJO	MUY ALTO	BAJO	1	MEDIO	Secretaria	MAÑANA	3	Satisfecho
3	M	Divorciado	B	construcción	minería	en Pago	SI	SI	MEDIO	PARTICUL...	ADULTO	BAJO	NINGUNO	ALTO	2	MEDIO	Administra...	NOCHE	3	Muy Satisf...
4	M	Casado	A	Operarios	profesional...	en Pago	NO	NO	MUY ALTO	NACIONAL	JOVEN	MUY BAJO	MUY ALTO	NINGUNO	2	MEDIO	Especialist...	NOCHE	2	Muy Insati...
5	M	Soltero	B	Operarios	otros	Propia	SI	SI	MEDIO	NACIONAL	ADULTO	MUY BAJO	BAJO	ALTO	3	ALTO	Administra...	TARDE	2	Muy Satisf...
6	M	Conviviente	C	artesanos	hogar	Alquilada	SI	NO	MEDIO	PARTICUL...	JOVEN	BAJO	ALTO	BAJO	2	BAJO	Tecnico en...	MAÑANA	3	Muy Insati...
7	M	Soltero	D	industria m...	minería	Propia	NO	SI	MUY ALTO	NACIONAL	MENOR	MEDIO	ALTO	BAJO	3	ALTO	Tecnico en...	NOCHE	3	Muy Insati...
8	M	Soltero	E	minería	Sin Ocupa...	Propia	NO	NO	MEDIO	PARTICUL...	JOVEN	MEDIO	NINGUNO	BAJO	4	MEDIO	Especialist...	MAÑANA	4	Muy Insati...
9	M	Casado	E	construcción	construcción	Alquilada	SI	SI	BAJO	PARTICUL...	JOVEN	ALTO	BAJO	BAJO	1	ALTO	Secretaria	MAÑANA	2	Insatisfecho
10	M	Soltero	C	Sin Ocupa...	hogar	en Pago	SI	SI	BAJO	PARTICUL...	MENOR	MUY BAJO	MUY ALTO	ALTO	1	MEDIO	Secretaria	NOCHE	2	Insatisfecho
11	M	Soltero	A	Sin Ocupa...	construcción	en Pago	SI	NO	MUY BAJO	NACIONAL	JOVEN	MUY BAJO	NINGUNO	BAJO	4	MEDIO	Computaci...	NOCHE	3	Muy Insati...
12	M	Soltero	C	hogar	construcción	Propia	NO	SI	ALTO	NACIONAL	MENOR	ALTO	ALTO	MUY ALTO	4	MEDIO	Secretaria	TARDE	2	Satisfecho
13	M	Soltero	D	artesanos	otros	Alquilada	SI	NO	MEDIO	NACIONAL	JOVEN	BAJO	MUY ALTO	NINGUNO	1	ALTO	Especialist...	MAÑANA	4	Muy Insati...
14	M	Viudo	C	minería	Operarios	Alquilada	SI	SI	MEDIO	PARTICUL...	ADULTO	MUY BAJO	ALTO	MUY ALTO	4	MEDIO	Asistente ...	NOCHE	2	Satisfecho
15	M	Conviviente	E	profesional...	Operarios	Alquilada	SI	NO	BAJO	NACIONAL	JOVEN	ALTO	ALTO	ALTO	1	ALTO	Administra...	MAÑANA	3	Satisfecho
16	M	Conviviente	A	industria m...	Sin Ocupa...	Propia	SI	SI	ALTO	PARTICUL...	JOVEN	MEDIO	ALTO	MUY ALTO	1	MEDIO	Especialist...	NOCHE	2	Satisfecho
17	M	Divorciado	C	Sin Ocupa...	Operarios	Propia	NO	NO	MEDIO	NACIONAL	MENOR	BAJO	NINGUNO	BAJO	2	BAJO	Contabilidad	MAÑANA	2	Satisfecho
18	M	Soltero	C	hogar	otros	Propia	NO	SI	MUY BAJO	NACIONAL	MENOR	MUY BAJO	ALTO	NINGUNO	3	MEDIO	Asistente ...	MAÑANA	3	Muy Satisf...
19	M	Soltero	C	Operarios	otros	Propia	NO	NO	MUY ALTO	PARTICUL...	MAYOR	MEDIO	MUY ALTO	NINGUNO	4	ALTO	Asistente ...	MAÑANA	2	Muy Satisf...
20	M	Soltero	A	minería	Operarios	Propia	NO	NO	MUY BAJO	PARTICUL...	MENOR	MEDIO	ALTO	NINGUNO	2	ALTO	Tecnico en...	MAÑANA	4	Muy Insati...
21	M	Conviviente	E	hogar	Operarios	en Pago	NO	SI	MEDIO	NACIONAL	MENOR	MEDIO	BAJO	ALTO	1	BAJO	Tecnico en...	TARDE	2	Muy Satisf...
22	M	Conviviente	E	construcción	hogar	Alquilada	SI	SI	MUY ALTO	PARTICUL...	ADULTO	BAJO	NINGUNO	BAJO	3	BAJO	Secretaria	TARDE	3	Muy Insati...
23	M	Divorciado	C	Operarios	hogar	Propia	NO	NO	ALTO	NACIONAL	MAYOR	ALTO	NINGUNO	BAJO	4	ALTO	Especialist...	MAÑANA	2	Muy Insati...
24	M	Casado	D	construcción	minería	Alquilada	SI	NO	MEDIO	NACIONAL	JOVEN	MUY BAJO	NINGUNO	MUY ALTO	4	ALTO	Especialist...	MAÑANA	2	Muy Satisf...
25	M	Conviviente	C	profesional...	minería	Alquilada	NO	SI	MUY ALTO	PARTICUL...	ADULTO	BAJO	ALTO	BAJO	3	ALTO	Contabilidad	TARDE	3	Muy Insati...
26	M	Soltero	C	profesional...	Sin Ocupa...	Propia	NO	SI	MEDIO	PARTICUL...	ADULTO	ALTO	MUY ALTO	BAJO	1	MEDIO	Tecnico en...	TARDE	3	Muy Satisf...
27	M	Conviviente	A	hogar	profesional...	Propia	SI	SI	ALTO	NACIONAL	MAYOR	MEDIO	MUY ALTO	MUY ALTO	4	BAJO	Especialist...	NOCHE	2	Insatisfecho

Total de registros: 599

4.3. MODELADO

4.3.3. Seleccionar la técnica de modelado:

En esta fase realizamos la selección de la técnica de Minería de datos más indicada para el tipo de problema a solucionar. Se seleccionó la técnica de acuerdo a las siguientes características:

1. Es una técnica útil en la selección de variables significativas entre una gran cantidad de variables, más aun cuando no se cuenta con suficiente información de expertos conocedores del problema que se está modelando.
2. Los Nodos arboles manejan datos no numéricos muy bien.
3. La estructura de condición y ramificación de un árbol es idónea para este tipo de problema de clasificación.
4. Nos permitirá generar un conjunto de reglas y condiciones que serán determinantes para la toma de nuestras decisiones.

De acuerdo a estas características se escogieron 4 Técnicas de modelado:

a. **Modelo Árbol C5.0:** Los modelos C5.0 dividen la muestra en función del campo que ofrece la máxima ganancia de información. Las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las submuestras de nuevo. Por último se vuelven a examinar las divisiones del nivel inferior, y se eliminan o podan las que no contribuyen significativamente con el valor del modelo.

b. **Modelo Árbol AS:** este método genera árboles de decisión mediante estadísticos de chi-cuadrado para identificar divisiones óptimas. Examina campos de entrada y los resultados para comprobar la significación mediante una comprobación de independencia de chi-cuadrado. Si varias de estas relaciones son estadísticamente importantes, AS selecciona el campo de

entrada de mayor relevancia. Si una entrada cuenta con más de dos categorías, se compararán estas categorías y se contraerán las que no presenten diferencias en los resultados. Para ello, se unirá el par de categorías que presenten menor diferencia, y así sucesivamente. Este proceso de fusión de categorías se detiene cuando todas las categorías restantes difieren entre sí en el nivel de comprobación especificado.

c. **Modelo Árbol C&R:** basado en árboles, similar a C5.0. Este método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos con valores de campo de salida similares. El nodo C&RT comienza por realizar un examen de los campos de entrada para buscar la mejor división, que se ha medido mediante la reducción del índice de impureza resultado de la división. La división define dos subgrupos, que se siguen dividiendo en otros dos subgrupos sucesivamente hasta que se activa un criterio de parada. Todas las divisiones son binarias (solamente se crean dos subgrupos).

d. **Modelo Red Bayesiana:** crea un modelo de probabilidad combinando pruebas con conocimiento del mundo real para establecer la probabilidad de instancias. Es un modelo gráfico que muestra en un conjunto de datos y las independencias probabilísticas o condicionales entre ellas.

4.3.2. Generación del plan de Prueba:

Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo.

Los datos ya preparados se encuentran en el archivo de IBM SPSS Statistics "Estudiantes.sav" que son llevados hacia la herramienta de IBM SPSS Modeler Subscription, para luego construir el modelo basado en el conjunto de entradas y medir la calidad del modelo y luego analizar los resultados.

A continuación procederemos a elaborar la distribución del proyecto en las siguientes partes:

- Preparación de datos, Exploración y Validación de los datos
- Modelado
- Evaluación del Modelo

4.3.3. CONSTRUCCIÓN DEL MODELO:

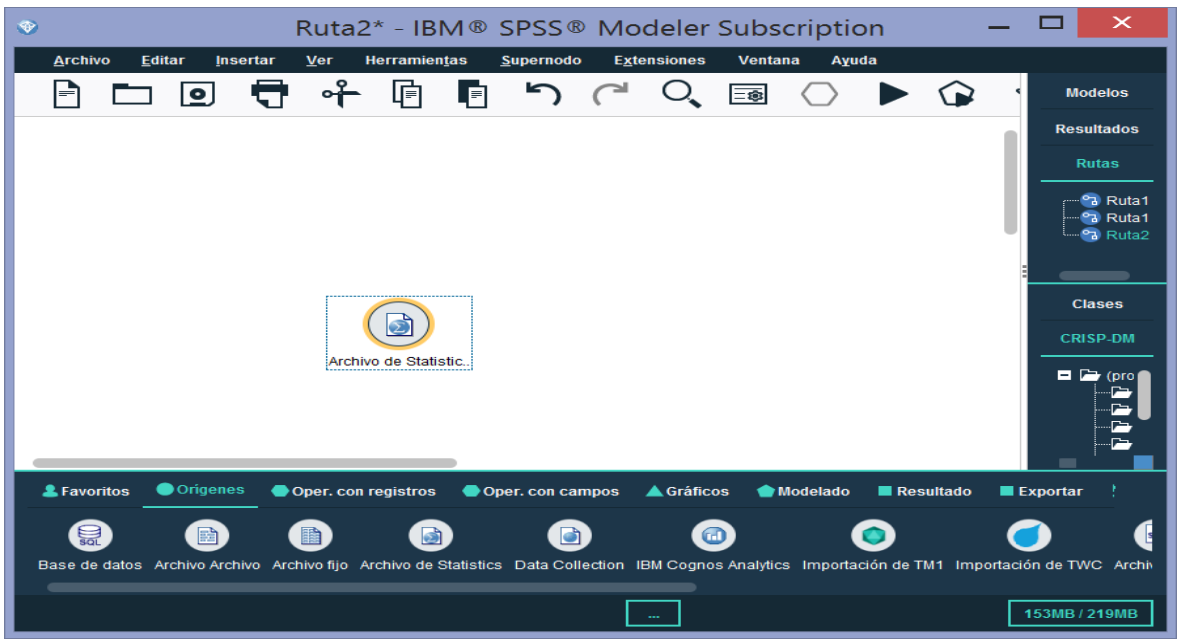
Para la construcción de los modelos a evaluar se utilizó la el software IBM SPSS Modeler y el archivo preparado IBM SPSS Statistics "Estudiantes.sav".

Primero realizamos la preparación de los datos.

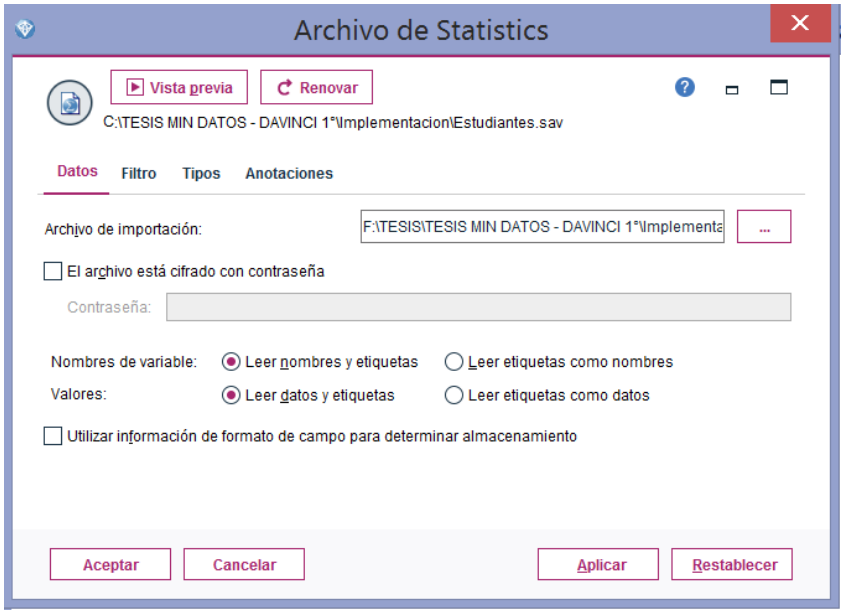
A continuación procederemos a elaborar la distribución de los datos en 2 grupos: Datos Comprobación y Datos Entrenamiento.

En la herramienta Spss Modeler vamos a implementar un nodo de fuente de datos al área de trabajo que nos sirve como conexión al archivo de .sav preparado. Como se muestra en la figura.

Creación de la ruta a implementar

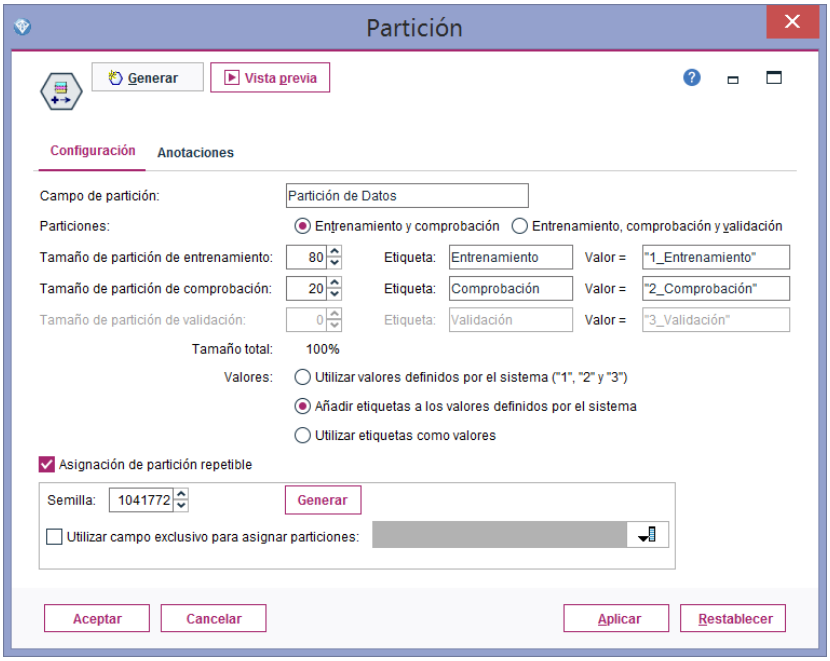


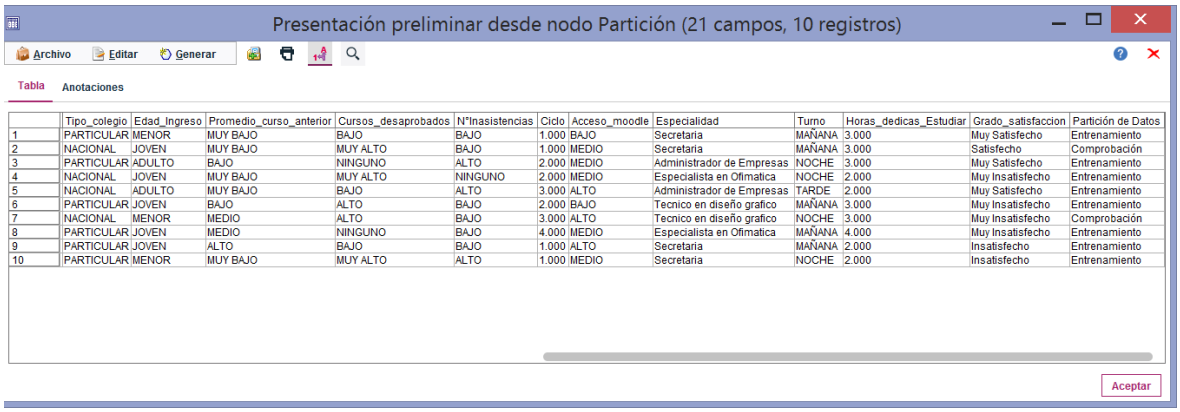
Luego conectamos el nodo con una fuente de datos. En la siguiente figura se muestra el archivo a utilizar.



Luego procederemos a formar el grupo de entrenamiento (80 %) y el grupo de Comprobación (20%), para ello usamos el nodo “Partición” y designamos a cada partición el porcentaje de los datos que se le asignara.

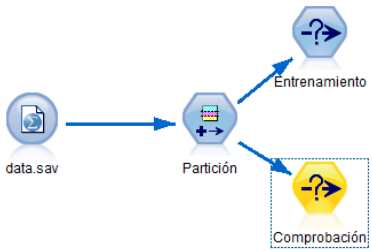
Partición de los datos





	Tipo_colegio	Edad_Ingreso	Promedio_curso_anterior	Cursos_desaprobados	N°Inasistencias	Ciclo	Acceso_moodle	Especialidad	Turno	Horas_dedicas_Estudiar	Grado_satisfaccion	Partición de Datos
1	PARTICULAR	MENOR	MUY BAJO	BAJO	BAJO	1.000	BAJO	Secretaria	MAÑANA	3.000	Muy Satisfecho	Entrenamiento
2	NACIONAL	JOVEN	MUY BAJO	MUY ALTO	BAJO	1.000	MEDIO	Secretaria	MAÑANA	3.000	Satisfecho	Comprobación
3	PARTICULAR	ADULTO	BAJO	NINGUNO	ALTO	2.000	MEDIO	Administrador de Empresas	NOCHE	3.000	Muy Satisfecho	Entrenamiento
4	NACIONAL	JOVEN	MUY BAJO	MUY ALTO	NINGUNO	2.000	MEDIO	Especialista en Ofimatica	NOCHE	2.000	Muy Insatisfecho	Entrenamiento
5	NACIONAL	ADULTO	MUY BAJO	BAJO	ALTO	3.000	ALTO	Administrador de Empresas	TARDE	2.000	Muy Satisfecho	Entrenamiento
6	PARTICULAR	JOVEN	BAJO	ALTO	BAJO	2.000	BAJO	Tecnico en diseño grafico	MAÑANA	3.000	Muy Insatisfecho	Entrenamiento
7	NACIONAL	MENOR	MEDIO	ALTO	BAJO	3.000	ALTO	Tecnico en diseño grafico	NOCHE	3.000	Muy Insatisfecho	Comprobación
8	PARTICULAR	JOVEN	MEDIO	NINGUNO	BAJO	4.000	MEDIO	Especialista en Ofimatica	MAÑANA	4.000	Muy Insatisfecho	Entrenamiento
9	PARTICULAR	JOVEN	ALTO	BAJO	BAJO	1.000	ALTO	Secretaria	MAÑANA	2.000	Insatisfecho	Entrenamiento
10	PARTICULAR	MENOR	MUY BAJO	MUY ALTO	ALTO	1.000	MEDIO	Secretaria	NOCHE	2.000	Insatisfecho	Entrenamiento

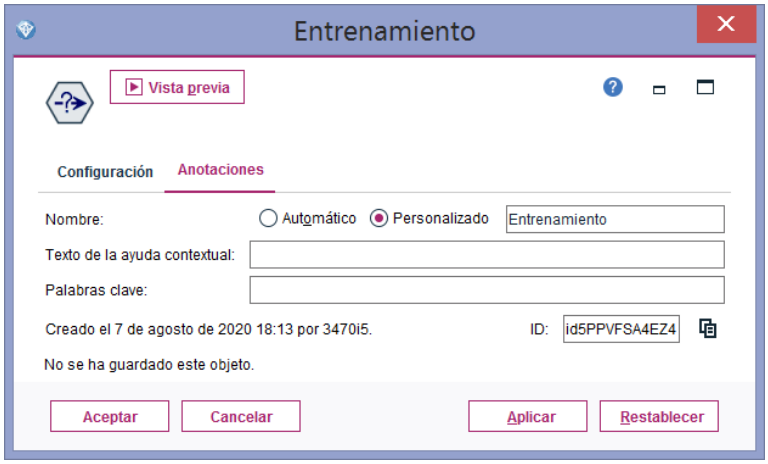
Luego para seleccionar los datos de cada partición los enlazamos con un nodo "Seleccionar" y usamos como condición el nombre de cada una de ellas:



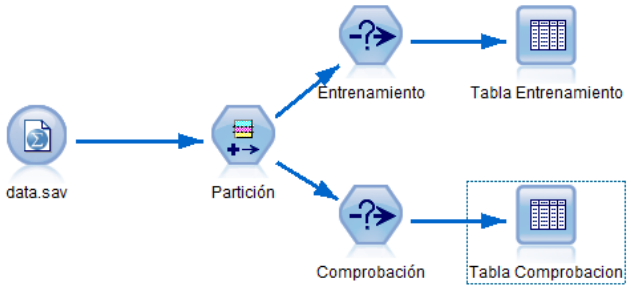
Partición de entrenamiento

Para diferenciar los nodos “Seleccionar” le asignaremos nombre a cada uno.

Asignación de nombre para el entrenamiento de datos

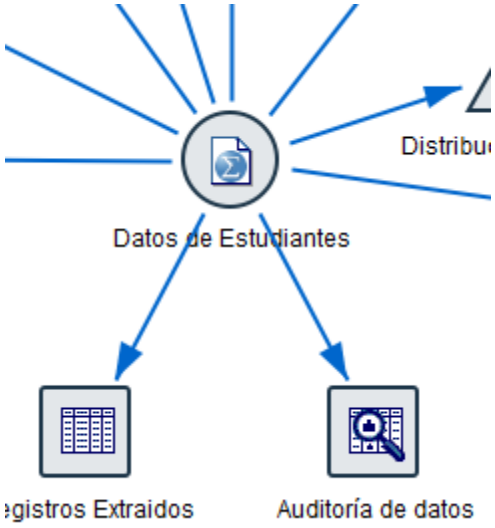


A continuación vinculamos cada nodo “Seleccionar” con los nodos “Tabla” para visualizar los datos de cada partición:

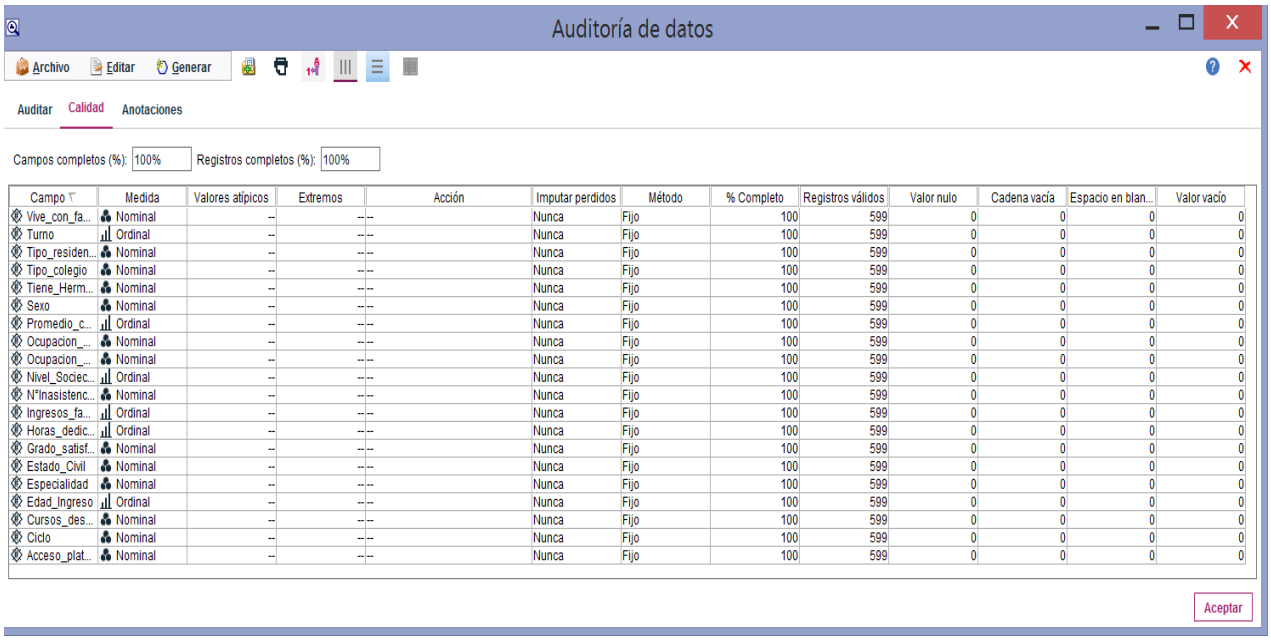


Luego se agregan el análisis de los datos en las gráficas de distribución de Datos x Estado civil, Datos x Estado Civil x Sexo y Datos x Ingresos, Datos x CursosDesaprobados. Datos x Ingresos Familiares.

También se procedió a realizar una auditoria de los datos como se muestra a continuación:



Auditoria de datos

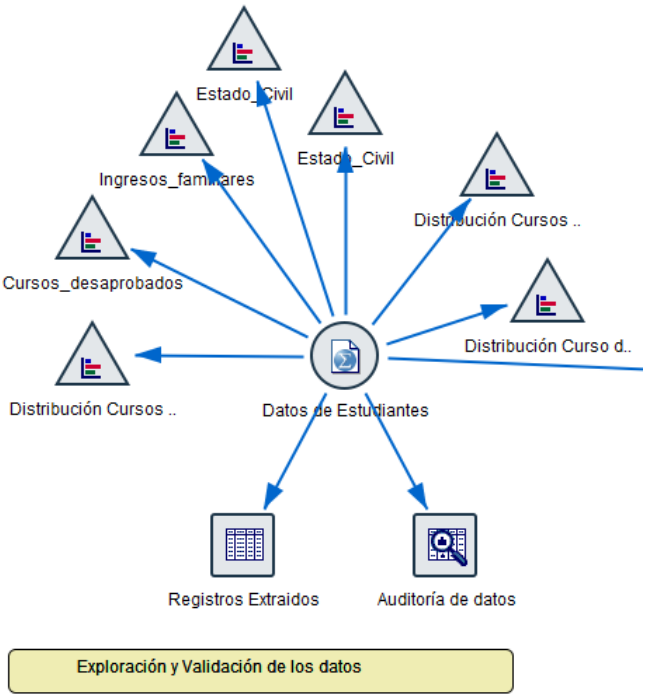


Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blan...	Valor vacío
Vive_con_fa...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Turno	Ordinal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Tipo_residen...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Tipo_colegio	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Tiene_Herm...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Sexo	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Promedio_c...	Ordinal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Ocupacion_...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Ocupacion_...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Nivel_Sociec...	Ordinal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
N°Inasistenc...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Ingresos_fa...	Ordinal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Horas_dedic...	Ordinal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Grado_satisf...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Estado_Civil	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Especialidad	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Edad_Ingreso	Ordinal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Cursos_des...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Ciclo	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0
Acceso_plat...	Nominal	--	--	--	Nunca	Fijo	100	599	0	0	0	0

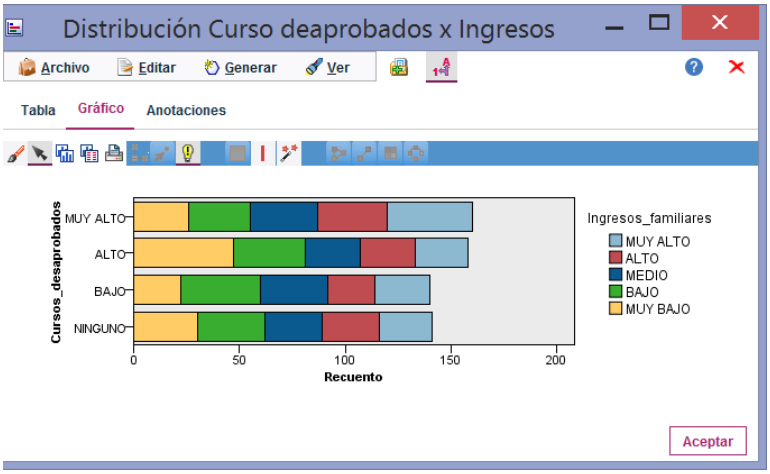
Para la aplicación del modelo utiliza la data cruda de los estudiantes correspondiente al periodo 2018-2019 de la Institución educativa.

A continuación añadimos un nodo “Distribución” en la categoría “Graficos” y lo enlazamos con la fuente de datos.

Distribución de los datos



Detalle de la Distribución de cursos desaprobados

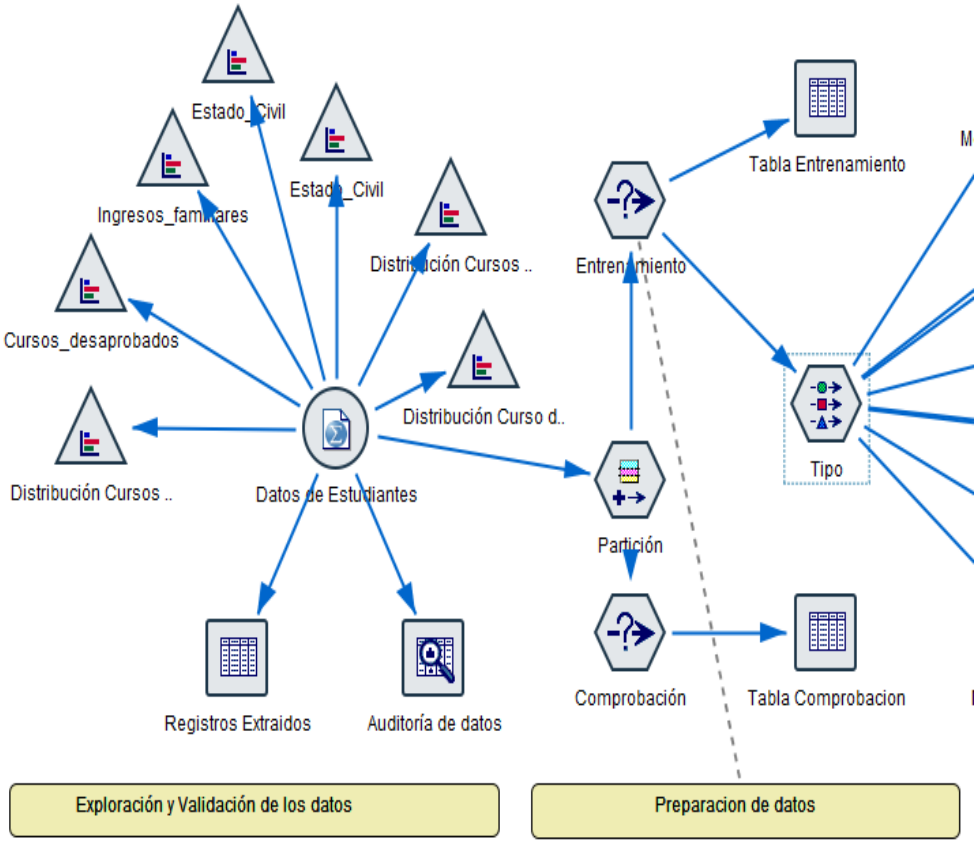


Ahora podemos ejecutar la "ruta" y ver la distribución de cursos desaprobados de los estudiantes.

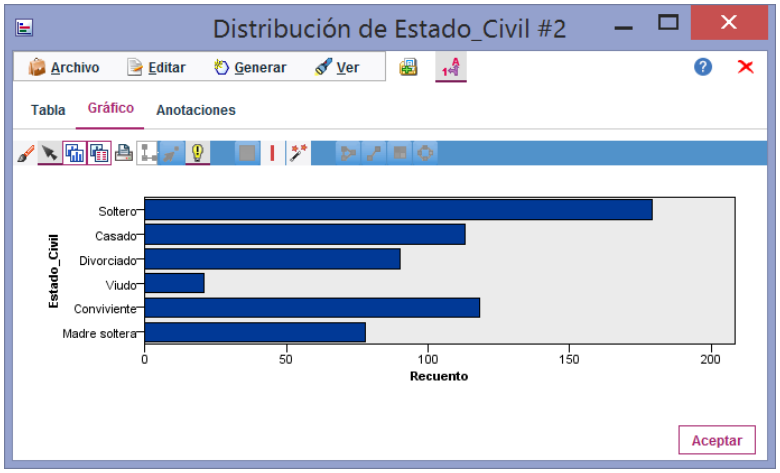
Podemos apreciar que una gran cantidad de los estudiantes han tenido cursos desaprobados, nuestro objetivo es determinar cual es el factor que pudo llevar a que los alumnos a desertar.

Del mismo modo podemos obtener la distribución de la variable “Sexo”, Estado Civil, Ingresos familiares, Nivel Socioeconómico :

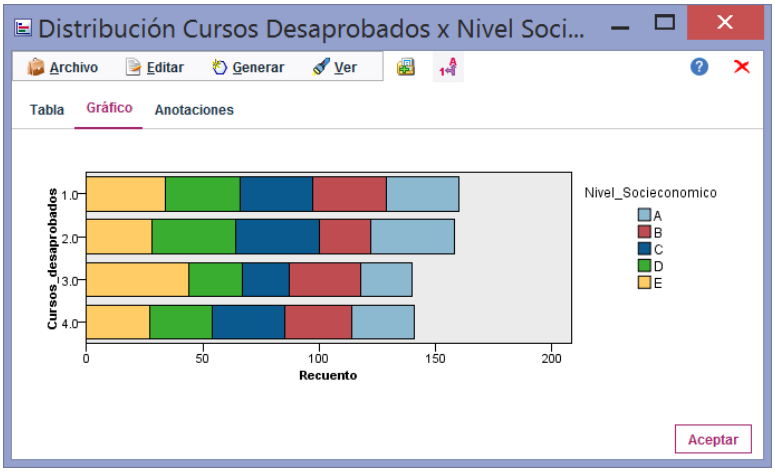
Preparación de datos



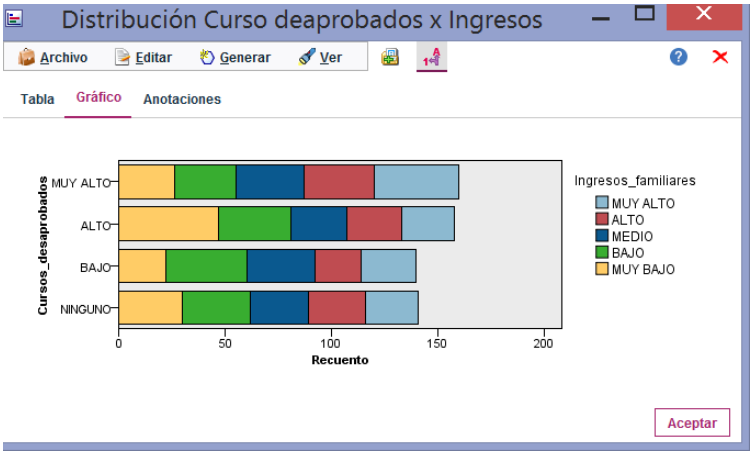
Distribución x estado civil



Distribución x Nivel Socioeconómico



Distribución x Ingresos



Datos de Entrenamiento ingresados al modelo (80%)

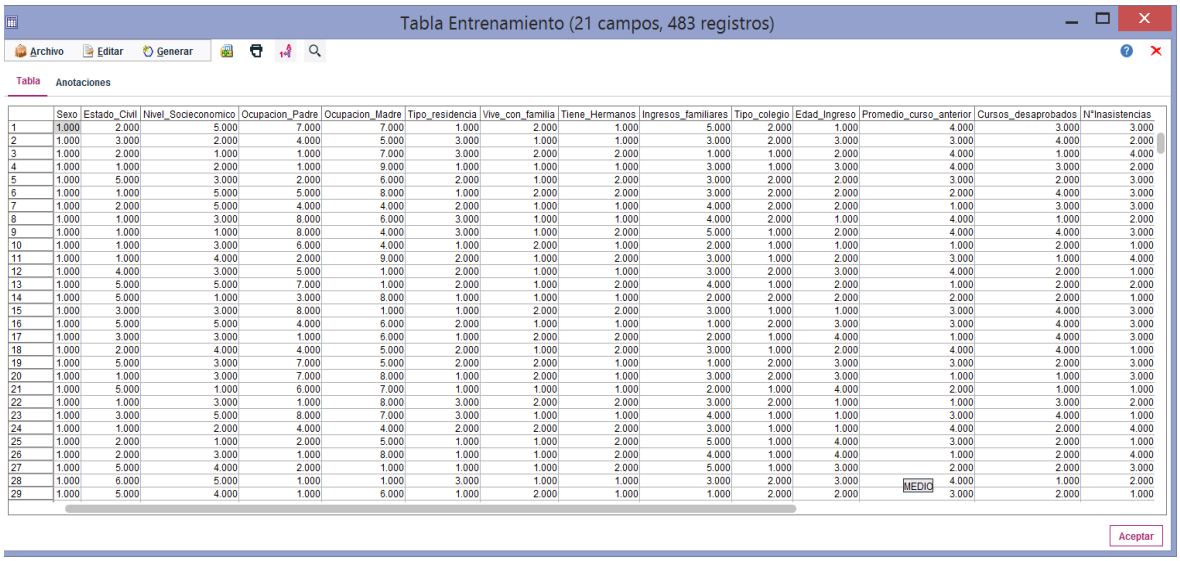


Tabla Entrenamiento (21 campos, 483 registros)

Archivo Editar Generar

Tabla Anotaciones

	Sexo	Estado_Civil	Nivel_Socieconomico	Ocupacion_Padre	Ocupacion_Madre	Tipo_residencia	Vive_con_familia	Tiene_Hermanos	Ingresos_familiares	Tipo_colegio	Edad_Ingreso	Promedio_curso_anterior	Cursos_desaprobados	N°Inasistencias
1	1.000	2.000	5.000	7.000	7.000	1.000	2.000	1.000	5.000	2.000	1.000	4.000	3.000	3.000
2	1.000	3.000	2.000	4.000	5.000	3.000	1.000	1.000	3.000	2.000	3.000	3.000	4.000	2.000
3	1.000	2.000	1.000	1.000	7.000	3.000	2.000	2.000	1.000	1.000	2.000	4.000	1.000	4.000
4	1.000	1.000	2.000	1.000	9.000	1.000	1.000	1.000	3.000	1.000	3.000	4.000	3.000	2.000
5	1.000	5.000	3.000	2.000	6.000	2.000	1.000	2.000	3.000	2.000	2.000	3.000	2.000	3.000
6	1.000	1.000	5.000	5.000	8.000	1.000	2.000	2.000	3.000	2.000	2.000	2.000	4.000	3.000
7	1.000	2.000	5.000	4.000	4.000	2.000	1.000	1.000	4.000	2.000	2.000	1.000	3.000	3.000
8	1.000	1.000	3.000	8.000	6.000	3.000	1.000	1.000	4.000	2.000	1.000	4.000	1.000	2.000
9	1.000	1.000	1.000	8.000	4.000	3.000	1.000	2.000	5.000	1.000	2.000	4.000	4.000	3.000
10	1.000	1.000	3.000	6.000	4.000	1.000	2.000	1.000	2.000	1.000	1.000	1.000	2.000	1.000
11	1.000	1.000	4.000	2.000	9.000	2.000	1.000	2.000	3.000	1.000	2.000	3.000	1.000	4.000
12	1.000	4.000	3.000	5.000	1.000	2.000	1.000	1.000	3.000	2.000	3.000	4.000	2.000	1.000
13	1.000	5.000	5.000	7.000	1.000	2.000	1.000	2.000	4.000	1.000	2.000	1.000	2.000	2.000
14	1.000	5.000	1.000	3.000	8.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	2.000	1.000
15	1.000	3.000	3.000	8.000	1.000	1.000	2.000	2.000	3.000	1.000	1.000	3.000	4.000	3.000
16	1.000	5.000	5.000	4.000	6.000	2.000	1.000	1.000	1.000	2.000	3.000	3.000	4.000	3.000
17	1.000	3.000	3.000	1.000	6.000	1.000	2.000	2.000	2.000	1.000	4.000	1.000	4.000	3.000
18	1.000	2.000	4.000	4.000	5.000	2.000	1.000	2.000	3.000	1.000	2.000	4.000	4.000	1.000
19	1.000	5.000	3.000	7.000	5.000	2.000	2.000	1.000	1.000	2.000	3.000	3.000	2.000	3.000
20	1.000	1.000	3.000	7.000	8.000	1.000	2.000	1.000	3.000	2.000	3.000	1.000	1.000	3.000
21	1.000	5.000	1.000	6.000	7.000	1.000	1.000	1.000	2.000	1.000	4.000	2.000	1.000	1.000
22	1.000	1.000	3.000	1.000	8.000	3.000	2.000	2.000	3.000	2.000	1.000	1.000	3.000	2.000
23	1.000	3.000	5.000	8.000	7.000	3.000	1.000	1.000	4.000	1.000	1.000	3.000	4.000	1.000
24	1.000	1.000	2.000	4.000	4.000	2.000	2.000	2.000	3.000	1.000	1.000	4.000	2.000	4.000
25	1.000	2.000	1.000	2.000	5.000	1.000	1.000	2.000	5.000	1.000	4.000	3.000	2.000	1.000
26	1.000	2.000	3.000	1.000	8.000	1.000	1.000	2.000	4.000	1.000	4.000	1.000	2.000	4.000
27	1.000	5.000	4.000	2.000	1.000	1.000	1.000	2.000	5.000	1.000	3.000	2.000	2.000	3.000
28	1.000	6.000	5.000	1.000	1.000	3.000	1.000	1.000	3.000	2.000	3.000	MEDIO	4.000	2.000
29	1.000	5.000	4.000	1.000	6.000	1.000	2.000	1.000	1.000	2.000	2.000	3.000	2.000	1.000

Aceptar

Datos de comprobación ingresados al modelo (20%)

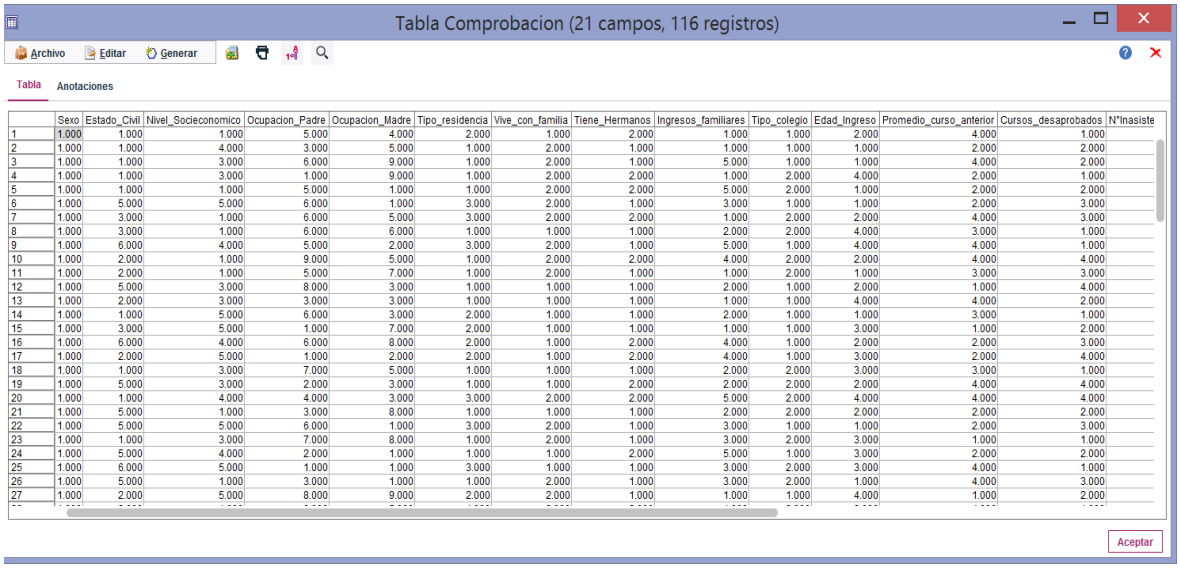


Tabla Comprobacion (21 campos, 116 registros)

Archivo Editar Generar

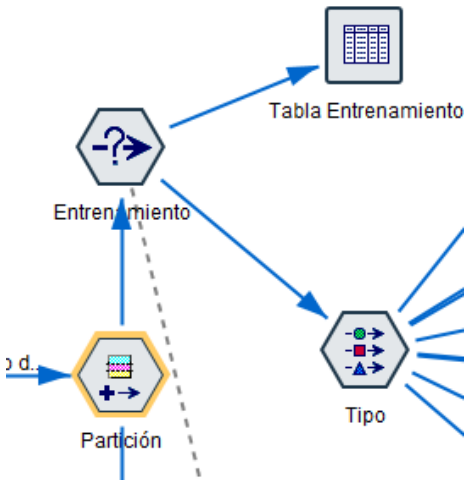
Tabla Anotaciones

	Sexo	Estado_Civil	Nivel_Socieconomico	Ocupacion_Padre	Ocupacion_Madre	Tipo_residencia	Vive_con_familia	Tiene_Hermanos	Ingresos_familiares	Tipo_colegio	Edad_Ingreso	Promedio_curso_anterior	Cursos_desaprobados	N°Inasiste
1	1.000	1.000	1.000	5.000	4.000	2.000	1.000	2.000	1.000	1.000	2.000	4.000	1.000	
2	1.000	1.000	4.000	3.000	5.000	1.000	2.000	1.000	1.000	1.000	1.000	2.000	2.000	
3	1.000	1.000	3.000	6.000	9.000	1.000	2.000	1.000	5.000	1.000	1.000	4.000	2.000	
4	1.000	1.000	3.000	1.000	9.000	1.000	2.000	2.000	1.000	2.000	4.000	2.000	1.000	
5	1.000	1.000	1.000	5.000	1.000	1.000	2.000	2.000	5.000	2.000	1.000	2.000	2.000	
6	1.000	5.000	5.000	6.000	1.000	3.000	2.000	1.000	3.000	1.000	1.000	2.000	3.000	
7	1.000	3.000	1.000	6.000	5.000	3.000	2.000	2.000	1.000	2.000	2.000	4.000	3.000	
8	1.000	3.000	1.000	6.000	6.000	1.000	1.000	1.000	2.000	2.000	4.000	3.000	1.000	
9	1.000	6.000	4.000	5.000	2.000	3.000	2.000	1.000	5.000	1.000	4.000	4.000	1.000	
10	1.000	2.000	1.000	9.000	5.000	1.000	2.000	2.000	4.000	2.000	2.000	4.000	4.000	
11	1.000	2.000	1.000	5.000	7.000	1.000	2.000	1.000	1.000	2.000	1.000	3.000	3.000	
12	1.000	5.000	3.000	8.000	3.000	1.000	1.000	1.000	2.000	1.000	2.000	1.000	4.000	
13	1.000	2.000	3.000	3.000	3.000	1.000	1.000	1.000	1.000	1.000	4.000	4.000	2.000	
14	1.000	1.000	5.000	6.000	3.000	2.000	1.000	1.000	2.000	1.000	1.000	3.000	1.000	
15	1.000	3.000	5.000	1.000	7.000	2.000	1.000	1.000	1.000	1.000	3.000	1.000	2.000	
16	1.000	6.000	4.000	6.000	8.000	2.000	1.000	2.000	4.000	1.000	2.000	2.000	3.000	
17	1.000	2.000	5.000	1.000	2.000	2.000	1.000	2.000	4.000	1.000	3.000	2.000	4.000	
18	1.000	1.000	3.000	7.000	5.000	1.000	1.000	1.000	2.000	2.000	3.000	3.000	1.000	
19	1.000	5.000	3.000	2.000	3.000	1.000	1.000	2.000	2.000	2.000	2.000	4.000	4.000	
20	1.000	1.000	4.000	4.000	3.000	3.000	2.000	2.000	5.000	2.000	4.000	4.000	4.000	
21	1.000	5.000	1.000	3.000	8.000	1.000	1.000	1.000	2.000	2.000	2.000	2.000	2.000	
22	1.000	5.000	5.000	6.000	1.000	3.000	2.000	1.000	3.000	1.000	1.000	2.000	3.000	
23	1.000	1.000	3.000	7.000	8.000	1.000	2.000	1.000	3.000	2.000	3.000	1.000	1.000	
24	1.000	5.000	4.000	2.000	1.000	1.000	1.000	2.000	5.000	1.000	3.000	2.000	2.000	
25	1.000	6.000	5.000	1.000	1.000	3.000	1.000	1.000	3.000	2.000	3.000	4.000	1.000	
26	1.000	5.000	1.000	3.000	1.000	1.000	2.000	1.000	3.000	2.000	1.000	4.000	3.000	
27	1.000	2.000	5.000	8.000	9.000	2.000	2.000	1.000	1.000	1.000	4.000	1.000	2.000	

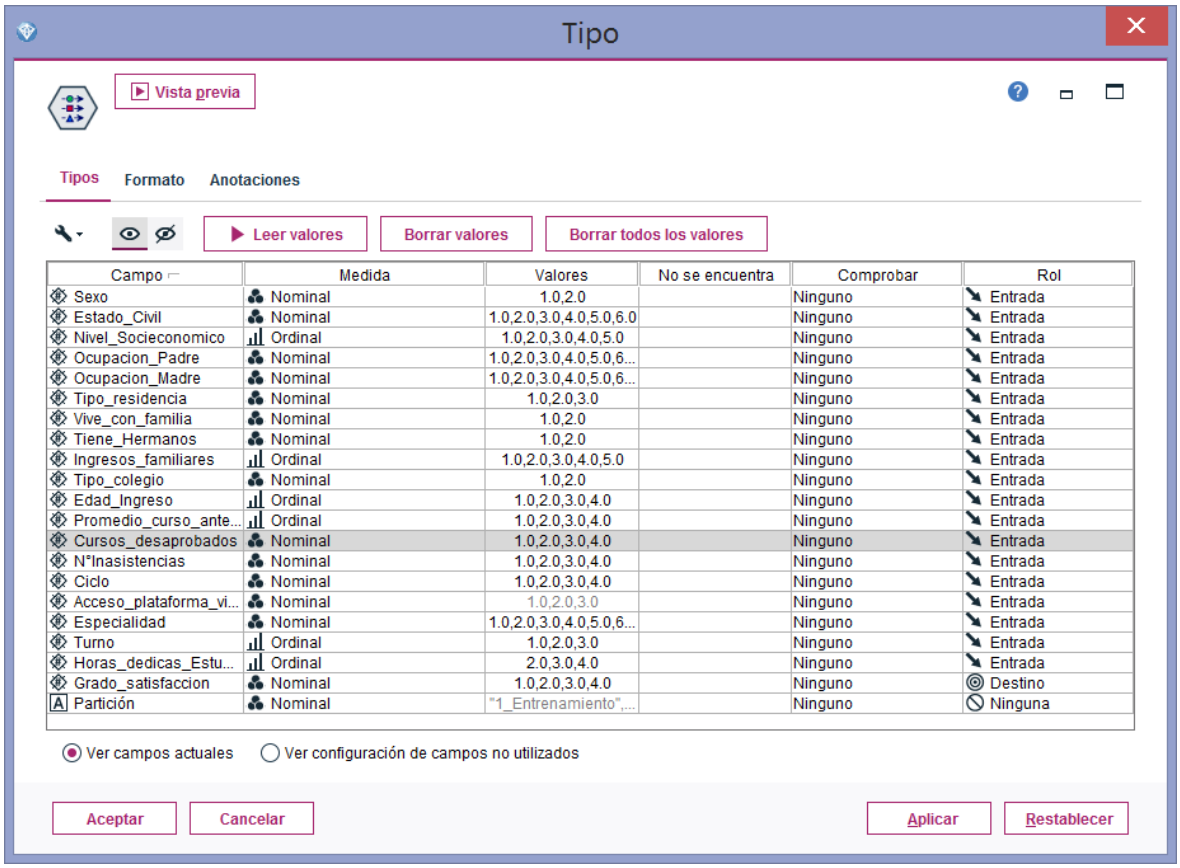
Aceptar

Ya estamos en disposición de intentar aprender un modelo a partir de los datos de entrenamiento, que dados unos determinados valores de los atributos de entrada nos dé un valor de salida.

Para ello añadimos un nodo “tipo”, que se encuentra en la categoría “Operaciones con Campos” y enlazamos el nodo seleccionar “Entrenamiento” “tipo”. Después editamos el nodo “tipo”. Como la salida va a ser “GradoDatisfacción” modificamos su dirección a salida “destino” y la categoría “Partición” en ninguno, como vemos en la siguiente figura:

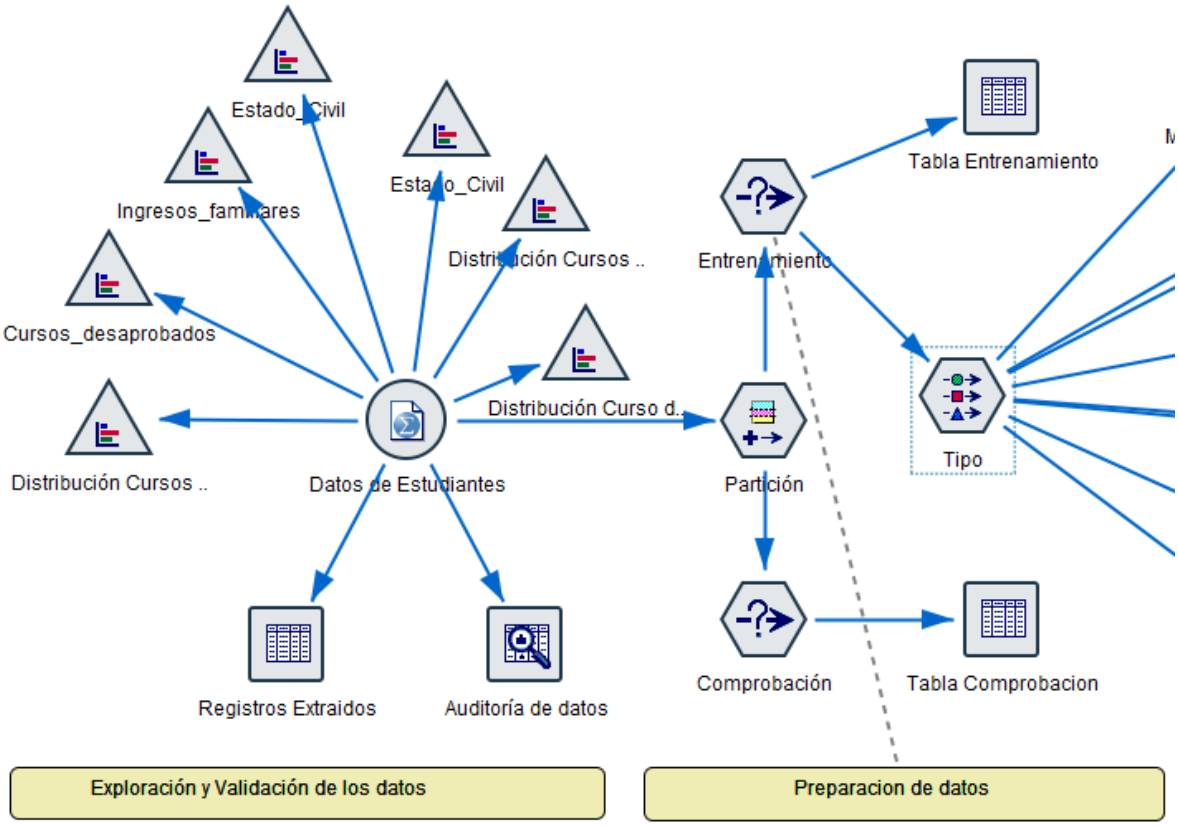


Configuraciones del nodo tipo



Campo	Medida	Valores	No se encuentra	Comprobar	Rol
Sexo	Nominal	1.0,2.0		Ninguno	Entrada
Estado_Civil	Nominal	1.0,2.0,3.0,4.0,5.0,6.0		Ninguno	Entrada
Nivel_Socieconomico	Ordinal	1.0,2.0,3.0,4.0,5.0		Ninguno	Entrada
Ocupacion_Padre	Nominal	1.0,2.0,3.0,4.0,5.0,6...		Ninguno	Entrada
Ocupacion_Madre	Nominal	1.0,2.0,3.0,4.0,5.0,6...		Ninguno	Entrada
Tipo_residencia	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Vive_con_familia	Nominal	1.0,2.0		Ninguno	Entrada
Tiene_Hermanos	Nominal	1.0,2.0		Ninguno	Entrada
Ingresos_familiares	Ordinal	1.0,2.0,3.0,4.0,5.0		Ninguno	Entrada
Tipo_colegio	Nominal	1.0,2.0		Ninguno	Entrada
Edad_Ingreso	Ordinal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Promedio_curso_ante...	Ordinal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Cursos_desaprobados	Nominal	1.0,2.0,3.0,4.0		Ninguno	Entrada
N°Inasistencias	Nominal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Ciclo	Nominal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Acceso_plataforma_vi...	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Especialidad	Nominal	1.0,2.0,3.0,4.0,5.0,6...		Ninguno	Entrada
Turno	Ordinal	1.0,2.0,3.0		Ninguno	Entrada
Horas_dedicas_Estu...	Ordinal	2.0,3.0,4.0		Ninguno	Entrada
Grado_satisfaccion	Nominal	1.0,2.0,3.0,4.0		Ninguno	Destino
Partición	Nominal	"1_Entrenamiento",...		Ninguno	Ninguna

Finalmente en este primer paso obtenemos la fase de la Preparación de los datos completa



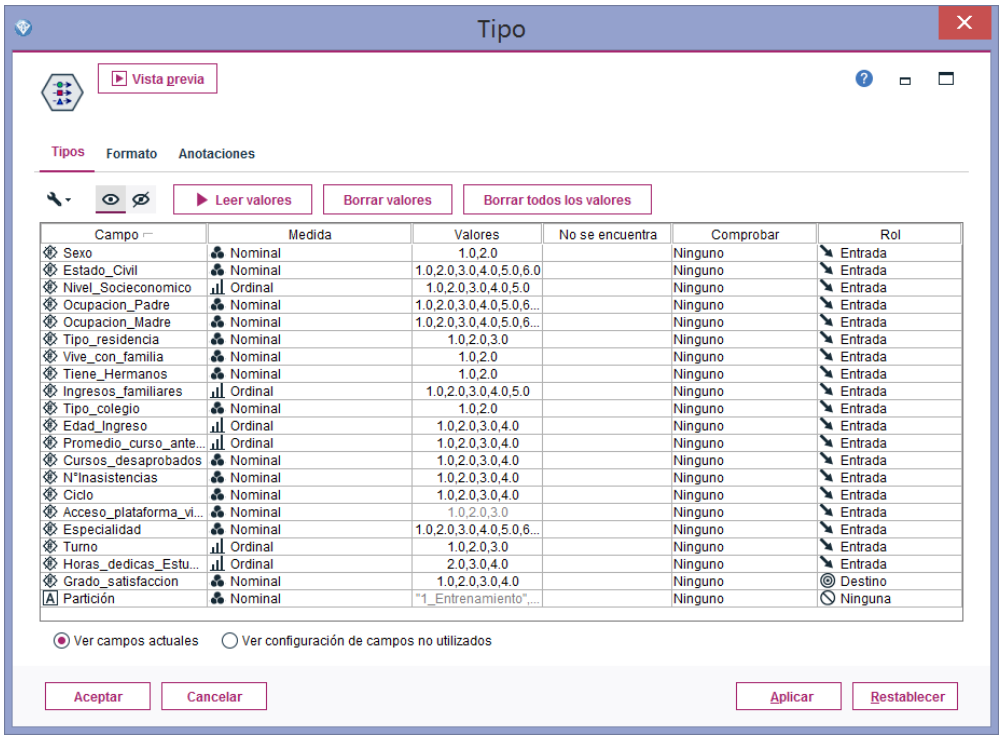
Segundo realizamos la construcción de cada modelo:

a. **Nodo Árbol C&R**
b. **Nodo Árbol C5.0**
c. **Nodo Árbol AS**
d. **Nodo Red Bayesiana**

a. Nodo Árbol C&R

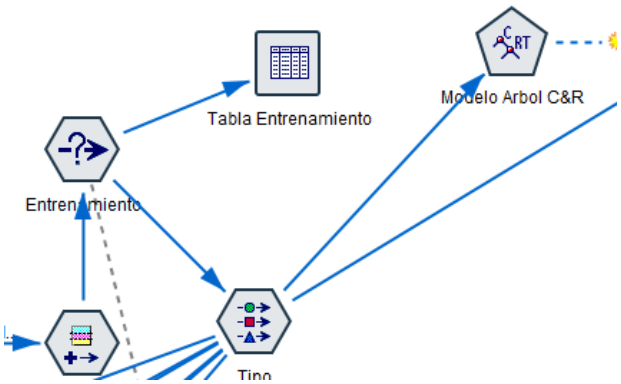
Para la creación de este Modelo usamos también los datos de entrenamiento conectando los datos desde el nodo tipo hacia un Nodo de Árbol C&R.

Datos a utilizar en el Nodo de Árbol C&R



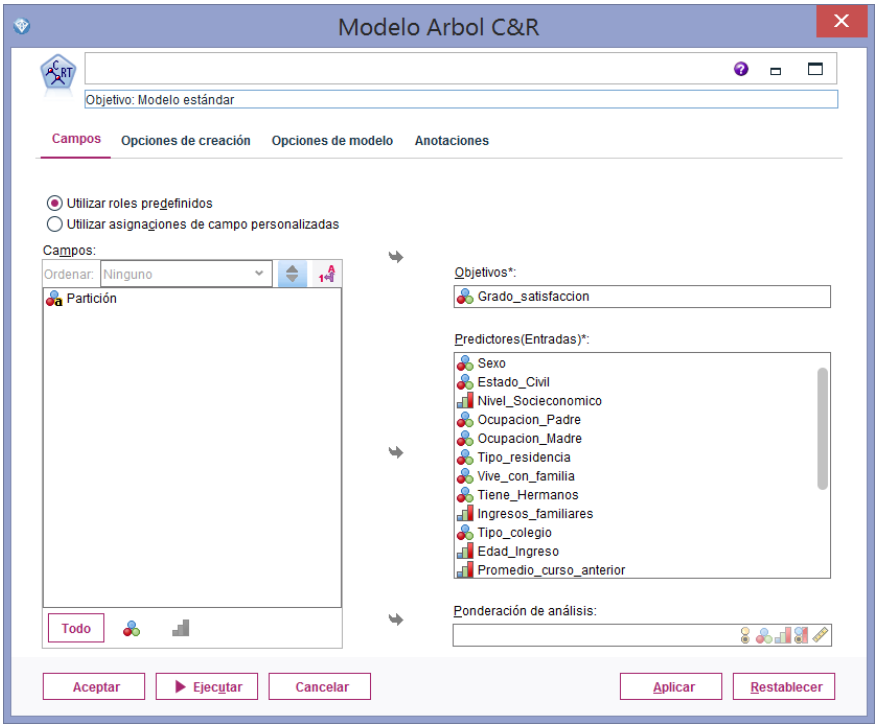
Campo	Medida	Valores	No se encuentra	Comprobar	Rol
Sexo	Nominal	1.0,2.0		Ninguno	Entrada
Estado_Civil	Nominal	1.0,2.0,3.0,4.0,5.0,6.0		Ninguno	Entrada
Nivel_Socieconomico	Ordinal	1.0,2.0,3.0,4.0,5.0		Ninguno	Entrada
Ocupacion_Padre	Nominal	1.0,2.0,3.0,4.0,5.0,6...		Ninguno	Entrada
Ocupacion_Madre	Nominal	1.0,2.0,3.0,4.0,5.0,6...		Ninguno	Entrada
Tipo_residencia	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Vive_con_familia	Nominal	1.0,2.0		Ninguno	Entrada
Tiene_Hermanos	Nominal	1.0,2.0		Ninguno	Entrada
Ingresos_familiares	Ordinal	1.0,2.0,3.0,4.0,5.0		Ninguno	Entrada
Tipo_colegio	Nominal	1.0,2.0		Ninguno	Entrada
Edad_Ingreso	Ordinal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Promedio_curso_ante...	Ordinal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Cursos_desaprobados	Nominal	1.0,2.0,3.0,4.0		Ninguno	Entrada
N°Inasistencias	Nominal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Ciclo	Nominal	1.0,2.0,3.0,4.0		Ninguno	Entrada
Acceso_plataforma_vi...	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Especialidad	Nominal	1.0,2.0,3.0,4.0,5.0,6...		Ninguno	Entrada
Turno	Ordinal	1.0,2.0,3.0		Ninguno	Entrada
Horas_dedicas_Estu...	Ordinal	2.0,3.0,4.0		Ninguno	Entrada
Grado_satisfaccion	Nominal	1.0,2.0,3.0,4.0		Ninguno	Destino
Partición	Nominal	"1_Entrenamiento",...		Ninguno	Ninguna

Nodo de Árbol C&R agregado a la ruta



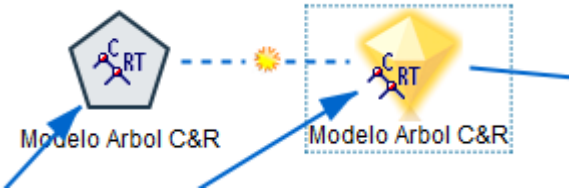
Luego configuramos el Nodo Árbol C&R , donde el campo escogido vienen desde la partición, el campo destino es el Grado_Satifaccion y los Predictores el resto de campos. Como se muestra en la imagen:

Configuración del Nodo de Árbol C&R



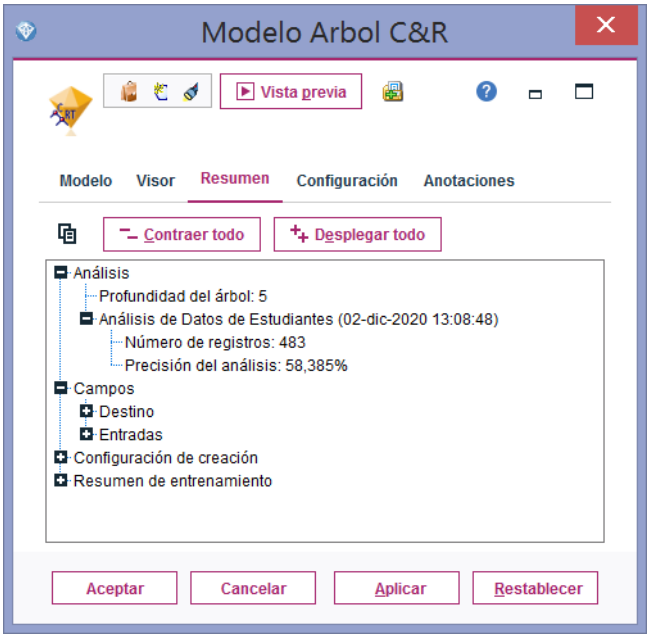
Luego se Ejecuta el Nodo árbol C&R para generar el Diamante (Nugget) correspondiente a este modelo.

Implementación del Nodo de Árbol C&R

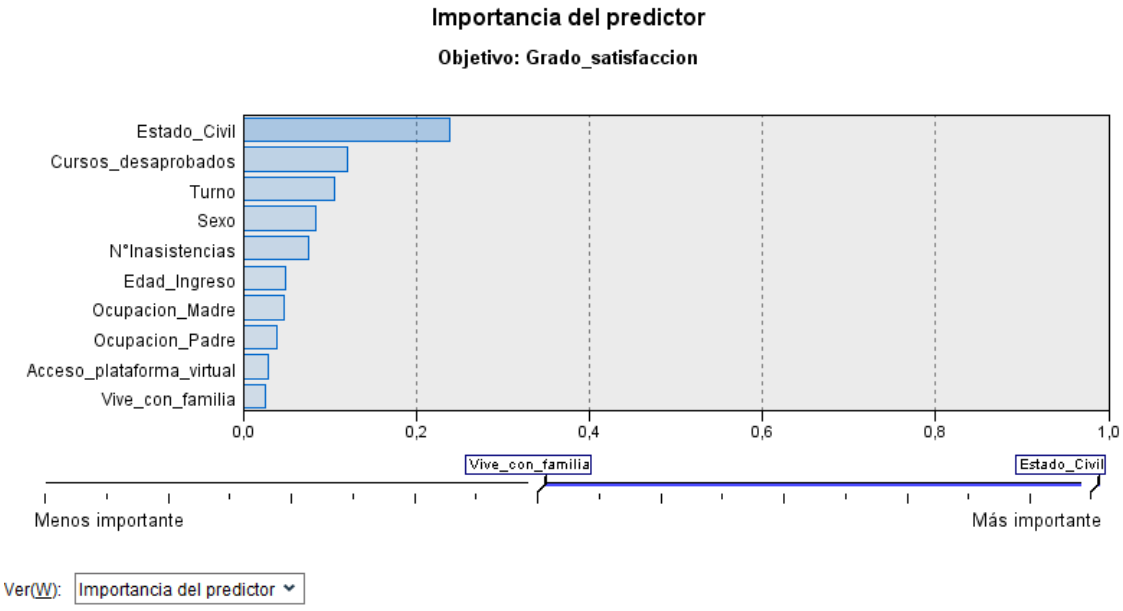


En el resumen del modelo podemos observar que la precisión del análisis es de 58,385%

Resumen del Nodo de Árbol C&R



También observamos que en la importancia del predictor nos da que también el campo Estado Civil es el de mayor importancia en el análisis



Obteniendo un árbol de profundidad de 5 como se aprecia en la imagen

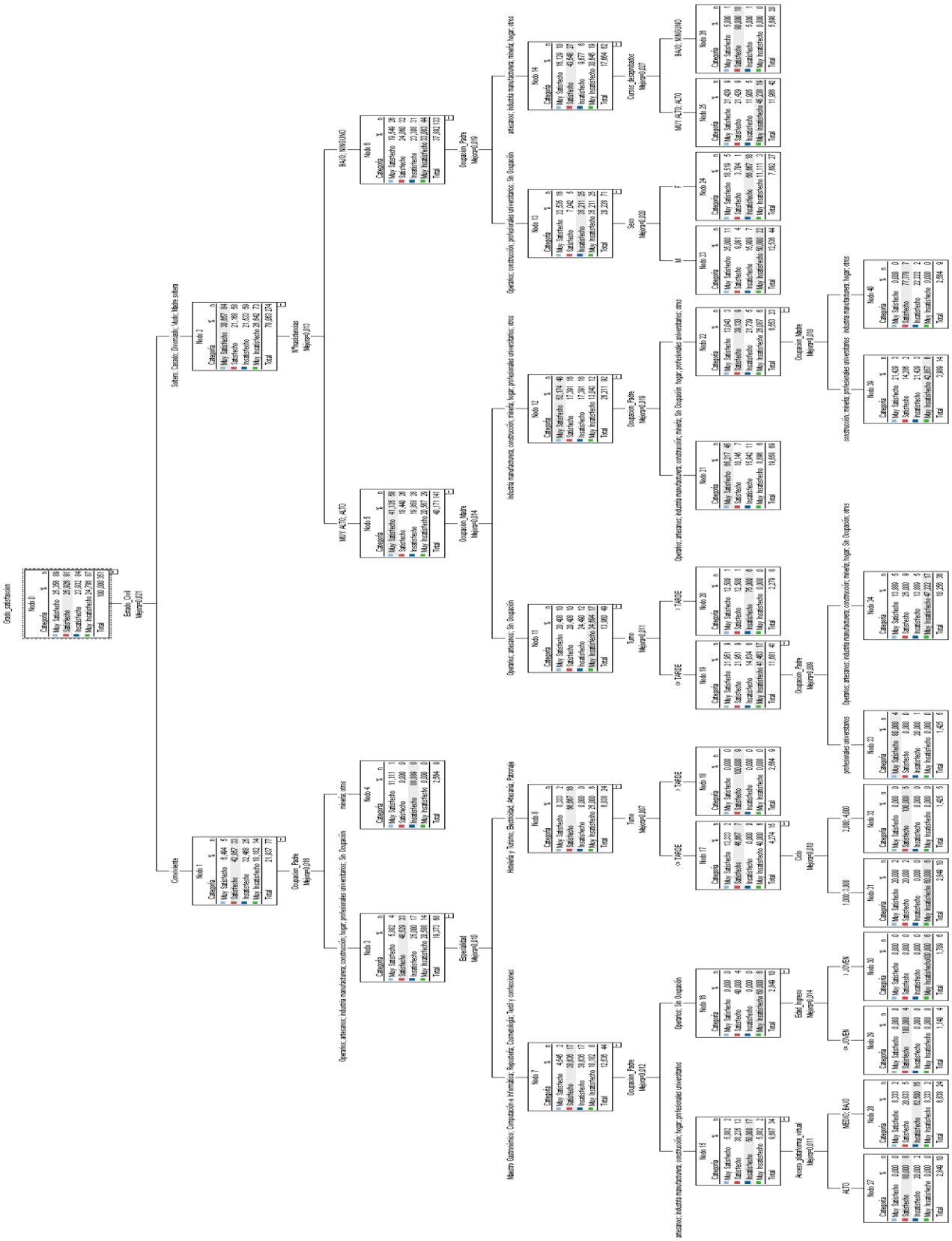
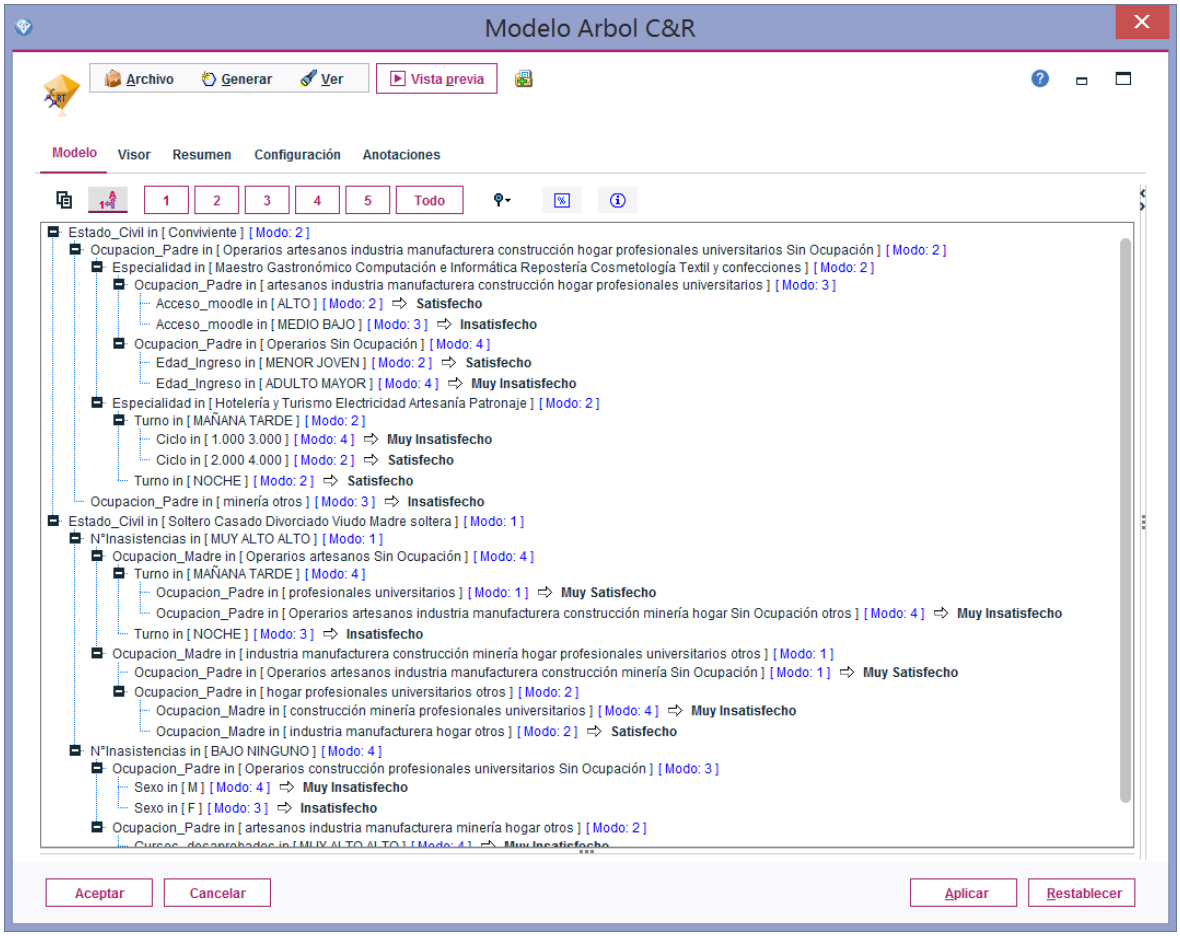


Figura 09: Árbol de decisión: Modelo Árbol C&R

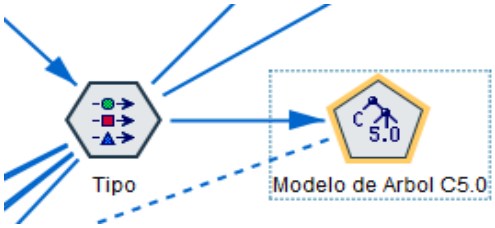
Modelo Generado por el Árbol C&R:



b. Nodo Árbol C5.0

Añadimos un nuevo nodo “C5.0” desde la categoría “Modelado” para construir un árbol de decisión sobre los datos. Conectamos el nodo “type” con el nodo “C5.0”, que pasa a llamarse Modelo Arbol C5.0, sin olvidar activar la opción de “Validacion cruzada” para la posterior evaluación, quedando el resultado como se muestra a continuación:

Nodo de árbol C5.0

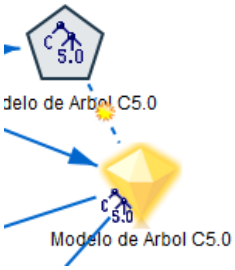


Configuración Modelo de entrenamiento del árbol

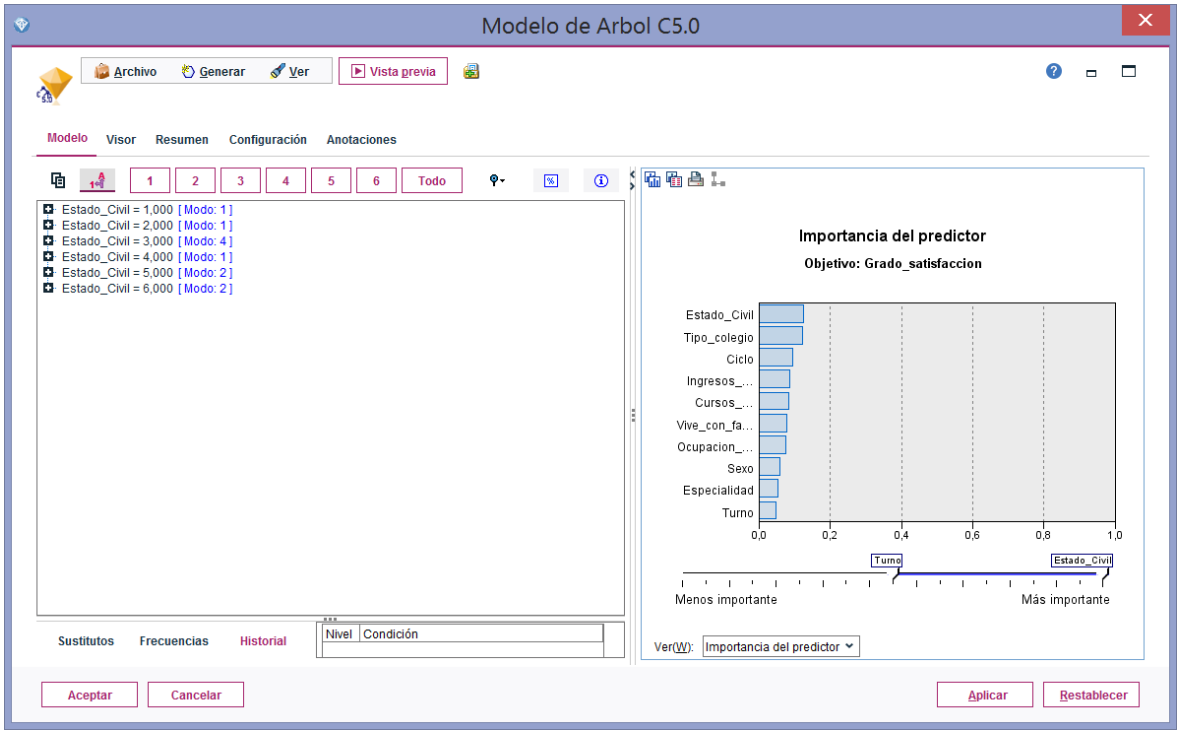
Ahora ya estamos en disposición de aprender un nuevo modelo (en este caso un árbol de decisión). Para ello ejecutamos el nodo Modelo Arbol C5.0.

Como se puede observar, se ha generado un nuevo icono en el área de trabajo de la derecha (pestaña de “Modelos”, con la forma de un diamante).

Modelo árbol C5.0 implementado



Verificamos reglas del árbol implementadas en el diamante obtenido. Se puede visualizar el modelo, así como la importancia del predictor:



¿Cómo interpretamos el árbol anterior? Si pulsamos en “Visor” tenemos una representación gráfica:

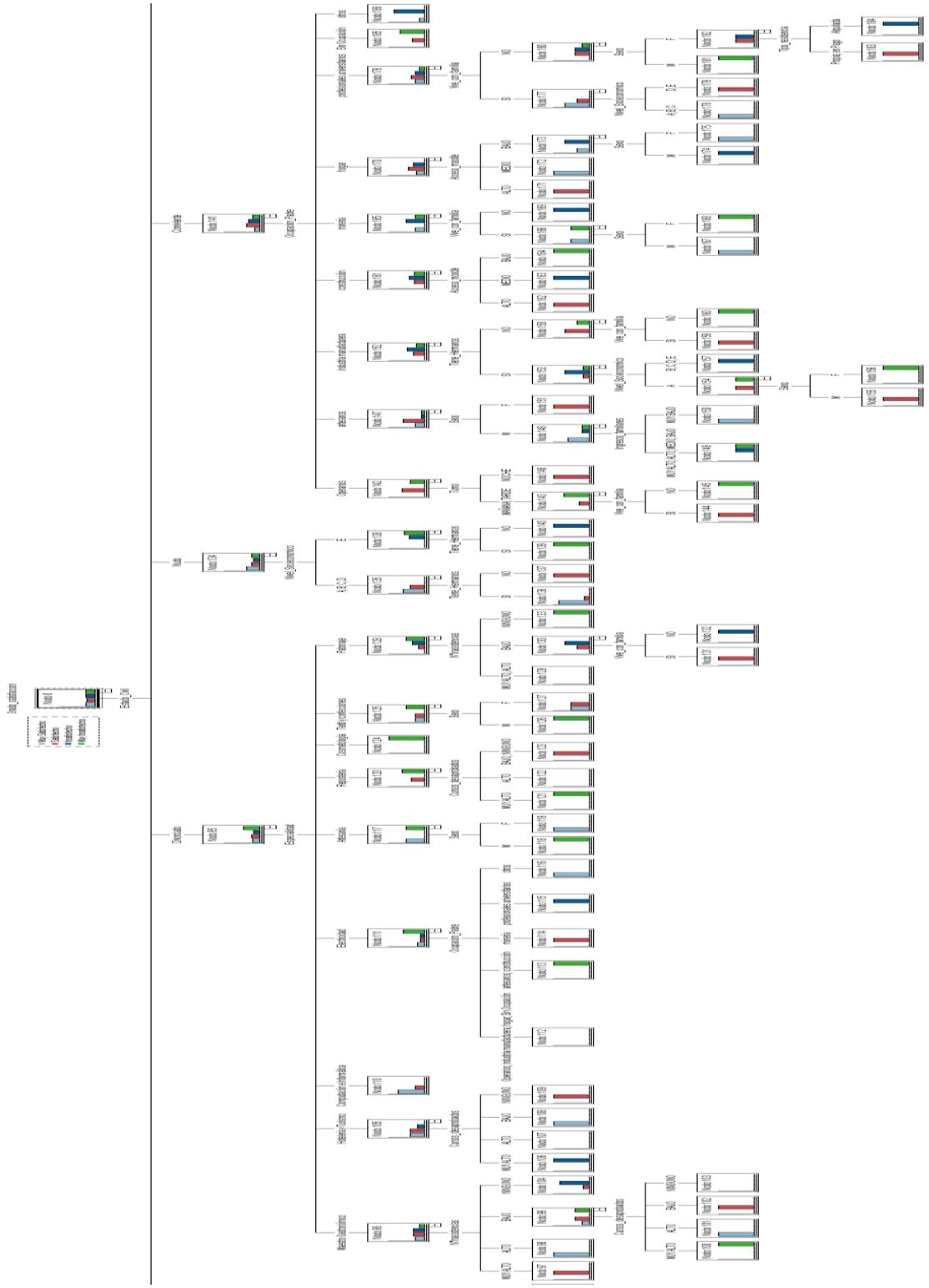
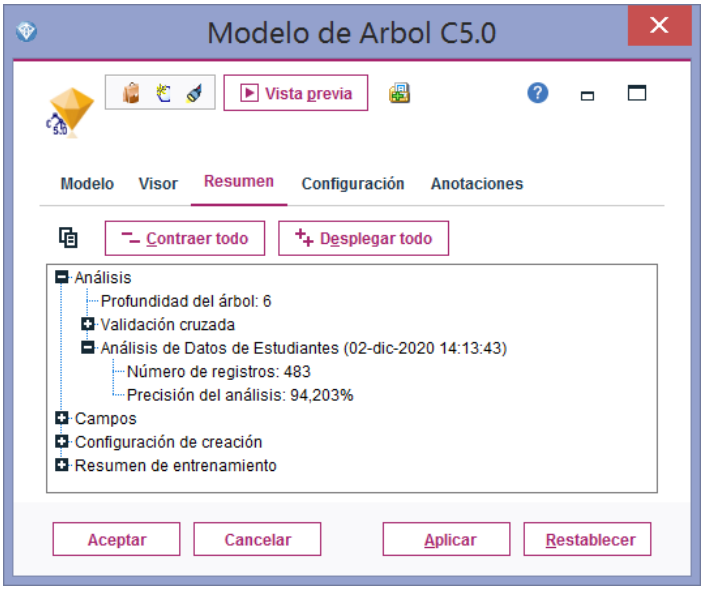


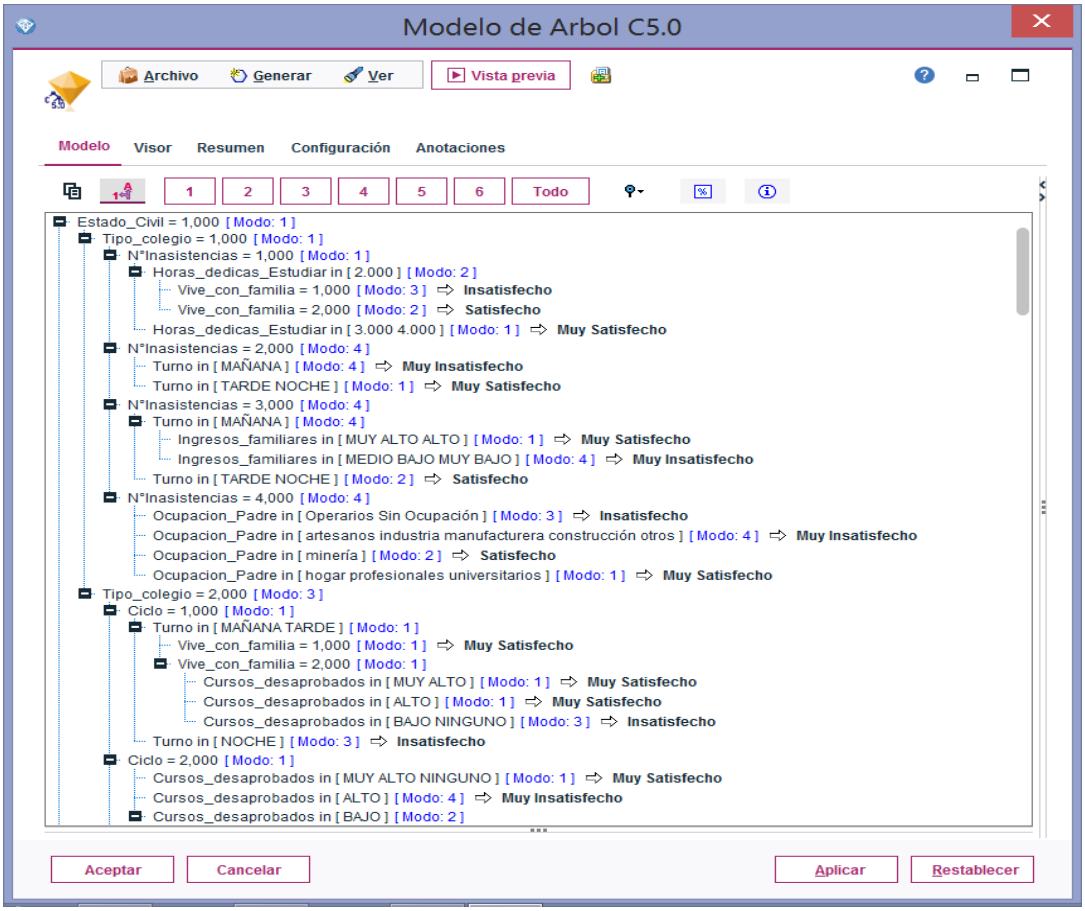
Figura 10: Árbol de decisión: Modelo Árbol C5.0

En el resumen de este modelo podemos observar que la precisión del análisis es de 94.203%

Resumen Modelo árbol C5.0



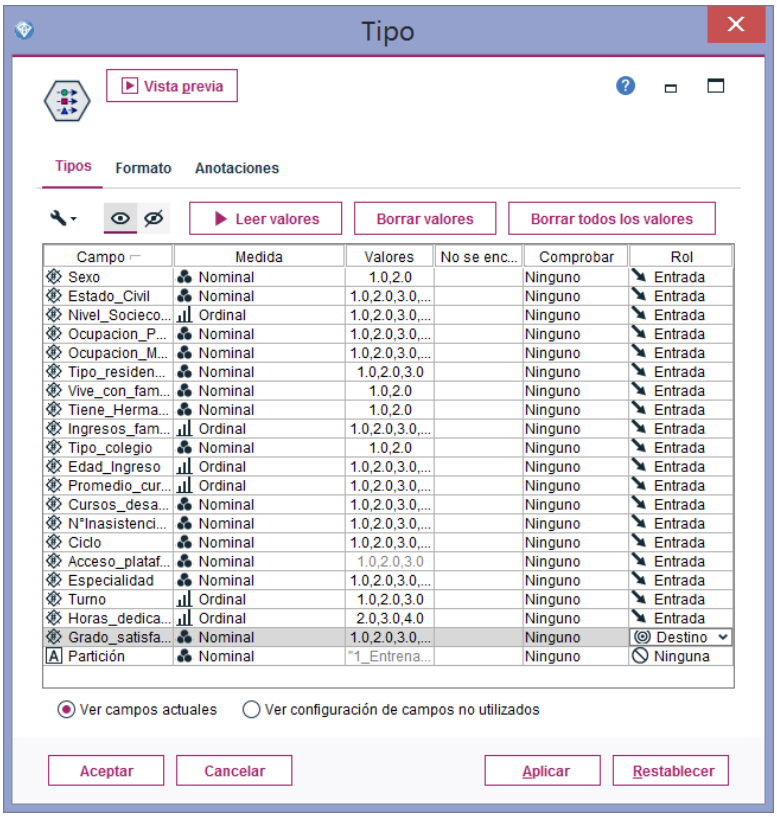
Modelo Generado por el Árbol C5.0:



c. Nodo Árbol AS (CHAID)

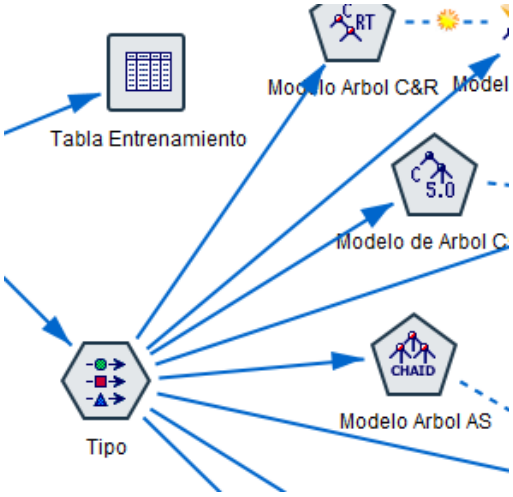
Para la creación de este Modelo usamos también los datos de entrenamiento conectando los datos desde el nodo tipo hacia un Nodo de Arbol AS.

Datos a ingresar al Nodo de Árbol AS



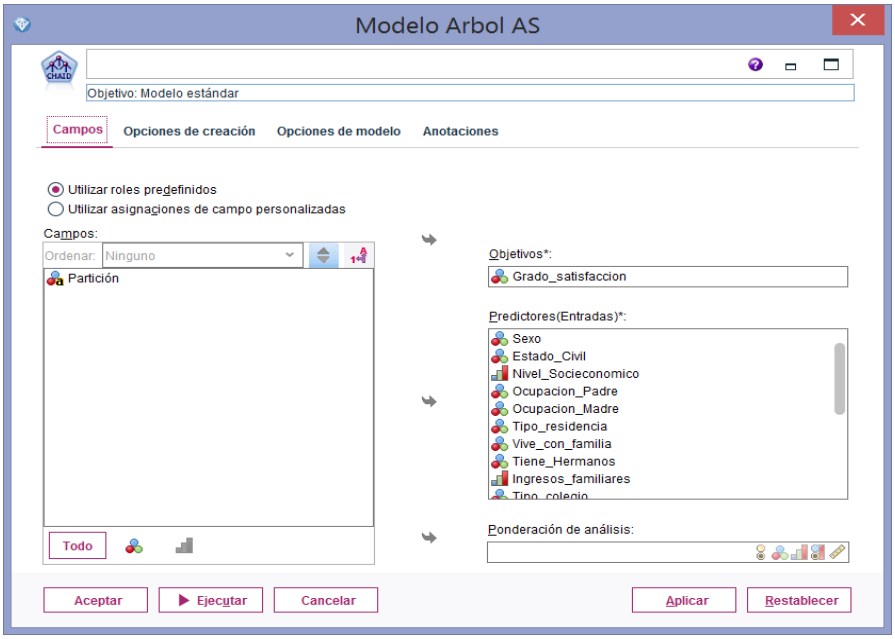
Campo	Medida	Valores	No se enc...	Comprobar	Rol
Sexo	Nominal	1.0,2.0		Ninguno	Entrada
Estado_Civil	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Nivel_Socieco...	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Ocupacion_P...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Ocupacion_M...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Tipo_residen...	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Vive_con_fam...	Nominal	1.0,2.0		Ninguno	Entrada
Tiene_Herma...	Nominal	1.0,2.0		Ninguno	Entrada
Ingresos_fam...	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Tipo_colegio	Nominal	1.0,2.0		Ninguno	Entrada
Edad_Ingreso	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Promedio_cur...	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Cursos_desa...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
N°Inasistenci...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Ciclo	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Acceso_plataf...	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Especialidad	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Turno	Ordinal	1.0,2.0,3.0		Ninguno	Entrada
Horas_dedica...	Ordinal	2.0,3.0,4.0		Ninguno	Entrada
Grado_satisfa...	Nominal	1.0,2.0,3.0,...		Ninguno	Destino
Partición	Nominal	"1_Entrena...		Ninguno	Ninguna

Nodo Árbol AS agregado a la ruta



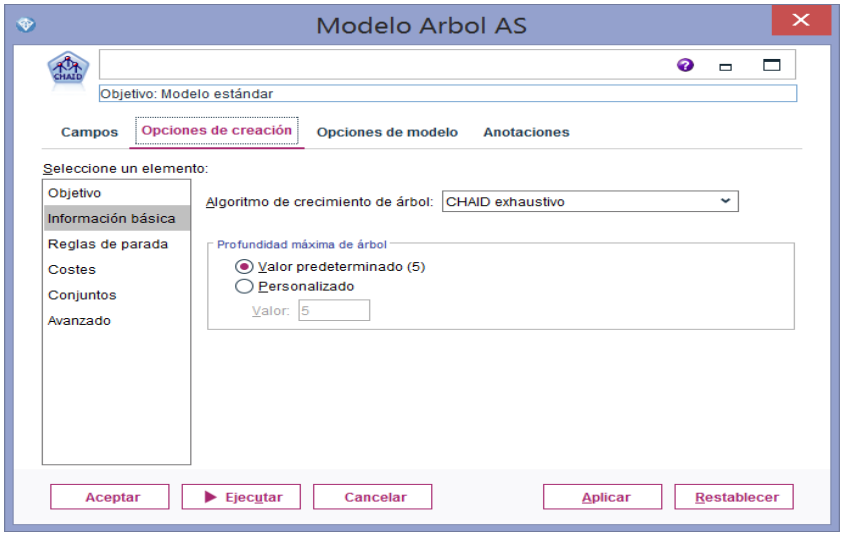
Luego configuramos el Nodo Árbol AS , donde el campo escogido vienen desde la partición, el campo destino es el Grado de Satisfacción y los Predictores el resto de campos. Como se muestra en la imagen:

Configuración del Nodo Árbol AS

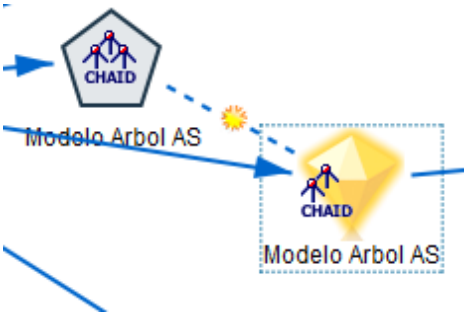


Después se ha escogido la opción de generación de un árbol AS (CHAID) Exhaustivo que examina con mayor precisión todas las divisiones posibles para cada predictor.

Opciones de generación de un árbol AS

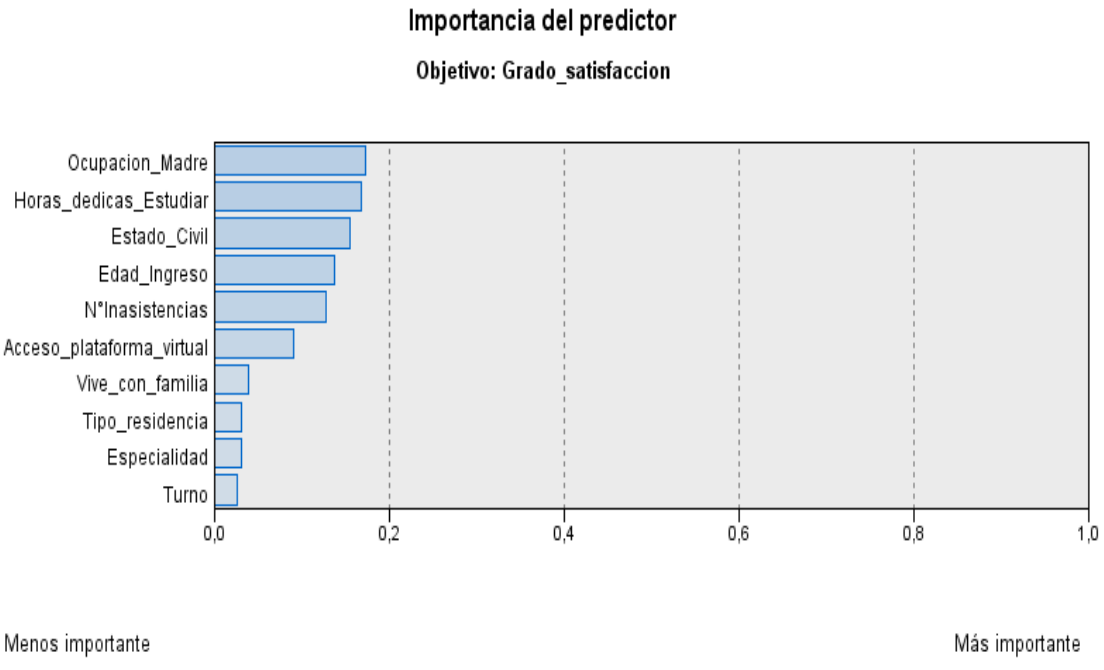


Luego se Ejecuta el Nodo árbol AS para generar el Nugget correspondiente a este modelo.



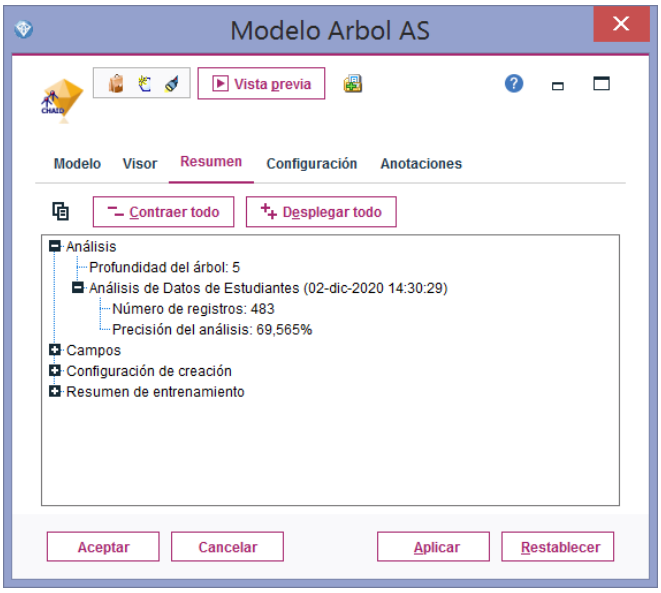
Modelo árbol AS implementado

Tenemos como resultado en importancia de predictor que Ocupación de la madre tiene mayor importancia.



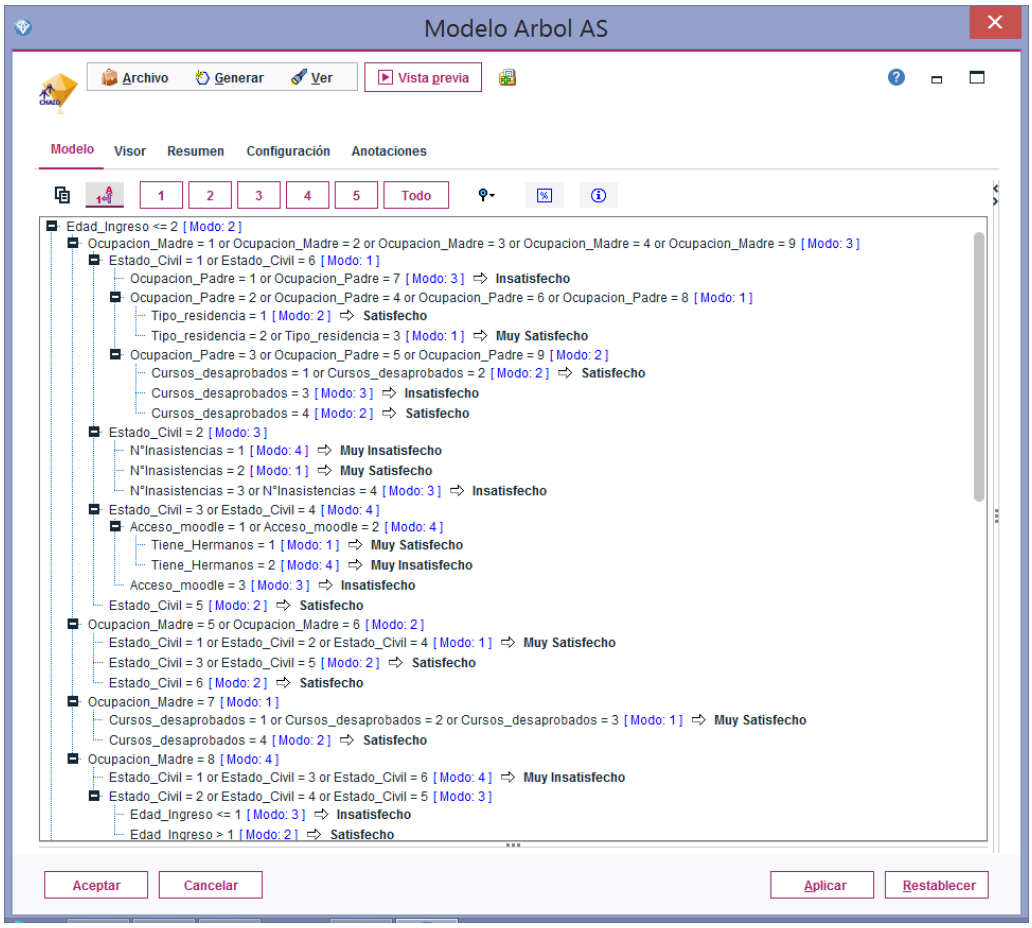
En el resumen del modelo podemos observar que la precisión del análisis es de 69.565%

Resumen del modelo Árbol AS



Obteniendo un árbol de profundidad de 5 como se aprecia en la imagen

Modelo Generado por el Árbol AS (CHAID)



Árbol de decisión: Modelo Árbol AS

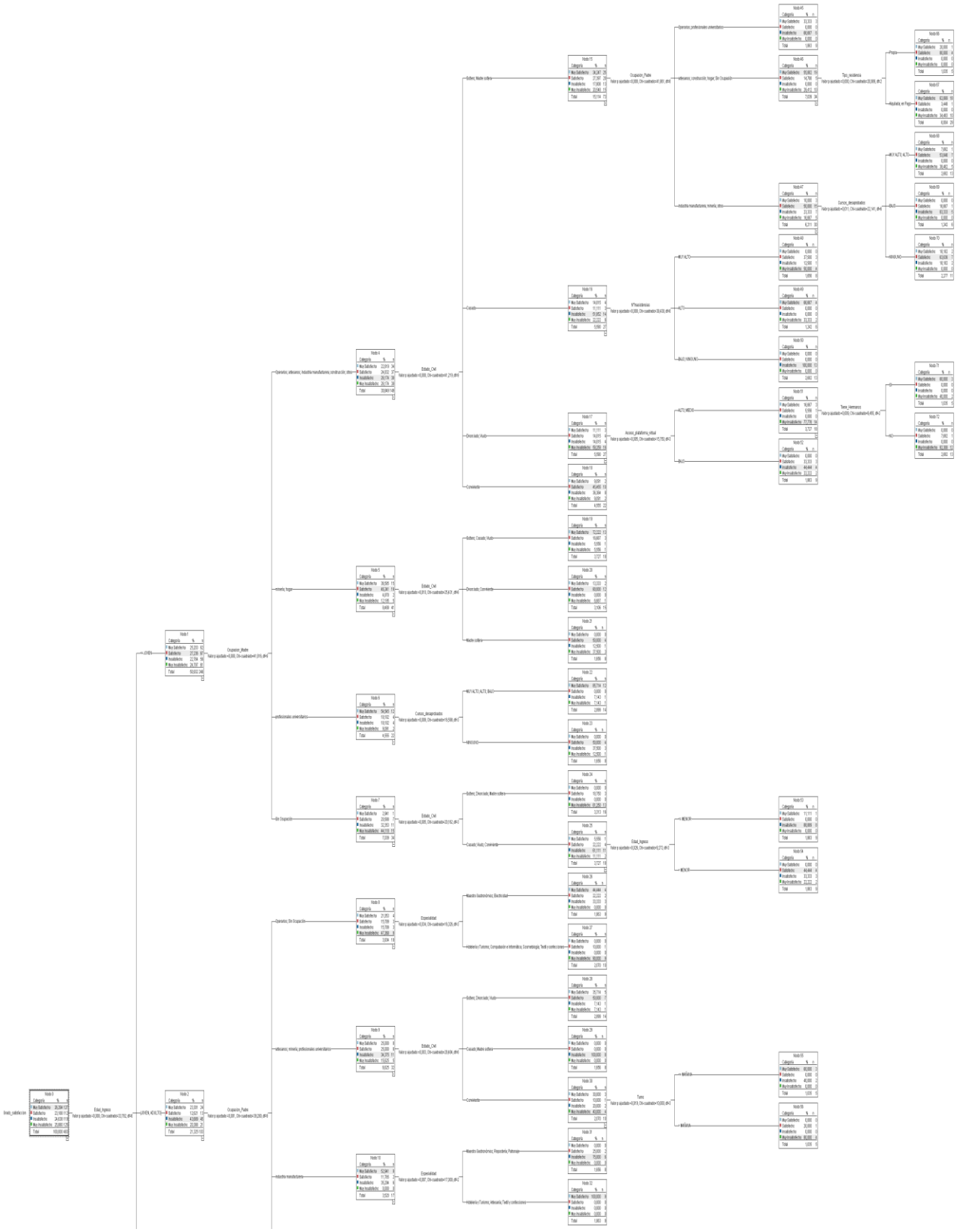
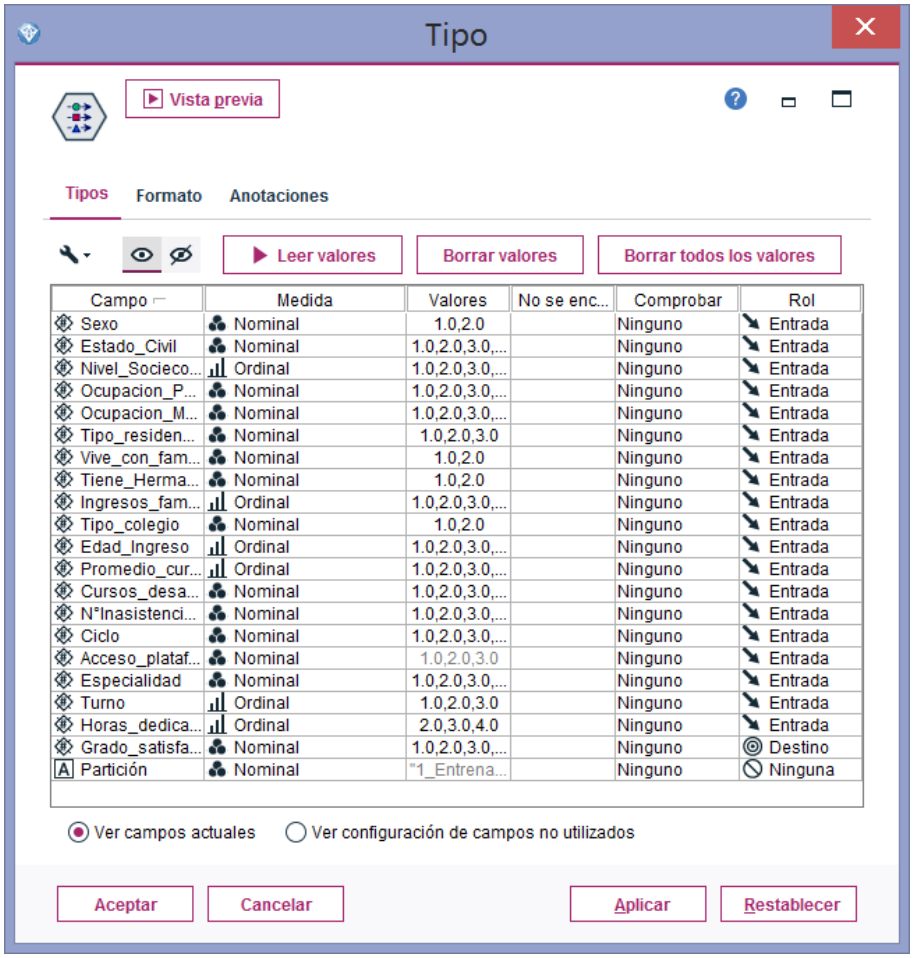


Figura 11: Árbol de decisión: Modelo Árbol AS

d. Nodo Red Bayesiana

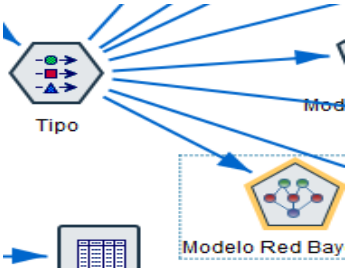
Para la construcción de este modelo utilizamos también el nodo tipo para el ingreso de los datos:

Datos de entrada al modelo



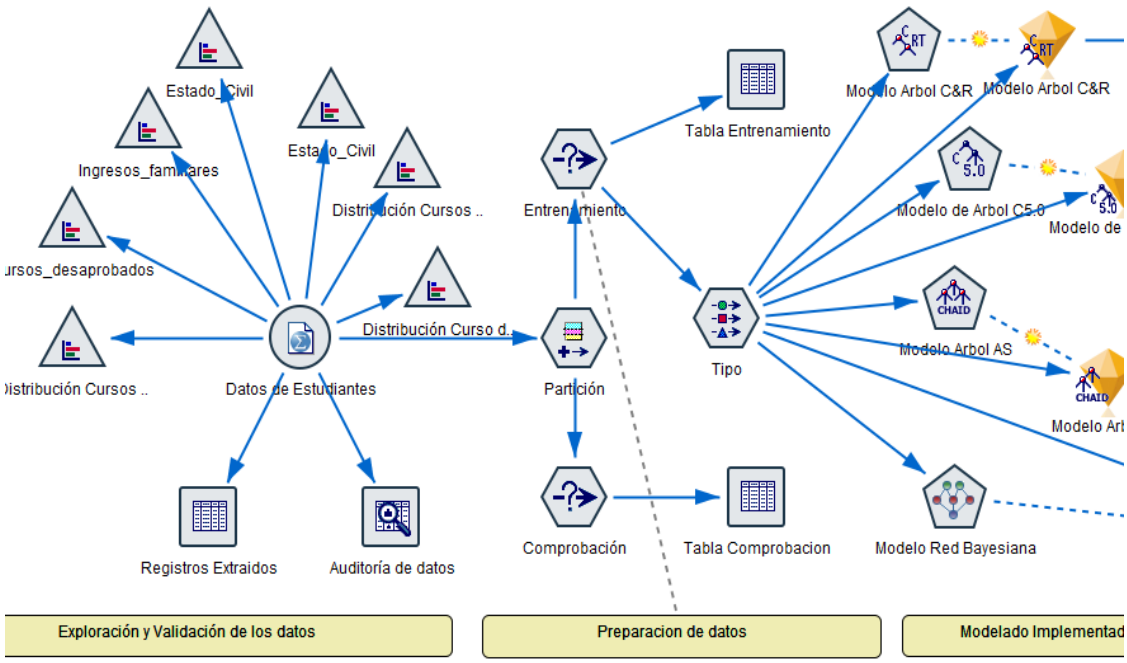
Campo	Medida	Valores	No se enc...	Comprobar	Rol
Sexo	Nominal	1.0,2.0		Ninguno	Entrada
Estado_Civil	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Nivel_Socieco...	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Ocupacion_P...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Ocupacion_M...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Tipo_residen...	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Vive_con_fam...	Nominal	1.0,2.0		Ninguno	Entrada
Tiene_Herma...	Nominal	1.0,2.0		Ninguno	Entrada
Ingresos_fam...	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Tipo_colegio	Nominal	1.0,2.0		Ninguno	Entrada
Edad_Ingreso	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Promedio_cur...	Ordinal	1.0,2.0,3.0,...		Ninguno	Entrada
Cursos_desa...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
N°Inasistenci...	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Ciclo	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Acceso_plataf...	Nominal	1.0,2.0,3.0		Ninguno	Entrada
Especialidad	Nominal	1.0,2.0,3.0,...		Ninguno	Entrada
Turno	Ordinal	1.0,2.0,3.0		Ninguno	Entrada
Horas_dedica...	Ordinal	2.0,3.0,4.0		Ninguno	Entrada
Grado_satisfa...	Nominal	1.0,2.0,3.0,...		Ninguno	Destino
Partición	Nominal	"1_Entrena...		Ninguno	Ninguna

Luego agregamos el "nodo Red bayesiana" obtenido desde la paleta de "Molelado"



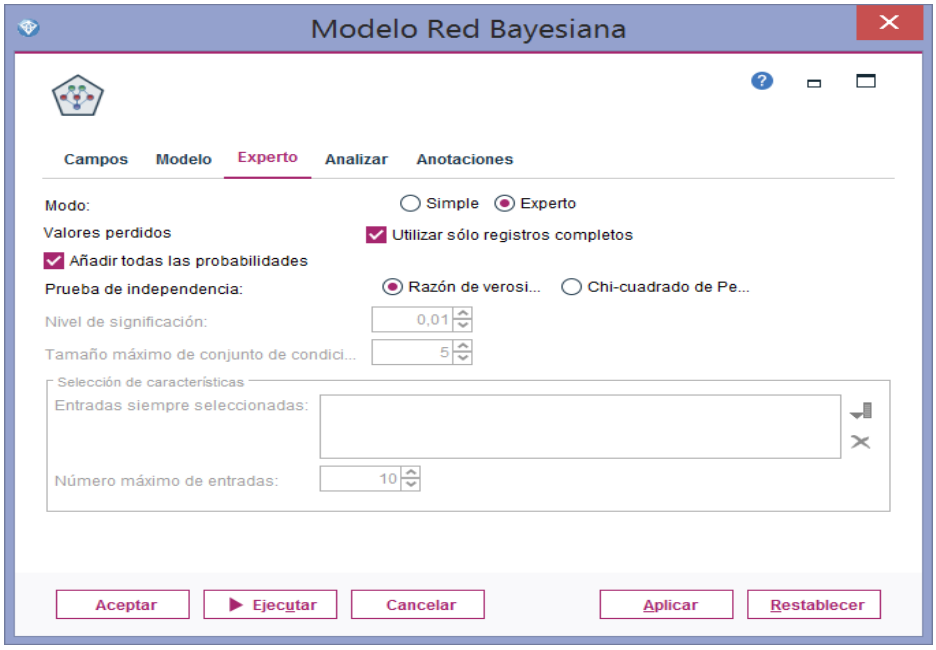
Conectamos el noto tipo con el nodo Red Bayesiana para luego configurar en Nodo de la Red Bayesiana como se muestra a continuación:

Red Bayesiana Agregada a la ruta



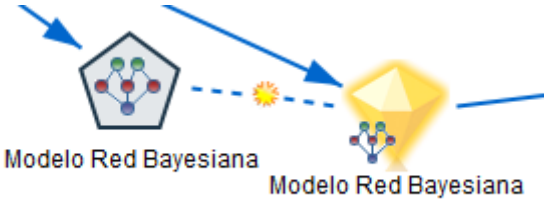
Se configura el Nodo de Red Bayesiana para que realice un análisis Experto permitiendo ajustar el proceso de generación de modelos

Configuración del Nodo de Red Bayesiana



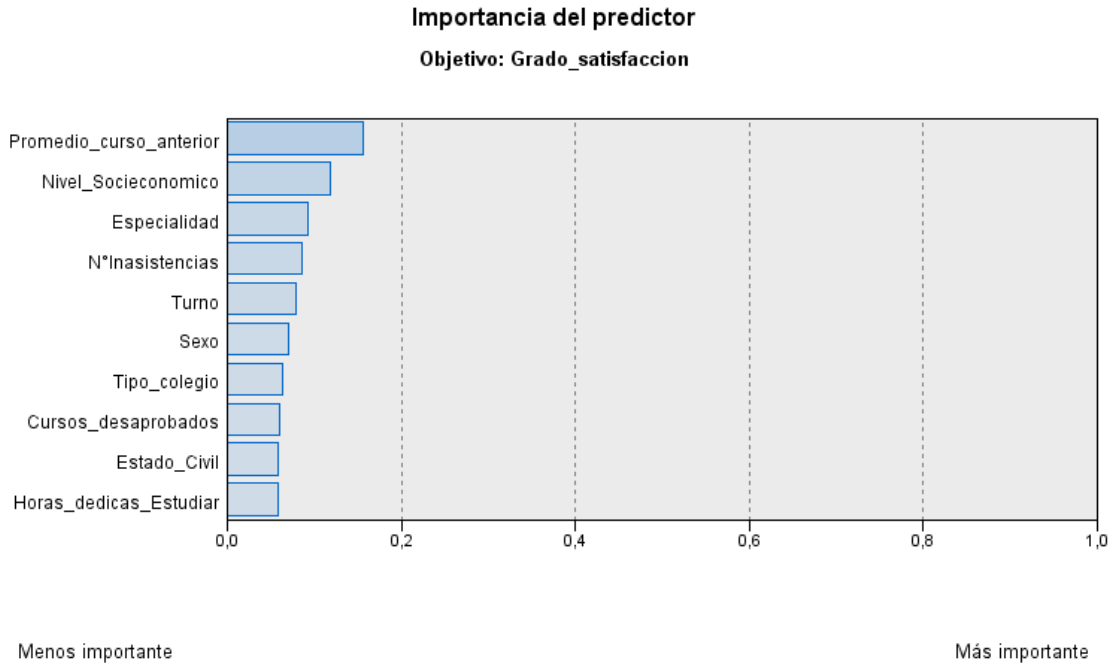
Luego ejecutamos el Nodo de Red Bayesiana para que genere el Nugget (Diamante) del modelo.

Implementación del modelo Red Bayesiana



Analizamos el resultado del modelo observando lo siguiente:

La importancia del Predictor nos dice que los de mayor importancia es el promedio del curso anterior, nivel socioeconómico, especialidad.



También se observa la red de gráficos de nodos que muestra la relación entre el objetivo Grado de satisfacción y sus predictores más importantes, y la relación entre los predictores. La importancia de cada predictor se muestra según la densidad del color; un color más fuerte muestra un predictor importante y viceversa.

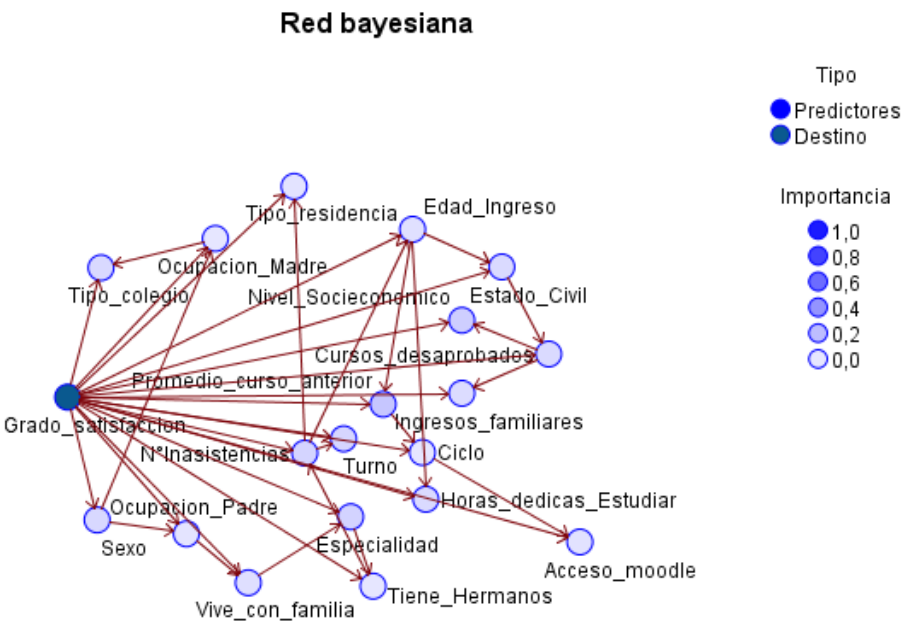
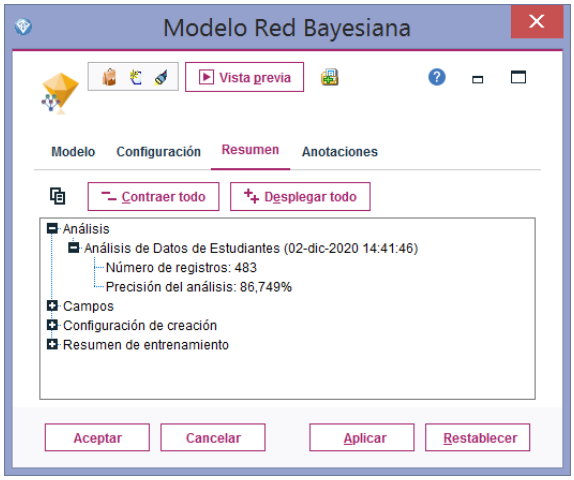


Figura 12: Red Bayesiana

En el resumen del modelo podemos observar que la precisión del análisis es de 86.749%

Resumen del modelo Red Bayesiana



4.4. EVALUACIÓN DEL MODELO:

En este paso interpretamos los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito y precisión que devuelve cada modelo.

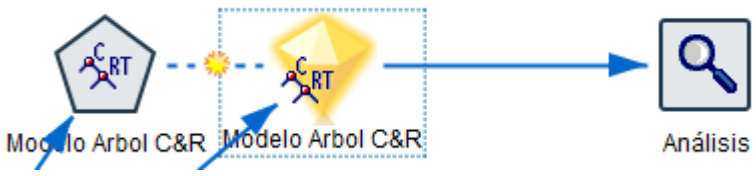
Se evaluaron el resultado de cada Nugget (diamante) generado de cada modelo donde se observa las siguientes diferencias:

a. Modelo Árbol C&R

La profundidad del árbol es de 5 y su precisión de análisis es de 58.385%

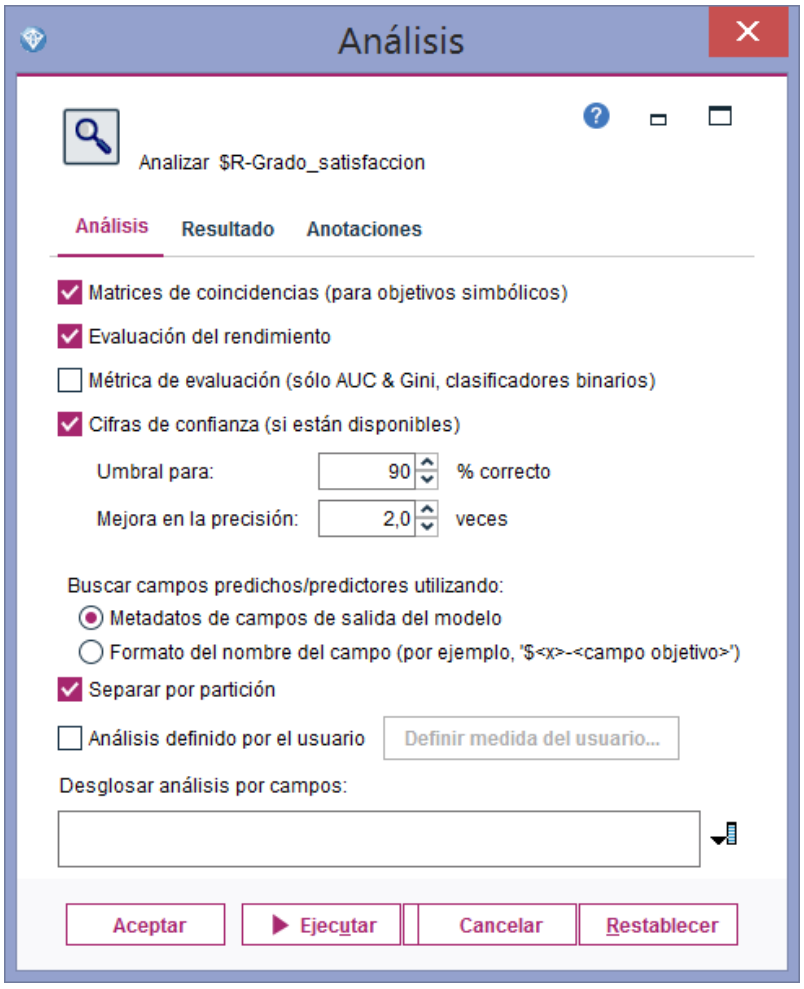
También se agregó un Nodo de Análisis donde se dan Resultados para el campo de salida.

Evaluación del modelo del árbol C&R



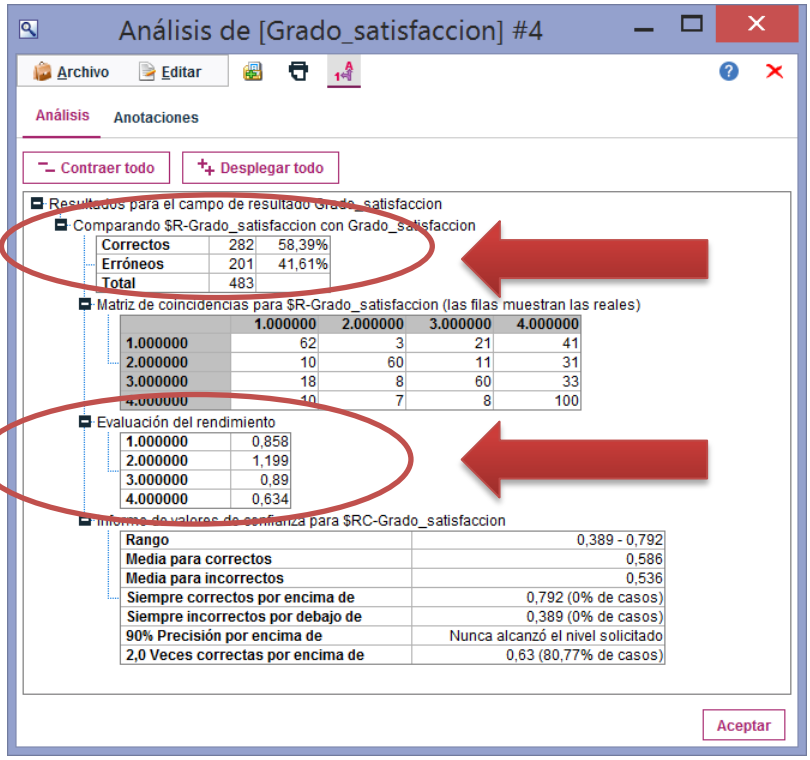
Se configura este análisis para el árbol C&R para que muestre: Matrices de coincidencias, evaluación de rendimiento, cifras de confianza.

Configuración del análisis de la evaluación

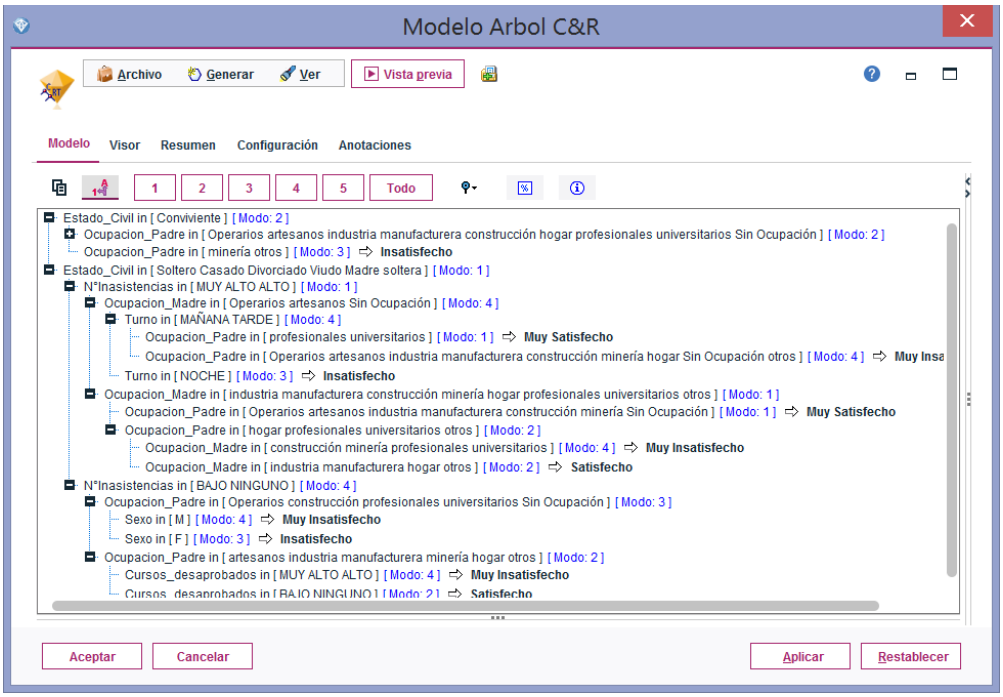


Obteniendo el siguiente resultado:

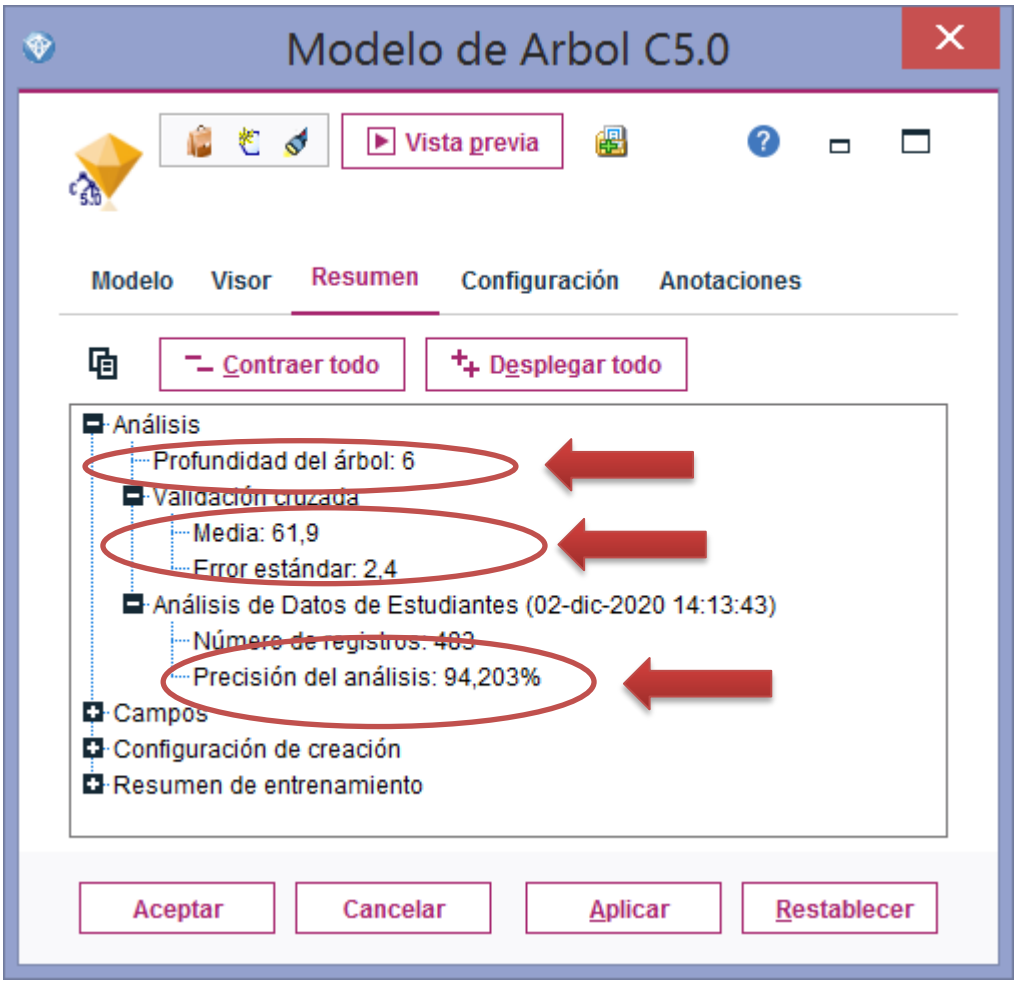
Resultados de la evaluación del modelo árbol C&R



Aquí también mostramos las Reglas de Clasificación usando la técnica o algoritmo árbol C&R



b. Modelo Árbol C5.0



La profundidad del árbol es de 6 y su precisión de análisis es de **94.203%**

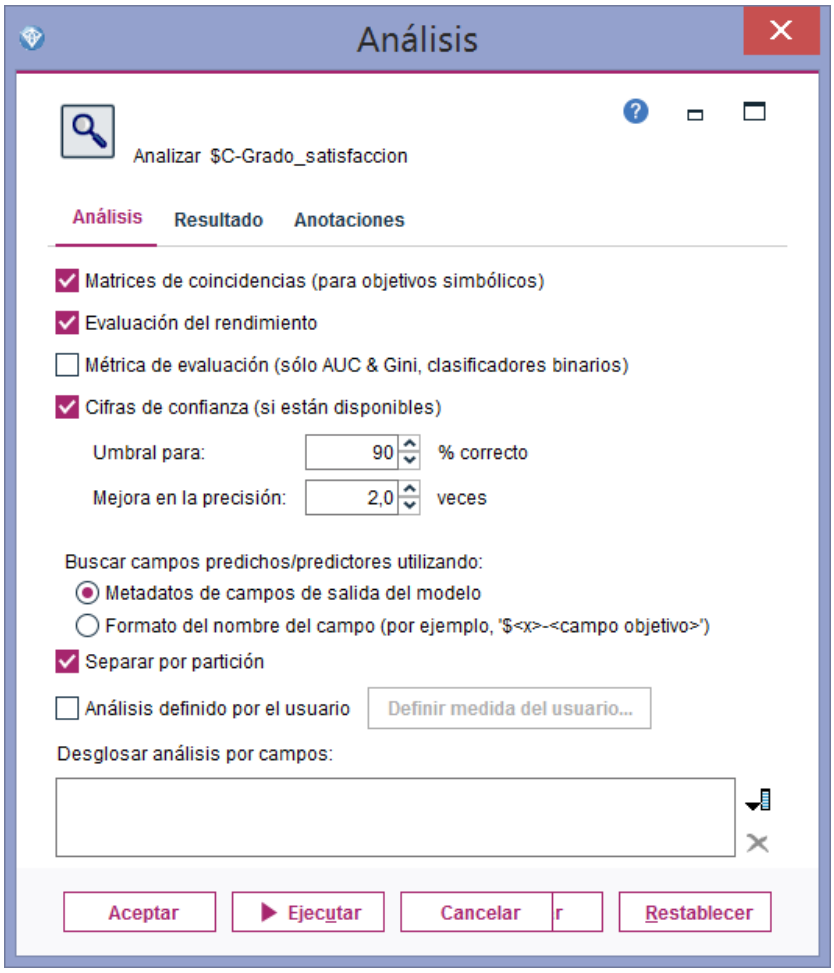
También se agregó un Nodo de Análisis donde se dan Resultados para el campo de salida. El resultado del análisis contiene una sección para cada campo de salida con un campo de predicción correspondiente creado por un modelo generado. Comparando dentro de la sección del campo de salida con una subsección para cada campo de predicción asociada con dicho campo de salida.

Para campos de salida categóricos, el nivel superior de esta sección contiene una tabla que muestra el número y el porcentaje de predicciones correctas e incorrectas y el número total de registros en la ruta.

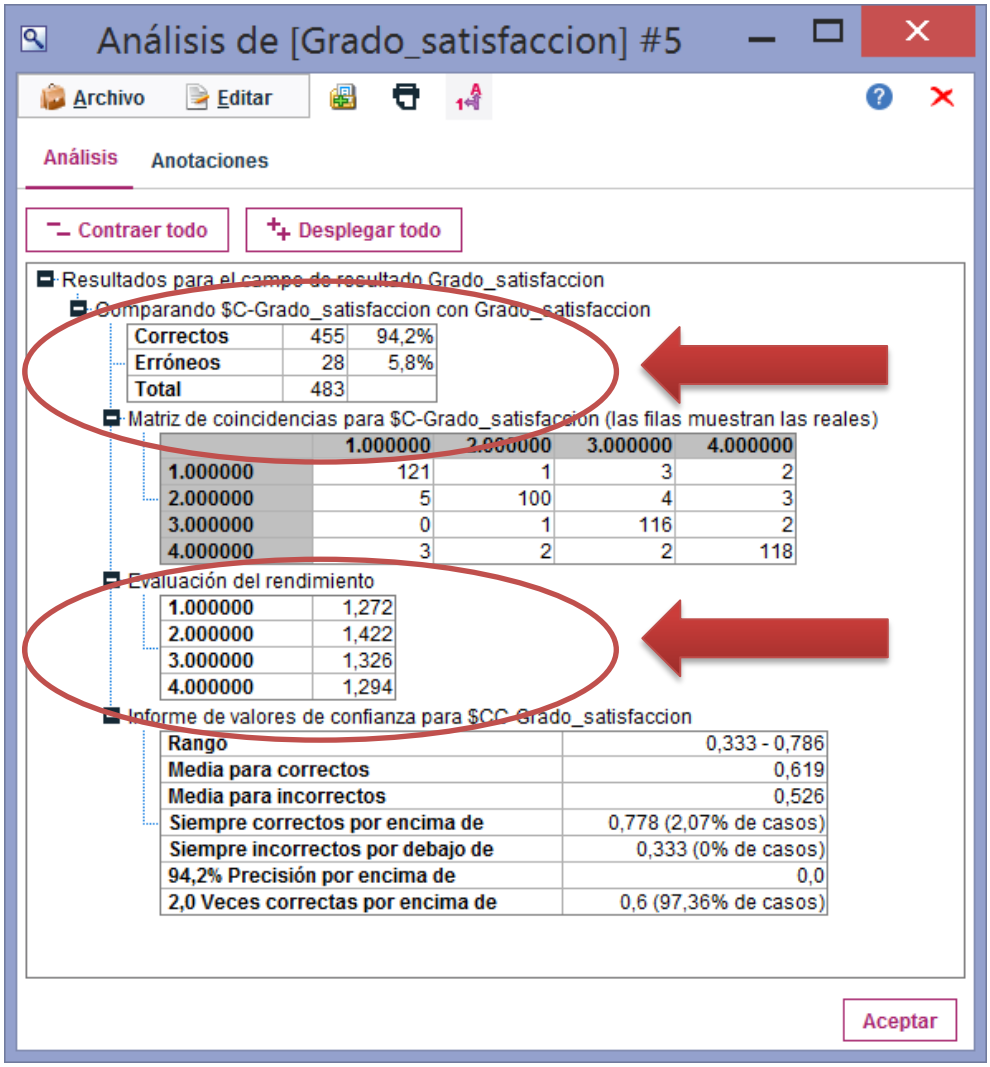
Evaluación del modelo del árbol C5.0

Se configura este análisis para que muestre: Matrices de coincidencias, evaluación de rendimiento, cifras de confianza.

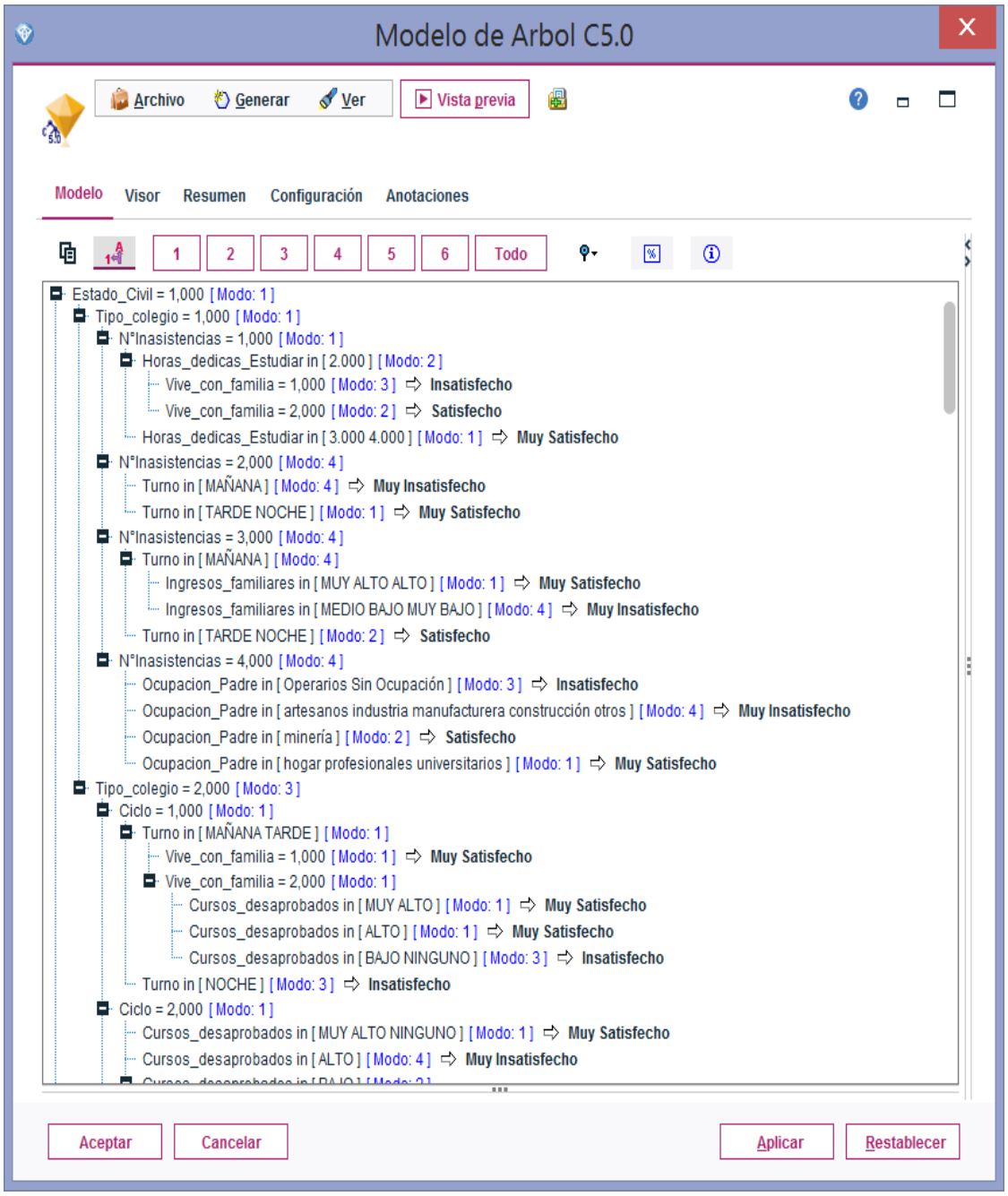
Configuración del análisis de la evaluación



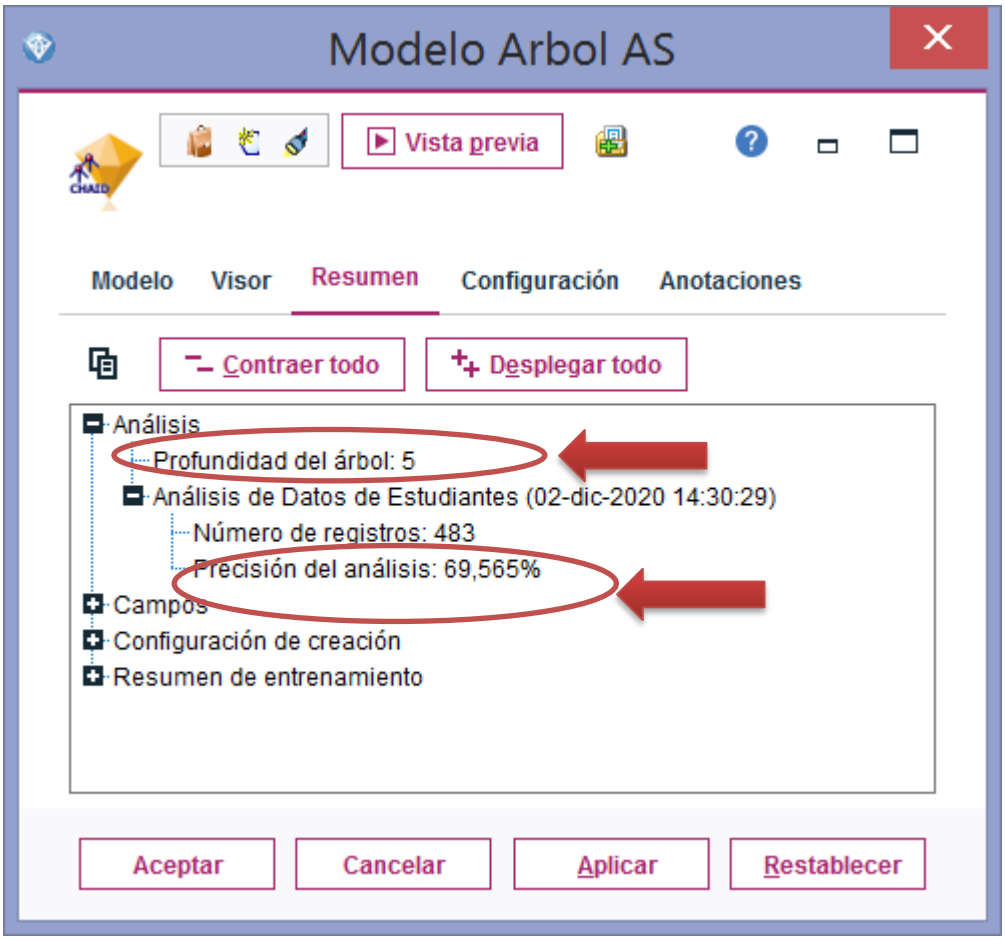
Resultados de la evaluación del modelo árbol C5.0:



En esta etapa también mostramos las Reglas de Clasificación usando la técnica o algoritmo C5.0



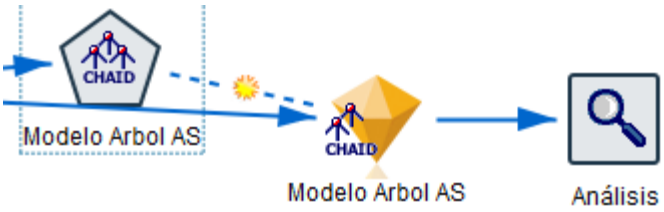
c. Modelo Árbol AS (CHAID)



La profundidad del árbol es de 5 y su precisión de análisis es de 69.565%

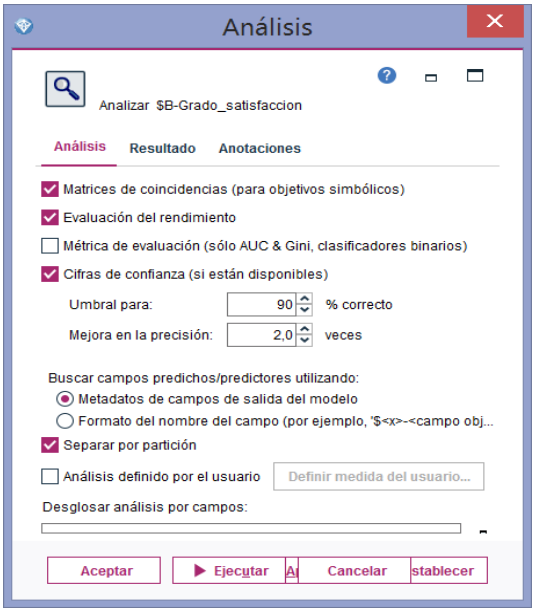
También se agregó un Nodo de Análisis donde se dan Resultados para el campo de salida.

Evaluación del modelo del árbol AS



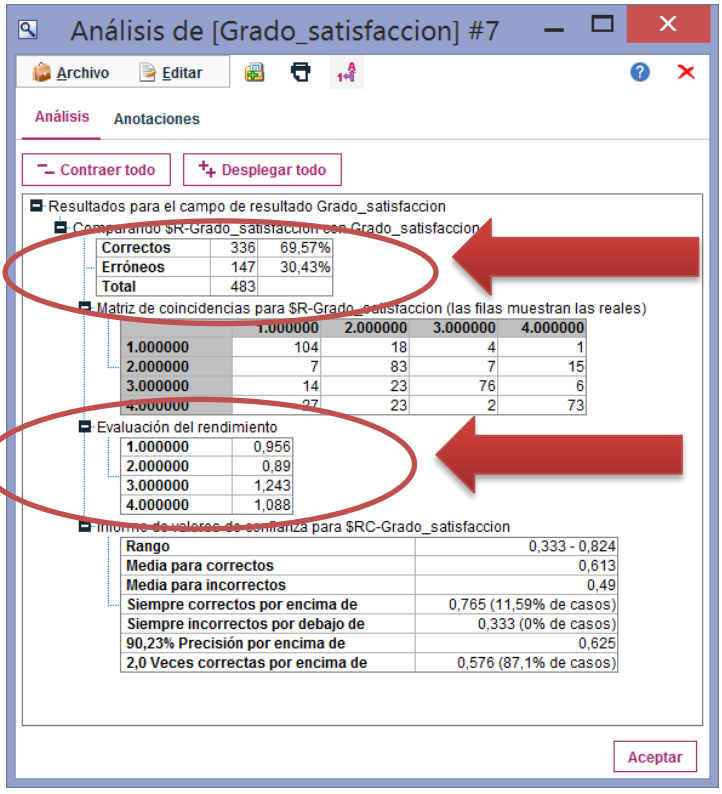
Se configura este análisis para el árbol AS para que muestre: Matrices de coincidencias, evaluación de rendimiento, cifras de confianza.

Configuración del análisis de la evaluación

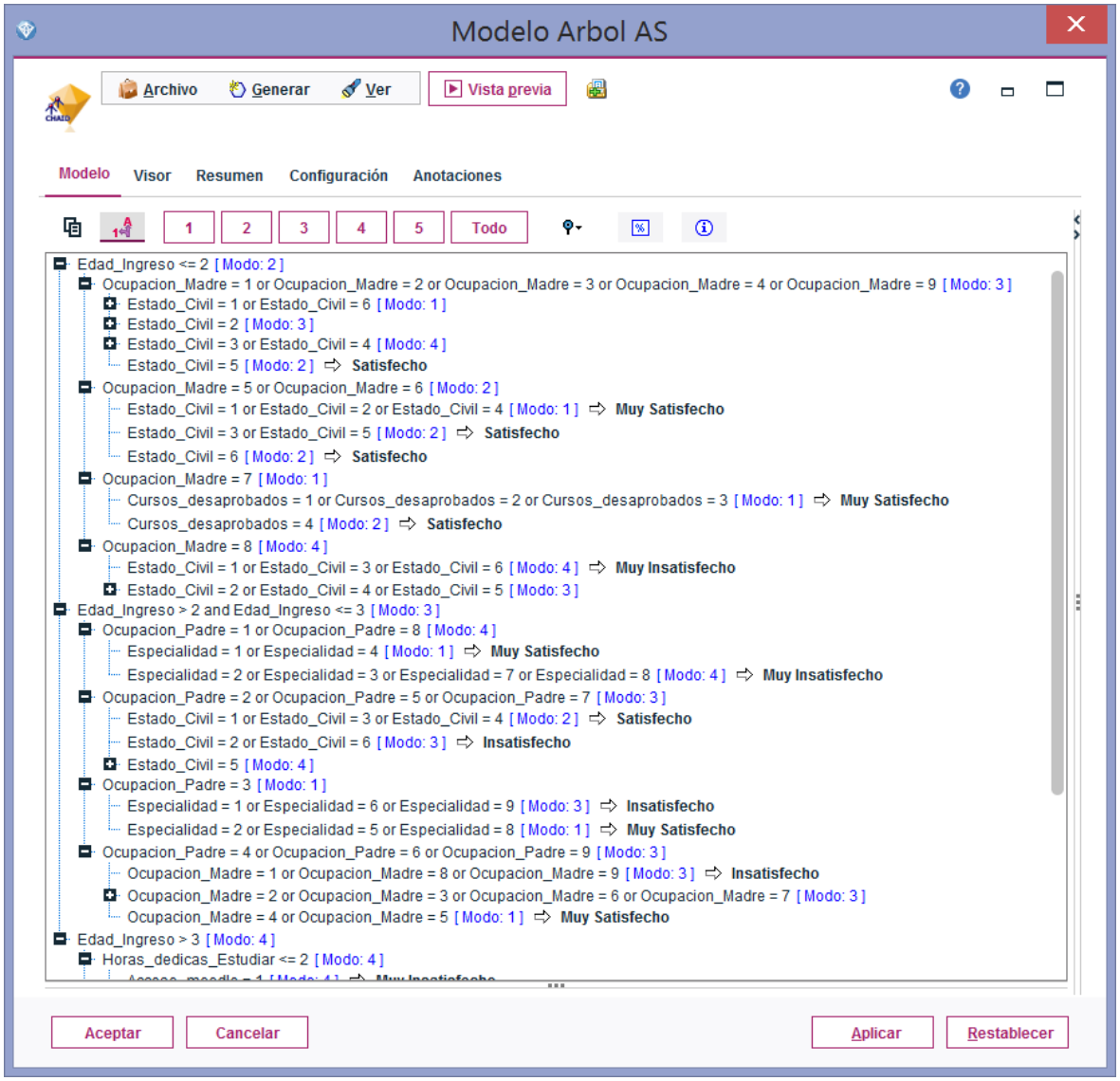


Obteniendo el siguiente resultado:

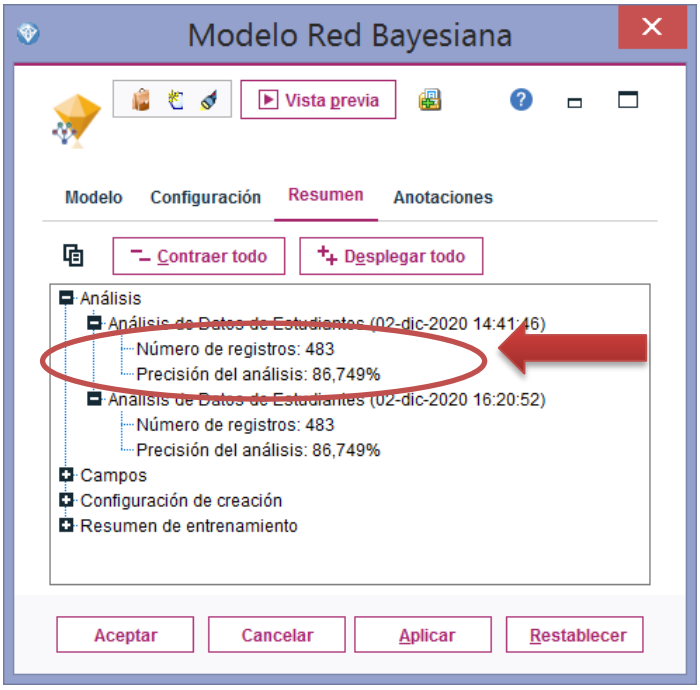
Resultados de la evaluación del modelo árbol AS



En esta etapa también mostramos las Reglas de Clasificación usando la técnica o algoritmo árbol AS



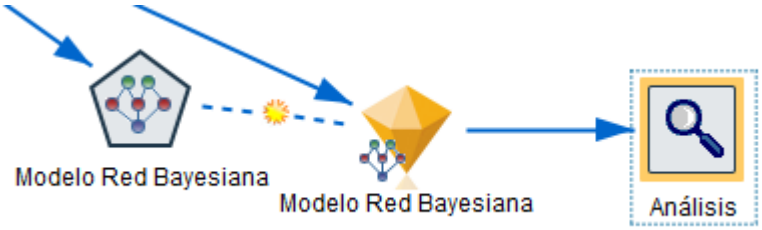
d. Modelo Red Bayesiana



Este Modelo en su Nugget nos devuelve una precisión de análisis de 86.749%

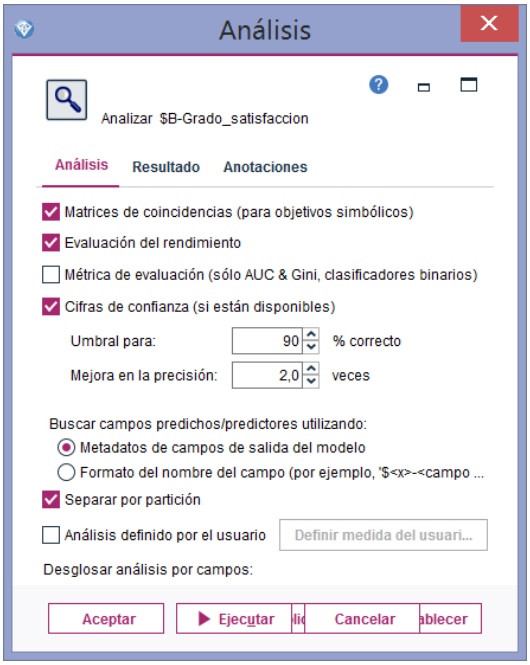
También se agregó un Nodo de Análisis donde se dan Resultados para el campo de salida.

Evaluación del modelo Red Bayesiana



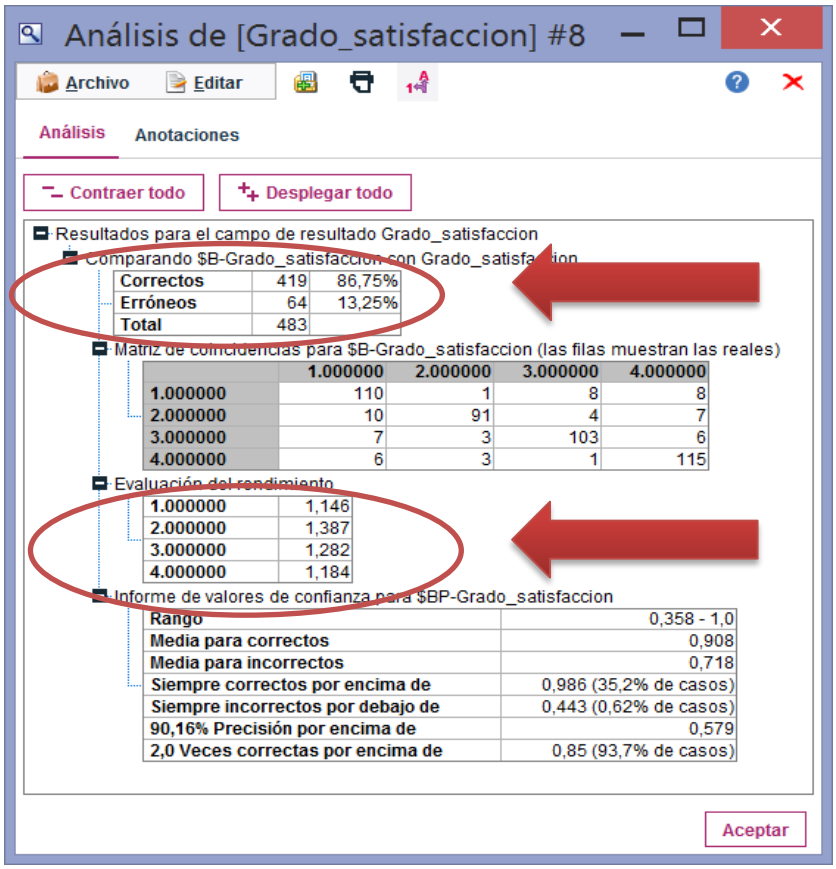
Se configura este análisis para el árbol red bayesiana para que muestre: Matrices de coincidencias, evaluación de rendimiento, cifras de confianza.

Configuración del análisis de la evaluación



Obteniendo el siguiente resultado:

Resultados de la evaluación del modelo Red Bayesiana



4.5. EXPLOTACION

En esta fase de explotación mostramos todos los modelos desplegados en la siguiente figura:

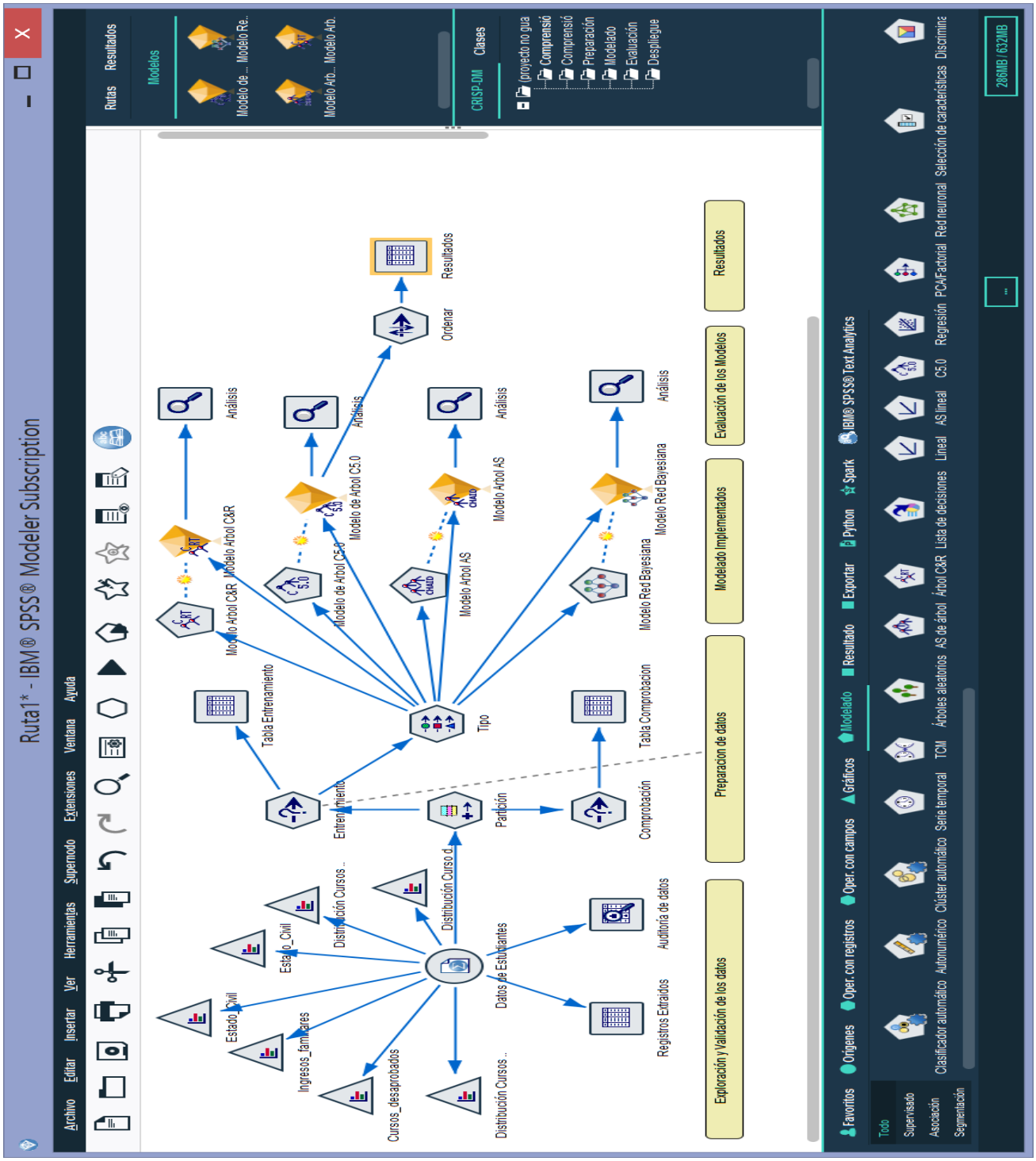


Figura 13: Modelos Implementados en IBM SPSS Modeler

A continuación se muestra los resultados obtenidos al aplicar el modelo del Árbol de decisión C5.0:

Evaluación de resultados modelo del Árbol de decisión C5.0

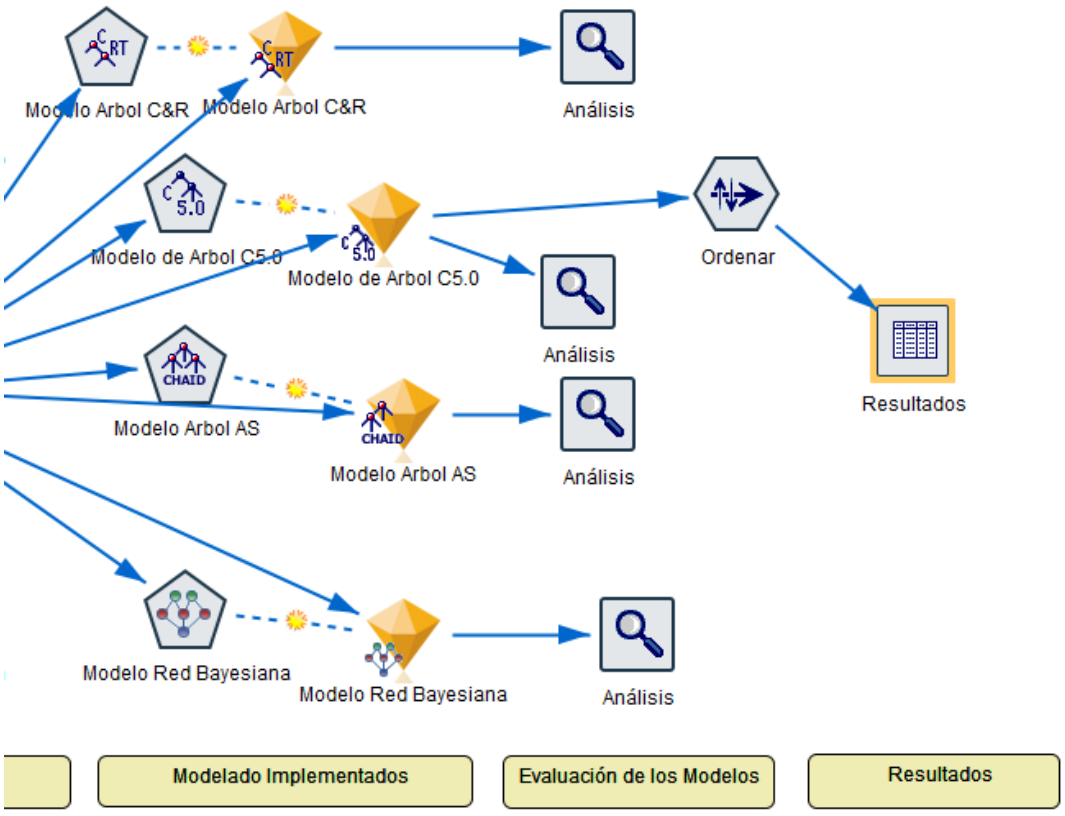


Tabla de resultados del modelo del Árbol de decisión C5.0

Resultados (23 campos, 483 registros) #8

Archivo | Editar | Generar

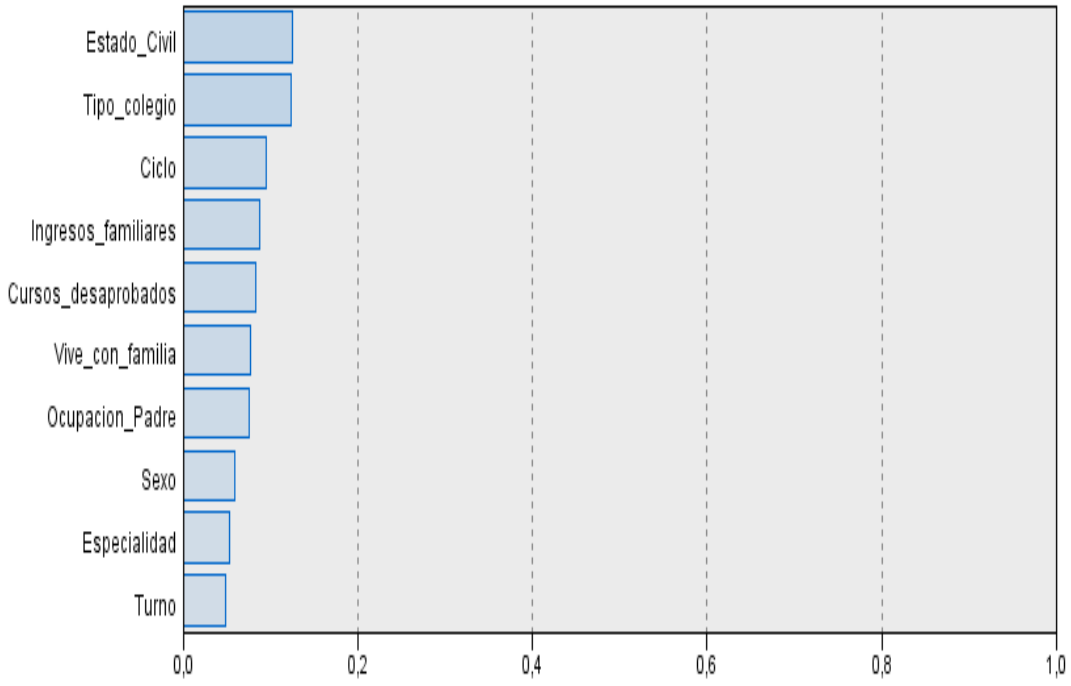
Tabla | Anotaciones

lia	Tiene_Hermanos	Ingresos_familiares	Tipo_colegio	Edad_Ingreso	Promedio_curso_anterior	Cursos_desaprobados	N°Inasistencias	Ciclo	Acceso_moodle	Especialidad	Turno	Horas_...	Partición	Grado_satisfaccion	$CC-Grado_sa
449	NO	BAJO	PARTICULAR MENOR	MEDIO	MUY ALTO	MUY ALTO	BAJO	4.000	MEDIO	Repostería	MAÑANA	4.000	Entrenamiento	Muy Insatisfecho	0.667
450	NO	BAJO	PARTICULAR MENOR	MUY BAJO	ALTO	NINGUNO	4.000	ALTO	Patronaje	TARDE	2.000	...	Entrenamiento	Muy Insatisfecho	0.700
451	SI	BAJO	PARTICULAR MENOR	MEDIO	BAJO	ALTO	2.000	MEDIO	Cosmetología	MAÑANA	2.000	...	Entrenamiento	Muy Insatisfecho	0.600
452	NO	MUY BAJO	PARTICULAR MAYOR	ALTO	BAJO	NINGUNO	1.000	ALTO	Cosmetología	NOCHE	3.000	...	Entrenamiento	Muy Insatisfecho	0.750
453	SI	BAJO	NACIONAL MAYOR	MUY BAJO	MUY ALTO	BAJO	4.000	ALTO	Maestro Gastronó...	NOCHE	3.000	...	Entrenamiento	Muy Insatisfecho	0.500
454	SI	BAJO	PARTICULAR ADULTO	BAJO	NINGUNO	NINGUNO	1.000	MEDIO	Cosmetología	NOCHE	4.000	...	Entrenamiento	Muy Insatisfecho	0.429
455	SI	BAJO	NACIONAL JOVEN	MEDIO	ALTO	ALTO	1.000	BAJO	Artesanía	MAÑANA	4.000	...	Entrenamiento	Muy Insatisfecho	0.625
456	NO	MUY ALTO	NACIONAL MENOR	MUY BAJO	MUY ALTO	BAJO	4.000	MEDIO	Repostería	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.667
457	SI	MEDIO	NACIONAL JOVEN	MEDIO	ALTO	MUY ALTO	3.000	BAJO	Electricidad	MAÑANA	2.000	...	Entrenamiento	Muy Insatisfecho	0.700
458	NO	BAJO	NACIONAL MAYOR	ALTO	ALTO	NINGUNO	4.000	BAJO	Computación e Info...	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.500
459	NO	ALTO	NACIONAL MENOR	MUY BAJO	NINGUNO	ALTO	2.000	MEDIO	Artesanía	TARDE	4.000	...	Entrenamiento	Muy Insatisfecho	0.500
460	NO	MEDIO	NACIONAL MENOR	MUY BAJO	ALTO	NINGUNO	4.000	BAJO	Maestro Gastronó...	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.750
461	NO	MUY BAJO	PARTICULAR JOVEN	ALTO	ALTO	NINGUNO	3.000	MEDIO	Repostería	NOCHE	3.000	...	Entrenamiento	Muy Insatisfecho	0.800
462	SI	MEDIO	PARTICULAR JOVEN	MUY BAJO	MUY ALTO	MUY ALTO	4.000	BAJO	Computación e Info...	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.571
463	SI	MUY BAJO	PARTICULAR MAYOR	MEDIO	BAJO	ALTO	4.000	BAJO	Textil y confecciones	MAÑANA	3.000	...	Entrenamiento	Muy Insatisfecho	0.667
464	SI	ALTO	PARTICULAR MAYOR	ALTO	MUY ALTO	NINGUNO	4.000	MEDIO	Hotelería y Turismo	MAÑANA	4.000	I...	Entrenamiento	Muy Insatisfecho	0.429
465	SI	MUY ALTO	PARTICULAR ADULTO	BAJO	ALTO	BAJO	3.000	ALTO	Hotelería y Turismo	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.500
466	SI	MEDIO	PARTICULAR MAYOR	BAJO	MUY ALTO	NINGUNO	4.000	ALTO	Computación e Info...	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.429
467	NO	ALTO	NACIONAL MAYOR	ALTO	NINGUNO	NINGUNO	4.000	ALTO	Cosmetología	MAÑANA	2.000	...	Entrenamiento	Muy Insatisfecho	0.750
468	SI	MUY ALTO	PARTICULAR ADULTO	BAJO	NINGUNO	BAJO	3.000	BAJO	Artesanía	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.500
469	NO	ALTO	PARTICULAR MAYOR	BAJO	BAJO	MUY ALTO	2.000	ALTO	Patronaje	MAÑANA	3.000	...	Entrenamiento	Muy Insatisfecho	0.500
470	SI	MEDIO	PARTICULAR MENOR	MEDIO	ALTO	ALTO	2.000	MEDIO	Textil y confecciones	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.500
471	SI	BAJO	NACIONAL JOVEN	MUY BAJO	BAJO	MUY ALTO	2.000	MEDIO	Maestro Gastronó...	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.625
472	NO	MUY BAJO	PARTICULAR MENOR	MUY BAJO	ALTO	MUY ALTO	2.000	MEDIO	Hotelería y Turismo	MAÑANA	4.000	I...	Entrenamiento	Muy Insatisfecho	0.500
473	NO	MEDIO	NACIONAL JOVEN	BAJO	MUY ALTO	NINGUNO	1.000	ALTO	Cosmetología	MAÑANA	4.000	...	Entrenamiento	Muy Insatisfecho	0.750
474	SI	BAJO	NACIONAL ADULTO	BAJO	BAJO	NINGUNO	2.000	BAJO	Cosmetología	TARDE	3.000	...	Entrenamiento	Muy Insatisfecho	0.750
475	SI	MUY ALTO	PARTICULAR MAYOR	ALTO	MUY ALTO	NINGUNO	4.000	MEDIO	Artesanía	MAÑANA	4.000	...	Entrenamiento	Muy Insatisfecho	0.625
476	NO	MEDIO	PARTICULAR MENOR	MUY BAJO	ALTO	BAJO	2.000	MEDIO	Artesanía	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.500
477	NO	MUY ALTO	NACIONAL JOVEN	MEDIO	MUY ALTO	MUY ALTO	1.000	ALTO	Hotelería y Turismo	TARDE	2.000	...	Entrenamiento	Muy Insatisfecho	0.500
478	NO	MEDIO	PARTICULAR JOVEN	MEDIO	NINGUNO	BAJO	4.000	MEDIO	Cosmetología	MAÑANA	4.000	...	Entrenamiento	Muy Insatisfecho	0.571
479	NO	ALTO	NACIONAL MENOR	MUY BAJO	MUY ALTO	NINGUNO	3.000	ALTO	Electricidad	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.750
480	NO	BAJO	NACIONAL MAYOR	ALTO	ALTO	NINGUNO	4.000	BAJO	Computación e Info...	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.500
481	NO	MUY ALTO	NACIONAL JOVEN	MUY BAJO	MUY ALTO	NINGUNO	2.000	MEDIO	Cosmetología	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.571
482	SI	MUY BAJO	PARTICULAR ADULTO	BAJO	NINGUNO	BAJO	3.000	ALTO	Textil y confecciones	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.500
483	SI	MUY BAJO	PARTICULAR MAYOR	ALTO	ALTO	MUY ALTO	4.000	ALTO	Electricidad	NOCHE	2.000	...	Entrenamiento	Muy Insatisfecho	0.700

Aceptar

4.6. PATRONES ENCONTRADOS:

De acuerdo a los resultados obtenidos por el Modelo de Minería de datos del Árbol C5.0 se listan a continuación la lista de patrones que influyen en la deserción académica en función a la prioridad de sus predictores:



5. DISCUSIÓN DE LA HIPOTESIS

Para la discusión de la hipótesis se ha considerado lo siguiente puntos:

5.1. Formulación del Problema

¿Cómo identificar de patrones que influye en la deserción académica en el Instituto Superior Leonardo Davinci?

5.2. Hipótesis

"El desarrollo de un Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler permite conocer los patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci"

Luego se definen las variables que intervienen en la veracidad o falsedad de la hipótesis:

- ✓ Independiente (VI): Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler.
- ✓ Dependiente (VD): patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci.

Tabla 04: Operacionalización de las variables

Variable	Dimensión	Indicador	Unidad de medida	Instrumento de Investigación
VI	Precisión	Porcentaje de precisión del modelo	% Precisión	Hoja de captura de datos
VD	Satisfacción	Grado de Satisfacción	Rango de satisfacción	Tabla de satisfacción

5.3. Población y muestra.

5.3.1. Población

Estudiantes del Instituto Leonardo Davinci.

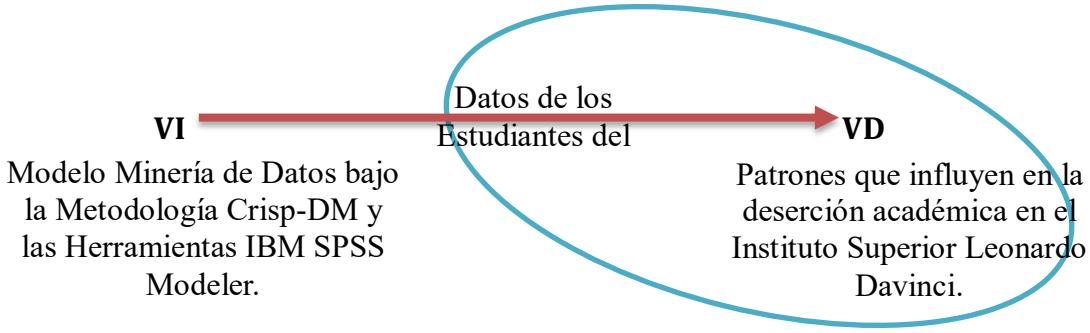
5.3.2. Muestra

Registros de estudiantes de los años 2018 -2019 almacenados en la base de datos del Instituto Leonardo Davinci.

5.3.3. Unidad de análisis

Datos de los Estudiantes del Instituto Leonardo Davinci

5.4. MANERA PRESENCIAL



5.5. DISEÑO PREEXPERIMENTAL PRE-PRUEBA Y POST-PRUEBA

- ✓ DISEÑO PREEXPERIMENTAL PRE-PRUEBA (O1): Es la medición previa de X(Tratamiento o estímulo) a G (Grupo de sujetos)
- ✓ DISEÑO PREEXPERIMENTAL POST-PRUEBA (O2): Corresponde a la nueva medición de X a G

Para esta parte se usó el Diseño PreExperimental Pre-Prueba y Post-Prueba, dado que nuestra hipótesis se adecua perfectamente a este diseño.

Este diseño contiene un solo grupo de sujetos, el cual es medido a través de un cuestionario antes y después de ser expuestos al estímulo (MD). Este diseño se presenta de la siguiente manera:

G **O_1** **X** **O_2**

Dónde: X: Tratamiento, estímulo (MD)

O: Medición

G: Grupo de sujetos (Empleados)

5.5.1. CÁLCULO DE LOS INDICADORES DE LA HIPÓTESIS

EVALUACION DE MODELOS

Modelo / Criterios de Evaluación	Árbol C&R	Árbol C5.0	Árbol AS (CHAID)	Red Bayesiana
Precisión de Análisis	58.385 %	**94.203 %**	69.565 %	86.749 %
Profundidad de Árbol	5	6	5	---
Comparando $_GradoSatisfaccion				
✓ % Datos Correctos	58.39 %	**94.2 %**	69.57 %	86.75 %
✓ % Datos Erróneos	**41.61 %**	**5.8 %**	**30.43 %**	13.25 %
Evaluación de Rendimiento				
✓ Muy Alto	0.858	1.272	0.956	1.146
✓ Alto	1.199	1.422	0.890	1.387
✓ Bajo	0.890	1.326	1.243	1.282
✓ Ninguno	0.634	1.294	1.088	1.184

Tabla 05: Tabla resumen de evaluación de modelos

Conclusión de la evaluación: De acuerdo a los resultados de la evaluación de cada módulo podemos concluir que el Modelo basado en el Nodo Árbol C5.0 es el más apropiado para clasificar los datos de los estudiantes y conocer los Patrones que influyen en la deserción académica de los estudiantes.

5.5.2. ANÁLISIS ESTADÍSTICO

Paso 1: Planteamiento de hipótesis.

Ho : $O_1 >= O_2$

H_1 : $O_2 >= O_1$

Dónde:

Ho es la hipótesis Nula: "El desarrollo de un Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler no permite conocer los patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci"

H_1 es la hipótesis Alternativa: "El desarrollo de un Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler permite conocer los patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci"

Paso 2: Nivel de significancia.

Para todo valor de probabilidad igual o menor que 0.05, se acepta H_1 y se rechaza Ho. $\alpha = 0,05$.

Paso 3: Prueba estadística.

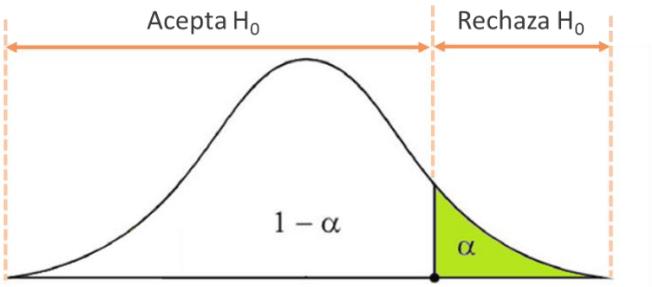
Debido a que la muestra es n= 2 (tomadores de decisión en la institución), y por ende menor a 30, se aplicó la prueba estadística t-student, en esta prueba estadística se exige dependencia entre ambas, en las que hay dos momentos uno antes y otro después. Con ello se da a entender que, en el primer período, las observaciones servirán de control o testigo, para conocer los cambios que se susciten después de aplicar una variable experimental.

Paso 4: Zona de rechazo.

Para todo valor de probabilidad mayor que 0.05, se acepta Ho y se rechaza H1.

Si la $t_c > t_t$ se rechaza Ho y se acepta H1.

Dónde: t_c es la t calculada y t_t es la t de tabla



Paso 5: Calculo de t_t y t_c

me

$$\overline{D} = \frac{\sum D}{n}, \delta = \sqrt{\frac{\sum(Di - \overline{D})^2}{n-1}}, t_c = \frac{\overline{D}}{\frac{\delta}{\sqrt{n}}}$$

Donde:

t_c	:	T calculado.
δ	:	Desviación estándar
n	:	Tamaño de la muestra
$\overline{D}$	:	Valor promedio o media aritmética de las diferencias entre los momentos antes y después.

Para el cálculo del valor de t calculado

Para el cálculo del valor T calculado se realizó un cuestionario donde se evaluó el grado de satisfacción a los usuarios del modelo de minería después de haber interactuado con él.

RANGO	GRADO DE SATISFACION
0 – 2.5	Insatisfecho
2.6 – 5.0	Medianamente Satisfecho
5.1 – 7.5	Satisfecho
7.6 – 10.0	Muy Satisfecho

Las respuestas obtenidas por los usuarios después de aplicar el cuestionario fueron evaluadas de acuerdo al rango de satisfacción que muestran en la tabla anterior.

Las respuestas obtenidas por los usuarios fueron:

N°	INDICADORES	Media Pre U_1	Media Post U_2	$D= (U_2- U_1)$	$(Di - \overline{D})$	$(Di - \overline{D})^2$
1	Se puede conocer los patrones que influyen en la deserción académica de los estudiantes.	4.0	9.0	5	-0.6	0.36
2	Se desea saber la cantidad de patrones por Nivel de Grado de satisfacción.	3.0	9.0	6	0.4	0.16
3	Se puede conocer los patrones con mayor incidencia por nivel socioeconómico.	3.0	9.0	6	0.4	0.16
4	Se puede conocer los patrones con mayor incidencia por estado civil.	3.0	8.0	5	-0.6	0.36
5	Se puede conocer los patrones con mayor incidencia por rendimiento académico.	3.0	9.0	6	0.4	0.16

N = 5 ; $\sum$D= 28 ; $\overline{D}$ = 5.6 ; $\sum(D_i - \overline{D})^2$ = 1.2 ; δ =0.5477 ; $\sqrt{n}$ = 2.24

$$t_c = \frac{\overline{D}}{\frac{\delta}{\sqrt{n}}}$$

$t_c = 4.89$

Interpretación: Como $t_c > t_t$, (4.89 > 2.92) se acepta la hipótesis alternativa, entendiéndose que el "El desarrollo de un Modelo Minería de Datos bajo la Metodología Crisp-DM y las Herramientas IBM SPSS Modeler **permite** conocer los patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci".

CONCLUSIONES

- ✓ Se identificaron los problemas y necesidades del área basándose en sus objetivos de negocio encontrando 2 problemas principales: Mejorar el proceso de toma de Decisiones del área académica y Conocer los patrones que repercuten en la deserción académica de los estudiantes del Instituto Leonardo Davinci orientándonos a una solución usando un modelo de minería de datos.

- ✓ En el análisis y preparación de datos su procesaron un total de 599 registros de los estudiantes recolectados en un archivo de datos de IBM SPSS Statistics llamado Estudiantes.sav, de donde se determinó usar un 80% del total de datos para el Entrenamiento que sirvió como entrada de datos al modelo de minería propuesto.

- ✓ Se diseñó, construyó y aplicó 04 Modelos de Minería de datos como son : Árbol C&R, Árbol C5.0, Árbol AS y Red Bayesiana utilizando IBM SPSS Modeler Subscription, analizando el modelo más apropiado para dar solución al problema y que nos permita obtener una mayor precisión en el análisis de datos, obteniendo que el modelo de Árbol C5.0 nos proporciona un mejor análisis de datos con una mayor precisión.

- ✓ Después de evaluar los 04 modelos implementados, obtuvimos que el modelo de árbol C5.0 nos da un 94.2% de datos correctos y con una precisión similar de 94.203% (ver pág. 91), siendo el modelo a utilizar para analizar los patrones que repercuten en la deserción académica de los estudiantes (Ver pág. 88)

RECOMENDACIONES

- ✓ Se recomienda actualizar los datos cada semestre y conocer el comportamiento del modelo y confirmar el porcentaje de precisión que se obtuvo.

- ✓ Recomendamos siempre realizar una auditoría de datos para despejar los datos nulos o anómalos, con esto se asegura que los datos están limpios para iniciar las pruebas dentro del modelo evaluado y así poder obtener buenos resultados.

- ✓ Se recomienda usar la herramienta IBM SPSS Modeler para implementar una solución de minería de datos debido a su facilidad de uso e interfaces entendibles (Ver Benchmarking Pág. 17).

- ✓ Se recomienda el uso de encuestas o formularios para obtener mayor información e ingresarla al modelo para obtener más indicadores o patrones que pueden estar afectando e incidiendo en la deserción estudiantil.

REFERENCIAS BIBLIOGRAFICAS

ConexionEsan. (03 de 11 de 2016). *Apuntes Empresariales*. Obtenido de http://www.esan.edu.pe/apuntes-empresariales/2016/11/el-proceso-de-la-toma-de-decisiones-en-la-organizacion/

Dataprix. (25 de febrero de 2010). *Inteligencia de Negocios*. Obtenido de http://www.dataprix.com/blogs/respinosamilla/qu-business-intelligence

Ecured. (06 de 06 de 2018). *Toma de decisiones*. Obtenido de https://www.ecured.cu/Toma_de_decisiones

Finlay, P. (1994). *Introducing decision support systems.* Oxford UK: Blackwell Publishers.

IBM. (10 de 02 de 2017). *spss-modeler*. Obtenido de http://www-03.ibm.com/software/products/es/spss-modeler

Itson. (01 de 05 de 2018). *Introducción a los sistemas de información*. Obtenido de https://biblioteca.itson.mx/oa/dip_ago/introduccion_sistemas/p3.htm

Kendall, K., & Kendall, J. (2005). *Análisis y diseño de sistemas.* Mexico: Pearson.

Microsoft. (08 de 08 de 2016). *www.microsoft.com*. Obtenido de Microsoft Azure: http://www.migesamicrosoft.com/que-se-puede-hacer-con-microsoft-azure/

Pérez López, C. (2007). *Minería de datos: técnicas y herramientas.* Editorial Paraninfo.

Sinexus. (15 de 08 de 2016). *Sistema de Soporte de Decisiones*. Obtenido de www.sinnexus.com: http://www.sinnexus.com/business_intelligence/sistemas_soporte_decisiones.aspx

Sinexus. (10 de 02 de 2017). *Datamining*. Obtenido de http://www.sinnexus.com/business_intelligence/datamining.aspx

Sinnexus. (20 de octubre de 2016). *¿Qué es Business Intelligence?* Obtenido de http://www.sinnexus.com/business_intelligence/index.aspx

Smartbase Group. (13 de 06 de 2016). *Metodología CRISP-DM*. Obtenido de http://smartbasegroup.com

Transformacion Digital. (19 de 08 de 2017). *Los 6 principales tipos de sistemas de información*. Obtenido de https://smarterworkspaces.kyocera.es/blog/los-6-principales-tipos-sistemas-informacion/

Turban, E. (2005). *Decision support and expert systems: management support systems.* Englewood Cliffs: Prentice Hall.

Universidad de Cadiz. (08 de 05 de 2017). *www.csintranet.org*. Obtenido de http://www.csintranet.org/competenciaslaborales/index.php?option=com_content&view=article&id=163:toma-de-decisiones&catid=55:com

Webyempresas. (01 de 07 de 2018). *Toma De Decisiones: Concepto De Vital Importancia En La Empresa*. Obtenido de https://www.webyempresas.com/toma-de-decisiones/

ANEXOS

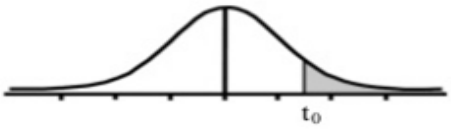
ANEXO A

CUESTIONARIO DIRIGIDO DL DIRECTOR ACADEMICO Y ASISTENTE

PREGUNTAS	Puntuación del 1 al 10										
		1	2	3	4	5	6	7	8	9	10
Se puede conocer los patrones que influyen en la deserción académica de los estudiantes.											
Se desea saber la cantidad de patrones por Nivel de Grado de satisfacción.											
Se puede conocer los patrones con mayor incidencia por nivel socioeconómico.											
Se puede conocer los patrones con mayor incidencia por estado civil.											
Se puede conocer los patrones con mayor incidencia por rendimiento académico.											

ANEXO B

Tabla t-Student



Grados de libertad	0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7970
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3007	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800