

UNIVERSIDAD PRIVADA ANTENOR ORREGO
FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN
Y DE SISTEMAS



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

**“ANALÍTICA PREDICTIVA PARA CONOCER EL PATRÓN DE
CONSUMO DE LOS CLIENTES EN LA EMPRESA CIENPHARMA
S.A.C. UTILIZANDO IBM SPSS MODELER Y LA METODOLOGÍA
CRISP-DM”**

Línea de Investigación:

Gestión de Datos y de Información.

Autor:

Br. ABEL ISAIAS CONTRERAS ARTEAGA

Br. FRANK WILLIAM SANCHEZ COTRINA

Asesor:

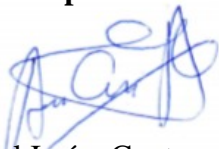
Ing. AGUSTIN EDUARDO ULLON RAMIREZ

2019

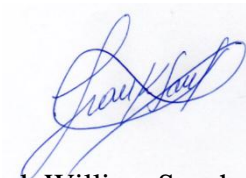
Fecha de Sustentación: 13/12/19

**“ANALÍTICA PREDICTIVA PARA CONOCER EL PATRÓN DE
CONSUMO DE LOS CLIENTES EN LA EMPRESA CIENPHARMA
S.A.C. UTILIZANDO IBM SPSS MODELER Y LA METODOLOGÍA
CRISP-DM”**

Elaborado por:



Br. Abel Isaías Contreras Arteaga

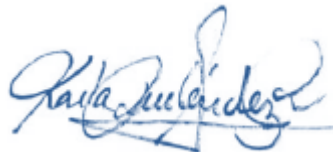


Br. Frank William Sanchez Cotrina

Aprobada por:



Ing. Edward Fernando Castillo Robles
Presidente
CIP: 192352



Ing. Karla Vanessa Meléndez Revilla
Secretario
CIP: 120097



Ing. Heber Gerson Abanto Cabrera
Vocal
CIP: 106421



Ing. Agustín Eduardo Ullón Ramírez
Asesor
CIP: 137602

PRESENTACIÓN

Señores Miembros del Jurado:

De acuerdo a los requisitos estipulados en el reglamento de grados y Títulos de la Universidad Privada Antenor Orrego y el Reglamento Interno de la Escuela Profesional de Ingeniería de Computación y Sistemas, ponemos a vuestra disposición el presente Trabajo de Tesis: **“ANALÍTICA PREDICTIVA PARA CONOCER EL PATRÓN DE CONSUMO DE LOS CLIENTES EN LA EMPRESA CIENPHARMA S.A.C. UTILIZANDO IBM SPSS MODELER Y LA METODOLOGÍA CRISP-DM”** para obtener el Título Profesional de Ingeniero de Computación y Sistemas.

El presente trabajo de tesis ha sido desarrollado conforme al marco de referencia los lineamientos establecidos por la Facultad de Ingeniería, la Escuela Profesional de Ingeniería de Computación y Sistemas y basado en los conocimientos adquiridos durante nuestra formación profesional, además de consulta de fuentes bibliográficas.

Los autores.

DEDICATORIA

A Dios y A mis padres quienes con su amor, paciencia y esfuerzo me han permitido llegar a cumplir hoy un sueño más, gracias por inculcar en mí el ejemplo de esfuerzo y valentía.

Br. Abel Isaías Contreras Arteaga

El presente trabajo está dedicado a mi familia por haber sido mi apoyo a lo largo de toda mi carrera universitaria y a lo largo de mi vida. A todas las personas especiales que me acompañaron en esta etapa, aportando a mi formación tanto profesional y como ser humano.

Br. Frank William Sanchez Cotrina

AGRADECIMIENTO

Un agradecimiento especial al personal que labora en la empresa CIENPHARMA quienes nos brindaron todas las facilidades para conocer más sobre la problemática, dándonos acceso a la información necesaria para el desarrollo de la presente tesis.

Agradecemos a nuestros docentes, por haber compartido sus conocimientos a lo largo de la preparación de nuestra profesión, de manera especial, al Ing. Agustín Ullón asesor de nuestro proyecto de investigación quien ha guiado con su paciencia, y conocimiento dando un aporte valioso a nuestro trabajo de tesis.

Y también a todos los familiares que estuvieron presente siempre con nosotros durante todo el camino apoyando en este trabajo de tesis.

Muchas Gracias.

Los autores.

RESUMEN

“ANALÍTICA PREDICTIVA PARA CONOCER EL PATRÓN DE CONSUMO DE LOS CLIENTES EN LA EMPRESA CIENPHARMA S.A.C. UTILIZANDO IBM SPSS MODELER Y LA METODOLOGÍA CRISP-DM”

Por:

Br. Abel Isaías Contreras Arteaga

Br. Frank William Sánchez Cotrina

En la actualidad explorar los datos contenidos en la base de datos transaccional de las empresas con la analítica predictiva de datos, nos facilita comprender reglas que determinan patrones de consumo y tendencias que siguen los datos, lo que a su vez genera un conocimiento; el cual nos permitirá asociar productos que tengan mayor rotación con aquellos que no la tienen, generando de esta forma venta cruzada que resulte en el incremento de ingresos.

La empresa Cienpharma S.A.C. está basando sus decisiones en datos aislados que no están debidamente procesados y que muchas veces no reflejan la realidad de lo que sucede con la información almacenada en su base de datos.

El presente proyecto de tesis se basa en obtener un modelo de analítica predictiva para determinar patrones de consumo de los clientes en la empresa Cienpharma S.A.C. utilizando técnicas de minería de datos. Identificando los requerimientos y necesidades del área mediante un análisis de modelo de negocio, realizando un análisis y preparación de datos de los clientes obtenidos desde el sistema transaccional de la empresa, construyendo un modelo de búsqueda de patrones de consumo de los clientes basado en las técnicas de modelado de minería de utilizando IBM SPSS Modeler y finalmente evaluando los resultados de los informes que muestran el modelo de la Minería de Datos.

Para el proceso de Minería de Datos se aplicaron los algoritmos de asociación, clústeres y redes neuronales haciendo búsqueda de patrones basado en las técnicas de modelado de minería de Datos utilizando IBM SPSS Modeler.

ABSTRACT

“PREDICTIVE ANALYTICS TO KNOW THE CUSTOMER CONSUMPTION PATTERN IN EMPRESA CIENPHARMA S.A.C. USING IBM SPSS MODELER AND THE CRISP-DM METHODOLOGY”

By:

Br. Abel Isaías Contreras Arteaga

Br. Frank William Sánchez Cotrina

Currently, exploring the data contained in the transactional database of companies with predictive data analytics, makes it easier for us to understand rules that determine consumption patterns and trends that follow the data, which in turn generates knowledge; which will allow us to associate products that have greater rotation with those that do not have it, thus generating cross-selling that results in increased revenues.

The company Cienpharma S.A.C. It is basing its decisions on isolated data that are not properly processed and that many times do not reflect the reality of what happens with the information stored in its database.

This thesis project is based on obtaining a predictive analytical model to determine customer consumption patterns in the company Cienpharma S.A.C. using data mining techniques. Identifying the requirements and needs of the area through a business model analysis, performing an analysis and preparation of customer data obtained from the transactional system of the company, building a model of customer consumption pattern search based on techniques of mining modeling using IBM SPSS Modeler and finally evaluating the results of the reports that show the Data Mining model.

For the Data Mining process the association algorithms, clusters and neural networks were applied by searching for patterns based on data mining modeling techniques using IBM SPSS Modeler.

ÍNDICE DE CONTENIDO

PRESENTACIÓN	ii
DEDICATORIA	iii
AGRADECIMIENTO	iv
RESUMEN	v
ABSTRACT	vi
ÍNDICE DE CONTENIDO	vii
INDICE DE IMÁGENES	ix
INDICE DE TABLAS	x
1. INTRODUCCION	01
1.1. Planteamiento del problema	01
1.2. Delimitación del problema	03
1.3. Formulación del problema.....	03
1.4. Formulación del hipótesis.....	03
1.5. Objetivos del estudio	03
1.6. Justificación del estudio.....	04
1.6.1. Importancia de la investigación.....	04
1.6.2. Viabilidad de la investigación.....	04
1.6.3. Aportes.....	05
2. MARCO TEÓRICO	06
2.1. ANTECEDENTES.....	06
2.2. DEFINICIONES.....	08
2.2.1. ANALÍTICA PREDICTIVA.....	08
2.2.2. MINERÍA DE DATOS	09
2.2.3 Algoritmos y Técnicas de Segmentación	11
2.2.3. Modelo de Minería de Datos	14
2.2.4. Toma de decisiones	15
2.2.5. Proceso de Toma de decisiones	16
2.3. METODOLOGIA PARA EL DESARROLLO DEL PROYECTO.....	21
3. MATERIALES Y METODOS	21
3.1. Material.....	21
3.2. Método.....	21

4. RESULTADOS: APLICACIÓN DE LA METODOLOGIA.....	24
4.1. ANALISIS DEL PROBLEMA.....	24
4.1.1. Objetivos Institucionales.....	24
4.1.2. Evaluación de la Situación.....	25
4.1.3. Recursos Computacionales.....	25
4.2. ANALISIS DE DATOS	26
4.2.1. Recolección de Datos Iniciales	26
4.2.2. Descripción de los Datos	27
4.2.3. Exploración y Validación de los datos	33
4.2.4. Selección y Limpieza de datos	34
4.2.5. Preparación y Construcción de datos	34
4.3. MODELADO.....	36
4.3.1. Selección de la técnica de modelado más apropiada	36
4.3.2. Generación del plan de prueba.....	38
4.3.3. Construcción del modelo	39
4.4. EVALUACION.....	64
4.5. EXPLOTACION.....	75
4.6. RESULTADOS DEL MODELO.....	77
5. DISCUSION DE RESULTADOS	80
6. CONCLUSIONES.....	89
7. RECOMENDACIONES.....	90
8. REFERENCIAS BIBLIOGRAFICAS.....	91
ANEXOS.....	91

INDICE DE FIGURAS

Figura 01: Infografía Data Mining.....	11
Figura 02: Metodología CRISP DM.....	19
Figura 03: Base de datos Cienpharma.....	27
Figura 04: Tabla Cliente	28
Figura 05: Tabla Factura	29
Figura 06: Tabla Articulo	30
Figura 07: Tabla Ubigeo	31
Figura 08: Número de clientes por Ubicación	32
Figura 09: Cantidad en monto de compras por cliente	33
Figura 10: Consulta SQL – Preparación de datos	34
Figura 11: Resultado de Consulta SQL – Preparación de datos	35
Figura 12: Ejecución del Modelo K-medias.....	43
Figura 13: Resultados del Modelo K-medias	44
Figura 14: Ejecución del Modelo Kohonen	51
Figura 15: Resultados del Modelo Kohonen	52
Figura 17: Ejecución del Modelo Árbol	61
Figura 18: Resultados del Modelo de árbol	62
Figura 19: Árbol de decisión: Modelo Árbol C5.0	63
Figura 20: Modelos de minería de datos construidos	64

INDICE DE TABLAS

Tabla 01: Diagrama de investigación.....	22
Tabla 02: Operacionalización de las variables	22
Tabla 03: Recursos Computacionales.....	26
Tabla 04: Tabla Cliente.....	29
Tabla 05: Tabla Factura.....	30
Tabla 06: Tabla Artículo.....	30
Tabla 07: Tabla Ubigeo.....	31
Tabla 08: Datos a utilizar.....	35
Tabla 09: Resumen de Evaluación de Modelos.....	74
Tabla 10: Rango de grado de valoración.....	82
Tabla 11: Evaluación de los indicadores de la hipótesis	83

1. INTRODUCCION

1.1. Planteamiento del problema

Actualmente las empresas buscan como guiar o dirigir su mercadería a clientes y para eso es necesario conocer sus gustos o preferencias y así ofrecerles productos o servicios adecuados. Para una empresa dedicada a las ventas conocer esta información de sus clientes le va permitir tomar mejores decisiones como por ejemplo promover productos a precios más bajo o a un mayor precio basado en el conocimiento del patrón de consumo de sus clientes. Para realizar esto se comienza recolectando y analizando información y termina actuando sobre la información de manera apropiada y efectiva logrando encontrar estrategias competitivas, que logren aumentar la innovación, productividad y a su vez generar cambios constantes en el producto, manejo de la publicidad, estrategias de promoción, definir las necesidades del cliente y satisfacerla al máximo

Existen herramientas especializadas como la analítica predictiva, que indaga como se puede incrementar la eficiencia, estimulen la innovación, a fin de definir argumentos para la toma de decisiones.

Explorar los datos contenidos en la base de datos transaccional de las empresas con la analítica predictiva de datos, nos facilita comprender reglas que determinan patrones de consumo y tendencias que siguen los datos, lo que a su vez genera un conocimiento; el cual nos permitirá asociar productos que tengan mayor rotación con aquellos que no la tienen, generando de esta forma venta cruzada que resulte en el incremento de ingresos.

La empresa Cienpharma SAC no quiere permanecer distante del uso de las TI, por la necesidad que tiene de obtener información confiable que permita a la gerencia tomar mejores decisiones.

La empresa Cienpharma SAC brinda servicios de Distribución y Logística en varios departamentos del territorio Peruano. La empresa ofrece a sus clientes un servicio

comercial, garantizando una rápida entrega de los productos farmacéuticos, en todos los segmentos del mercado a donde estén dirigidos. A sus clientes les ofrecen productos de calidad, a precios justos y respaldados por el Servicio Post-Venta de la compañía. Con la analítica predictiva sobre sus datos se puede ofrecer beneficios importantes para la empresa, y así aprovechar la información oculta en los registros de los clientes.

Cienpharma actualmente basa sus decisiones en datos aislados, muchos de ellos no están procesados adecuadamente y no reflejan la realidad de cómo esta almacenada la información en su base de datos.

La aplicación de la analítica predictiva en una empresa de comercialización de productos como es la empresa Cienpharma, les va a permitir descubrir patrones de comportamiento de clientes, que la empresa podrá utilizar para elaborar estrategias de marketing dirigidas hacia los distintos tipos de clientes que poseen.

La empresa desea conocer el patrón de consumo de sus clientes, además se busca generar paquetes de productos que puedan venderse de forma conjunta, aumentando de esta forma su accionar en sus ventas, es así que en este escenario en donde el análisis de la Base de Datos es preciso para lograr conocer mejor a sus clientes y sobre los resultados poder aplicar acciones que logre incrementar el nivel de ventas de la empresa.

- **Características problemáticas**

- ✓ No existe una preparación de datos adecuada para el análisis de la información del patrón de comportamiento de los clientes.
- ✓ El sistema transaccional de ventas con que cuenta la empresa recolectar, almacenar, modificar y recuperar todo tipo de información, pero no está brindando información que de soporte en la toma de decisiones sobre sus clientes.
- ✓ Falta de información para clasificar o diferenciar los clientes para campañas de marketing.
- ✓ Existen altos tiempos de ejecución de consultas para obtener información sobre las transacciones de ventas de los clientes.

- ✓ Costos excedidos en tomar malas decisiones por no contar con la información en el momento oportuno.

1.2. Delimitación del problema

El siguiente proyecto se realizará analizando la realidad en que se encuentran la información de los clientes en el proceso de ventas de la empresa Cienpharma S.A.C. y de cómo se lleva a cabo el proceso de toma de decisiones.

1.3. Formulación del problema

Dada la problemática de la investigación se formula lo siguiente:

¿Cómo conocer el patrón de consumo de clientes de grupos que siguen comportamientos similares en la empresa Cienpharma S.A.C.?

1.4. Formulación del hipótesis

El desarrollo de una solución de analítica predictiva permitirá conocer el patrón de consumo de clientes en la empresa Cienpharma S.A.C.

1.5. Objetivos del estudio

El Objetivo general es:

Realizar una analítica predictiva para determinar patrones de consumo de los clientes en la empresa Cienpharma S.A.C. utilizando técnicas de minería de datos.

Los objetivos específicos son los siguientes:

- ✓ Identificar necesidades del área y Realizar el análisis y preparación de datos de los clientes obtenidos desde el sistema transaccional de la empresa.
- ✓ Construir el modelo de búsqueda de patrones de consumo de los clientes basado en las técnicas de modelado de minería de utilizando IBM SPSS Modeler.
- ✓ Evaluar los resultados de los informes que muestran los modelos de la Minería de Datos utilizados.

1.6. Justificación del estudio

1.6.1. Importancia de la investigación

- La principal importancia será que el proyecto brindará información sobre el patrón de comportamiento de sus clientes basado en un Modelo de Minería de Datos creando una mejor relación con ellos y también con esta información se podrá elaborar estrategias de marketing.
- Con el patrón de comportamiento de sus clientes la empresa podrá seleccionar mejor los canales de distribución y comunicación con sus clientes.
- Se justifica por que la aplicación de minería de datos mejoraría su rentabilidad y procesos teniendo un panorama veraz de la información basada en patrones de comportamientos e información de sus clientes permitiendo aumentar la retención de clientes, potenciar las ventas cruzadas y posicionarnos dentro de mercados con necesidades específicas.
- La empresa podrá mejorar la obtención y análisis de información fiable, precisa, a tiempo para tener mejores decisiones en futuros proyectos de marketing.

1.6.2. Viabilidad de la investigación

- Es viable porque se cuenta con el acceso directo a la información de la base de datos transaccional de la empresa y una buena comunicación los responsables del área, siendo de gran ayuda para el desarrollo del modelo.
- Es factible porque se cuenta con las herramientas para el desarrollo de este proyecto y de los cuales se ha seleccionado IBM SPSS Modeler teniendo en cuenta su nivel de dificultad y el rápido manejo y aprendizaje por parte de nosotros los autores.

1.6.3. Aportes

El desarrollo de esta investigación generará considerables beneficios a la empresa, entre los cuales tenemos:

- Información correcta y oportuna para que en la empresa puedan tomar decisiones acertadas.
- Implementación de un modelo ajustado a la identificación de perfiles y patrones de comportamiento sus clientes.
- Con el conocimiento de sus patrones de comportamiento de sus clientes obtendremos una mayor habilidad en la obtención de información de los clientes que sirvan en mejorar la toma de decisiones en el área de ventas.
- La minería de datos no va a permitir descubrir patrones escondidos en la base de datos transaccional de la empresa.

El presente trabajo de tesis está organizado en diferentes puntos que facilitarán el uso y entendimiento del mismo dando a continuación una breve descripción del mismo:

Marco teórico: Fundamento teórico y metodología, en esta parte del trabajo brindamos la información necesaria sobre los temas y el modelo de referencia a utilizar para la solución del problema planteado. Dando los conocimientos básicos de que es una Minería de datos, Algoritmos y Técnicas, Modelos de minería de datos, Toma de decisiones, indicadores, y la metodología utilizada.

Resultados: Desarrollo del trabajo, en éste punto se muestra el desarrollo de los pasos enunciados en el Esquema de la Metodología. Mostramos los resultados obtenidos con relación a los objetivos planteados al inicio de este proyecto.

Discusión de resultados: En este capítulo se verifica si la Hipótesis planteada es aceptada a esto se le llama Contrastación de la Hipótesis.

Conclusiones y Recomendaciones, en éste último capítulo se encuentra las conclusiones que se llegó después de haber culminado todo el proyecto y las recomendaciones.

2. MARCO TEÓRICO

2.1. Antecedentes

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

Título de Investigación: Modelo predictivo de minería de datos de apoyo a la gestión hospitalaria sobre la morbilidad de pacientes hospitalizados. Repositorio Académico Univ. Privada Antenor Orrego, Trujillo 2016.

Descripción y análisis del trabajo:

En este trabajo de tesis propone la aplicación de un modelo de minería de datos como apoyo a la Gestión Hospitalaria sobre la morbilidad con pacientes hospitalizados basado en el algoritmo de análisis de serie de tiempo con información histórica de los pacientes del Hospital Víctor Ramos Guardia realizando la extracción de los datos, transformación de los datos y carga de datos que sirvieron como datos de entrada para el modelo. Luego procedieron a crear el modelo de pronósticos. Como resultado obtuvieron mayor información de los casos de morbilidad en pacientes hospitalizados para los próximos tres años.

- **Autor:** Johanna Denise Flores Coaguila

Título de la investigación: Propuesta de modelo de detección de fraudes de Energía eléctrica en clientes residenciales de Lima Metropolitana aplicando minería de datos.

Repositorio Académico USMP , Lima 2014

Descripción y análisis del trabajo:

En el presente trabajo, desarrollaron un modelo como propuesta para predecir potenciales situaciones de fraudes de energía eléctrica en clientes residenciales basado en aprender el comportamiento de clientes que anteriormente hurtaron para ello se aplicó el proceso de Minería de Datos basado para analizar, extraer y almacenar información de la base de datos, la cual contiene el historial de los consumos de energía. El modelo se propone para apoyar a las empresas de distribución eléctrica en especial a los técnicos eléctricos a examinar y verificar, acertadamente, de manera rápida y oportuna los resultados obtenidos y contribuir,

de esta forma, en la toma de decisiones. El modelo desarrollado permite evaluar y detectar los fraudes de energía eléctrica con un alto grado de sensibilidad 55.03% y especificidad es de 89% en clientes residenciales, siendo su valor de éxito el 86.99% mientras que el proceso tradicional según la memoria anual publicada por la empresa eléctrica Edelnor en el 2011 se realizaron 193,516 inspecciones se logra regularizar el 17419 consumos no registrados equivalente al 9% porcentaje de éxito.

- **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:

En las últimas décadas las compañías y organizaciones vienen presentando necesidades de mayor escala y mayor complejidad; esto se presenta debido a las innovaciones y/o mejoras constantes que deben aplicar a sus procesos de negocios en los diversos niveles de su estructura organizacional a efectos de liderar el mercado de su rubro, mantenerse competitivas y/o brindar un servicio de calidad y eficiente a sus clientes y/o usuarios. Estas organizaciones convencionales así como las modernas compañías offshore de hoy en día presentan una estructura funcional global; distribuidas físicamente en amplios espacios geográficos que pueden abarcar varios continentes inclusive. El presente trabajo realiza una revisión bibliográfica de las técnicas clustering k-means, elabora una propuesta concreta, desarrolla un prototipo de aplicación y concluye fundamentando los beneficios que obtendrían las organizaciones con su implementación.

- **Autores:** 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 UNT, Trujillo 2014

Descripción y análisis del trabajo:

El objetivo principal de este trabajo de tesis es Predecir los clientes rentables para las pymes de Trujillo mediante la aplicación de técnicas de minería de datos, analizando las técnicas de minería de datos para predicción de datos, teniendo una comprensión del negocio y los datos; construyendo el modelo de minería de datos para predecir clientes rentables. Luego evaluación del modelo de minería de datos y su aplicación del modelo en una determinada pyme de Trujillo.

- **Autores:** Camacho Gonzales, Anthony José

Título de la investigación: Clustering y árboles de decisión como técnicas de minería de procesos aplicadas al proceso de admisión de pregrado.

Repositorio Académico UPAO, Trujillo 2015

Descripción y análisis del trabajo:

El objetivo principal de este trabajo de tesis es aplicar las técnicas de minería de procesos de Clustering y árboles de decisión, que tiene como objetivo optimizar y administrar los procesos operacionales reales de un negocio; usando la metodología CRISP-DM y la herramienta especializada en minería de datos como SPSS Clementine, así como en minería de procesos se ha utilizado la herramienta PROM 6 (Open Source). Logrando de esta manera disminuir el tiempo de ejecución del proceso de admisión en un 16.48%.

2.2. DEFINICIONES

2.2.1. ANALÍTICA PREDICTIVA

La analítica predictiva proporciona herramientas para estimar aquellos datos de negocio que son desconocidos o inciertos, o que requieren de un proceso manual o costoso para su obtención. (Big Data Magazine, 2018)

Implica la aplicación de técnicas de análisis estadístico, consultas analíticas y algoritmos automáticos de aprendizaje automático a conjuntos de datos para crear modelos predictivos que sitúen un valor numérico o puntuación en la probabilidad de que ocurra un evento particular. (Big Data Magazine, 2018)

Más allá del puro análisis de la información histórica que realiza la analítica descriptiva, las predicciones de datos que realiza la analítica predictiva

fortalecen las decisiones de negocio y permiten por ejemplo:

- ✓ Anticipar demandas de clientes en distintos puntos de venta, teniendo en cuenta factores controlables, como el precio de venta, y factores externos, como calendarios laborales o efectos meteorológicos.
- ✓ Detectar si una transacción bancaria es susceptible de haber sido realizada como parte de un fraude.
- ✓ Descubrir grupos afines entre los clientes de un CRM (Customer Relationship Management), que compartan características demográficas o preferencias por determinados servicios o productos.

Se utilizan variables que pueden medirse y analizarse para predecir el comportamiento probable de individuos, maquinaria y otras con un modelo predictivo eficaz de evaluar probabilidades futuras con nivel de fiabilidad. El software se basa en gran medida en algoritmos avanzados y metodologías tales como regresiones logísticas, análisis de series de tiempo y árboles de decisión. (Big Data Magazine, 2018)

2.2.2. MINERÍA DE DATOS

La minería de datos es una forma innovadora de obtener información comercial valiosa mediante el análisis de los datos contenidos en la base de datos de la empresa. Esta información sirve de ayuda para una adecuada toma de decisiones empresariales. Esencialmente, la minería de datos es un método innovador de aprovechar la información ya existente en la empresa a fin de, por ejemplo, mejorar procesos, mejorar el rendimiento de la inversión u optimizar el uso de recursos. La minería de datos revela información comercial exhaustiva utilizando técnicas avanzadas de análisis y creación de modelos. Mediante la minería de datos, puede hacer consultas mucho más complejas de sus datos que utilizando métodos de consulta convencionales. La información que la minería proporciona puede mejorar notablemente la calidad y fiabilidad de la toma de decisiones empresariales. (IBM Knowledge Center, 2017)

La minería de datos 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. (Sinexus, 2017)

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. (Pérez López, 2007)

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. (Prieto, 2012)

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. (Braga & Paulo, 2009)

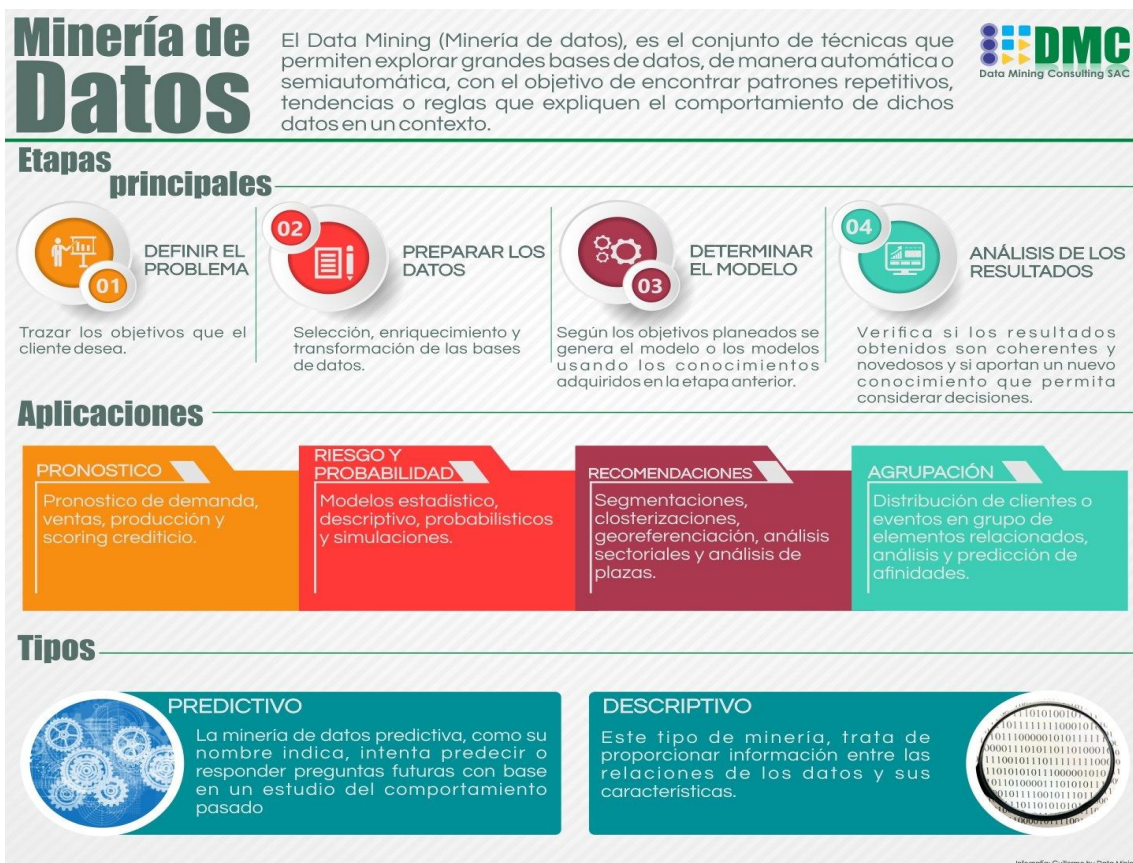


Figura 01: Infografía Data Mining

Fuente: (Data Mining Consulting, 2014)

2.2.3. Algoritmos y Técnicas de Segmentación

Las técnicas de la minería de datos provienen de la Inteligencia Artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados. Los algoritmos y técnicas más apropiadas para la Segmentación son: (IBM Knowledge Center, 2017)

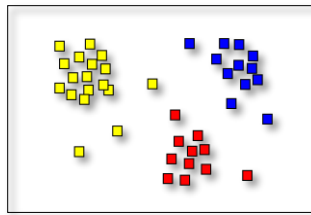
✓ Agrupamiento (clustering)

Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes (IBM Knowledge Center, 2017). Ejemplos:

- Algoritmo K-mean.
- Algoritmo K-medoids.

No hay que confundir clustering con segmentación. La segmentación se usa para identificar grupos que tienen características comunes. El clustering es un modo de segmentar datos en grupos que no están previamente definidos (IBM Knowledge Center, 2017).

A diferencia de la clasificación, no se sabe dónde habrá clusters o con que atributos de los datos se harán los clusters.

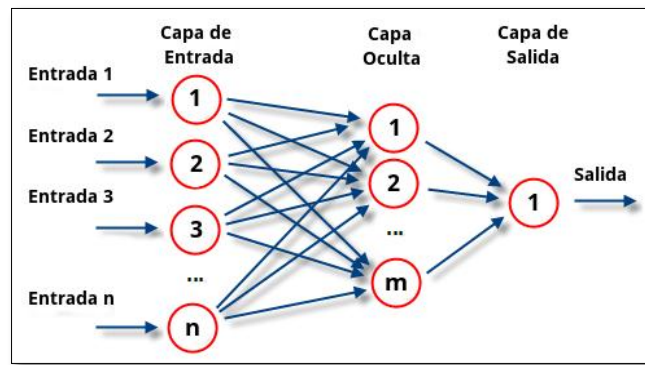


“El resultado del encadenamiento (clustering) se muestra como el color de los cuadros en 3 grupos (cadenas)”

✓ **Redes Neuronales**

Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida (IBM Knowledge Center, 2017). Algunos ejemplos de red neuronal son:

- El Perceptrón.
- El Perceptrón multicapa.
- Los Mapas Autoorganizados, también conocidos como redes de Kohonen.

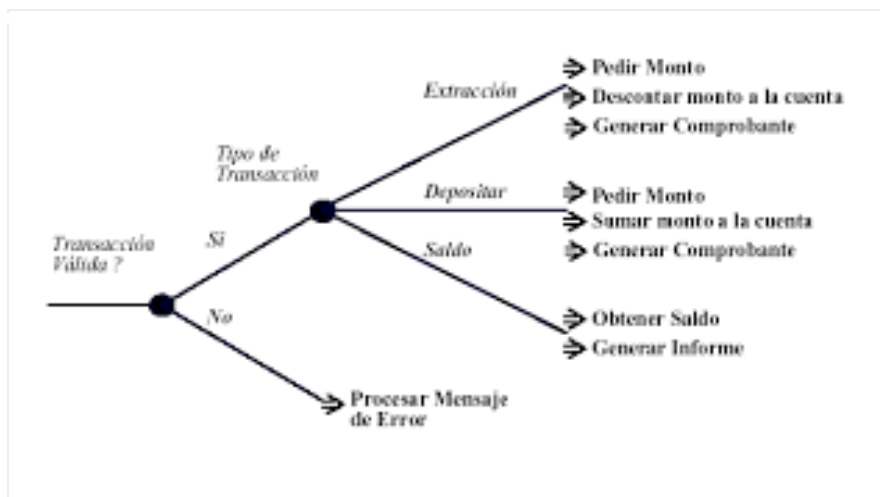


“Red neuronal artificial perceptrón simple con n neuronas de entrada, neuronas en su capa oculta y una neurona de salida”

✓ **Árboles de decisión**

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema (IBM Knowledge Center, 2017). Ejemplos:

- § Algoritmo ID3.
- § Algoritmo C4.5.



2.2.4. Modelo de Minería de Datos

Un modelo de minería de datos se crea mediante la aplicación de un algoritmo a los datos, pero es algo más que un algoritmo o un contenedor de metadatos: es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones. (MSDN, 2017)

Un modelo de minería de datos recibe los datos de una estructura de minería de datos y, a continuación, los analiza utilizando un algoritmo de minería de datos. La estructura y el modelo de minería de datos son objetos independientes. La estructura de minería de datos almacena la información que define el origen de datos. Un modelo de minería de datos almacena la información derivada del procesamiento estadístico de los datos, como los patrones encontrados como resultado del análisis. (MSDN, 2017)

Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- **Pronóstico:** cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
- **Riesgo y probabilidad:** elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
- **Recomendaciones:** determinación de los productos que se pueden vender juntos y generación de recomendaciones.
- **Búsqueda de secuencias:** análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
- **Agrupación:** distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

2.2.5. Toma de decisiones:

La toma de decisiones es el núcleo de la planeación, y se define como la selección de un curso de acción entre varias alternativas. No puede decirse que exista un plan a menos que se haya tomado una decisión: que se hayan comprometido los recursos, la dirección o la reputación; hasta ese momento sólo existen estudios de planeación y análisis. Algunas veces los gerentes consideran que la toma de decisiones es su principal tarea, pues constantemente deciden qué hacer, quién debe hacerlo y cuándo, dónde, e incluso, cómo se ha de hacer; sin embargo, la toma de decisiones es sólo un paso en el sistema de planeación. Así, incluso cuando se actúa rápido y sin pensarlo mucho, o cuando una acción tiene influencia sólo unos minutos, la planeación está presente: es parte de la vida diaria de todos. Raras veces puede juzgarse un curso de acción aislado, porque virtualmente cada decisión debe orientarse hacia otros planes. (Harold Koontz, 2012)

Supone un análisis que requiere de un objetivo y una comprensión clara de las alternativas mediante las que se puede alcanzar dicho objetivo. Además de comprender la situación que se presenta, se debe analizar, evaluar, reunir alternativas y considerar las variables, comparar varios cursos de acción y finalmente seleccionar la acción que se va a realizar. La calidad de las decisiones tomadas marca la diferencia entre el éxito o el fracaso. 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)

2.2.6. IBM SPSS MODELER

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, 2017)

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 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 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.

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 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.

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.
- Combinar modelos predictivos, reglas de negocio y técnicas de optimización para mejorar la toma de decisiones.
- Ofrecer recomendaciones a personas y sistemas, que redundan en una mejora de las decisiones y las acciones.
- Integrar resultados analíticos en procesos empresariales existentes y aplicaciones operativas.

Extraer valor de los datos

- Descubrir información y modelos atrapados en datos con algoritmos estadísticos y análisis de texto.
- 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.

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.
- Habilitar un flujo de trabajo dinámico a partir de la integración con IBM SPSS Statistics, Cognos Business Intelligence, Cognos TM1 e InfoSphere Streams.

- 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.

2.3. METODOLOGIA PARA EL DESARROLLO DEL PROYECTO

2.3.1. METODOLOGÍA CRISP-DM

Crisp-DM es una organización europea creada por tres grandes jugadores en proyectos de minería de datos que son SPSS, NCR Y Daimler Chrysler. Lo que trata esta metodología es desarrollar los proyectos de minería de datos bajo un proceso estandarizado de definición y validación de tal forma que se desarrollen proyectos minimizando los costos que impliquen y con un alto impacto en el negocio.

La metodología CRISP-DM proporciona dos documentos como herramienta de ayuda en el desarrollo del proyecto de minería de datos: el modelo de referencia y la guía del usuario.

El documento modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de minería en general.

La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de minería de datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de minería de datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.



. Fases del proceso de modelado metodología CRISP-DM.

Figura 02: Metodología CRISP DM
(IBM, 2012)

En la figura, las flechas indican relaciones más habituales entre las fases, aunque podamos establecer relaciones entre cualquier fase. El círculo exterior simboliza la naturaleza cíclica del proceso de modelado.

La primera fase análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

La segunda fase de análisis de datos comprende la recolección inicial de datos en orden a que se establezca un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

Una vez realizado el análisis de datos, la metodología establece que se proceda **la preparación de los datos**, de tal forma que sean tratados por las técnicas de modelado. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

La fase de preparación de los datos, se encuentra muy relacionada con **la fase de modelado**. Independientemente de la técnica de modelado, los datos necesitan ser procesados en diferentes formas. Por lo tanto las fases de preparación y modelado interactúan de forma sistemática.

En la **fase de modelado se seleccionan las técnicas de modelado** más apropiadas para el proyecto de minería de datos específico.

En la fase de evaluación se evalúa el modelo escogido, no desde el punto de vista general, sino del cumplimiento de los objetivos del negocio. Se revisa el proceso teniendo en cuenta los resultados obtenidos, para repetir alguna fase en caso que se hayan cometido errores. El modelo generado es válido en función a su precisión y número de clústeres.

3. MATERIALES Y METODOS

3.1. Material

3.1.1. Población

Registros de las ventas de la base de datos transaccional de la empresa.

3.1.2. Muestra

Registros de las ventas de los años 2014-2015 de la empresa.

3.1.3. Unidad de análisis

Información de las ventas.

3.2. Método

3.2.1. Tipo de investigación

Se tomará información general para aplicar a un caso puntual por tanto se utilizará el método hipotético - deductivo, posteriormente del resultado de la investigación y aplicación se obtendrá una base para la implementación de la minería de datos en este tipo de empresas que puede ser guía para la sociedad, surgiendo así una idea aplicable a un universo. El proyecto también estará bajo un método de investigación de tipo inductivo, en conclusión el método será **Hipotético Deductivo – Inductivo**.

3.2.2. Diseño de Investigación

Se realizará un diseño cuasi-experimental en el cual se evaluará la eficiencia del nuevo modelo encontrado a través de la Minería de Datos.

Se realizarán observaciones después del tratamiento.

Diagrama de investigación cuasi-experimental	G -> X -> O1
G (Grupo a investigar)	Información de las ventas
X (Tratamiento)	Aplicación del Modelo
O (Observación)	O1: Observación post-test

Tabla 1. Diagrama de investigación

3.2.3. Variables de estudio y Operacionalización

- ✓ Independiente (VI): Solución de analítica predictiva.
- ✓ Dependiente (VD): Conocer el Patrón de consumo de clientes en la empresa Cienpharma S.A.C.

Tabla 2: Operacionalización de las variables

Variable	Dimensión	Indicador	Unidad de medida	Instrumento de Investigación
VI	Modelo de analítica predictiva	Técnicas aplicadas al modelo	Nº Técnicas usadas	Lista de Técnicas de minería de datos usadas
VD	Patrón de consumo	Nº de patrones encontrados	Nº de patrones encontrados	Hoja resumen de Nº de patrones encontrados

3.2.4. Técnicas e instrumentos de recolección de datos

3.2.4.1. Técnicas

Se utilizará como técnica la encuesta que se les aplicara a las personas encargadas en la toma de decisiones para saber cuáles son las necesidades del área en específica.

3.2.4.2.Instrumentos

Se empleará un cuestionario para saber cuáles son los puntos necesarios a conocer y saber cuáles son los requerimientos que tiene el área en específico.

El Tipo de cuestionarios de encuesta a utilizar será de entrevista personal siendo mixta, es decir parte del cuestionario será un cuestionario abierto y otra parte será cerrado.

3.2.5. Técnicas de procesamiento y análisis de datos

3.2.5.1.Procesamiento de datos

Una vez realizada la recojo de datos a través de las entrevistas y cuestionarios, así como la obtención de la base de datos transaccional, comienza 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 esquematiza en describir el tratamiento estadístico de los datos a través de gráficos, tablas y cuadros generados por el análisis de los datos, describir datos, valores y puntuación y distribución de frecuencia para cada variable.

La función estadística a emplear está basada en la “Distribución T de Student”.

4. RESULTADOS : APLICACIÓN DE LA METODOLOGIA

4.1. ANALISIS DEL PROBLEMA

4.1.1. Objetivos Institucionales:

Cienpharma S.A.C. es una empresa que se dedica a la distribución y logística de productos farmacéuticos en varios departamentos del país, cuyo rango de influencia es el norte peruano, esforzándose día a día para ofrecer a sus clientes un servicio comercial de primera calidad, garantizando el mejor trato posible así como una rápida entrega de los productos farmacéuticos, en todos los segmentos del mercado a donde estén dirigidos.

Cienpharma S.A.C. es una empresa con más de 14 años en el sector de Venta al por Mayor de Distribución de Medicamentos, Tiene distribución en varios departamentos a nivel nacional y dando la preferencia a sus clientes que han apostado por sus productos y calidad de servicio.

CienPharma como distribuidor de medicamentos inicia sus operaciones el 14 de Noviembre del 2005, aperturando la primera sucursal en la ciudad de Trujillo a principios del mes de marzo del 2006, la cual por ser la primera distribuidora CienPharma en abrir sus puertas a la atención de nuestros clientes se formó como una de las pioneras en cuanto a lo que la empresa considera la Farmacia Ideal! Acumulando su labor durante las 24 horas sin descanso hasta el momento. Seguidamente en poco tiempo se aperturaron 4 agencias/sucursales CienPharma en varias zonas de la Cajamarca, La Libertad y Lambayeque a finales del mes de marzo del 2007, con una amplia visión de la importancia del mercado farmacéutico en esa zona tan poblada, se postuló rápidamente como la número 1 de entre la competencia local.

Actualmente CienPharma es la distribuidora de medicamentos más renombrada de la Libertad y Cajamarca que cuenta con más de 200 clientes.

Visión de la empresa

“Ser el más grande proveedor de farmacéutico de la industria.”

Misión de la empresa

“Contribuir con el bienestar general de la humanidad, brindando – a través de un óptimo servicio – salud a las personas y solvencia a nuestros clientes.”

Valores de la empresa

- ✓ Servicio
- ✓ Dedicación
- ✓ Integridad

El objetivo institucional para la empresa es disminuir la problemática descrita en el enunciado del problema, conocer el patrón de consumo de los clientes usando técnicas de Minería de Datos y la Metodología CRISP-DM.

4.1.2. Evaluación de la Situación:

- El sistema transaccional de ventas con que cuenta la empresa fue desarrollada con el fin de recolectar, almacenar, modificar y recuperar todo tipo de información y no de brindar información del patrón de comportamiento de los clientes.
- No existe una preparación de datos adecuada para el análisis de la información.
- Altos tiempos de ejecución de consultas para obtener información sobre las transacciones de ventas de los clientes.
- Falta de información para clasificar o diferenciar los clientes para temas de marketing.

4.1.3. Recursos Computacionales:

HARDWARE Y SOFTWARE: 10 PCs	
Resumen Software:	
Sistema operativo:	Microsoft Windows 8 Pro
Ofimática:	Microsoft Office 2016
Software para Minería de Datos:	IBM SPSS Modeler 18
Software para Base de datos:	MS SQL SERVER 2014

Resumen Hardware:	
Procesador:	Intel Core i5 4570 3.2GHz
Case:	Dell OptiPlex 9020 SFF
Placa base:	Intel Q87
Memoria:	8 GBytes
Disco duro:	500 GB Sata 6Gb/s 7200 rpm
Monitor:	Monitor Dell E1916H de 18.5"

Tabla 3: Recursos Computacionales

4.2. ANALISIS DE DATOS

4.2.1. Recolección de Datos Iniciales:

En este primer paso se procedió a la recopilación de los datos necesarios para llevar a cabo el estudio en este trabajo de la empresa, donde la empresa proporciono los registros de clientes y ventas realizadas en los años 2013-2014.

Los datos recolectados para la investigación se han categorizado de la siguiente manera:

- **Clientes:** Codigo_cliente, nombre_cliente, Departamento, provincia, urbanización de procedencia.
- **Ventas (Pedidos):** Se tomó como referencia las compras realizadas por los clientes en los periodos antes mencionados.
- **Artículos:** Se tomó como referencia los productos comprados por los clientes en los periodos antes mencionados.

Las tablas involucradas en esta recolección se encuentran en la base de datos transaccional de la empresa como por ejemplo:

- Clientes
- FacArt (Factura)
- Arti (Articulo)
- Ubigeo

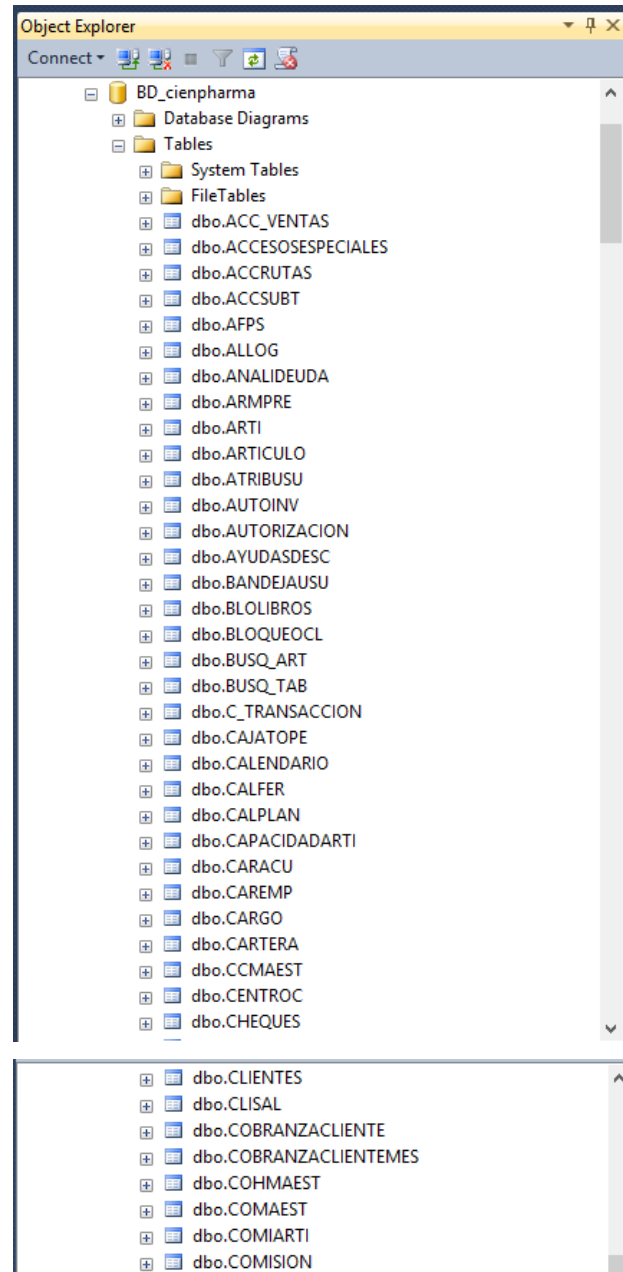


Figura 03: Base de datos Cienpharma

4.2.2. Descripción de los Datos:

La base de datos que tiene la empresa se almacena a través de Transact-SQL en el

gestor de base de datos de MS SQL Server 2014. Las tablas que utilizaremos en este trabajo serán las siguientes:

Tabla 1 CLIENTE: En esta tabla se almacena información básica de los clientes que tiene la empresa en todos sus años de funcionamiento. Esta tabla contiene un total de 2338 registros.

Column Name	Data Type	Allow Nulls
CLI_CODCLIE	numeric(8, 0)	<input type="checkbox"/>
CLI_CODCIA	char(2)	<input type="checkbox"/>
CLI_CP	char(1)	<input type="checkbox"/>
CLI_NOMBRE	char(100)	<input checked="" type="checkbox"/>
CLI_NOMBRE_ESPOSO	char(100)	<input checked="" type="checkbox"/>
CLI_NOMBRE_ESPOSA	char(100)	<input checked="" type="checkbox"/>
CLI_NOMBRE_EMPRESA	char(100)	<input checked="" type="checkbox"/>
CLI_123	numeric(1, 0)	<input checked="" type="checkbox"/>
CLI_TELEF1	char(12)	<input checked="" type="checkbox"/>
CLI_TELEF2	char(12)	<input checked="" type="checkbox"/>
CLI_CASA_DIREC	char(100)	<input checked="" type="checkbox"/>
CLI_CASA_NUM	numeric(4, 0)	<input checked="" type="checkbox"/>
CLI_CASA_ZONA	int	<input checked="" type="checkbox"/>
CLI_CASA_SUBZONA	int	<input checked="" type="checkbox"/>
CLI_TRAB_DIREC	char(80)	<input checked="" type="checkbox"/>
CLI_TRAB_NUM	numeric(4, 0)	<input checked="" type="checkbox"/>
CLI_TRAB_ZONA	int	<input checked="" type="checkbox"/>
CLI_TRAB_SUBZONA	int	<input checked="" type="checkbox"/>
CLI_TRAB_PROV	int	<input checked="" type="checkbox"/>
CLI_RUC_ESPOSO	char(15)	<input checked="" type="checkbox"/>
CLI_RUC_ESPOSA	char(15)	<input checked="" type="checkbox"/>
CLI_RUC_EMPRESA	char(15)	<input type="checkbox"/>

Figura 04: Tabla Cliente

CAMPOS	DESCRIPCIÓN	EJEMPLO
CLI_CODCLIE	Código del cliente	12 , 13, 14
CLI_NOMBRE	Nombres y Apellidos de los clientes	ALARCON BARRETO

Tabla 04:
Tabla Cliente

		IVAN ARTURO
CLI_RUC	RUC del cliente	20181712028
CLI_TRAB_DIREC	Dirección exacta del cliente	JR INCA ROCA 129 BAÑOS DEL INCA
CLI_LIMCRE	Línea de crédito del cliente	25000, 15000
CLI_DIA_VISITA	Número de veces que visito el negocio	42, 17
CLI_IDUBIGEO	Ubicación	Chachapoyas, Nuevo Cajamarca

Tabla 2 FACTURA: En esta tabla se almacena información de las compras realizadas por los clientes. Esta tabla contiene un total de 50732 registros.

Column Name	Data Type	Allow Nulls
FAR_TIPMOV	int	<input type="checkbox"/>
FAR_CODCIA	char(2)	<input type="checkbox"/>
FAR_NUMSER	char(3)	<input type="checkbox"/>
FAR_FBG	char(1)	<input type="checkbox"/>
FAR_NUMFAC	numeric(9, 0)	<input type="checkbox"/>
FAR_NUMSEC	int	<input type="checkbox"/>
FAR_FECHA	datetime	<input checked="" type="checkbox"/>
FAR_NUMOPER	int	<input type="checkbox"/>
FAR_CODCLIE	numeric(8, 0)	<input type="checkbox"/>
FAR_CODART	numeric(8, 0)	<input type="checkbox"/>
FAR_TRANSITO	char(1)	<input checked="" type="checkbox"/>
FAR_ESTADO	char(1)	<input type="checkbox"/>
FAR_NUMGUIA	numeric(9, 0)	<input checked="" type="checkbox"/>
FAR_DIAS	int	<input type="checkbox"/>
FAR_SIGNO_ARM	int	<input checked="" type="checkbox"/>
FAR_PRECIO	numeric(13, 4)	<input type="checkbox"/>
FAR_STOCK	numeric(13, 4)	<input type="checkbox"/>
FAR_COSPRO	numeric(13, 4)	<input type="checkbox"/>

Figura 05: Tabla Factura

CAMPOS	DESCRIPCIÓN	EJEMPLO
FAR_TIPMOV	Código del tipo de movimiento	101
FAR_CODCIA	Código del cia	15
FAR_NUMFAC	Numero de factura	1
FAR_FECHA	Fecha de factura	2014-01-04

Tabla 05:
Tabla Factura

FAR_CODCLIE	Código del cliente	1868
FAR_CODART	Código del artículo	290512
FAR_ESTADO	Estado	N
FAR_PRECIO	Precio	22.7429
FAR_STOCK	Stock	66.0000
FAR_BRUTO	Costo Bruto	318.40
FAR_NUM_LOTE	Número de lote	0
FAR_CANTIDAD	Cantidad	14.0000
FAR_CONCEPTO	Concepto	CANJE X VCTO OPHTHA
FAR_DESCRI	Descripción del producto	CJA
FAR_SUBTOTAL	Subtotal	318.40

Tabla 3 ARTICULO: En esta tabla se almacena información sobre los artículos vinculados a las compras realizadas por los clientes. Esta tabla contiene un total de 6493 registros.

Column Name	Data Type	Allow Nulls
ART_KEY	numeric(8, 0)	<input type="checkbox"/>
ART_CODCIA	char(2)	<input type="checkbox"/>
ART_NOMBRE	char(100)	<input checked="" type="checkbox"/>
ART_COSTO	numeric(11, 2)	<input checked="" type="checkbox"/>
ART_MARGEN	numeric(11, 2)	<input checked="" type="checkbox"/>
ART_CASH	numeric(11, 2)	<input checked="" type="checkbox"/>
ART_TIPO	char(1)	<input checked="" type="checkbox"/>
ART_ESTADO	char(1)	<input checked="" type="checkbox"/>
ART_NUMERO	int	<input checked="" type="checkbox"/>
ART_LINEA	int	<input checked="" type="checkbox"/>
ART MARCA	int	<input checked="" type="checkbox"/>

Figura 06: Tabla Articulo

Tabla N° 06:
Tabla Articulo

CAMPOS	DESCRIPCIÓN	EJEMPLO
ART_KEY	Código del Artículo	107
ART_NOMBRE	Nombre del Artículo	ATIFLAM 15 MG X 10 CAP
STOCK	Stock	0.000000000000000000
PROM_VENTA	Promedio de venta	0.0000000000000000
TIPO	tipo	T

Tabla 4 UBIGEO: En esta tabla se almacena información sobre el lugar de procedencia de los clientes. Esta tabla contiene un total de 2073 registros.

Column Name	Data Type	Allow Nulls
ID	float	<input checked="" type="checkbox"/>
C_PAIS	float	<input checked="" type="checkbox"/>
C_DEPART	nvarchar(4)	<input checked="" type="checkbox"/>
C_PROVIN	nvarchar(4)	<input checked="" type="checkbox"/>
C_DISTRI	nvarchar(4)	<input checked="" type="checkbox"/>
NOMBRE	nvarchar(180)	<input checked="" type="checkbox"/>
F_UBIGEO_EST	nvarchar(1)	<input checked="" type="checkbox"/>
		<input type="checkbox"/>

Figura 07: Tabla Ubigeo

ID	C_DEPART	DEPARTAMENTO	C_PROVIN	PROVINCIA	C_DISTRI	DISTRITO
0	4028					A
93	2	ANCASH	01	HUARAZ	01	HUARAZ
94	2	ANCASH	01	HUARAZ	02	COCHABAMBA
95	2	ANCASH	01	HUARAZ	03	COLCABAMBA
96	2	ANCASH	01	HUARAZ	04	HUANCHAY
97	2	ANCASH	01	HUARAZ	05	INDEPENDENCIA
98	2	ANCASH	01	HUARAZ	06	JANGAS
99	2	ANCASH	01	HUARAZ	07	LA LIBERTAD
100	2	ANCASH	01	HUARAZ	08	OLLEROS
101	2	ANCASH	01	HUARAZ	09	PAMPAS
102	2	ANCASH	01	HUARAZ	10	PARIACOTO
103	2	ANCASH	01	HUARAZ	11	PIRA
104	2	ANCASH	01	HUARAZ	12	TARICA
106	2	ANCASH	02	AJJA	01	AJJA

Tabla 07: Tabla Ubigeo

4.2.3. Exploración y Validación de los datos:

De acuerdo a los objetivos planteados en el trabajo sobre conocer la información de los clientes y sus compras realizadas en la empresa y descubrir el patrón de comportamiento de los clientes, se realizó un análisis desde la base de datos transaccional, para conocer la actividad de compra de los clientes.

Se realizaron algunas consultas previas:

✓ **Número de clientes por Ubigeo**

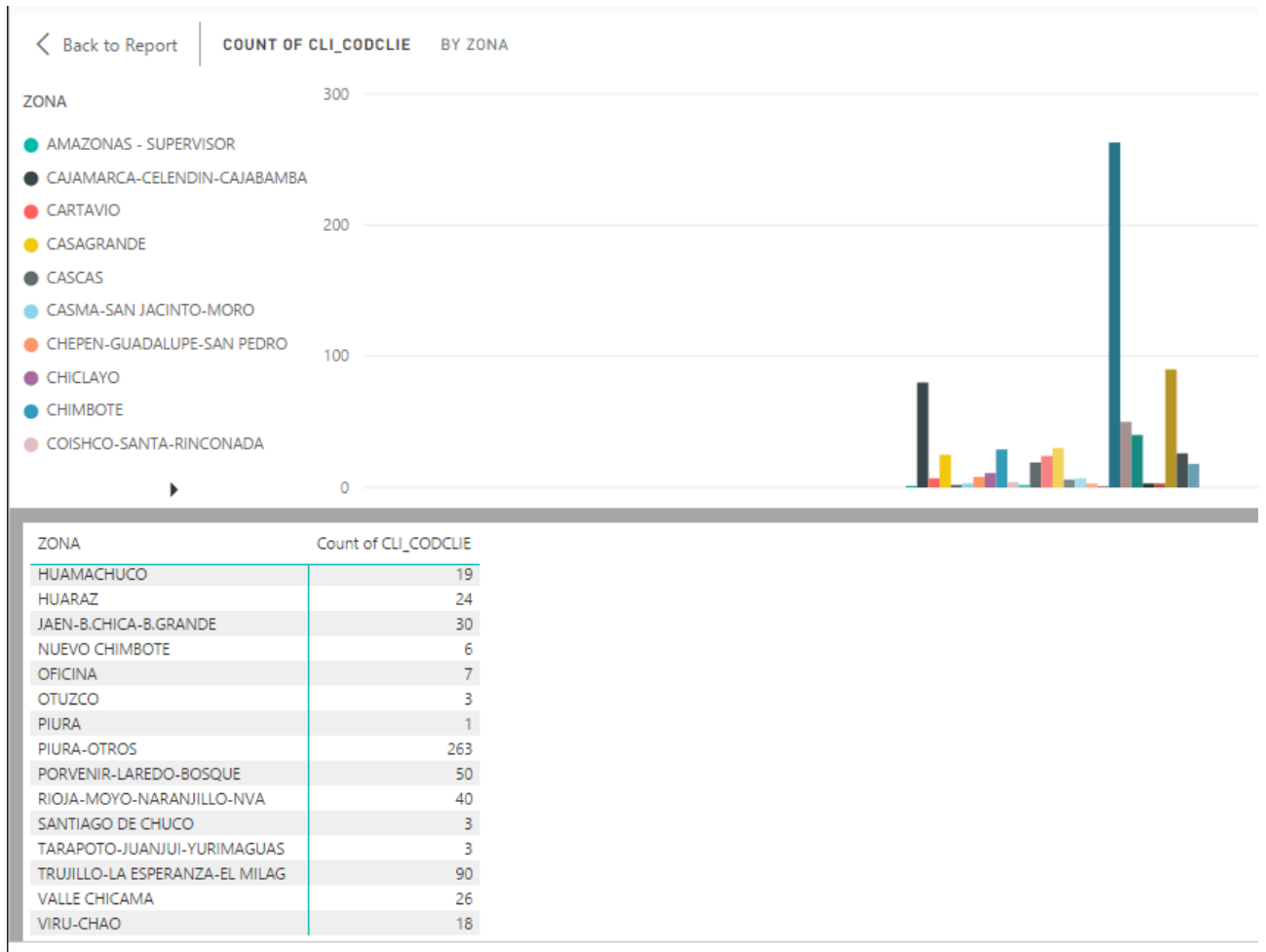


Figura 08: Número de clientes por Ubicación

Esta consulta nos muestra la ubicación de los clientes, evidenciando las zonas donde existe mayor cantidad de clientes de la empresa.

✓ **Cantidad en monto de compras por cliente**

FAR_CODCLIE	FAR_SUBTOTAL	Year	Quarter	Month	Day
56,00	2.061,24	2014	Qtr 1	February	3
56,00	0,00	2014	Qtr 1	February	28
56,00	1.479,30	2014	Qtr 1	March	20
56,00	2.275,39	2014	Qtr 2	April	23
56,00	6.634,20	2014	Qtr 2	May	23
56,00	432,00	2014	Qtr 2	June	3
57,00	612,50	2014	Qtr 1	January	22
57,00	120,00	2014	Qtr 1	February	26
57,00	989,52	2014	Qtr 1	March	4
57,00	148,14	2014	Qtr 1	March	5
57,00	738,95	2014	Qtr 1	March	7
57,00	262,14	2014	Qtr 1	March	10
57,00	280,55	2014	Qtr 1	March	15
57,00	612,50	2014	Qtr 1	March	27
57,00	248,94	2014	Qtr 2	April	5
57,00	353,06	2014	Qtr 2	April	8
57,00	118,84	2014	Qtr 2	April	10
57,00	543,32	2014	Qtr 2	April	12
57,00	461,04	2014	Qtr 2	April	21
57,00	246,90	2014	Qtr 2	April	22
57,00	155,97	2014	Qtr 2	April	24
57,00	86,04	2014	Qtr 2	April	26
57,00	64,65	2014	Qtr 2	April	28
57,00	51,98	2014	Qtr 2	April	29
Total	6.489.673,44				

Figura 09: Cantidad en monto de compras por cliente

En esta consulta se muestra un resumen del monto en compras por cliente.

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. Preparación y Construcción de datos:

En esta etapa se tiene como objetivo preparar y construir un conjunto de datos para que pueda ser minable.

Para este paso se seleccionó los datos tomando en cuenta los objetivos planteados y de acuerdo a la experiencia sobre criterios de éxito. Para lo cual se realizó una consulta a las tablas involucradas en la compra que los clientes realizaron en la empresa.

```
SELECT distinct dc.[CLI_NOMBRE]
,Estado_Civil
,c.[CLI_NOMBRE_EMPRESA] ,c.[CLI_LIMCRE] ,dc.[ZONA]
,dc.[DISTRITO] ,dc.[PROVINCIA] ,F.[FAR_FECHA]
,F.[FAR_CODART],a.ART_NOMBRE,F.[FAR_DESCRI]
,F.[FAR_CANTIDAD],F.[FAR_SUBTOTAL],c.[CLI_DIA_VISITA]

FROM [BD_cienpharma].[dbo].[FACART] F,
[BD_cienpharma].[dbo].[DireccionCliente] dc, [BD_cienpharma].[dbo].[ARTI] a,
BD_cienpharma.dbo.CLIENTES c

WHERE F.[FAR_CODCLIE]=dc.CLI_CODCLIE and dc.CLI_CODCLIE=c.CLI_CODCLIE
and F.FAR_CODART=a.ART_KEY
Group by dc.[CLI_NOMBRE], c.CLI_RUC_ESPOSA ,c.[CLI_NOMBRE_EMPRESA]
,c.[CLI_LIMCRE],dc.[ZONA] ,dc.[DISTRITO] ,dc.[PROVINCIA]
,F.[FAR_FECHA] ,F.[FAR_CODART] ,a.ART_NOMBRE ,F.[FAR_DESCRI]
,F.[FAR_CANTIDAD] ,F.[FAR_SUBTOTAL] ,c.[CLI_DIA_VISITA]
```

Figura 10: Consulta SQL – Preparación de datos

CLI_NOMBRE	Estado_Civil	CLI_NOMBRE_EMPRESA	CL...	ZONA	DISTRITO	PROVINCIA	FAR_FECHA	FAR_CO...	ART_NOMBRE
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	64070	CLARITROMICINA 500 MG X 10 TAB
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	100649	AZITROMICINA 200MG/5ML X15ML S
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	155405	ALBENDAZOL 200MG X 2 TAB
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	200014	GENTAMICINA VITALINE 0.3% GOTA'
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	228191	DICLOXACILINA 125MG/5ML JBE X 8
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	246191	VICK VAPORUB 12G X 12LATAS
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	275747	ASEPXIA JABON AZUFRE X 100 GR
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2013-12-17 00:00:00.000	275961	ASEPXIA JABON HERBAL X 100 GR
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2014-03-14 00:00:00.000	13884	FALEXIM 500 MG X 100 CAP
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2014-03-14 00:00:00.000	108491	TERBOCLOXIL 500MG X 100 CAP
1... GONZALES RODRI...	C	FARMACIA VIRGEN DEL CAR...	50...	PORVENIR-LAREDO-BO...	TRUJILLO	LA LIBERT...	2014-03-14 00:00:00.000	175198	ERITROMICINA 500MG X 100 COMP
1... GRANADOS RAMI...	C	BOTICAS G & R	40...	PORVENIR-LAREDO-BO...	EL PORV...	TRUJILLO	2014-01-16 00:00:00.000	315384	GUMMIES MULTIVITAMIN X 60 GOMI
1... GRANADOS RAMI...	C	BOTICAS G & R	40...	PORVENIR-LAREDO-BO...	EL PORV...	TRUJILLO	2014-01-16 00:00:00.000	315598	GUMMIES VITAMIN C X 60 GOMITAS
1... GRANADOS RAMI...	C	BOTICAS G & R	40...	PORVENIR-LAREDO-BO...	EL PORV...	TRUJILLO	2014-04-08 00:00:00.000	299903	EVA TEST-LAPICERO CJA X 1 PRUEE
1... GRUPO LIVES S.A.	S	LIVES	0.00	CASAGRANDE	PALERMO	TRUJILLO	2014-01-16 00:00:00.000	58312	NAPROXENO SODICO 550 MG X 100
1... GRUPO LIVES S.A.	S	LIVES	0.00	CASAGRANDE	PALERMO	TRUJILLO	2014-01-16 00:00:00.000	143056	COMPLEJO-B X 300 CAP

Figura 11: Resultado de Consulta SQL – Preparación de datos

Total de registros: 19092

Descripción de datos a utilizar:

CAMPOS	DESCRIPCIÓN	EJEMPLO
CLI_NOMBRE	Nombres y Apellidos de los clientes	ACOSTA TORRES ANASTACIO
Estado_Civil	Estado civil	C
CLI_NOMBRE_EMPRESA	Nombres de la empresa clientes	BOTICA TERESA DE CALCUTA
CLI_LIMCRE	Límite de crédito del cliente	3000.00
ZONA	Zona de procedencia	CAJAMARCA-CELENDIN-CAJABAMBA
DISTRITO	Distrito	CELENDIN
PROVINCIA	Provincia	CAJAMARCA
FAR_FECHA	Fecha de compra	2014-01-07
FAR_CODART	Código de artículo comprado	13135
ART_NOMBRE	Nombre del artículo	VITAGEN FCO X 345 ML
FAR_DESCRIP	Descripción del Artículo	UND
FAR_CANTIDAD	Cantidad comprada	2.0000
FAR_SUBTOTAL	Monto total comprado	30.10
CLI_DIA_VISITA	Días de visita del cliente	6

Tabla 08: Datos a utilizar

4.3.MODELADO

4.3.1. SELECCIÓN DE LA TÉCNICA DE MODELADO MÁS APROPIADA:

En esta tarea se realiza la selección de la técnica de Minería de datos más apropiada de acuerdo al tipo de problema que se desea resolver. Para ello se debe considerar el objetivo del proyecto y la relación que pueda existir con la herramienta de Minería de Datos escogida.

Las técnicas de modelado evaluadas se basan en conocer el patrón de consumo de los clientes en la empresa cienpharma. De las cuales se utilizaron 4 principales técnicas de modelado basado en **Clasificación y Agrupación** y usando la herramienta IBM SPSS Modeler:

1. **Nodo K-medias:** este modelo agrupa conjuntos de datos en grupos distintos (o clústeres). El método define un número fijo de clústeres, de forma iterativa asigna registros a los clústeres y ajusta los centros de los clústeres hasta que no se pueda mejorar el modelo. En lugar de intentar predecir un resultado, los modelos de k-medias utilizan un proceso conocido como aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

Los modelos de agrupación en clústeres se centran en la identificación de grupos de registros similares y en el etiquetado de registros según el grupo al que pertenecen. A menudo se hace referencia a estos modelos como modelos de aprendizaje no supervisado, ya que no hay ningún estándar externo con el que juzgar el rendimiento de la clasificación del modelo. No hay respuestas correctas o incorrectas para estos modelos. Su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones.



Los métodos de agrupación en clústeres se basan en la medición de distancias entre registros y entre clústeres. Los registros se asignan a los clústeres de un modo que tiende a minimizar la distancia entre los registros pertenecientes al mismo clúster.

- 2. Nodo Kohonen:** este modelo genera un tipo de red neuronal que se puede usar para agrupar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte. Puede observar el número de observaciones capturadas por cada unidad en el nuget de modelo para identificar unidades fuertes. Esto le proporcionará una idea del número apropiado de clústeres.



3. Nodo Árbol de decisión:

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema (IBM Knowledge Center, 2017).



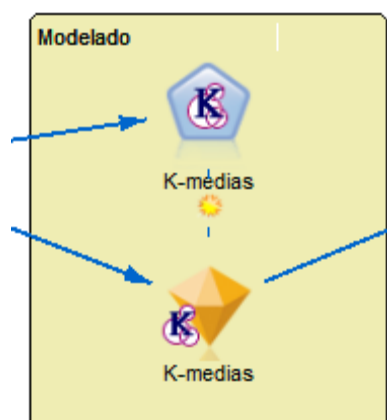
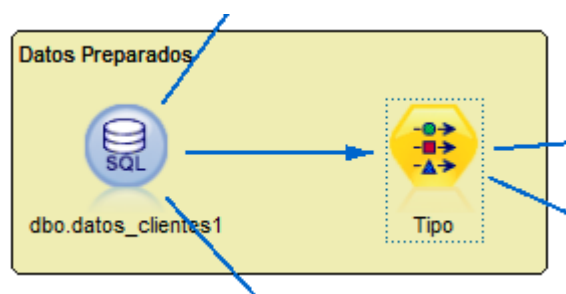
4.3.2. GENERACIÓN DEL PLAN DE PRUEBA:

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

Los datos ya preparados con anterioridad se encuentran en el servidor de SQL Server de la empresa que son llevados hacia la herramienta de IBM SPSS Modeler, para luego construir el modelo basado en el conjunto de entradas y medir la calidad del modelo y luego analizar los resultados en excel.

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

- Preparación de datos
- Modelo utilizado
- Resultados



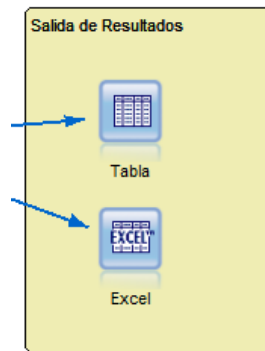


Figura N° 14: Resultado de la minería

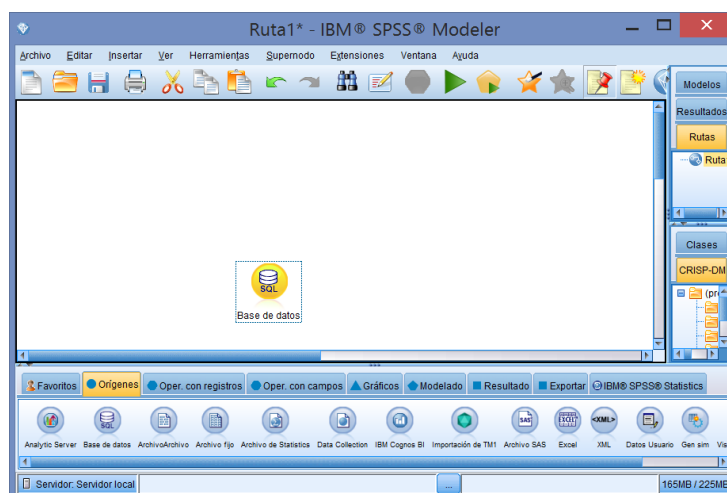
4.3.3. CONSTRUCCIÓN DEL MODELO:

Para la construcción de los modelos a evaluar se utilizó la el software IBM SPSS Modeler y la base de datos (Datos de las compras de los clientes) que están almacenadas en MS SQL Server de la empresa Cienpharma.

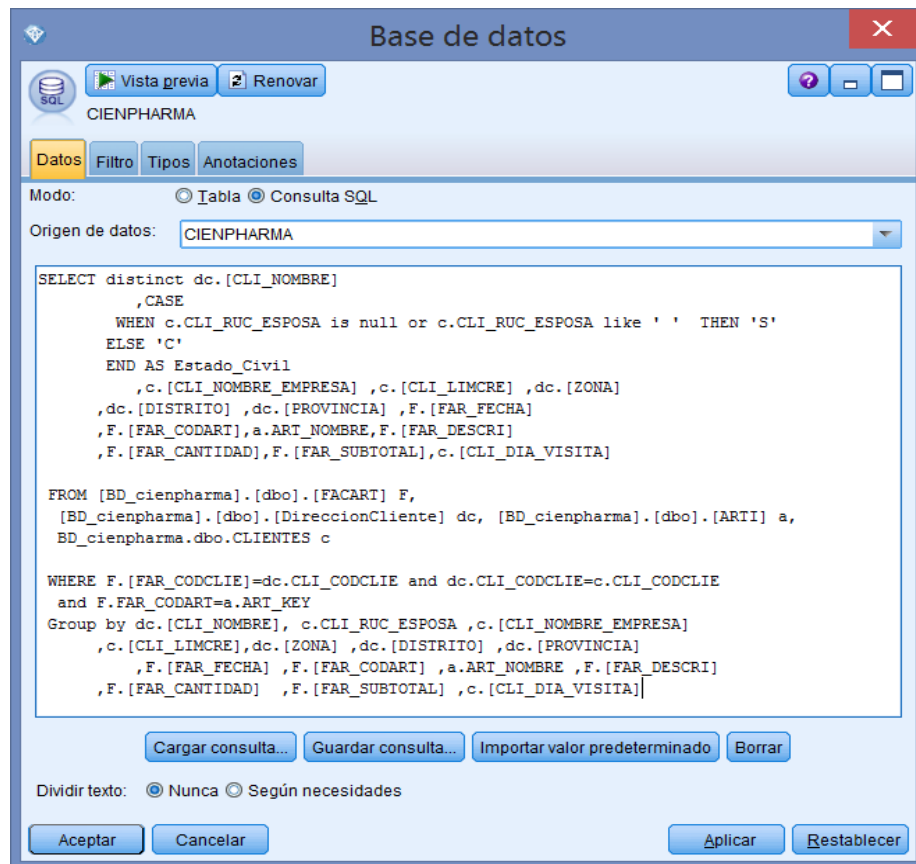
A continuación se describen los pasos a seguir para la construcción de cada modelo:

a. MODELO BASADO EN NODO K-MEDIAS

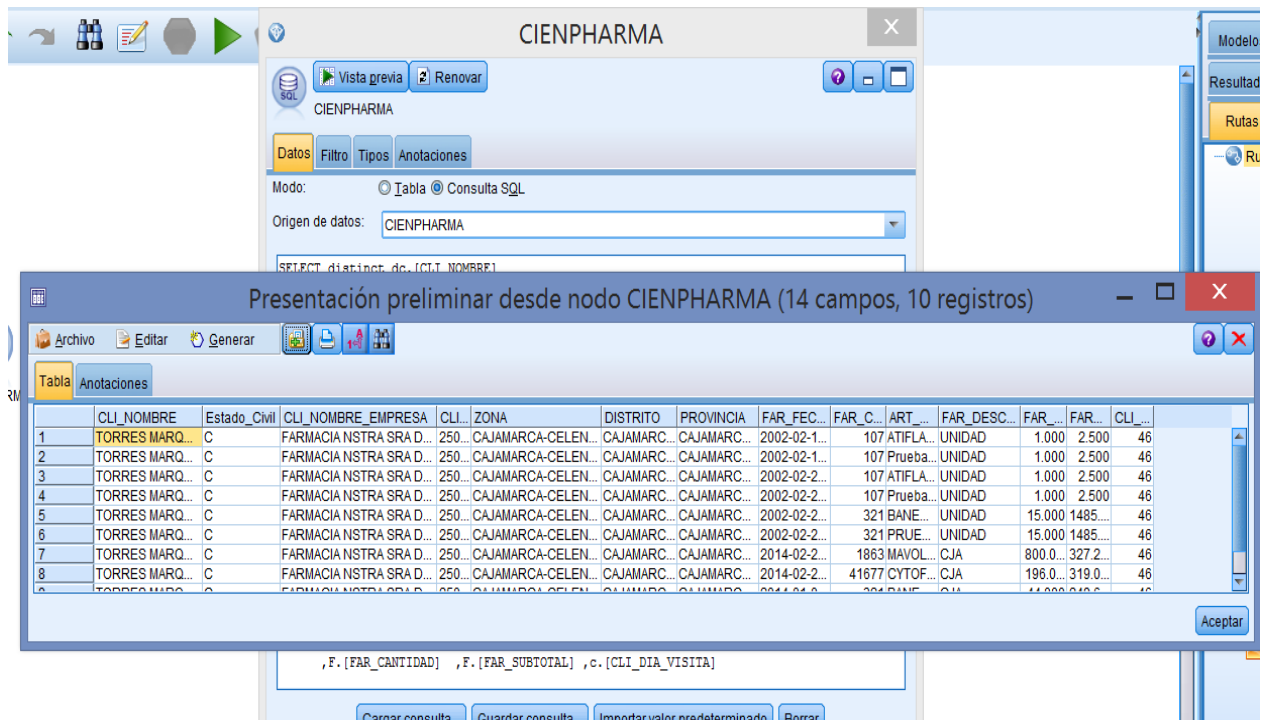
Paso 1: Abrimos el programa IBM SPSS MODELER y se procedió a establece la conexión con el origen de los datos.



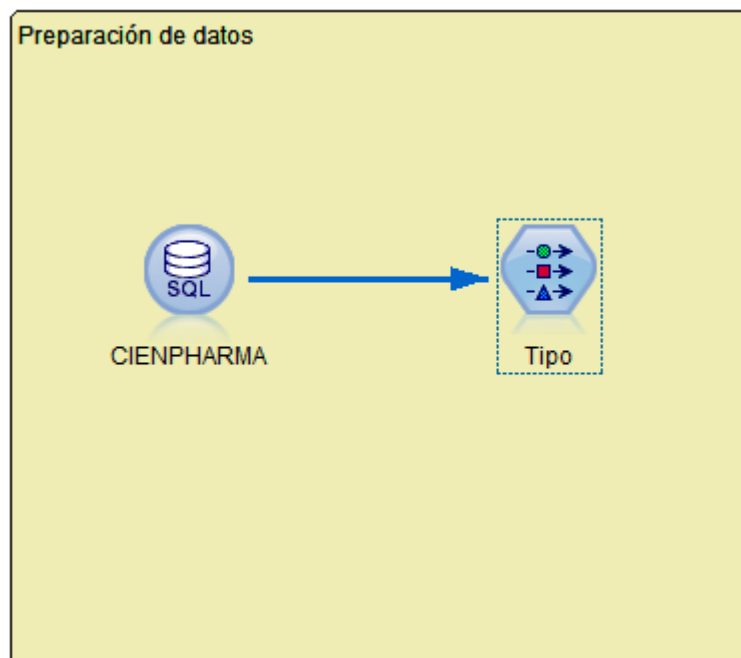
Procediendo a la selección de la tabla previamente preparada de los datos de los clientes:

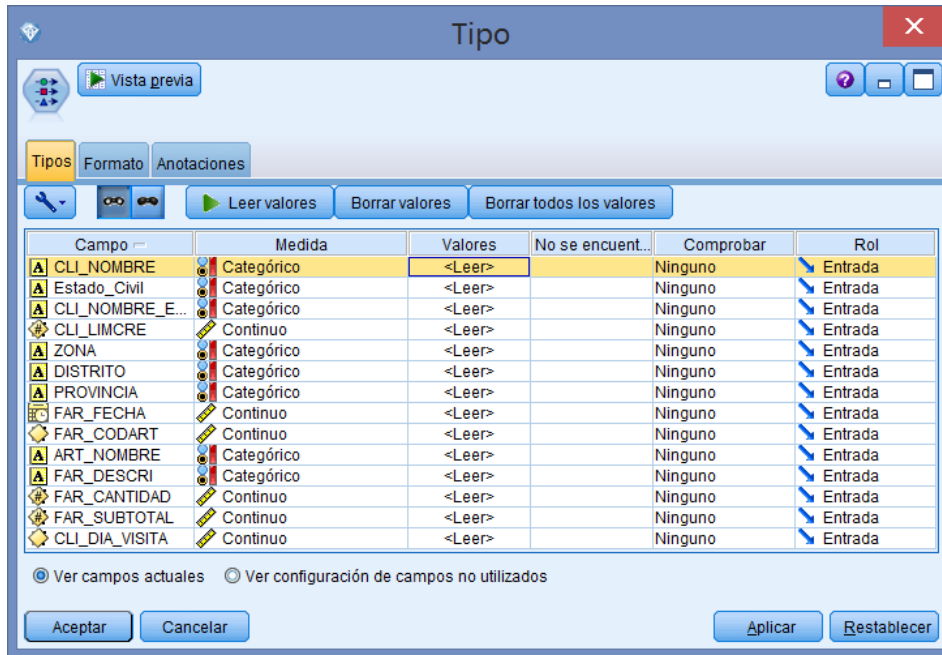


Donde se puede verificar los datos dando clic a la vista previa de los datos:

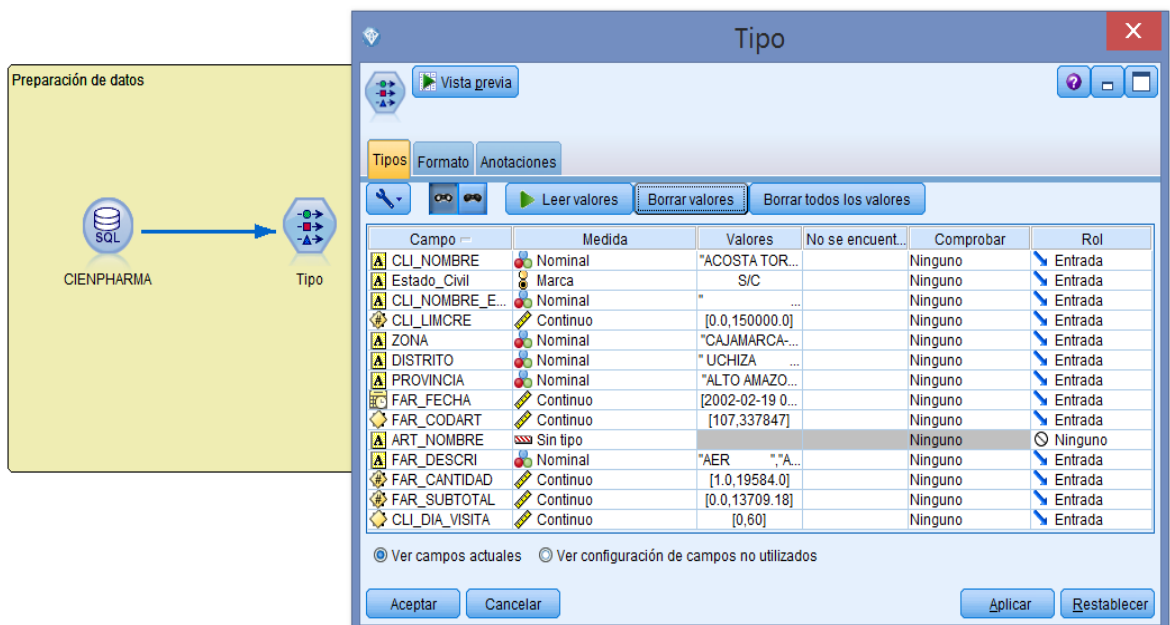


Paso 2: Se seleccionó un nodo “Tipo” de “Operaciones con campos”, para determinar y controlar los metadatos de los campos





Luego en esta configuración se da clic sobre el botón de “Leer valores” para leer los valores de los campos seleccionados cambiando la medida de acuerdo al valor del campo.



Paso 3: Se agrega al modelo el nodo del Modelo K-medias

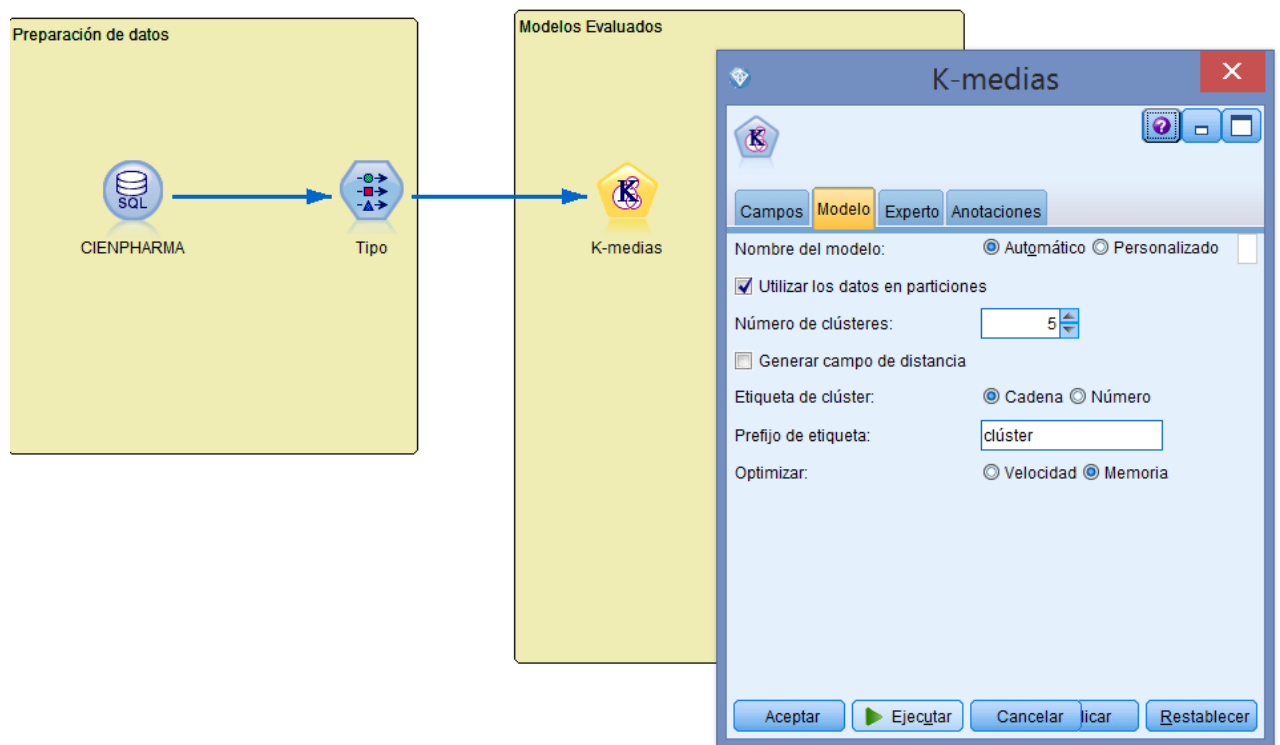
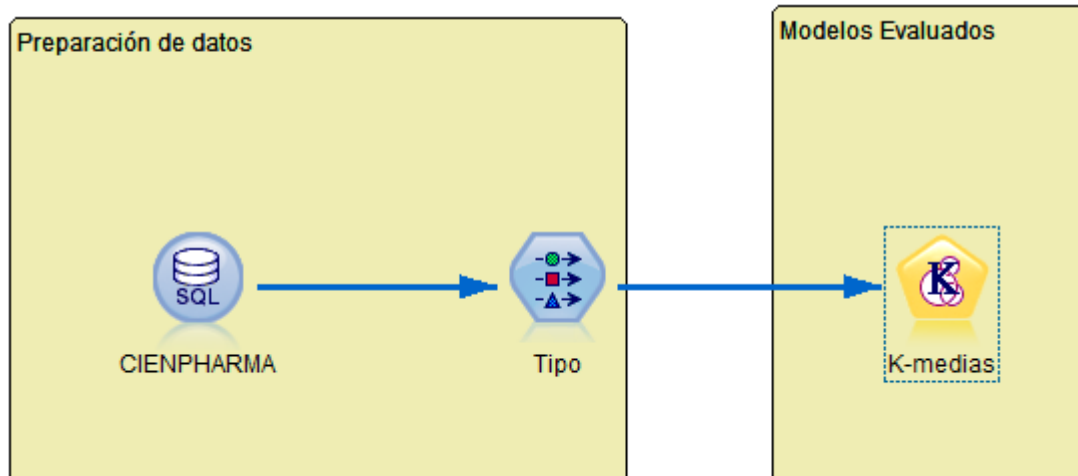


Figura 12: Ejecución del Modelo K-medias

Paso 4: Se genera el Nuget del modelo k-medias mostrando los clústeres encontrados.

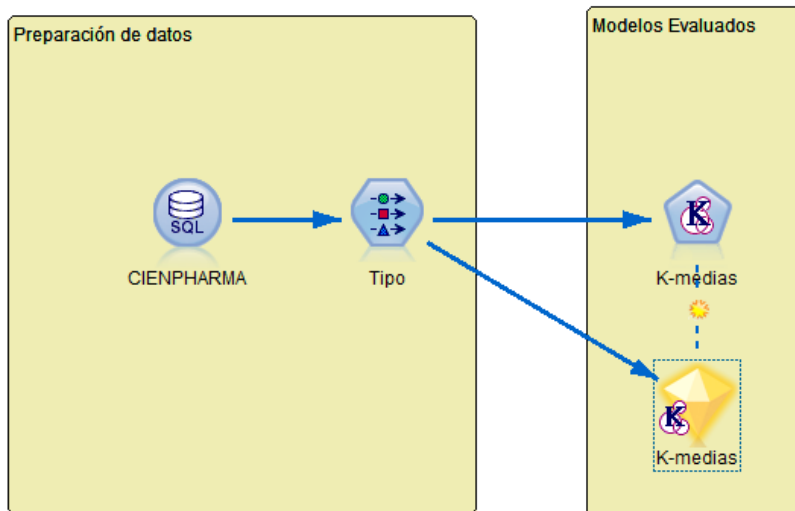
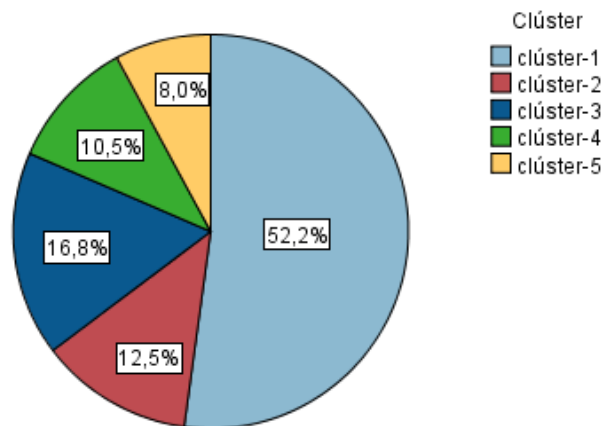


Figura 13: Resultados del Modelo K-medias

El modelo k-medias también generó 5 clústeres mostrando una visualización gráfica de estadísticas y distribuciones de resúmenes para campos entre clústeres

Tamaños de clúster



Tamaño del clúster más pequeño	1526 (8%)
Tamaño del clúster más grande	9960 (52,2%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	6,53

Como se aprecia el cluster 1 es el que agrupa una mayor cantidad de clientes con una cantidad de 9960 personas con una mayor importancia de entrada.

Clústeres

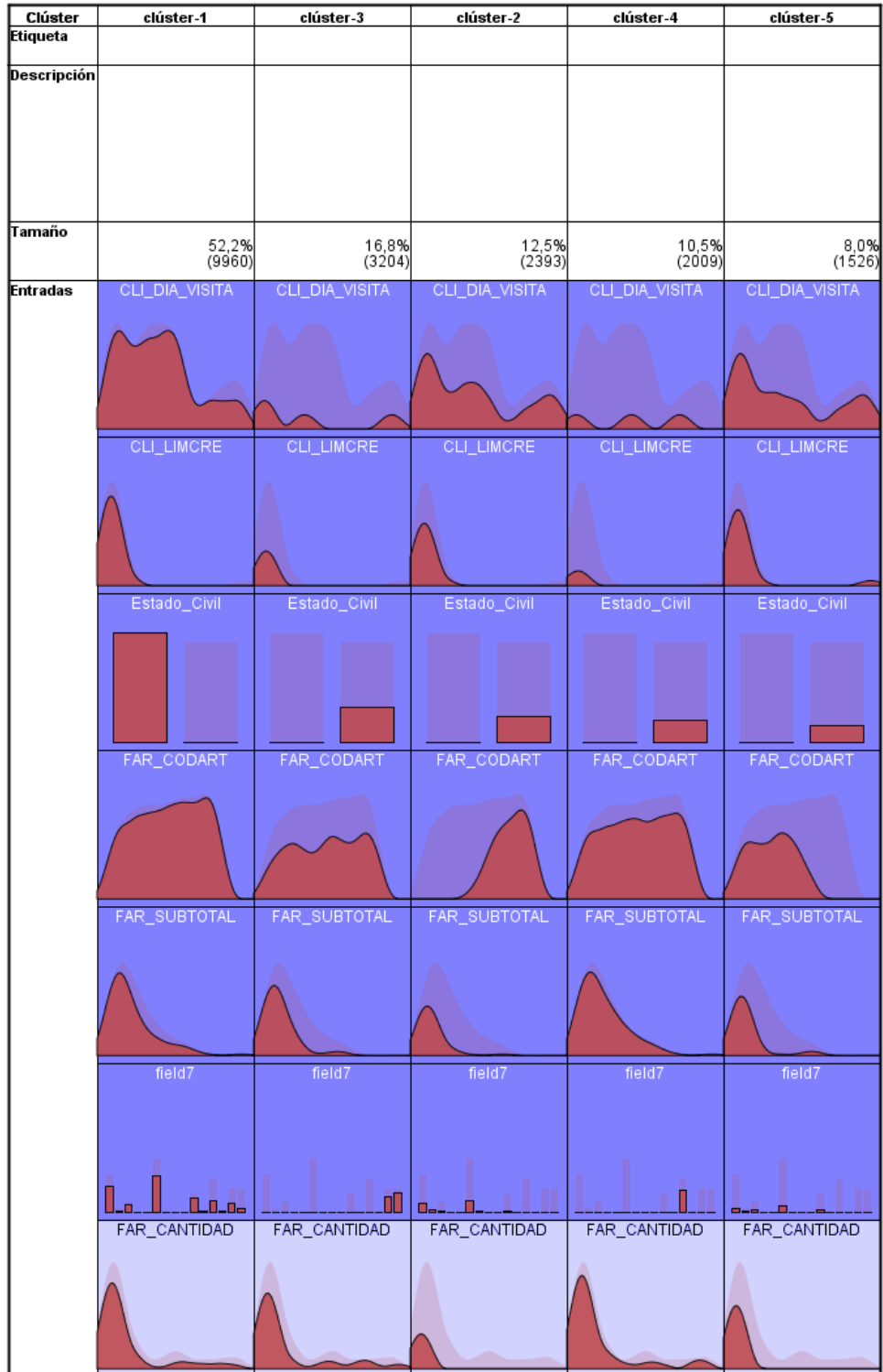
Importancia de entrada (predictor)
 1,0 0,8 0,6 0,4 0,2 0,0

Clúster	clúster-1	clúster-3	clúster-2	clúster-4	clúster-5
Etiqueta					
Descripción					
Tamaño	52,2% (9960)	16,8% (3204)	12,5% (2393)	10,5% (2009)	8,0% (1526)
Entradas	CLI_DIA_VISITA 18,35	CLI_DIA_VISITA 0,83	CLI_DIA_VISITA 3,20	CLI_DIA_VISITA 3,06	CLI_DIA_VISITA 8,79
	CLI_LIMCRE 7.480,80	CLI_LIMCRE 1.558,58	CLI_LIMCRE 10.917,84	CLI_LIMCRE 323,25	CLI_LIMCRE 10.634,01
	Estado_Civil C (100,0%)	Estado_Civil S (100,0%)	Estado_Civil S (100,0%)	Estado_Civil S (100,0%)	Estado_Civil S (100,0%)
	FAR_CODART 186.521,95	FAR_CODART 184.797,39	FAR_CODART 284.487,12	FAR_CODART 173.381,79	FAR_CODART 62.045,63
	FAR_SUBTOTAL 184,70	FAR_SUBTOTAL 560,99	FAR_SUBTOTAL 126,13	FAR_SUBTOTAL 688,80	FAR_SUBTOTAL 200,92
	field7 CAJAMARCA (32,9%)	field7 TRUJILLO (56,5%)	field7 CAJAMARCA (43,6%)	field7 LIMA (100,0%)	field7 CAJAMARCA (40,2%)
	FAR_CANTIDAD 31,94	FAR_CANTIDAD 278,79	FAR_CANTIDAD 8,40	FAR_CANTIDAD 129,15	FAR_CANTIDAD 36,52

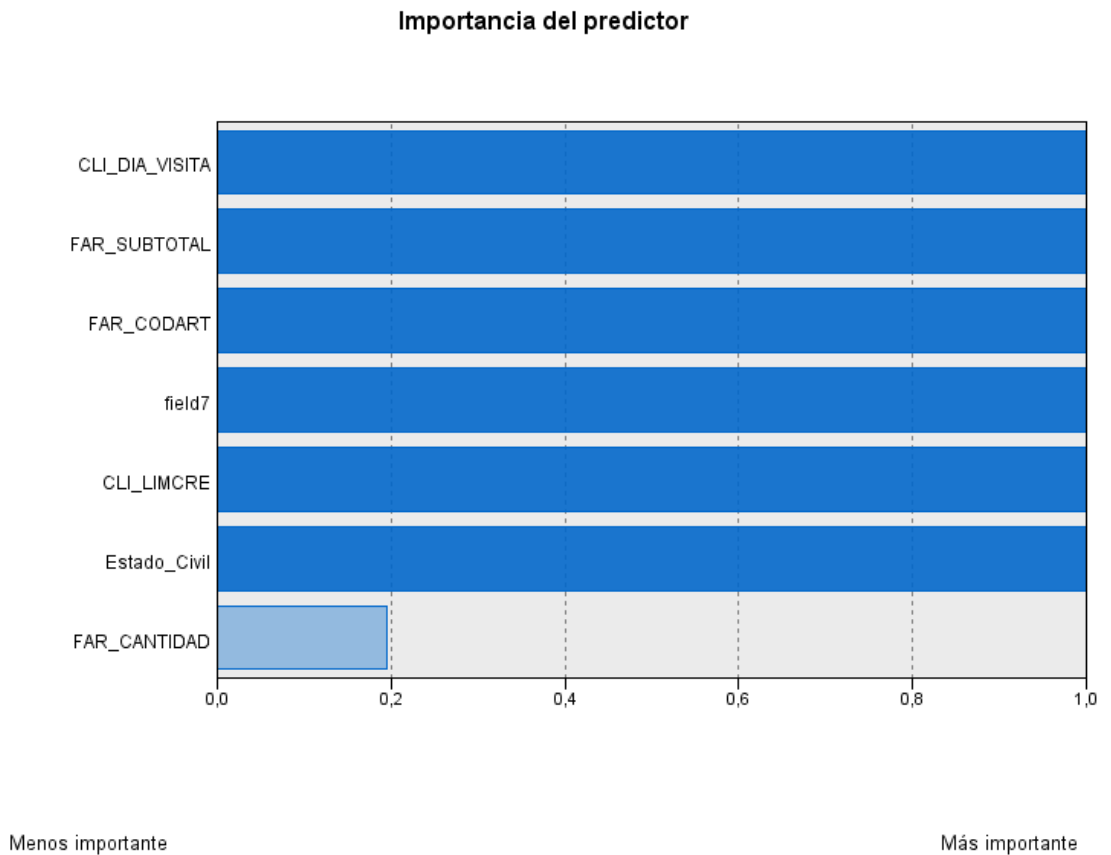
También se pueden apreciar la información de la distribución absoluta de las entradas con respecto a cada clúster.

Clústeres

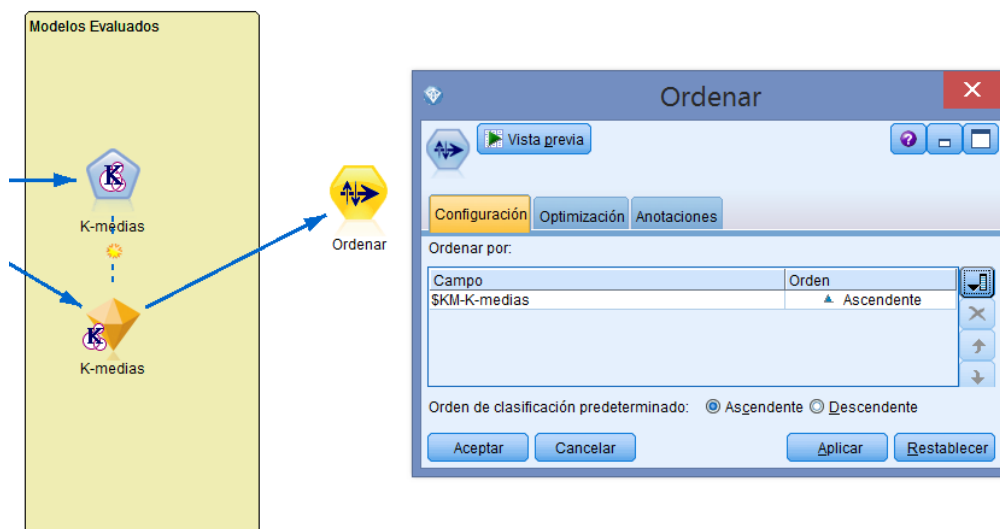
Importancia de entrada (predictor)
 ■ 1,0 ■ 0,8 ■ 0,6 ■ 0,4 ■ 0,2 ■ 0,0



Se puede apreciar en el visor también la importancia del predictor:



Paso 5: Luego se Agrega un nodo para ordenar la información resultante del nugget de k-medias para ser luego exportados a un archivo de Excel.



Presentación preliminar desde nodo Ordenar (15 campos, 10 registros) #1

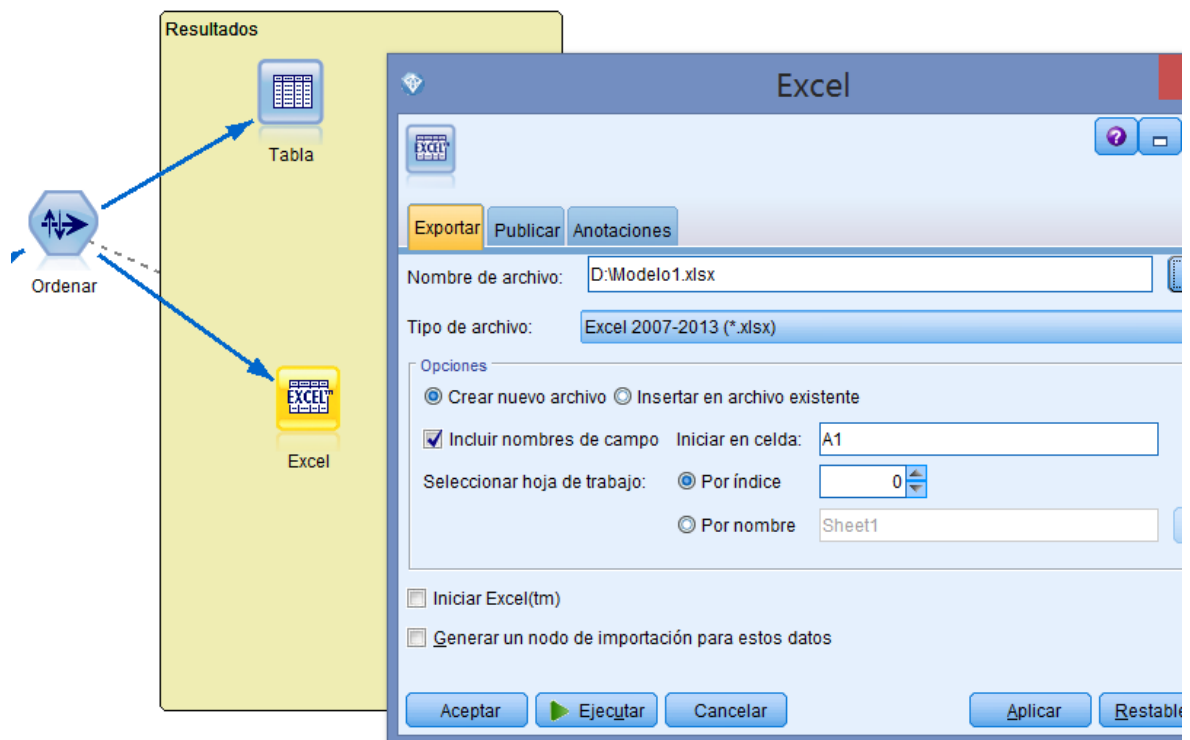
Archivo Editar Generar

Tabla Anotaciones

	Nombre_Cliente	Estado_Civil	Empresa	Limite_Credito	Zona	distrito	Provincia	FAR_FECHA	FAR_CODART	Nombre_Articulo	D
1	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-06-09 00:00:00	210.128	CLOTRIMAZOL 500MG X 50 TAB VAGINALES	C
2	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-26 00:00:00	314.919	PROMIZEM 550 MG X 30 TAB	C
3	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-26 00:00:00	294.433	DOLOL 500 MG X 100 TAB	C
4	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-15 00:00:00	135.063	ENJOY LOVE RETARDANTE SOB X 3UNID	SI
5	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-15 00:00:00	108.926	NOVOXACIL 500MG X 100 COMP.	C
6	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-15 00:00:00	45.170	DICLOFENACO 75 MG/3ML X 10 AMP.	UI
7	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-15 00:00:00	41.356	TAPSIN NOCHE LIMONADA X 60 SOBR	UI
8	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-05-15 00:00:00	41.142	TAPSIN DIA LIMONADA X 60 SOBR	UI
9	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-04-07 00:00:00	135.063	ENJOY LOVE RETARDANTE SOB X 3UNID	SI
10	ZEGARRA JARA JOSE FERNANDO	C	FARMACIA CONFIABLE	200.000	PORVENIR-LAREDO-BOSQUE	TRUJ...	LA LIBERTAD	2014-04-03 00:00:00	135.063	ENJOY LOVE RETARDANTE SOB X 3UNID	SI

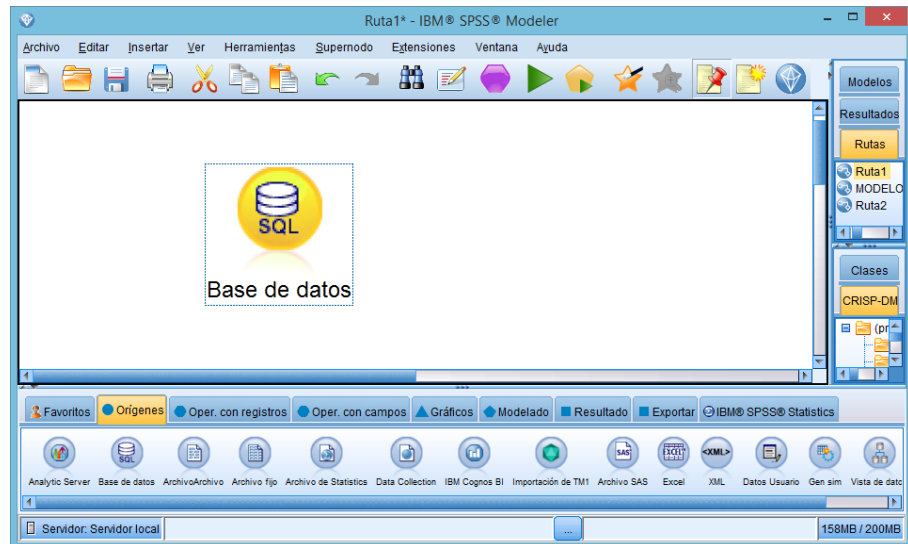
Aceptar

Paso 6: los datos son llevados a Excel para su posterior análisis y uso.

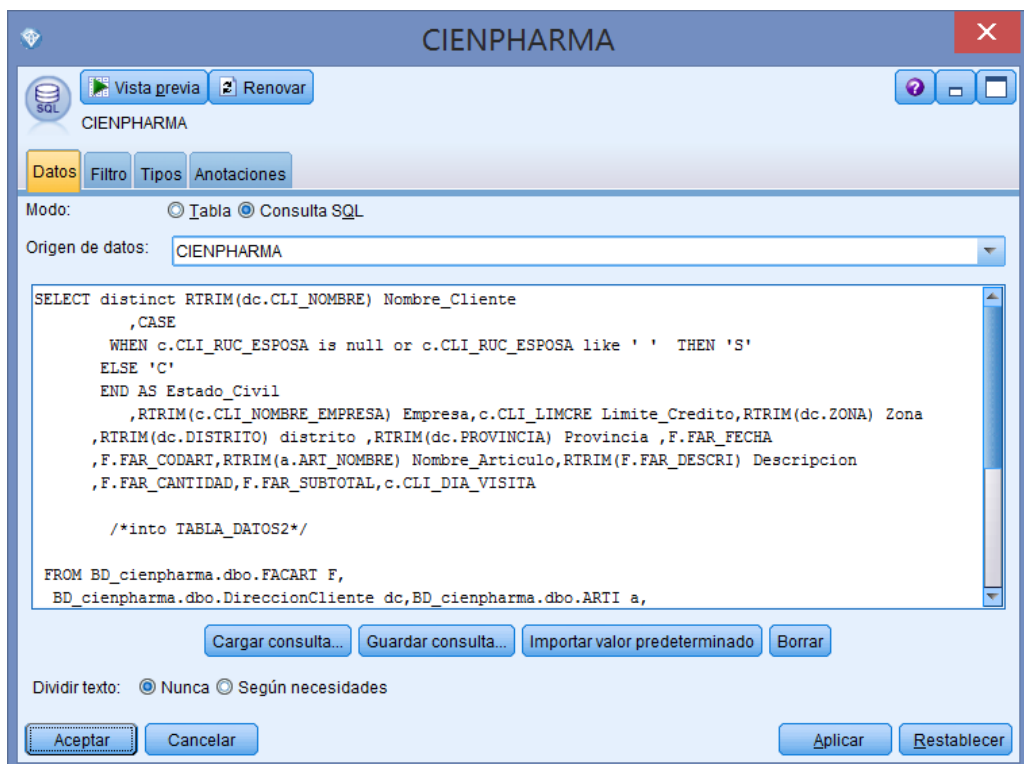


b. MODELO BASADO EN NODO KOHONEN

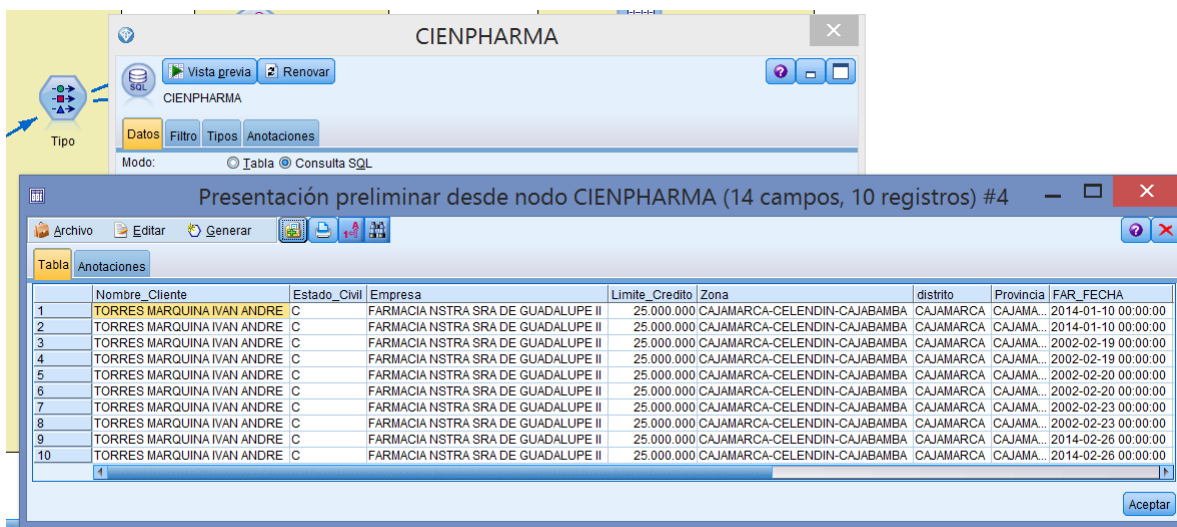
Paso 1: Igual como en la técnica anterior se abre el programa IBM SPSS MODELER y se procede a establecer la conexión con el origen de los datos.



Luego se procede a la selección de la consulta previamente preparada de los datos de los clientes:



Donde se puede verificar los datos dando clic a la vista previa de los datos:

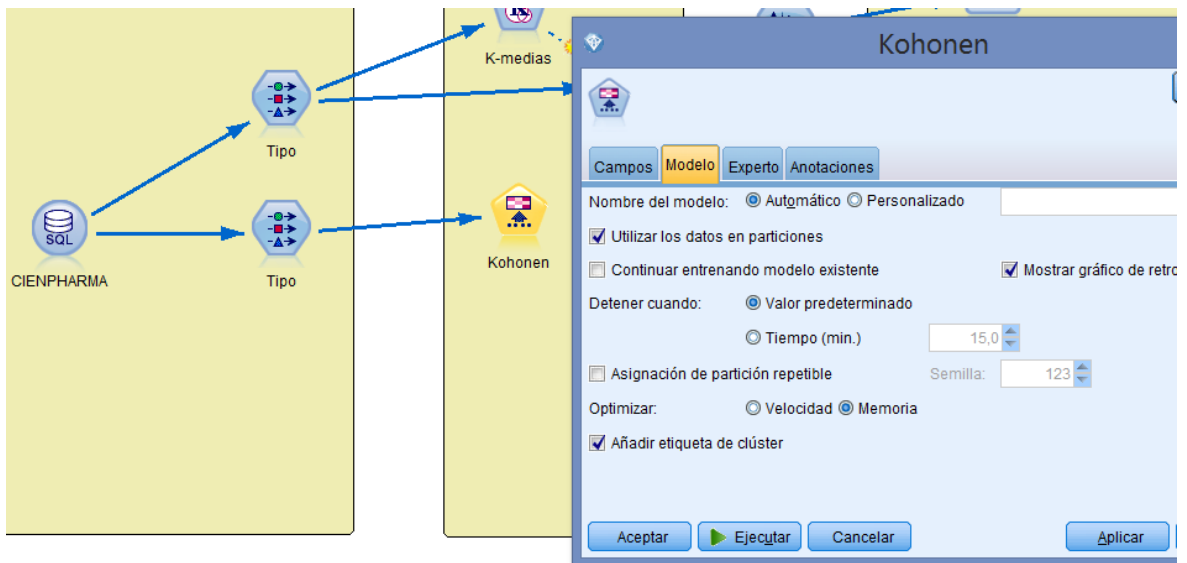


Paso 2: Se seleccionó un nodo “Tipo” de “Operaciones con campos”, para determinar y controlar los metadatos de los campos



Luego en esta configuración se da clic sobre el botón de “Leer valores” para leer los valores de los campos seleccionados cambiando la medida de acuerdo al valor del campo.

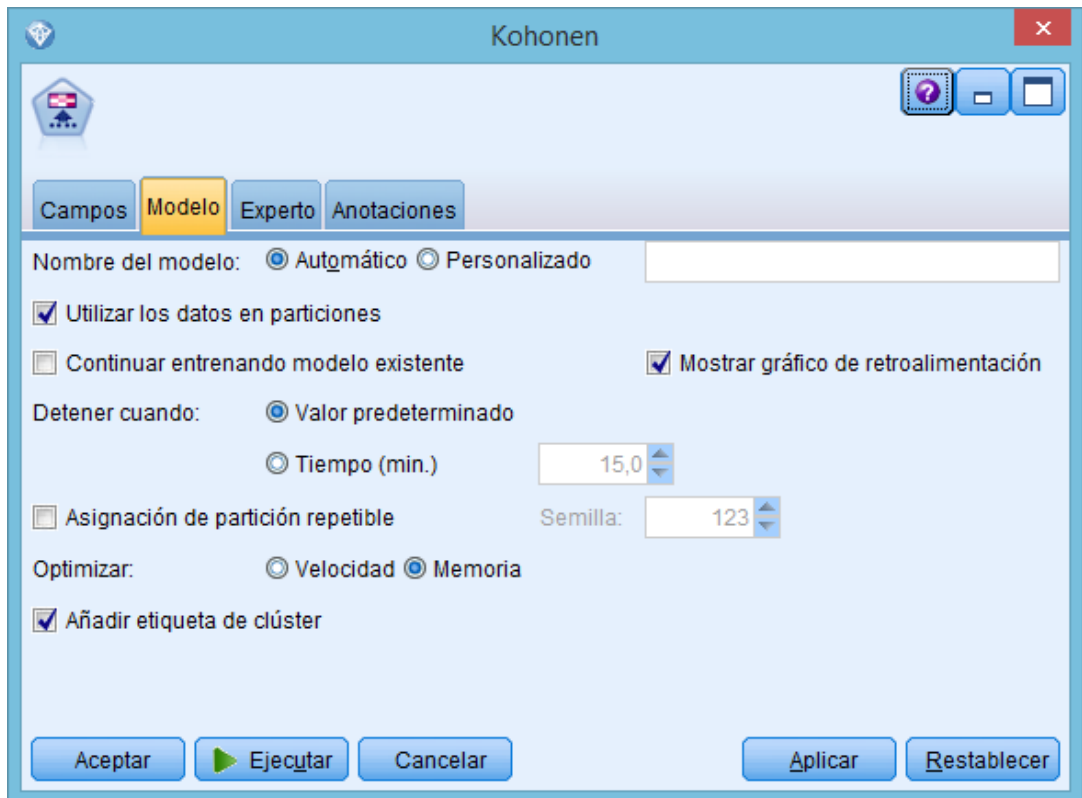
Paso 3: Se agrega al modelo el nodo del Modelo Kohonen



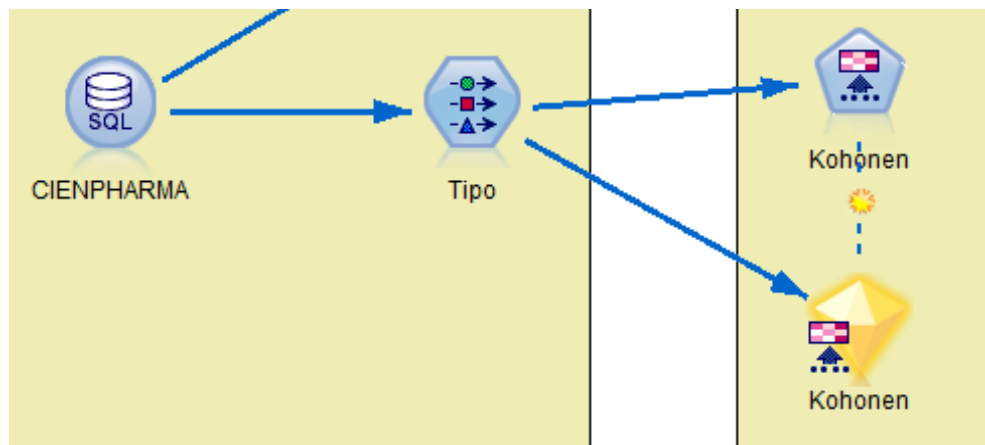
Paso 4: Luego se configura el modelo Kohonen con los campos de entrada la ser evaluados.



Figura 14: Ejecución del Modelo Kohonen



Para finalmente “Ejecutar” el modelo y generar el Nuget.



Visualizador de clusters resultante del modelo:

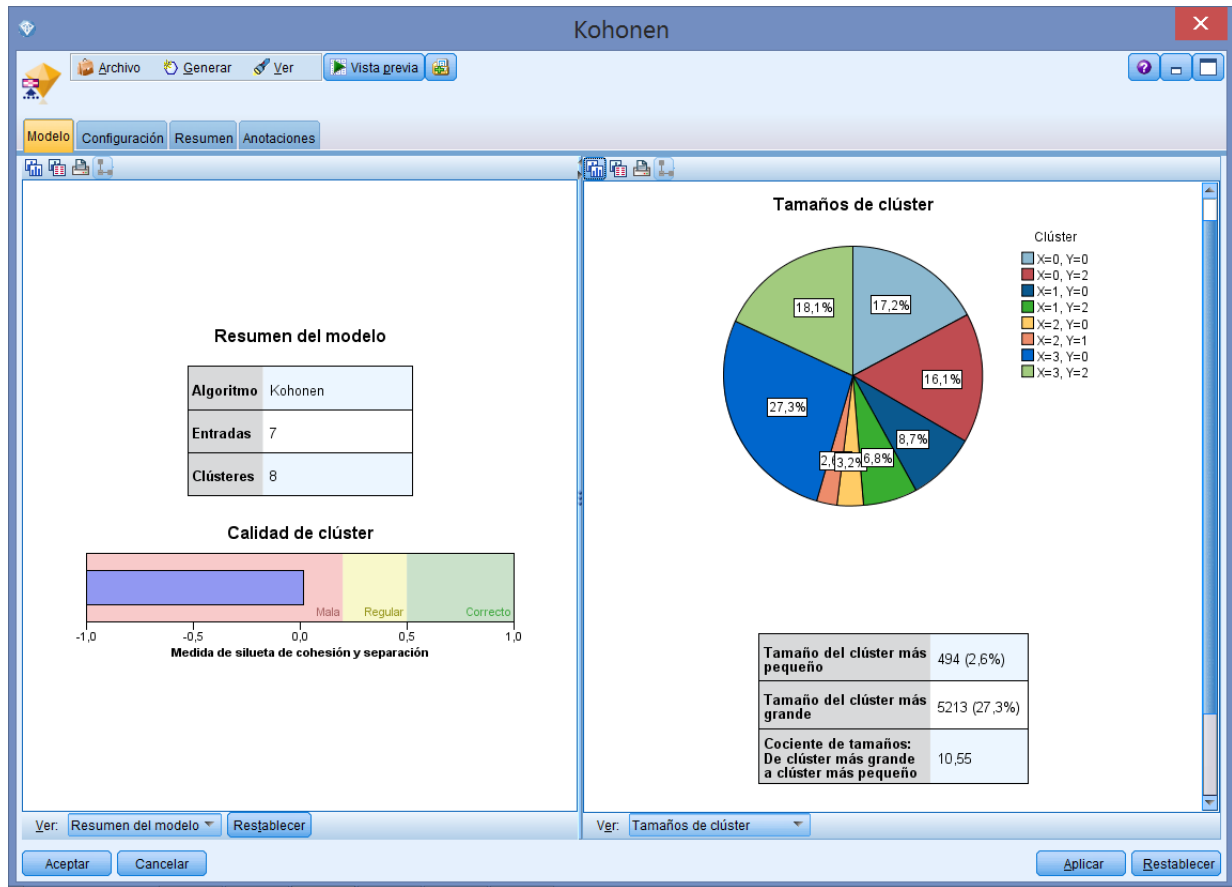
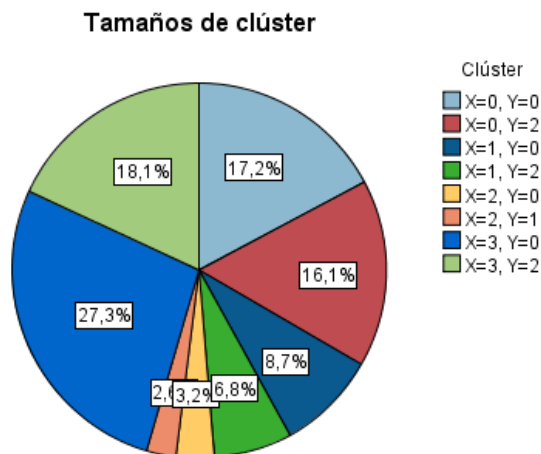


Figura 15: Resultados del Modelo Kohonen

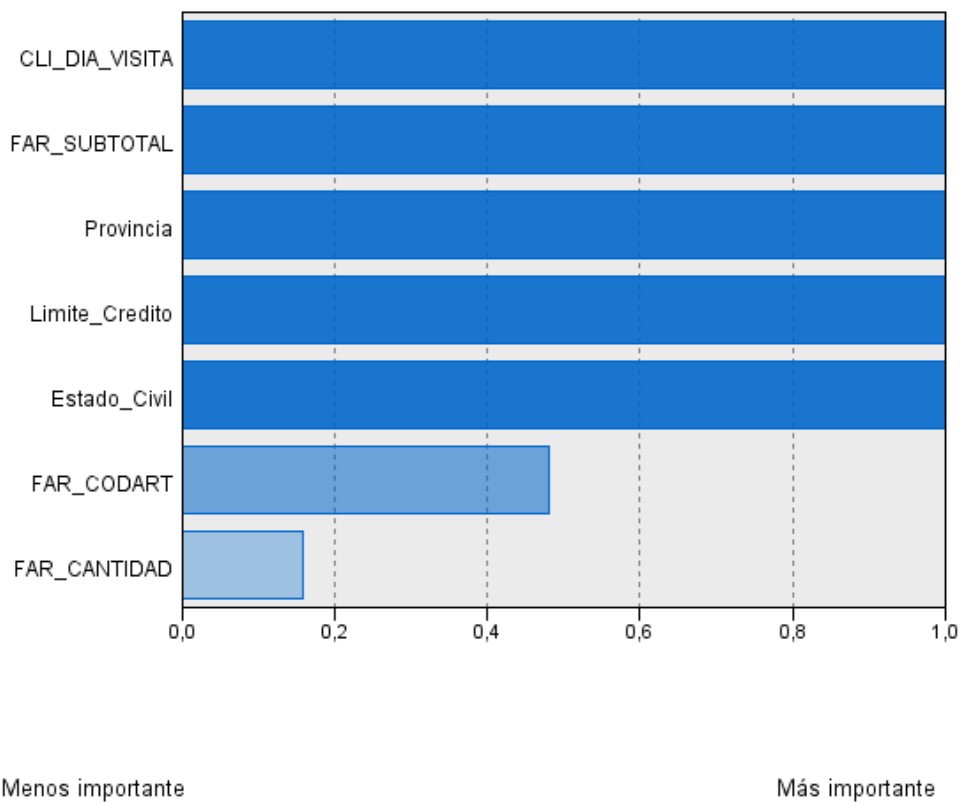
El modelo Kohonen generó 8 clústeres mostrando una visualización gráfica de estadísticas y distribuciones entre clústeres



Tamaño del clúster más pequeño	494 (2,6%)
Tamaño del clúster más grande	5213 (27,3%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	10,55

Siendo el cluster X=3, Y=0 el que agrupa una mayor cantidad de clientes con una cantidad de 5213 personas con una mayor importancia de entrada.

Importancia del predictor



En este cuadro se aprecia las entradas con respecto a los clústeres y su importancia de entrada (predictor).

Clústeres

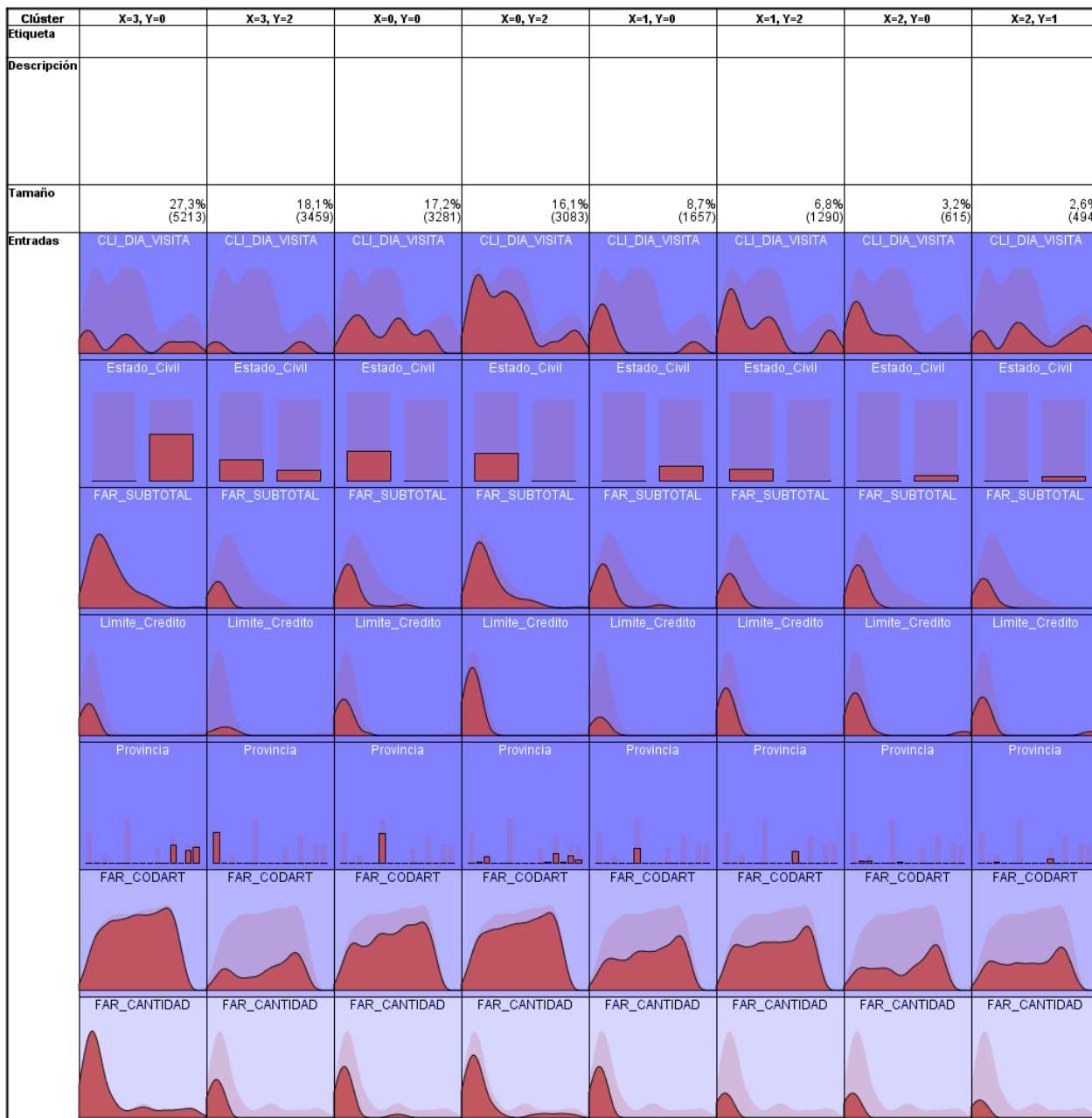
Importancia de entrada (predictor)
 ■ 1,0 ■ 0,8 ■ 0,6 ■ 0,4 ■ 0,2 ■ 0,0

Clúster	X=3, Y=0	X=3, Y=2	X=0, Y=0	X=0, Y=2	X=1, Y=0	X=1, Y=2	X=2, Y=0	X=2, Y=1
Etiqueta								
Descripción								
Tamaño	27,3% (5213)	18,1% (3459)	17,2% (3281)	16,1% (3083)	8,7% (1657)	6,8% (1290)	3,2% (615)	2,6% (494)
Entradas								
	CLI_DIA_VISITA 1,69	CLI_DIA_VISITA 16,00	CLI_DIA_VISITA 20,66	CLI_DIA_VISITA 13,81	CLI_DIA_VISITA 3,37	CLI_DIA_VISITA 14,13	CLI_DIA_VISITA 4,90	CLI_DIA_VISITA 22,88
	Estado_Civil S (100,0%)	Estado_Civil C (66,7%)	Estado_Civil C (100,0%)	Estado_Civil C (100,0%)	Estado_Civil S (100,0%)	Estado_Civil C (100,0%)	Estado_Civil S (100,0%)	Estado_Civil S (100,0%)
	FAR_SUBTOTAL 610,24	FAR_SUBTOTAL 103,38	FAR_SUBTOTAL 135,43	FAR_SUBTOTAL 330,28	FAR_SUBTOTAL 180,12	FAR_SUBTOTAL 107,47	FAR_SUBTOTAL 196,28	FAR_SUBTOTAL 141,85
	Limite_Credito 1.082,51	Limite_Credito 13.333,33	Limite_Credito 9.369,86	Limite_Credito 4.049,97	Limite_Credito 10.796,02	Limite_Credito 1.903,26	Limite_Credito 9.101,29	Limite_Credito 3.183,40
	Provincia LIMA (38,5%)	Provincia	Provincia	Provincia LIMA (34,1%)	Provincia	Provincia	Provincia AMAZONAS (43,7%)	Provincia
	FAR_CODART 180.398,01	FAR_CODART 218.714,95	FAR_CODART 199.891,55	FAR_CODART 163.284,48	FAR_CODART 204.328,11	FAR_CODART 150.505,18	FAR_CODART 177.211,07	FAR_CODART 153.288,45
	FAR_CANTIDAD 221,12	FAR_CANTIDAD 15,33	FAR_CANTIDAD 21,87	FAR_CANTIDAD 64,73	FAR_CANTIDAD 25,35	FAR_CANTIDAD 8,87	FAR_CANTIDAD 17,99	FAR_CANTIDAD 10,32

También se pueden apreciar la información de la distribución absoluta de las entradas con respecto a cada clúster.

Clústeres

Importancia de entrada (predictor)
 ■ 1,0 ■ 0,8 ■ 0,6 ■ 0,4 ■ 0,2 ■ 0,0



En esta tabla muestra nombres/etiquetas de características y distribuciones absolutas de las características de cada clúster. En el caso de las características categóricas, la visualización muestra gráficos de barras superpuestas con las categorías ordenadas en orden ascendente de valores de datos. En las características continuas, la

visualización muestra un gráfico de densidad suave que utiliza los mismos puntos finales e intervalos para cada clúster.

La visualización en color rojo oscuro muestra la distribución de clústeres, mientras que la más clara representa los datos generales.

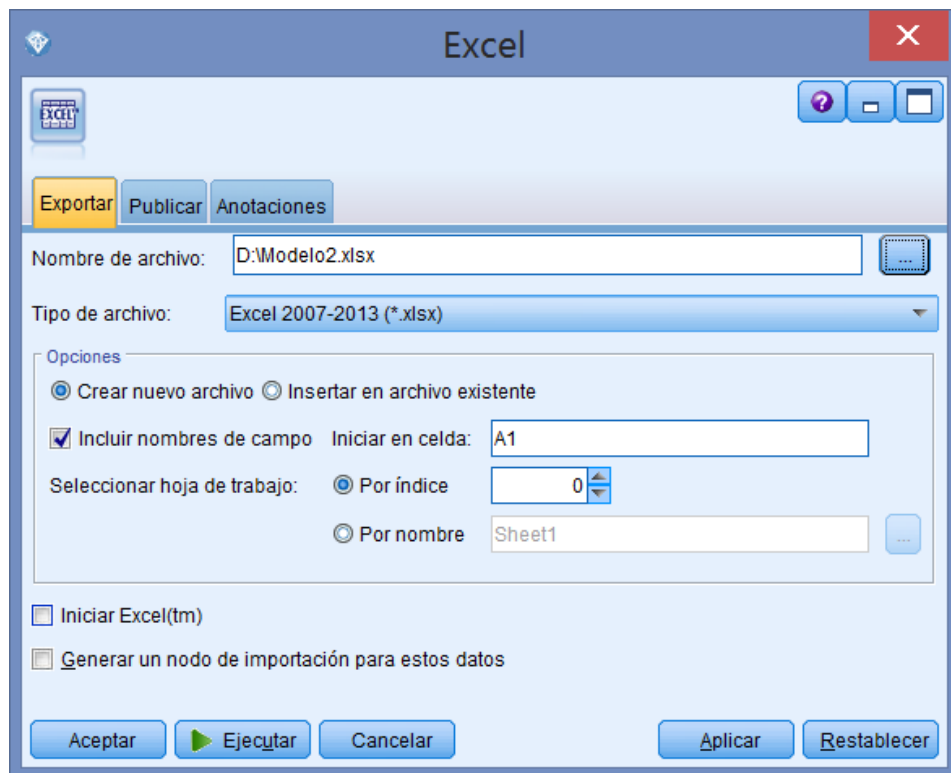
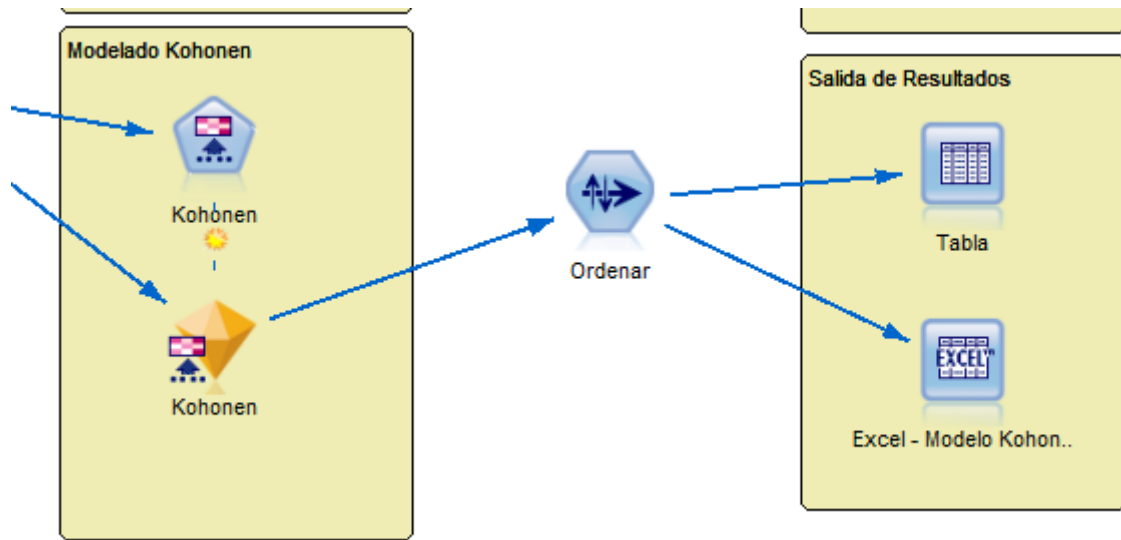
Paso 5: Luego se Agrega un nodo para ordenar la información resultante del nuget de Kohonen para ser luego exportados a un archivo de Excel.



Presentación preliminar desde nodo Ordenar (17 campos, 10 registros)

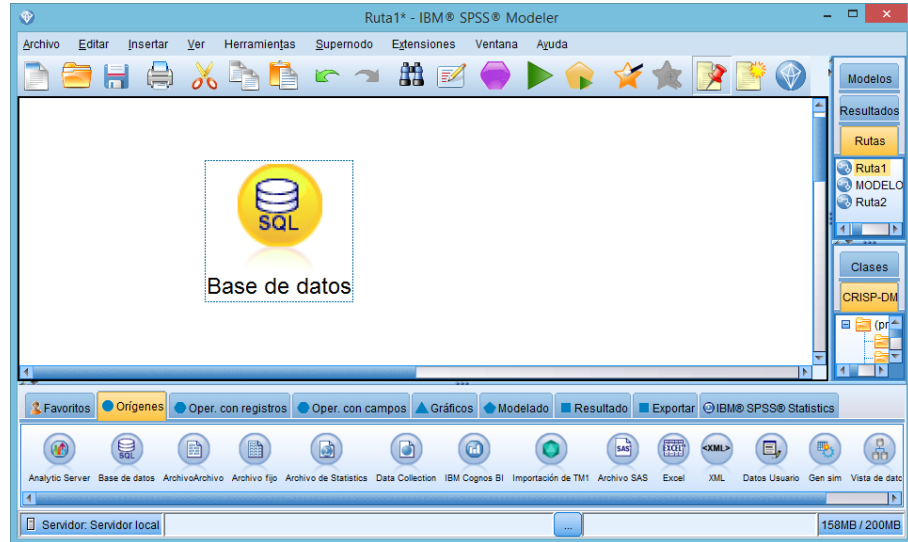
	Provincia	FAR_FECHA	FAR_CODART	Nombre_Articulo	Descripcion	FAR_CANTIDAD	FAR_SUBTOTAL	CLI_DIA_VISITA	\$KX-Kohonen	\$KY-Kohonen	\$KXY-Kohonen
1	CAJAMA...	2014-01-04 00:00:00	316.498	HEPAOXICOHOL 100 MG X 100 CAP	CAJA	2.000	110.860	1	0	0	X=0, Y=0
2	CAJAMA...	2014-02-14 00:00:00	280.826	FARMADOL 500 MG X 100 CAP BLANDAS	CJA	2.000	79.990	46	0	0	X=0, Y=0
3	CAJAMA...	2014-02-14 00:00:00	280.826	FARMADOL 500 MG X 100 CAP BLANDAS	CJA	1.000	40.000	46	0	0	X=0, Y=0
4	CAJAMA...	2014-02-14 00:00:00	263.512	ALERDIF 5MG/5ML JBE X 60ML	CJA	11.000	132.000	46	0	0	X=0, Y=0
5	CAJAMA...	2014-02-14 00:00:00	261.598	OMEGRAN-20 X 10 SACHETS	CJA	1.000	12.000	46	0	0	X=0, Y=0
6	CAJAMA...	2014-02-14 00:00:00	251.270	MARYLIN 1.5 X 1TAB	CJA	20.000	74.010	46	0	0	X=0, Y=0
7	CAJAMA...	2014-02-14 00:00:00	238.205	CEPRALER 10 MG X 100 TAB	CJA	1.000	19.090	46	0	0	X=0, Y=0
8	CAJAMA...	2014-02-14 00:00:00	238.091	PREXANDOL 550MG X 100 TAB	CJA	1.000	29.000	46	0	0	X=0, Y=0
9	CAJAMA...	2014-02-14 00:00:00	235.498	COLON FIBER 3.5GR X 30 SOB	CJA	4.000	124.780	46	0	0	X=0, Y=0
10	CAJAMA...	2014-02-14 00:00:00	235.498	COLON FIBER 3.5GR X 30 SOB	CJA	3.000	93.580	46	0	0	X=0, Y=0

Paso 6: los datos son llevados a Excel para su posterior análisis y uso.

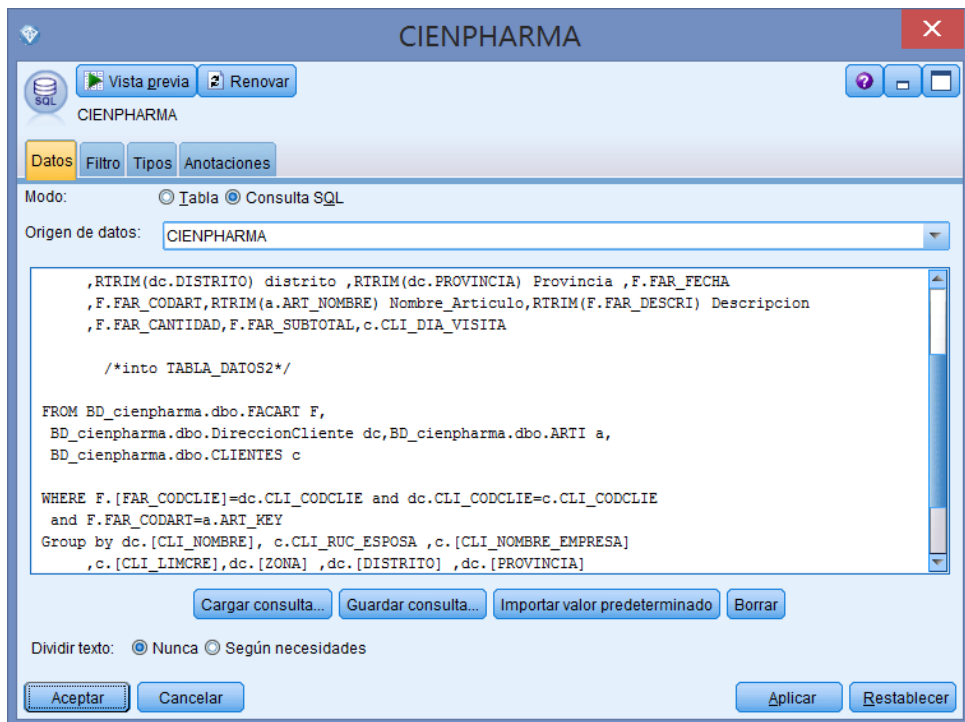


c. MODELO BASADO EN ARBOLES DE DECISION

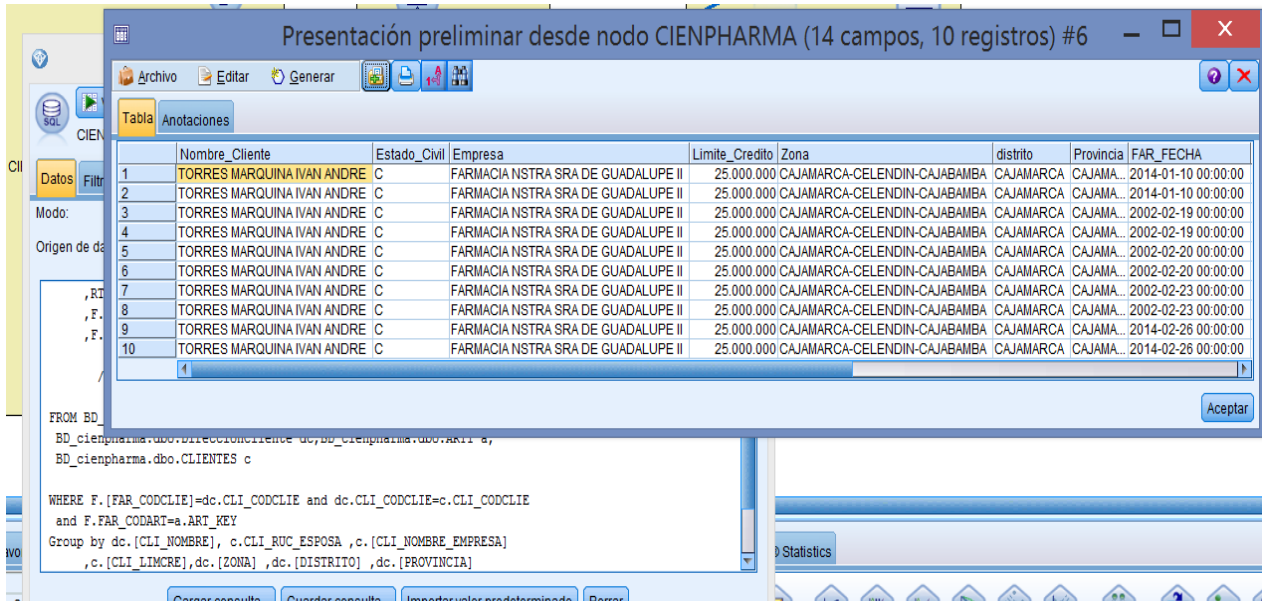
Paso 1: Abrimos el programa IBM SPSS MODELER y se procedió a establecer la conexión con el origen de los datos.



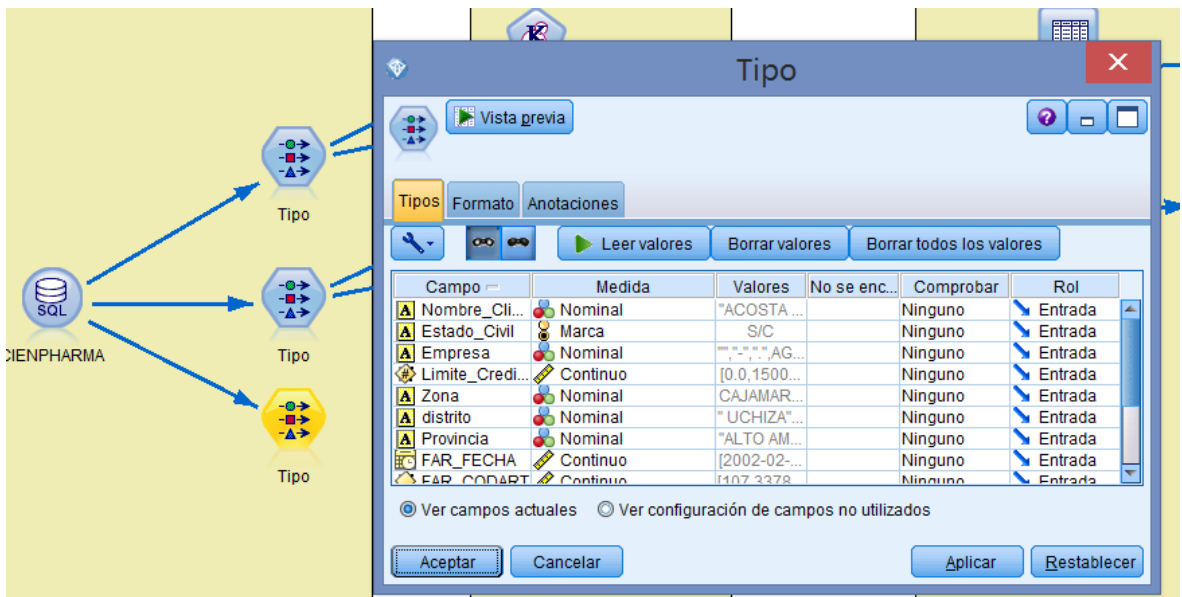
Procediendo a la selección de la tabla previamente preparada de los datos de los clientes:



Donde se puede verificar los datos dando clic a la vista previa de los datos:



Paso 2: Se seleccionó un nodo “Tipo” de “Operaciones con campos”, para determinar y controlar los metadatos de los campos



Luego en esta configuración se da clic sobre el botón de “Leer valores” para leer los valores de los campos seleccionados cambiando la medida de acuerdo al valor del campo.

Paso 3: Se agrega al modelo el nodo del Árbol

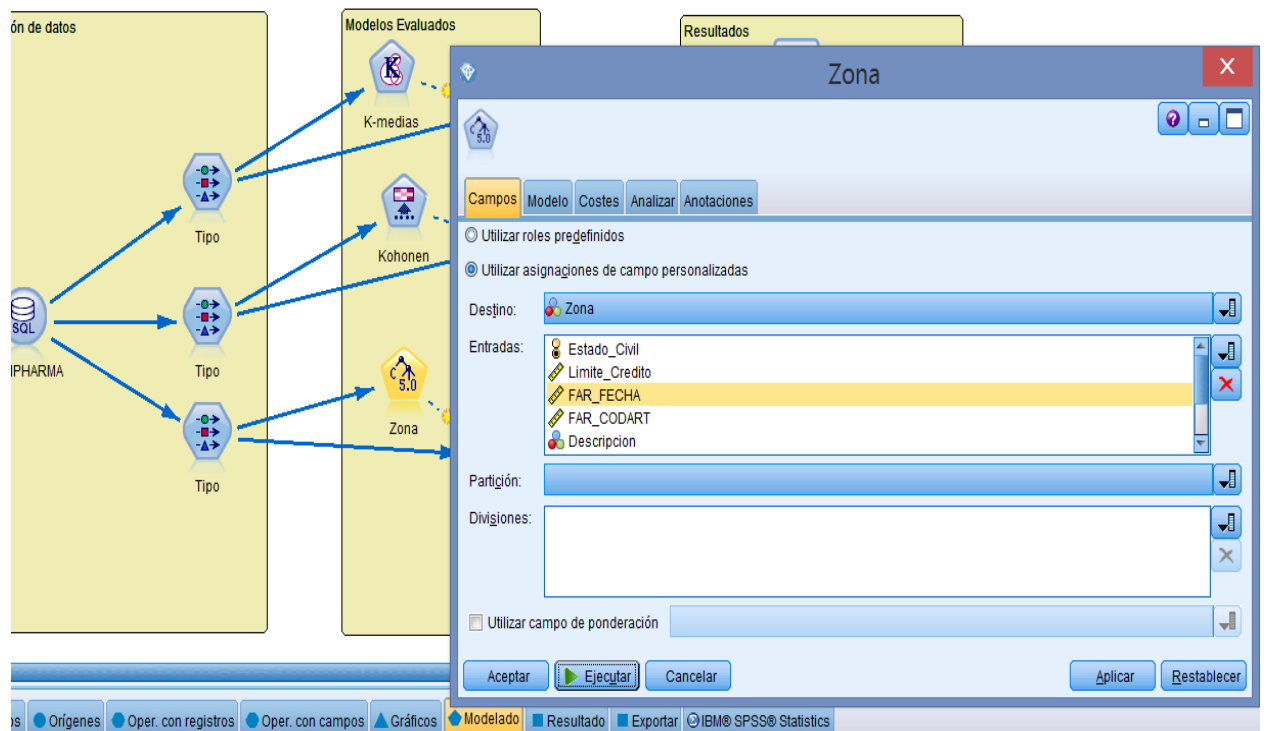
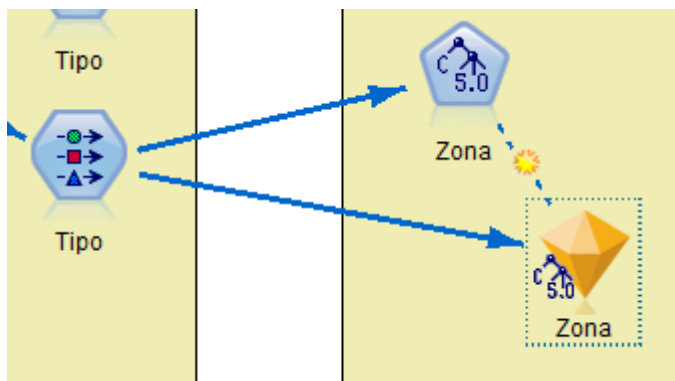


Figura 17: Ejecución del Modelo Árbol

Paso 4: Se genera el Nugget del modelo mostrando los resultados encontrados.



Verificamos reglas del árbol implementadas en el diamante obtenido. Para verlas todas pulsamos en “Todas”, con lo que tendremos ya todas las reglas. Además en el % , con lo que veremos el número de elementos que caen por cada regla y el

porcentaje de aciertos. Ahora tienes el árbol etiquetado como se muestra a continuación

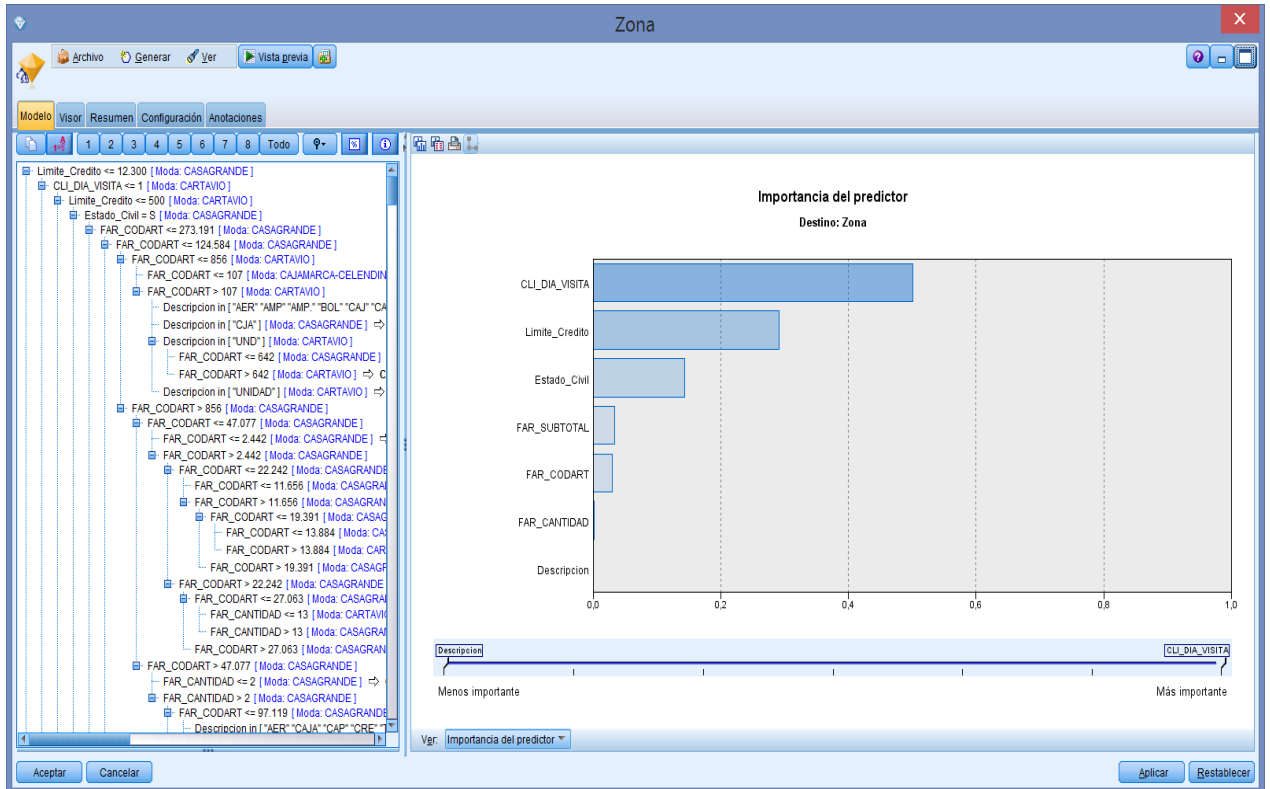


Figura 18: Resultados del Modelo de árbol

El modelo generó un árbol mostrando una visualización gráfica de acuerdo a la importancia del predictor

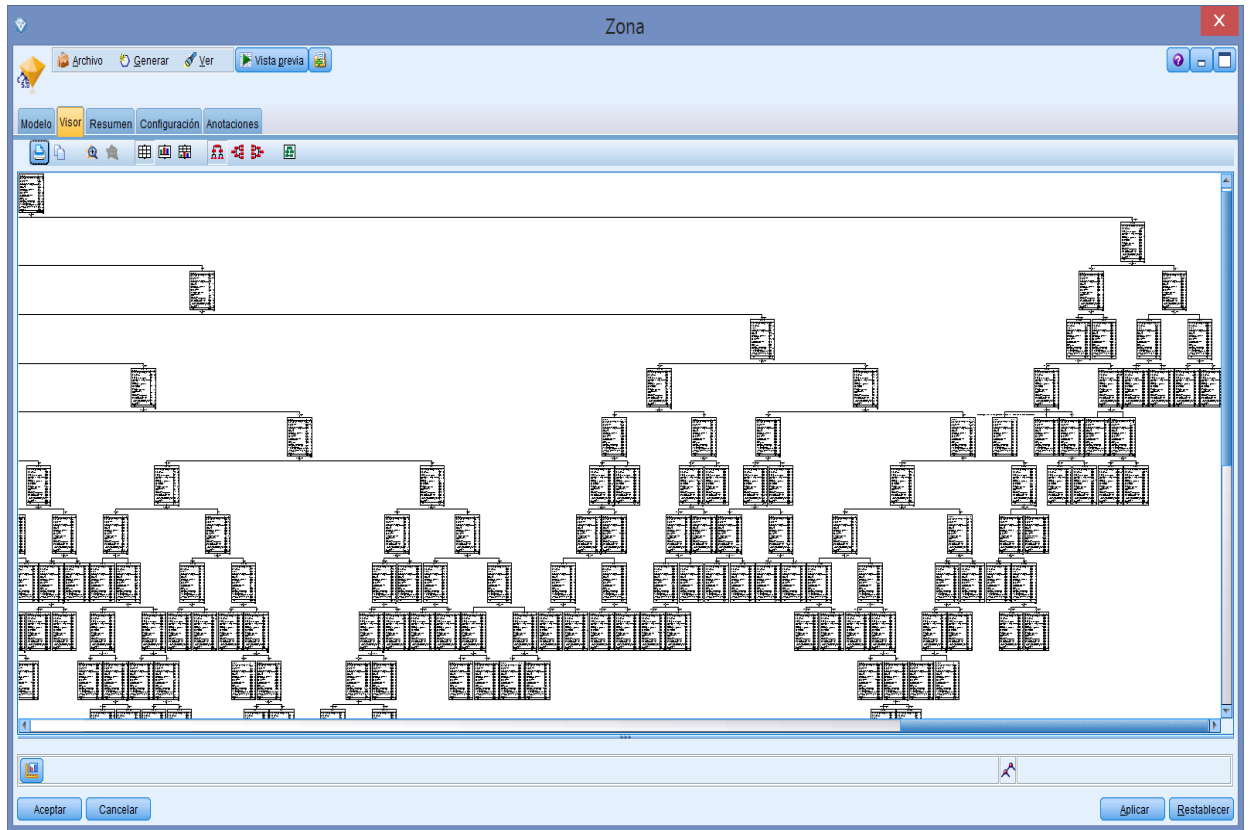
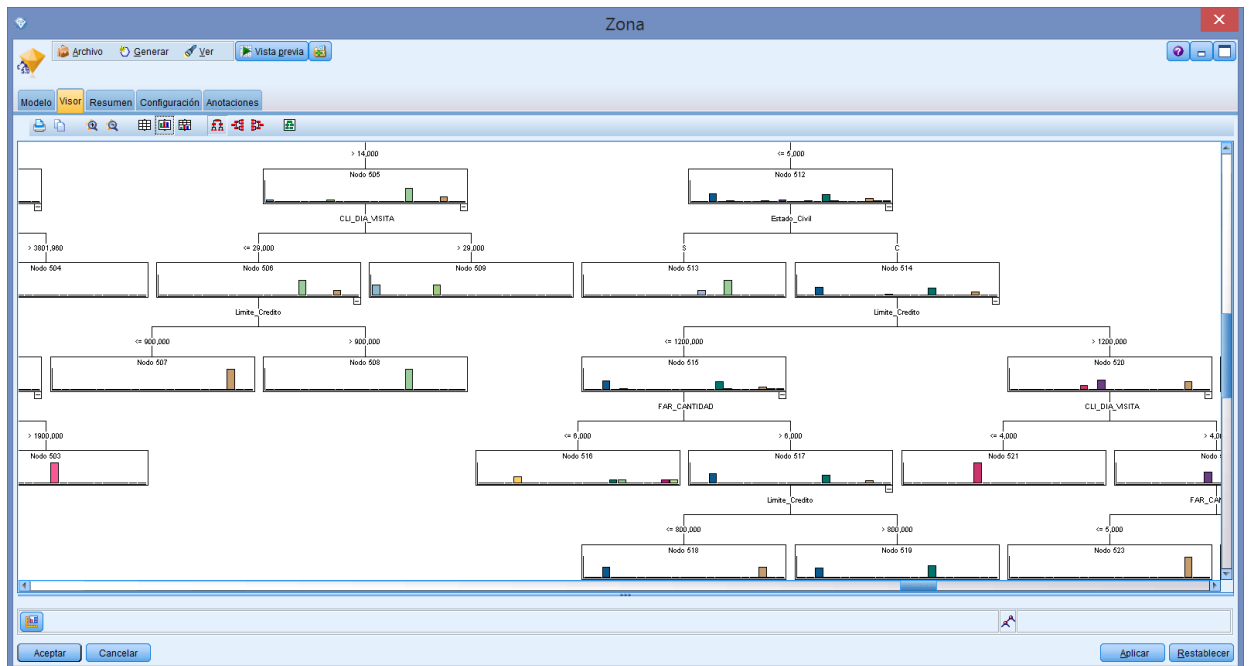


Figura 19: Árbol de decisión: Modelo Árbol C5.0



4.4.EVALUACION DE LOS MODELOS:

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

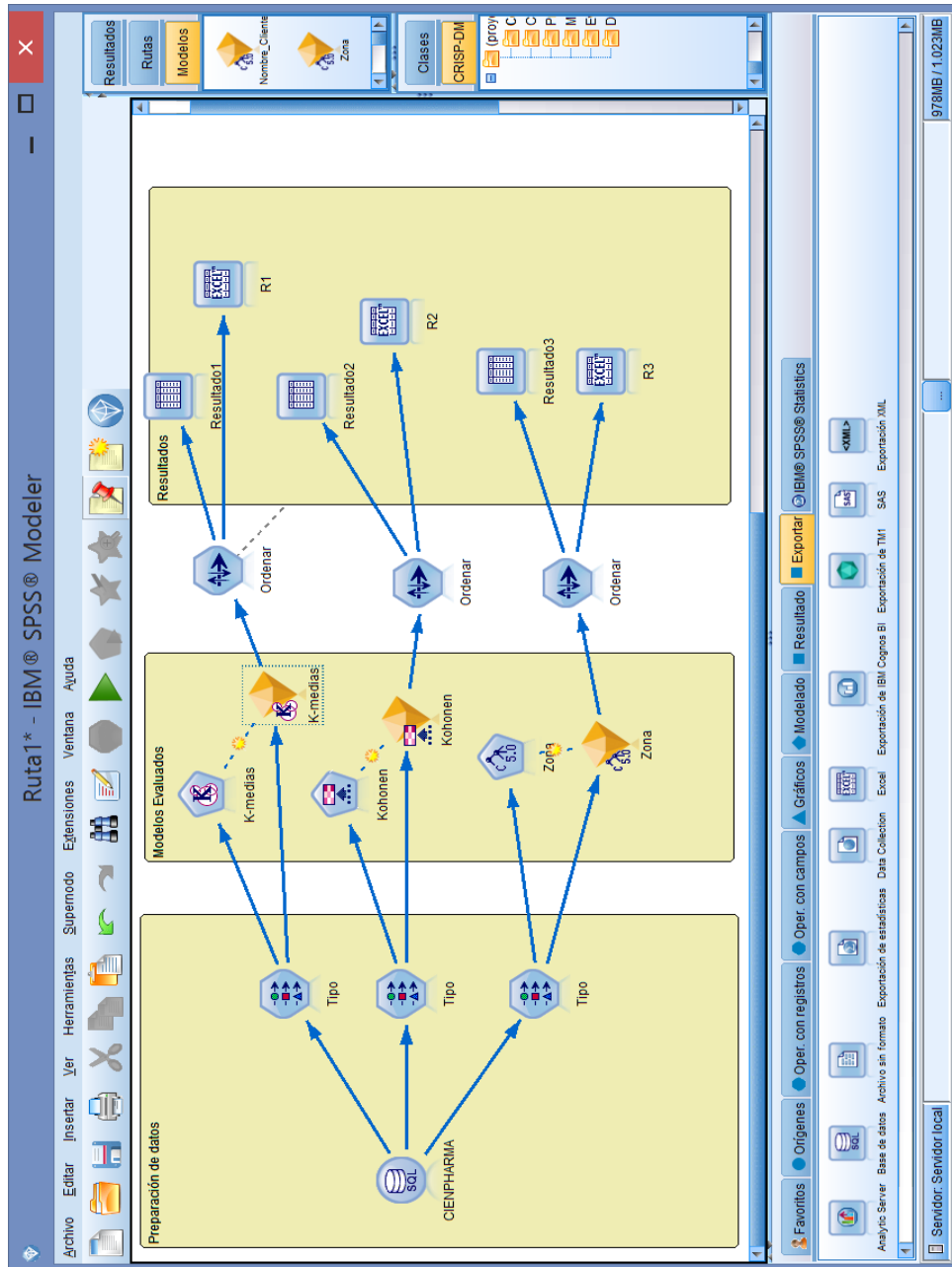


Figura 20: Modelos de minería de datos construidos

En esta tarea, se interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc...).

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

a. Modelo K-Medias:

K-medias empieza definiendo un conjunto de centros de clústeres iniciales derivados de datos. Después asigna cada registro al clúster de registros más similares, basándose en los valores de los campos de entrada de registros. Una vez asignados todos los casos, los centros de clústeres se actualizan para reflejar el nuevo conjunto de registros asignados a cada clúster. Los registros se vuelven a comprobar para ver si se deben reasignar a otro clúster, y el proceso de iteración de clúster/asignación continúa hasta que se alcanza el número máximo de iteraciones o el cambio entre una iteración y otra no sobrepasa el umbral especificado.

El modelo resultante depende, hasta cierto punto, del orden de los datos de entrenamiento. Reordenar los datos y regenerar el modelo puede dar como resultado un modelo de clústeres final distinto.

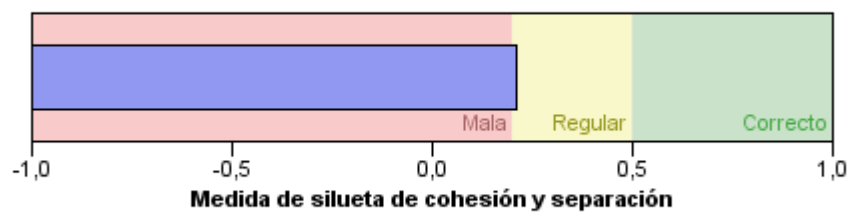
Para entrenar el modelo K-Means se necesita uno o más campos con su rol establecido como Entrada. Se ignorarán los campos con el rol establecido como Resultado, Ambos o Ninguno.

No es necesario tener los datos en pertenencia a grupos para crear un modelo de K-medias. Este modelo suele ser el método más rápido de agrupación en clústeres para conjuntos de datos grandes.

Resumen del modelo

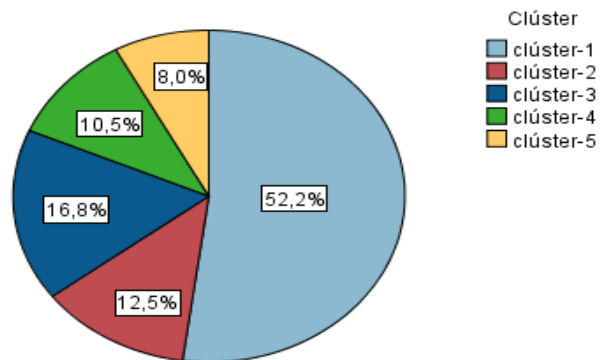
Algoritmo	K-medias
Entradas	7
Clústeres	5

Calidad de clúster



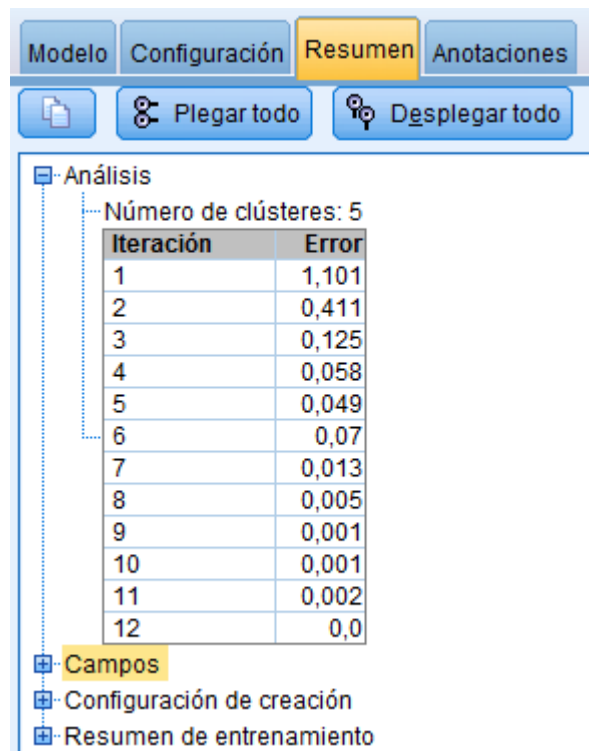
Este Modelo solo tiene 7 entradas, pero generan 5 clústeres. También nos muestra los tamaños mínimos (1526) y máximo (9960) de cada clúster

Tamaños de clúster



Tamaño del clúster más pequeño	1526 (8%)
Tamaño del clúster más grande	9960 (52,2%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	6,53

Se realizaron para este modelo 12 iteraciones disminuyendo el margen de error:



Iteración	Error
1	1,101
2	0,411
3	0,125
4	0,058
5	0,049
6	0,07
7	0,013
8	0,005
9	0,001
10	0,001
11	0,002
12	0,0

b. Modelo Kohonen:

Con este modelo se realizó agrupaciones en clústeres, también conocidas como knet o como un mapa autoorganizativo. Se utilizó para agrupar el conjunto de datos en grupos distintos. Los registros se agrupan de manera que los de un mismo grupo o clúster tiendan a ser similares entre ellos y que los de otros grupos sean distintos.

Las unidades básicas son neuronas y se organizan en dos capas: la capa de entrada y la capa de salida. Todas las neuronas de entrada están conectadas a todas las neuronas de salida, y estas conexiones tienen fuerzas o ponderaciones asociadas a ellas. Durante el entrenamiento, cada unidad compete con las demás para "ganar" cada registro. El mapa de resultados es una red de neuronas bidimensional sin conexiones entre las unidades.

Los datos de entrada se presentan en la capa de entrada y los valores se propagan a la capa de salida. La neurona de salida con la respuesta más fuerte se considera la ganadora y constituye la respuesta para dicha entrada.

Cuando la red se termina de entrenar, los registros que son similares se cierran juntos en el mapa de resultados, mientras que los registros que son muy diferentes aparecerían aparte.

Las redes de Kohonen no utilizan un campo objetivo. Este tipo de aprendizaje, sin campo objetivo, se denomina aprendizaje no supervisado. En lugar de intentar predecir un resultado, las redes de Kohonen intentan revelar los patrones en el conjunto de campos de entrada. Por lo general, una red de Kohonen termina con unas pocas unidades que resumen muchas observaciones (unidades fuertes) y varias unidades que no corresponden realmente con ninguna de las observaciones (unidades débiles).

Para entrenar a la red de Kohonen, se necesitó uno o más campos con su rol establecido como Entrada. Se ignorarán los campos con el rol establecido como Objetivo, Ambos o Ninguno.

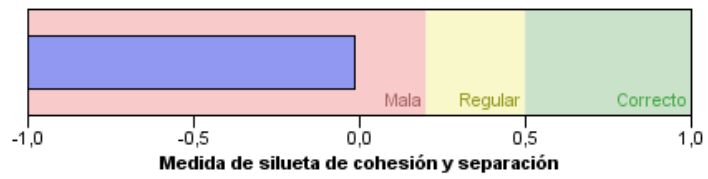
No es necesario tener los datos en pertenencia a grupos. Las redes de Kohonen comienzan con un número elevado de unidades y, según avanza el entrenamiento, las unidades se dejan atraer por los clústeres naturales de los datos.

Este Modelo tiene una mayor cantidad de entradas (7), generando 9 clústeres.

Resumen del modelo

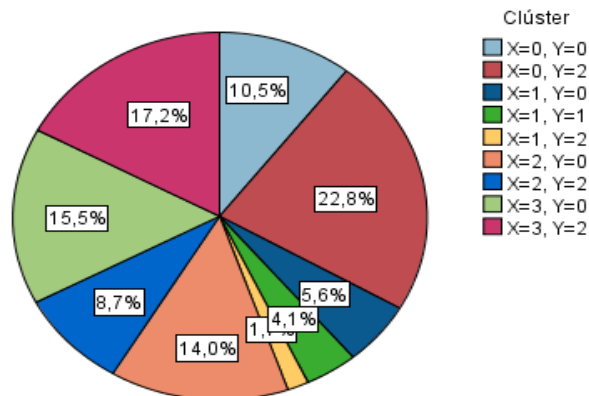
Algoritmo	Kohonen
Entradas	7
Clústeres	9

Calidad de clúster

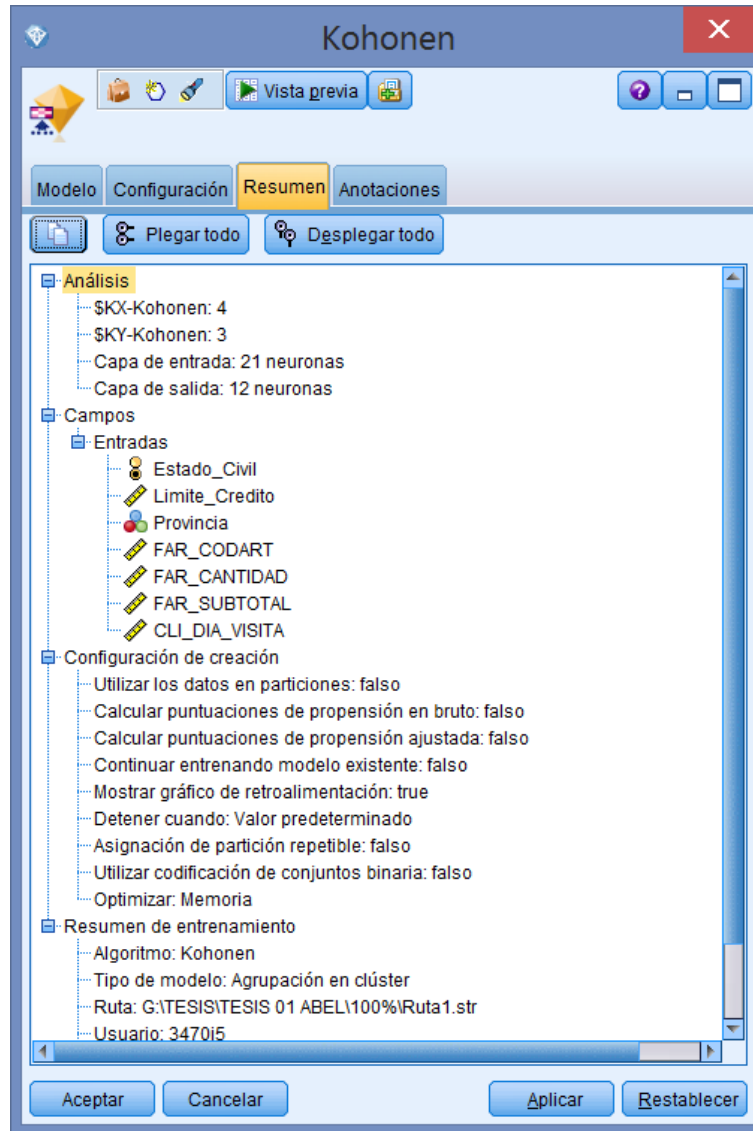


También nos muestra los tamaños mínimos (318) y máximo (4357) de cada clúster

Tamaños de clúster



Tamaño del clúster más pequeño	318 (1,7%)
Tamaño del clúster más grande	4357 (22,8%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	13,70

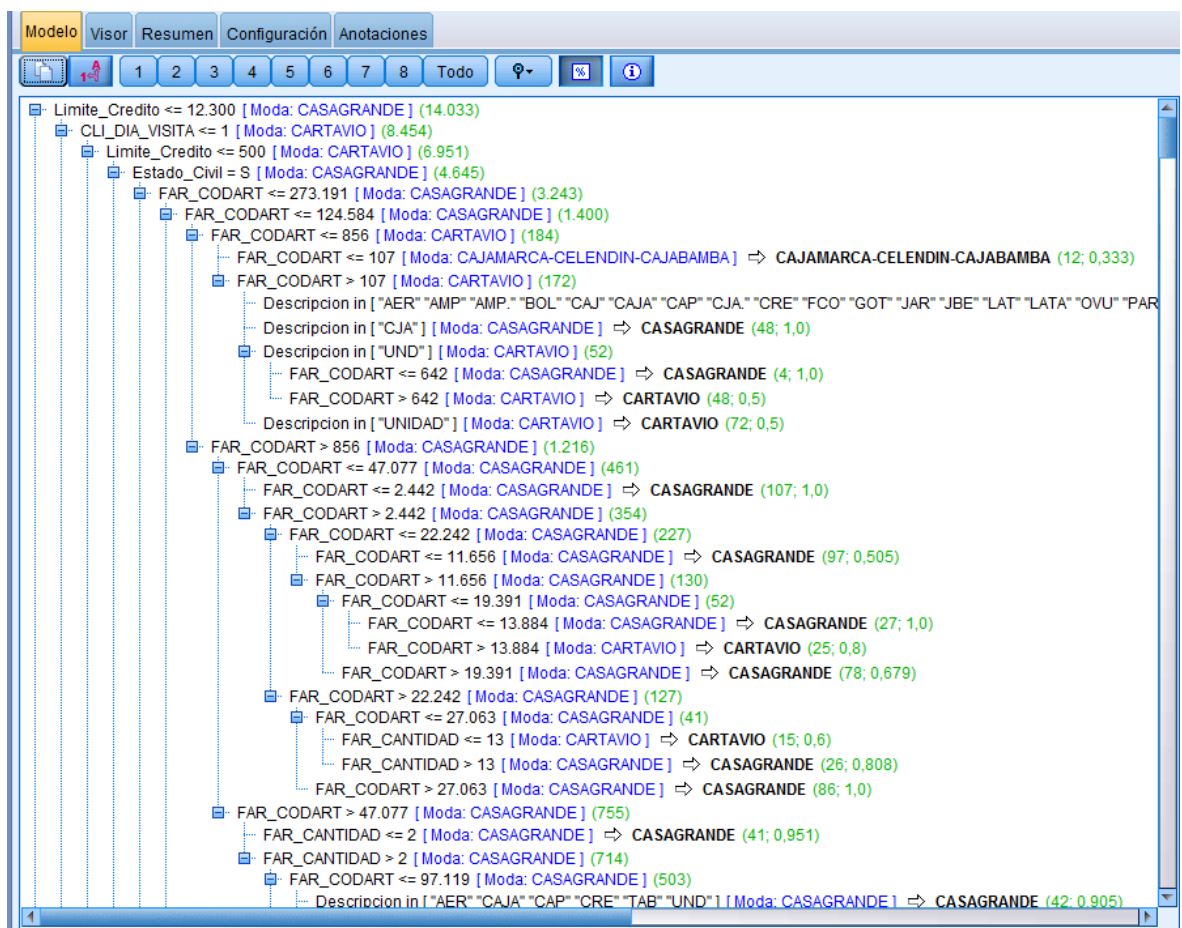


c. Modelo Árbol de decisión C5.0:

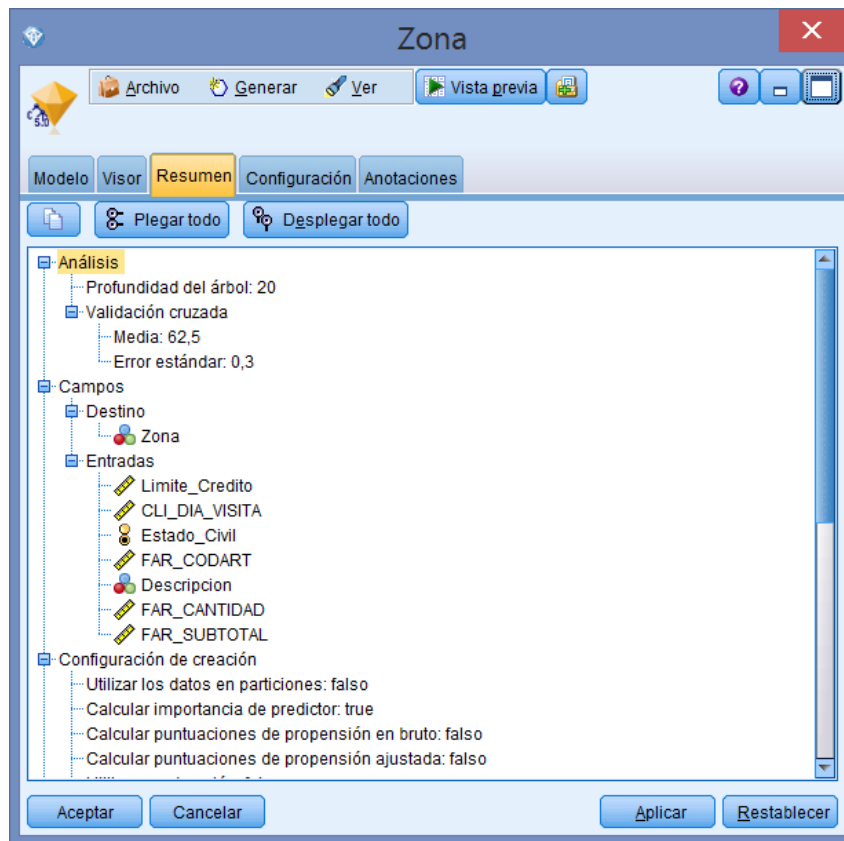
Se dividió la muestra en función del campo que ofrece la máxima ganancia de información. Las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las submuestras de nuevo. Por último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

C5.0 genera un conjunto de reglas que intenta realizar predicciones de registros individuales. Los conjuntos de reglas derivan de los árboles de decisión y, en cierto modo, representan una versión simplificada de la información que se incluye en estos árboles. Por lo general, los conjuntos de reglas pueden retener la mayor parte de la información significativa de un árbol de decisión completo, aunque utilizan un modelo menos complejo.

Para entrenar un modelo C5.0, debe existir un campo categórico (por ejemplo, nominal u ordinal) Objetivo y uno o más campos Entrada de cualquier tipo. Se ignorarán los campos establecidos en Ambos o Ninguno. Los tipos de los campos utilizados en el modelo deben estar completamente instanciados. También se puede especificar un campo de ponderación.

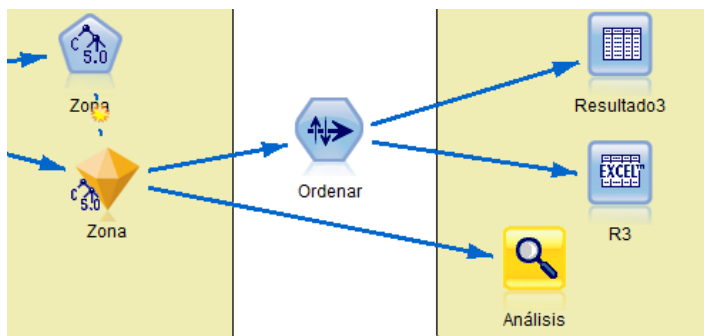


Este Modelo tiene una mayor cantidad de entradas.



La profundidad del árbol es de 20 y su validación cruzada media es de 62.5%

Finalmente agregamos un nodo de análisis al modelo para evaluar la capacidad del modelo para generar predicciones precisas. Los nodos Análisis realizan varias comparaciones entre los valores predichos y los valores (su campo de objetivo) reales para uno o más nuget de modelo. Los nodos Análisis también se pueden utilizar para comparar modelos predictivos con otros modelos predictivos.



Cuando ejecuta un nodo Análisis, se añade automáticamente un resumen de los resultados de análisis a la sección Análisis de la pestaña Resumen para cada nugget de modelo de la ruta ejecutada. Los resultados de análisis detallados aparecen en la pestaña Resultados de la ventana del gestor o se pueden escribir directamente en un archivo.

Archivo Editar

Análisis Anotaciones

Plegar todo Desplegar todo

Resultados para el campo de resultado Zona

Comparando \$C-Zona con Zona

Correctos	12.867	67,39%
Erróneos	6.225	32,61%
Total	19.092	

Evaluación del rendimiento

CAJAMARCA-CELENDIN-CAJABAMBA	0,959
CARTAVIO	0,643
CASAGRANDE	1,492
CASCAS	6,117
CASMA-SAN JACINTO-MORO	6,193
CHEPEN-GUADALUPE-SAN PEDRO	5,272
CHICLAYO	3,916
CHIMBOTE	3,881
COISHCO-SANTA-RINCONADA	5,85
CUTERVO-CHOTA	6,05
HUAMACHUCO	3,716
HUARAZ	4,462
JAEN-B.CHICA-B.GRANDE	4,257
NUEVO CHIMBOTE	4,24
OFICINA	7,911
PIURA-OTROS	4,883
PORVENIR-LAREDO-BOSQUE	4,217
RIOJA-MOYO-NARANJILLO-NVA	1,882
SANTIAGO DE CHUCO	5,945
TARAPOTO-JUANJUI-YURIMAGUAS	5,293
TRUJILLO-LA ESPERANZA-EL MILAG	3,198
VALLE CHICAMA	4,757
VIRU-CHAO	6,388

Aceptar

TABLA RESUMEN DE EVALUACION DE MODELOS

Modelo	Modelo K-medias	Modelo Kohonen
Criterios de Evaluación		
Número de Entradas	7	7
Numero de Clústeres	5	9
Tamaño de clúster más pequeño	1526	318
Tamaño de clúster más grande	9960	4357

Tabla 09: Resumen de Evaluación de Modelos

Modelo Árbol de decisión:

Resultados para el campo de resultado Zona

Comparando SC-Zona con Zona

Correctos	12.867	67,39%
Erróneos	6.225	32,61%
Total	19.092	

Evaluación del rendimiento

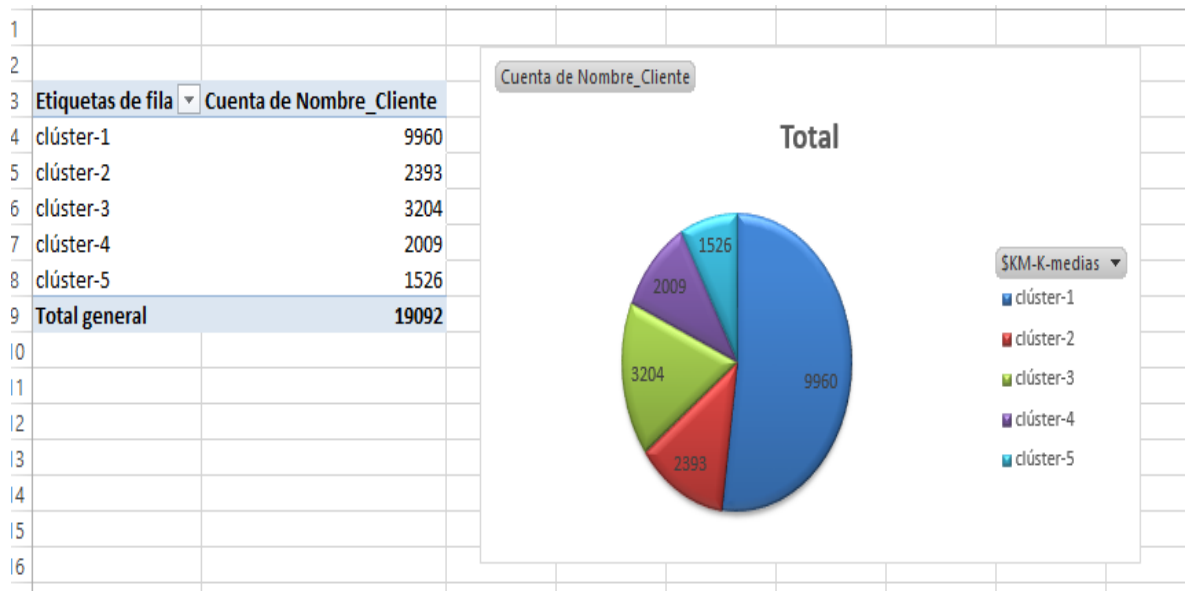
Conclusión de la evaluación: De acuerdo a los resultados de la evaluación de cada módulo podemos concluir que el Modelo basado en el Nodo Kohonen en combinación con un Modelo de Árbol de decisión C5.0 son los más apropiados para conocer el patrón de consumo de los clientes en la empresa Cienpharma.

4.5.EXPLOTAION

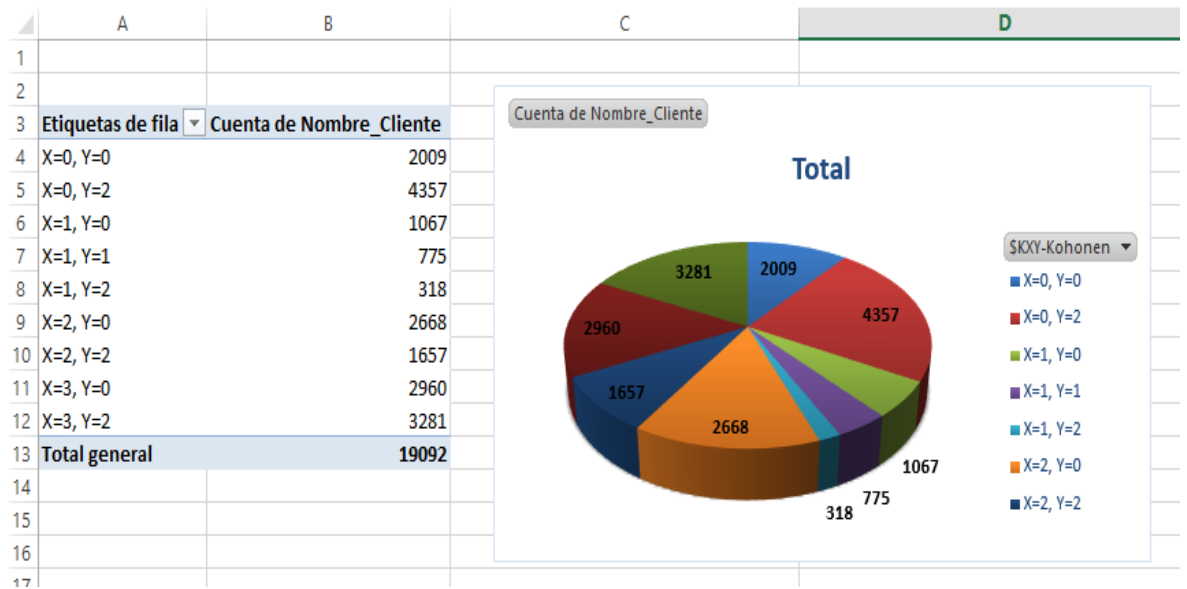
Una de las ventajas de la implementación de la minería de datos con aplicaciones como IBM SPSS Modeler, es que comprenden el proceso de punta a punta, iniciando desde el procedo de Análisis del Problema, Análisis de los Datos, Preparación de los Datos, Modelamiento, Evaluación y Explotación.

El sistema implementado abarca desde la toma de la información del Sistema de Información hasta la entrega de los informes de resultados.

Los datos de los clústeres del modelo K-medias en Excel:



Los datos de los clústeres del Modelo Kohonen en Excel:

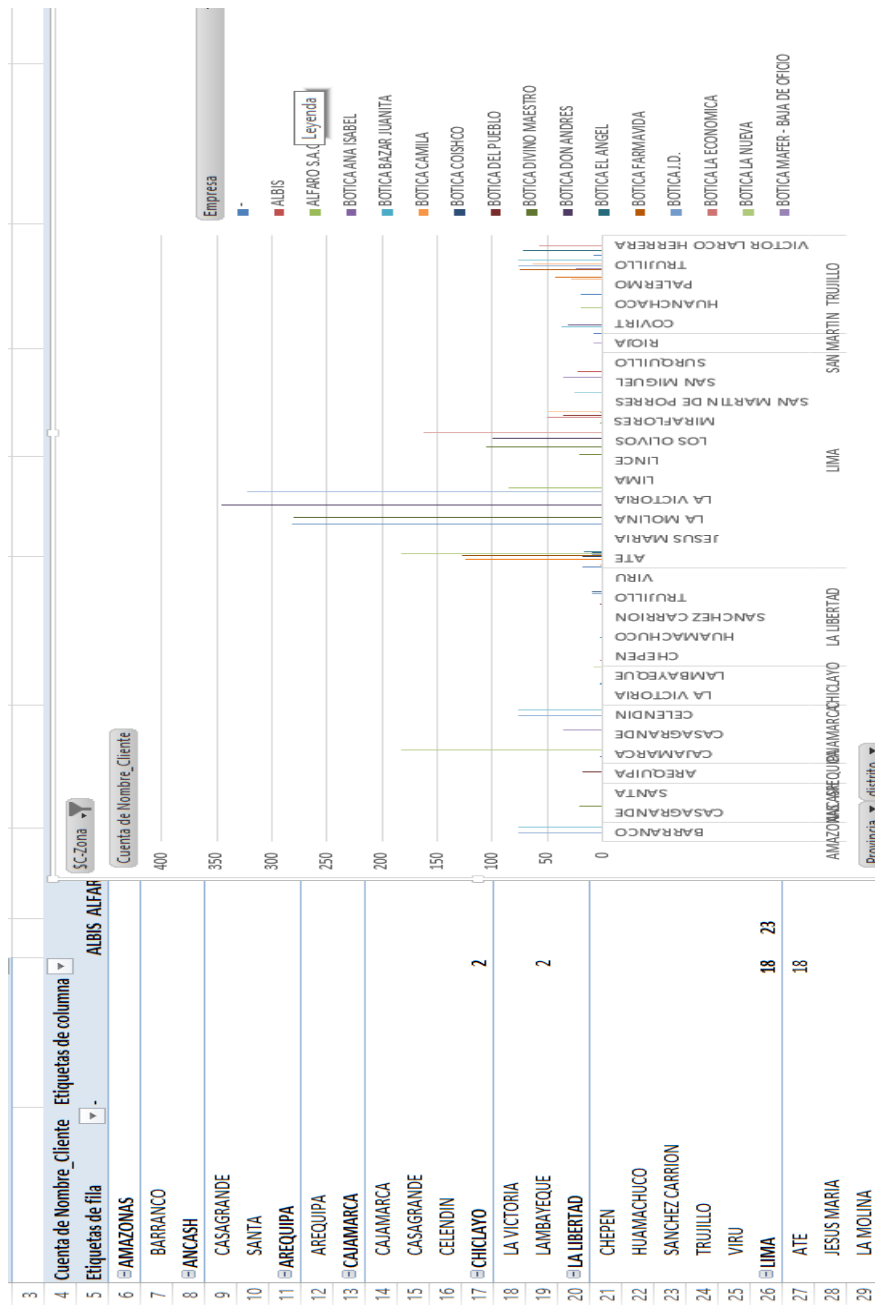


La definición del plan de ejecución del modelo debe ser una decisión de las directivas de acuerdo a la conveniencia administrativa que tiene la empresa. Cada cierto tiempo se debe ejecutar el proceso, de tal forma que se actualice el modelo a las nuevas condiciones.

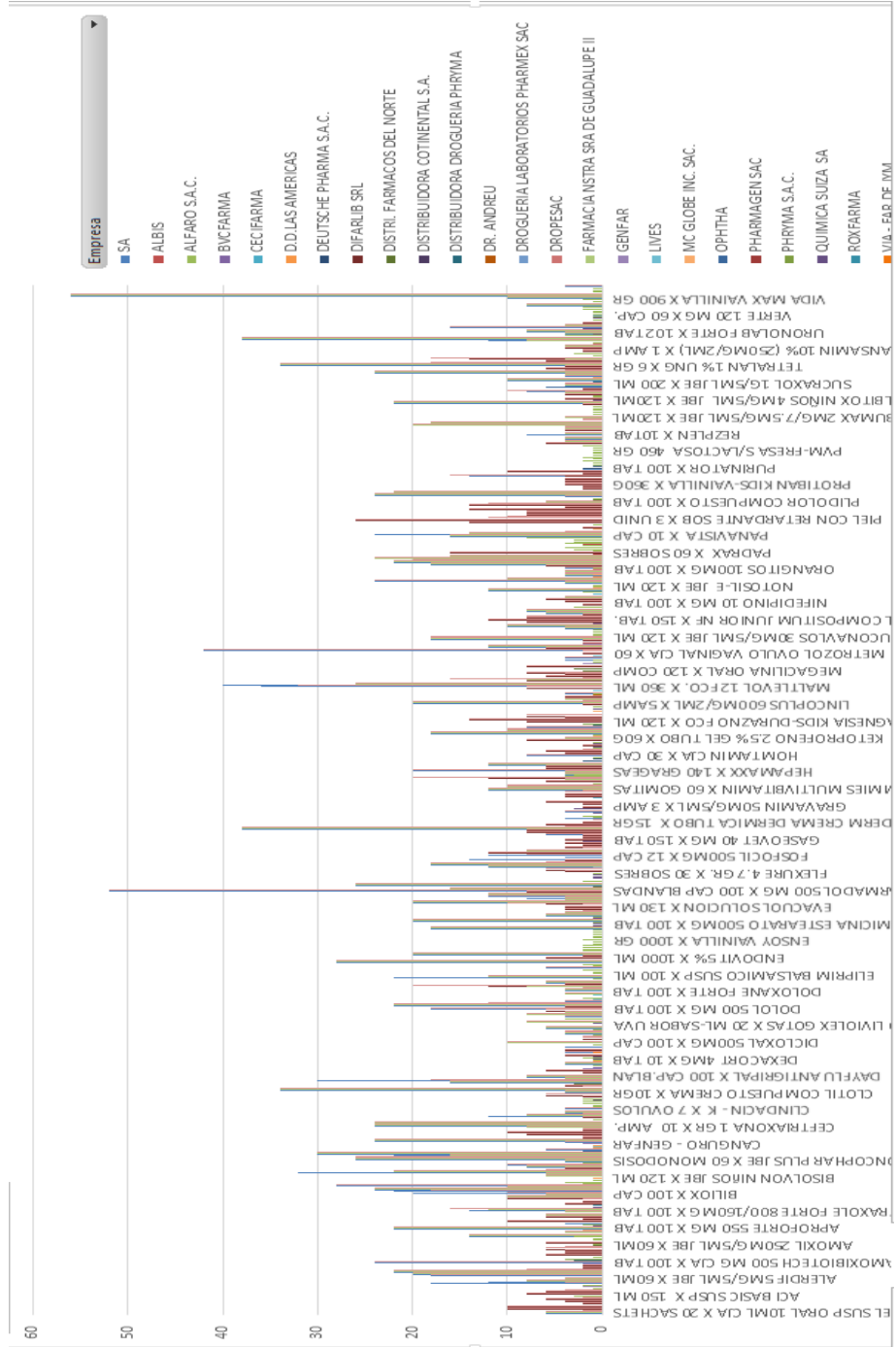
4.6. RESULTADOS DE PATRONES HALLADOS

De acuerdo a los modelos seleccionados (MODELO KOHONEN y ARBOL DE DECISION) se obtuvieron los siguientes patrones:

✓ **Cientes que más compran por Zonas:**



✓ **Productos comprados por empresas y por zona**



5. DISCUSION DE RESULTADOS

Para la contrastación de la hipótesis se ha considerado lo siguiente:

5.1. Formulación del Problema

¿Cómo conocer el patrón de consumo de clientes de grupos que siguen comportamientos similares en la empresa Cienpharma S.A.C?

5.2. Hipótesis

“El desarrollo de una solución de analítica predictiva permitirá conocer el patrón de consumo de clientes en la empresa Cienpharma S.A.C.”

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

- ✓ Independiente (VI): Analítica predictiva utilizando la Metodología CRISP-DM y IBM SPSS Modeler.
- ✓ Dependiente (VD): Patrón de consumo de clientes en la empresa Cienpharma S.A.C.

5.3. Población y muestra.

5.3.1. Población

Registros de las ventas de la base de datos transaccional de la empresa.

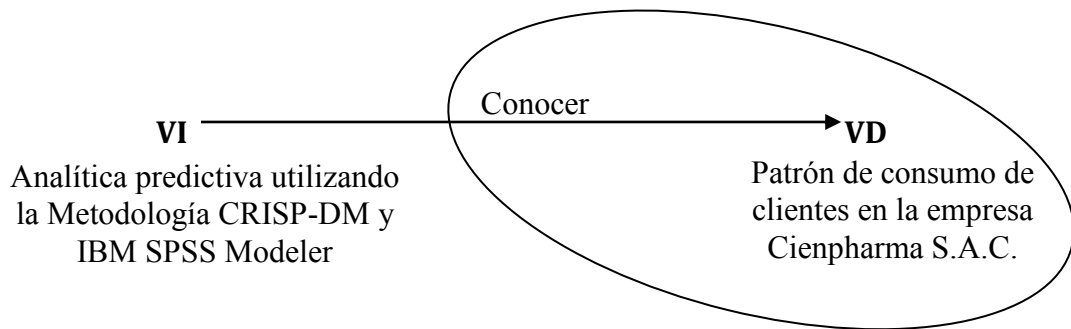
5.3.2. Muestra

Registros de las ventas de los años 2014-2015 de la empresa.

5.3.3. Unidad de análisis

Información de las ventas.

5.4.MANERA PRESENCIAL



5.5.DISEÑO PREEXPIMENTAL PRE-PRUEBA Y POST-PRUEBA

PRE-PRUEBA (O₁): Es la medición previa de X a G

POST-PRUEBA (O₂): Corresponde a la nueva medición de X a G

Se determinó usar el Diseño PreExperimental Pre-Prueba y Post-Prueba, porque nuestra hipótesis se adecua a este diseño. Este diseño experimenta con un solo grupo de sujetos el cual es medido a través de un cuestionario antes y después de presentar el estímulo (MD). Este diseño se presenta de la siguiente manera:

G O₁ X O₂

Donde:

X: Tratamiento, estímulo (MD)

O: Medición a sujetos (Cuestionario)

G: Grupo de sujetos (Empleados)

El espacio de la muestra que se tomó para la medición de los indicadores de la hipótesis, correspondió al total de personas que operarán el Modelo de minería de datos, a estas personas se le aplicó un cuestionario, antes de interactuar con el Modelo (O₁) y después de interactuar con el mismo (O₂).

Al concluir la investigación se establecen las diferencias entre O₁ y O₂ para determinar si hay o no incremento en los resultados obtenidos.

5.5.1. CÁLCULO DE LOS INDICADORES DE LA HIPÓTESIS

Para el cálculo de los indicadores de la hipótesis en el Modelo de Minería de datos Propuesto (MDP) propuesto y el Sistema Actual (SA), se realizó un cuestionario (Ver Anexo A) donde se evaluó a los usuarios luego de haber interactuado con el Modelo.

Los valores que los usuarios dieron a las respuestas del cuestionario fueron aplicados según el siguiente Rango de valoración:

RANGO	GRADO DE VALORACIÓN
1	Desacuerdo
2	Regular
3	Bueno
4	Muy Bueno
5	Excelente

Tabla 10: Rango de grado de valoración

ESCALA DE VALORACION TOTAL	
Inadecuado	15-45
Adecuado	46- 75

5.5.2. APLICACIÓN DEL RANGO DE VALORACIÓN A LOS INDICADORES DE LA HIPÓTESIS

Los valores aplicados a los indicadores de la hipótesis tanto para el sistema Actual como para el BI propuesto se muestran en la siguiente tabla:

Evaluación de los indicadores de la hipótesis:

N°	INDICADORES	VALORACION					\bar{X}
		1	2	3	4	5	
1	Se puede conocer el nivel de detalle de los clientes y su comportamiento o patrón en los diferentes clústeres					2	5
2	El volumen de información que devuelve el modelo cumple con las expectativas					2	5
3	La programación para el modelo de minería es lo recomendable					2	5
4	La respuesta del modelo evaluado es óptimo				1	1	4.5
5	La información proporcionada por el modelo cubre con las expectativas de los stakeholders.					2	5
6	La creación de este modelo cree que no afecta el rendimiento de los sistemas OLTP					2	5
7	La construcción del modelo de minería de datos es una decisión acertada.					2	5
8	La información presentada por el modelo de minería de datos cumple con el criterio de exactitud				1	1	4.5
9	El personal se encuentra satisfecha con el cambio del modo anterior de proceso de toma de decisiones al actual proporcionada por el modelo.					2	5
10	La información presentada apoya al proceso de toma de decisiones del área.					2	5
						$\sum \bar{X}$	49

Dónde: $X = (\text{Valor Valoración} * \text{Número de empleados respondieron en nivel valoración}) / 2$

Tabla 11: Evaluación de los indicadores de la hipótesis.

Interpretación: De acuerdo a la escala de valoración definida para esta ficha de observación, se determina que los Modelos de minería de datos (Kohonen y C5.0) propuesto son los **adecuados**, por ser sumatoria de los promedios 49 y

superior a 45.

5.5.3. ANÁLISIS ESTADÍSTICO PARA LA PRUEBA PRESENCIAL DE LA HIPÓTESIS

Paso 1: Planteamiento de hipótesis.

$$H_0 : O_1 \geq O_2$$

$$H_1 : O_2 \geq O_1$$

Dónde:

H₀ es la hipótesis Nula: “El desarrollo de una solución de analítica predictiva no permite conocer el patrón de consumo de clientes en la empresa Cienpharma S.A.C.”

H₁ es la hipótesis Alternativa: “El desarrollo de una solución de analítica predictiva permite conocer el patrón de consumo de clientes en la empresa Cienpharma S.A.C.”

Paso 2: Nivel de significancia.

Para todo valor de probabilidad igual o menor que 0.05, se acepta H₁ y se rechaza H₀. $\alpha = 0,05$.

Paso 3: Prueba estadística.

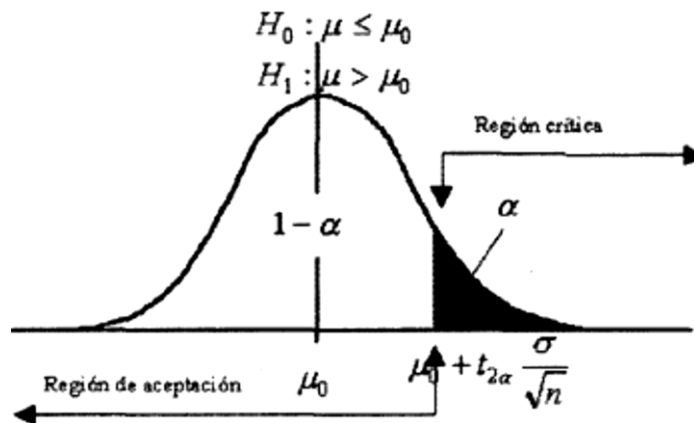
Debido a que la muestra es $n = 2$, y por ende menor a 30, se aplicó la prueba estadística t-student, en esta prueba estadística se exige dependencia entre ambas, en las que hay dos momentos uno antes y otro después. Con ello se da a entender que, en el primer período, las observaciones servirán de control 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 H_0 y se rechaza H_1 .

Si la $t_c > t_t$ se rechaza H_0 y se acepta H_1 .

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



Paso 5: Calculo de t_t y t_c

Calculo de la t de tabla t_t

$$t_t(95\%, 2) = 2,92 \quad \rightarrow \text{Ver Anexo C.}$$

Calculo de la t calculado t_c

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

Donde:

- t_c : T calculado.
- δ : Desviación estándar
- n : Tamaño de la muestra
- \bar{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 (Ver Anexo C) donde se evaluó el grado de satisfacción a los usuarios luego de haber interactuado con el Modelo de minería.

Los valores que los usuarios dieron a las respuestas del cuestionario fueron aplicados según el rango de satisfacción que muestran en la siguiente tabla:

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

N°	INDICADORES	Media Pre U ₁	Media Post U ₂	D= (U ₂ - U ₁)	(D _i - \bar{D})	(D _i - \bar{D}) ²
1	Se puede conocer la cantidad de clientes por clúster.	1.0	9.0	8	1	1
2	Se desea saber la cantidad de clientes que compran por Zona y clúster.	1.0	8.0	7	0	0
3	Se puede conocer los clientes por estado civil y clúster.	1.0	9.0	8	1	1
4	Se puede conocer los clientes por artículo comprado y clúster.	1.0	7.0	6	-1	1
5	Se puede conocer los clientes por monto de compra, por zona y clúster.	1.0	7.0	6	-1	1

$$N = 5 ; \sum D = 35 ; \bar{D} = 7 ; \sum (D_i - \bar{D})^2 = 4 ; \delta = 1 ; \sqrt{n} = 2.27$$

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

$$t_c = 15.9$$

Interpretación: Como $t_c > t_t$, ($15.9 > 2.92$) se acepta la hipótesis alternativa, entendiéndose que el desarrollo de una solución de analítica predictiva permite conocer el patrón de consumo de clientes en la empresa Cienpharma S.A.C.

5.6. CUADRO DE LA COMPARACIÓN DE TIEMPO DE DEMORA EN LA EJECUCIÓN DE LAS CONSULTAS.

NRO	CONSULTAS	CON EL SISTEMA OLTP	CON LA SOLUCION DE MINERIA
1	Se puede conocer la cantidad de clientes por clúster.	No existe la consulta	03 seg.
2	Se desea saber la cantidad de clientes que compran por Zona y clúster.	No existe la consulta	03 seg.
3	Se puede conocer los clientes por estado civil y clúster.	No existe la consulta	03 seg.
4	Se puede conocer los clientes por artículo comprado y clúster.	No existe la consulta	04 seg.
5	Se puede conocer los clientes por monto de compra, por zona y clúster.	No existe la consulta	04 seg
Fuente: La empresa			

6. CONCLUSIONES

- ✓ Se realizó un análisis de la herramienta IBM SPSS Modeler y la metodología CRISP-DM para el desarrollo de las soluciones mostrando los beneficios de la herramienta y fases de la metodología.
- ✓ Se logró identificar la necesidad de conocer el patrón de comportamiento de los clientes respecto a las compras realizadas mediante un análisis del problema que permitieron iniciar el desarrollo del proyecto.
- ✓ Se analizaron los datos de los clientes desde 04 tablas de la base de datos transaccional, siendo estos datos validados para ser usados como entrada de información al modelo desde una consulta preparada, estableciendo una conexión con la base de datos y creando su metadato para ser usado por el modelo.
- ✓ Se construyeron 03 modelos de minería para evaluar y entrenar los datos: K-medias, Kohonen y Árbol de decisión (C5.0), obteniendo diferente cantidad de clústeres por modelo de acuerdo a las entradas de datos evaluada, todo este proceso se realizó utilizando IBM SPSS Modeler.
- ✓ Se evaluaron los resultados de cada modelo determinando que para conocer el patrón de comportamiento de los clientes es necesario utilizar dos modelos de minería de datos (Kohonen + C5.0) por la cantidad de entradas y el número de clústeres devuelto por el modelo.

7. RECOMENDACIONES

- ✓ La base de datos transaccional debe de ser analizada y refinada antes de aplicar técnicas de minería de datos o mejor aún antes de iniciar un proyecto de Data Mining se debe realizar un planeamiento en la recolección de datos.
- ✓ Es importante emplear el número adecuado de registros que sean significativos, ya que los algoritmos que utiliza minería de datos deben de detectar tendencias para mostrar resultados.
- ✓ Recomendamos aplicar el modelo de la técnica de Kohonen evaluado a todos los datos de los clientes para ver el comportamiento del modelo y obtener un mayor volumen de información en los clústeres y así hacer uso de estos segmentos para los objetivos de la empresa.
- ✓ Para la obtención y comprensión de la lógica del negocio se debe mantener una relación estrecha con el usuario involucrado en el área de donde se realizara el proyecto , en cada una de las fases que comprende la metodología aplicada , para obtener información e identificar los puntos necesarios para el desarrollo del proyecto.
- ✓ Se recomienda realizar estudios de minería de datos en los demás procesos claves de la empresa, permitirán conocer aspectos adicionales y poder tomar medidas correctivas.

8. REFERENCIAS BIBLIOGRAFICAS

- Big Data Magazine. (01 de 08 de 2018). ¿Qué es la analítica predictiva? Obtenido de <https://bigdatamagazine.es/bigdatapeda-analitica-predictiva>
- Braga, V., & Paulo, L. (2009). Introducción a la Minería de Datos. Mexico: E-papers, .
- Data Mining Consulting. (07 de 07 de 2014). dataminingperu.com. Obtenido de http://www.dataminingperu.com/blog_dmc/13-blog/63-para-que-sirve-la-mineria-de-datos
- Harold Koontz, H. W. (2012). Administración "Una perspectiva global y empresarial". U.S.: McGraw Hill .
- IBM. (2012). CRISP DM. Obtenido de <ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf>
- IBM. (10 de 02 de 2017). spss-modeler. Obtenido de <http://www-03.ibm.com/software/products/es/spss-modeler>
- IBM Knowledge Center. (20 de 05 de 2017). Objetivos de la minería de datos. Obtenido de https://www.ibm.com/support/knowledgecenter/es/SSEPGG_10.1.0/com.ibm.im.overview.doc/c_dm_goals.html
- MSDN. (05 de 05 de 2017). msdn.microsoft.com. Obtenido de [msdn.microsoft.com: https://msdn.microsoft.com/es-es/library/cc645779.aspx](https://msdn.microsoft.com/es-es/library/cc645779.aspx)
- Pérez López, C. (2007). Minería de datos: técnicas y herramientas. Editorial Paraninfo.
- Prieto, A. (2012). Minería de datos con SAS Enterprise Miner a través de ejemplos.

U.S.: Create Space.

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

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

ANEXOS

ANEXO A

CUESTIONARIO DIRIGIDO: JEFE DE VENTAS Y ADMINISTRADOR DE TIENDA

PREGUNTAS	VALORES					
	0	1	2	3	4	5
Se puede conocer el nivel de detalle de los clientes y su comportamiento o patrón en los diferentes clústeres						
El volumen de información que devuelve el modelo cumple con las expectativas						
La programación para el modelo de minería es lo recomendable						
La respuesta del modelo evaluado es óptimo						
La información proporcionada por el modelo cubre con las expectativas de los stakeholders.						
La creación de este modelo cree que no afecta el rendimiento de los sistemas OLTP						
La construcción del modelo de minería de datos es una decisión acertada.						
La información presentada por el modelo de minería de datos cumple con el criterio de exactitud						
El personal se encuentra satisfecha con el cambio del modo anterior de proceso de toma de decisiones al actual proporcionada por el modelo.						
La información presentada apoya al proceso de toma de decisiones del área.						

Tabla A1. JEFE DE VENTAS Y ADMINISTRADOR DE TIENDA

ANEXO B

CUESTIONARIO DIRIGIDO AL JEFE DE VENTAS Y ADMINISTRADOR DE TIENDA

PREGUNTAS	VALORES										
	0	1	2	3	4	5	6	7	8	9	10
Se puede conocer la cantidad de clientes por clúster.											
Se desea saber la cantidad de clientes que compran por Zona y clúster.											
Se puede conocer los clientes por estado civil y clúster.											
Se puede conocer los clientes por artículo comprado y clúster.											
Se puede conocer los clientes por monto de compra, por zona y clúster.											

Tabla B1. Cuestionario Dirigido al JEFE DE VENTAS Y ADMINISTRADOR DE TIENDA

ANEXO C

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