IDENTIFICACION DE LAS VENTAJAS DE APLICAR MINERIA DE DATOS EN EL NEGOCIO AUTOMOTRIZ

Laura Consuelo Caro Martínez

Código: 201420025441

Fundación Universitaria Los Libertadores

Departamento de Ciencias Básicas

Especialización en estadística aplicada

Bogotá D.C.

2016

IDENTIFICACION DE LAS VENTAJAS DE APLICAR MINERIA DE DATOS EN EL NEGOCIO AUTOMOTRIZ

Laura Consuelo Caro Martínez
Código: 201420025441

Asesor:
Msc. Juan Carlos Borbon Arias

Fundación Universitaria Los Libertadores
Departamento de Ciencias Básicas
Especialización en estadística aplicada
Bogotá D.C.
2016

Nota de Aceptación

_____

_____

_____

_____

_____

_____

_____

Firma del presidente del jurado

_____

Firma del Jurado

_____

Firma del Jurado

Bogotá, D.C 21 agosto del 2016

Las Directivas de la Universidad de

Los Libertadores, los jurados calificadores y el cuerpo

Docente no son responsables por los

Criterios e ideas expuestas En el presente documento.

Estos corresponden únicamente a los autores

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

La minería de datos es un conjunto de técnicas cuyo fin es encontrar la información contenida en bases datos, este proyecto tiene como fin presentar un estudio de la viabilidad y beneficios de las técnicas descriptivas que puede ofrecer esta metodología aplicada en un concesionario.

Para descubrir el conocimiento de la información se utilizan varias formas de análisis con el objetivo de identificar patrones y reglas en los datos para luego representar la información en modelos matemáticos que ayuden en la toma de decisiones.

Palabras clave: minería de datos, máquinas de aprendizaje, análisis predictivo y descriptivo.

1 INTRODUCCIÓN

La minería de datos es un conjunto de procesos y técnicas o algoritmos que permiten extraer el conocimiento a partir de la información almacenada en grandes bases de datos. Las bases de datos son herramientas que nos permiten almacenar información útil de las cuales se puede extraer conocimiento.
Este proyecto se encuentra orientado a la exploración de los datos de un concesionario, para extraer información clave contenida en ellos utilizando herramientas computacionales como son SSPS, R y Matlab para así conocer las ventajas que estos datos puede tener a la hora de tomar de decisiones.

El principal objetivo es analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento inmerso en los datos, utilizando las herramientas de minería de datos, ya que existe un gran interés comercial por explotar los grandes volúmenes de información, pero no conoce la forma de transformar toda esa información en conocimiento que apoye efectivamente la toma de decisiones, especialmente, a nivel gerencial.

2 FORMULACIÓN O PREGUNTA PROBLEMA

¿Cuál es la utilidad práctica de la minería de datos en las ventas de automóviles?

3 OBJETIVOS

3.1 OBJETIVOS GENERALES

Identificar la utilidad práctica de las técnicas de minería de datos en un concesionario.

3.2 OBJETIVOS ESPECÍFICOS

- ✓ Determinar la utilización de la minería de datos en un concesionario
- ✓ Explorar los elementos que aporta la minería de datos en un concesionario

- ✓ Conocer la viabilidad de las técnicas minería de datos en un concesionario

4 JUSTIFICACIÓN

Actualmente el almacenamiento de datos es una tarea rutinaria que se realiza tanto en grandes, medianas y pequeñas empresas, por tanto es un desafío encontrar información a partir de un conjunto grande de datos. Existen programas como son SSPS, R, SAS, Matlab y otros softwares que permiten procesar datos y extraer información utilizando técnicas de minería de datos, este proyecto se realiza con el fin de utilizar las herramientas de minería de datos para llegar a conocer la viabilidad de la información en los datos de un concesionario e inclusive encontrar conocimiento oculto en los datos que puedan ayudar en la toma de decisiones.

Se realizar un pre procesamiento de los datos para corregir los valores erróneos o inconsistencias que se presenten en la base de datos y proceder aplicar las técnicas de minería de datos utilizando herramientas tecnológicas para analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento oculto en los datos y determinar qué tan útil pueden ser estas técnicas en un concesionario.

5 MARCO DE REFERENCIA

5.1 MINERÍA DE DATOS

5.1.1 Historia

Los avances tecnológicos han facilitado los procesos administrativos de las empresas, ya que éstas permiten almacenar los datos referentes a las funciones que se desempeñan, entre las cuales se encuentran, las interacciones basadas con los clientes, la contabilidad de sus procesos internos, entre otras muchas funciones que se llevan a cabo a diario en las empresas. Para estos datos almacenados surge la necesidad de extraer información oculta que contribuya en la toma de decisiones, siendo este el objetivo de la Minería de Datos.

La Minería de Datos tiene sus inicios básicamente en dos áreas del conocimiento:

- Como base principal se encuentra la estadística clásica, la cual cuenta con diversos conceptos como la distribución estándar, la varianza, análisis de clustering, entre muchos otros, los cuales juegan un papel muy Importante en el proceso de la misma, ya que éstos, brindan gran parte de la fundamentación bajo la cual muchos de sus modelos han sido construidos.

- La segunda área de conocimiento que hace parte de la fundamentación de la Minería de Datos es la inteligencia artificial, ésta disciplina procura aplicar procesamiento lógico a través de algoritmos genéticos, redes neuronales, árboles de decisión, entre otros, a diversos problemas estadísticos; para poder aplicar dicho procesamiento, es necesario contar con gran capacidad de poder de cómputo lo cual no fue posible hasta comienzos de los 80's cuando los computadores empezaron a ofrecer mayor capacidad de procesamiento a precios más asequibles, permitiendo que se empezaran a generar diferentes aplicaciones de éste tipo, que en un principio tuvieron fines científicos y de investigación, como se enuncia en A Brief History of Data Mining. Data mining S software". (2006).

A pesar que las técnicas de análisis estadístico permiten conocer información que puede ser útil, no permiten identificar relaciones cualitativas entre los datos, que podrían llegar a ser bastante significativas para las empresas.

Para poder obtener de los datos cierto tipo de información que aporte conocimiento altamente valioso para las organizaciones, se requiere disponer también de técnicas y métodos de análisis inteligente que aunque todavía no han sido perfectamente establecidos, están siendo desarrollados dentro de la inteligencia artificial con el fin de descubrir dicha información que se encuentra oculta en las bases de datos de la organizaciones.

El concepto de Minería de Datos fue usado por primera vez en los años sesenta, cuando los estadísticos manejaban términos como data fishing, data mining o data

archaealogy con la idea de encontrar correlaciones entre los datos sin una hipótesis previa, en bases de datos imprecisas e inconsistentes. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de la Minería de Datos, según Advances in knowledge and data mining (1996).

Actualmente el proceso de Minería de Datos, al estar compuesto por varias etapas, hace el uso de diferentes disciplinas, como la visualización, la computación de alto rendimiento, la estadística, modelos matemáticos y la inteligencia artificial, los cuales le permiten obtener mejores resultados a la hora de extraer información de las bases de datos 20, al igual que existen gran variedad de aplicaciones o herramientas comerciales que además de ser muy poderosas ya que cuentan con un sinfín de utilerías que facilitan el desarrollo de un proyecto, éstas pueden complementarse entre sí para poder arrojar resultados satisfactorios que entreguen información altamente significativa para la toma de decisiones en una organización. La minería de datos revela patrones ocultos, y llegamos a ver estos patrones como modelos, los cuales dan información estadística, tanto descriptiva como predictiva.

Caber resaltar que algunas de las aplicaciones de minería de datos son: web mining, el cual se enfoca en analizar las páginas más visitadas por un cliente, y respecto a esta información se le pueden sugerir otras relacionadas con sus intereses. El text mining para la síntesis y presentación de la información encontrada en la web, lo cual se podrá implementar con voz o imágenes. Y otra de las nuevas vías de investigación es fuzzy mining, la cual se utiliza con objetos simbológicos, que representa más verazmente la incertidumbre que tiene de los objetos que se estudia.

Actualmente se pretende integrar la estadística y la inteligencia artificial con el fin de aprovechar los puntos fuertes de ambas disciplinas y así interpretar y aplicar resultados a una pregunta de negocios en los diferentes problemas típicos como son: generación de recomendaciones, detección de anomalías, análisis de “separación”, administración de riesgos, pronostico, etc…

5.2 TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de minería de datos se clasifican en:

- ✓ Predictivas, las cuales se conocen como técnicas de modelado originado por la teoría, en el que las variables pueden ser dependientes e independientes.
- ✓ Descriptivas o técnicas de modelado originado por los datos, donde no se supone la existencia de variables dependientes ni independientes y tampoco la existencia de un modelo previo para los datos.
- ✓ Auxiliares son herramientas más superficiales y limitadas. Son métodos nuevos basados en técnicas estadísticas.

A continuación se ilustra la clasificación de las técnicas de minería

Figura 1. Técnicas de minería de datos

5.2.1 Fase de Exploración y limpieza

Antes de aplicar técnicas de minería de datos se deben tener en cuenta ciertos requisitos, por tanto es necesario realizar un análisis previo de la información, examinando las variables individuales y la relación entre ellas, para esto se puede utilizar herramientas de exploración visual como gráficos y para una exploración formal se utiliza estadísticos robustos son apropiados cuando los datos no se ajustan a una distribución normal.

5.2.2 Limpieza de Datos

La información puede contener valores atípicos, valores faltantes o valores erróneos, por la presencia de estos datos puede ser necesario llegar a utilizar algoritmos robustos (como son arboles de decisión), a filtrar información, a filtrar valores mediante técnicas de imputación y transformar datos continuos de discretización.

- ✓ Los valores atípicos son datos aislados cuyo comportamiento se diferencia claramente del comportamiento medio del resto de datos. Estos datos pueden ser detectados mediante el diagrama de cajas o el diagrama de control.

- ✓ Los datos desaparecidos, la presencia de esta información puede deberse a un registro defectuoso, a la ausencia natural de la información o a una falta de respuesta. Es vital averiguar si los datos ausentes obedecen a un proceso completamente aleatorio y por lo tanto pueden realizarse análisis estadísticos fiables imputando los datos ausentes.

5.3 FASE DE TRANSFORMACIÓN

Después del análisis exploratorio, los datos originales pueden necesitar ser transformados. Se consideran cuatro tipos de transformaciones; de acuerdo a Pérez, M. (2015).

- **Transformaciones lógicas**: Se unen categorías del campo de definición de las variables para reducir así su amplitud. De esta forma se pueden eliminar categorías sin respuesta. También pueden convertir variables de intervalo en ordinales o nominales y crear variables ficticias (dummy)

- **Transformaciones lineales:** se obtienen de sumar, restar, multiplicar o dividir las observaciones originales por una constante para mejorar su interpretación. Estas transformaciones no cambian la forma de la distribución, ni las distancias entre los valores ni el orden, y por lo tanto no provocan cambios considerables en las variables.

- **Transformaciones algebraicas:** se obtienen al aplicar transformaciones no lineales monotónicas a las observaciones originales (raíz cuadrada, logaritmos, etc.) por una constante para mejorar su interpretación. Estas transformaciones cambian la forma de la distribución al cambiar la distancia entre los valores, pero mantienen el orden.

- **Transformaciones no lineales no monotónicas:** cambian las distancias y el orden entre los valores. Puede cambiar demasiado la información original.

5.4 FASE DE ANÁLISIS DE MINERÍA DE DATOS

Las fases de selección, exploración y transformación optimizan la información para ser analizada

Se inicia realizando una clasificación de datos en técnicas predictivas, donde se clasifican las variables como dependientes e independientes, y en técnicas descriptivas donde todas las variables inicialmente tienen el mismo estatus. Estas dos técnicas están enfocadas a descubrir el conocimiento embebido en los datos.

- **Técnicas Descriptivas:** Las variables no tienen ningún rol específico. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. En este caso se incluyen las técnicas de clustering y segmentación, las cuales hacen parte de técnicas de clasificación, también se utilizan técnicas de dependencia, las técnicas de análisis exploratorio de datos, las técnicas de reducción de dimensión como es factorial, componentes principales, correspondencias, etc. y técnicas de escalamiento multidimensional

- ***Técnicas Predictivas:*** en base a un conocimiento básico previo se especifica el modelo para los datos y este debe contrastarse después del proceso de minería de datos para aceptarlo como válido.

Entre las técnicas predictivas se tienen:

Modelos de regresión, series temporales, análisis de varianza y covarianza, algoritmos genéticos y técnicas de clasificación o segmentación como son análisis discriminante, arboles de decisión, redes neuronales, y modelos logit y probit cuyo objetivo es construir un modelo que permita clasificar cualquier nuevo dato.

A continuación se presentara la utilidad práctica a nivel de gerencia o de negocios de las técnicas de minería de datos.

5.5 MINERÍA DE DATOS Y TOMA DE DECISIONES

La minería de datos es una herramienta que ayuda a las compañías a enfocarse en la información más importante en sus bases de datos o almacenes de datos. Las herramientas de minería de datos predicen comportamientos, permitiendo a los gerentes y empresarios ser más eficientes en la toma de decisiones y el manejo del conocimiento. Además puede responder a preguntas gerenciales que antes consumían demasiado tiempo en obtener respuesta.

5.5.1 Toma de decisiones

La toma de decisiones es un proceso que se lleva acabo a diario en diferentes contextos: en el ámbito laboral, familiar, sentimental, empresarial, etc. Este proceso es sistemático y racional a través del cual se selecciona una alternativa de entre varias, siendo la seleccionada la optimizadora (la mejor para nuestro propósito)[1].
Tomar la correcta decisión en un negocio o empresa es parte fundamental del administrador ya que sus decisiones influirán en el funcionamiento de la organización, generando repercusiones positivas o negativas según su elección.

En el proceso de tomar decisiones se tienen en cuenta las siguientes etapas:

- Identificar y analizar el problema
- Investigación y obtención de la información
- Determinación de parámetros

- Construcción de una alternativa
- Aplicación de la alternativa
- Especificación y evaluación de las alternativas
- Implantación

5.5.2 Los datos y la toma de decisiones

Piatesky-Shapiro (1991) destacan la Minería de Datos como el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos, a través de grandes cantidades de datos, posibilitando de esta manera el encuentro de relaciones o patrones entre los datos procesados, como lo ilustra Marcano y Talavar en el libro Minería de datos como soporte a la toma de decisiones empresariales.

Por su parte, Molina y García (2004) explican que los datos tal cual se almacenan en las bases de datos no suelen proporcionar beneficios directos; su valor real reside en la información que podamos extraer de ellos, es decir, información que nos ayude a tomar decisiones o a mejorar la comprensión de los fenómenos que nos rodean. Lo cual se logra definiendo medidas cuantitativas para los patrones obtenidos (precisión, utilidad y beneficio obtenido), para establecer medidas de interés que consideren la validez y simplicidad de los patrones obtenidos mediante alguna de las técnicas de Minería de Datos, con el fin de tomar decisiones a través de la información oculta en los datos.

Barreiro, Diez y Ruzo (2003) describen la importancia que tiene la Minería de Datos en la implementación de las actividades de negocio: tales como la bondad, aplicabilidad, la relevancia y la novedad; indicadores que aportan una idea de las implicaciones y utilidades que proporciona esta práctica, ayudando a obtener resultados confiables utilizando software de apoyo en menor tiempo.

A. Indicadores de la bondad del resultado

Los índices de bondad de resultado tratan de aportar una idea acerca del error que se comete al emplear un modelo para realizar una tarea. Tal como manifiestan Padmanabhan y Tuzhilin (1999), ésta es una medida de la fortaleza estadística del

resultado. Para este indicador se utilizan las siguientes medidas: Precisión, Ratio de error, Varianza y Matriz de confusión, siendo las dos últimas derivaciones de las anteriores. La precisión se utiliza cuando el resultado se presenta en forma de clasificación o estimación, la cual se mide a través del porcentaje de predicciones que son correctas. Para efectos de la clasificación, se emplea el porcentaje de casos bien clasificados y para la estimación del porcentaje de registros, se emplea una estimación que el decisor considere correcta. Para medir la precisión se puede emplear el coeficiente de confianza, el cual no es más que la probabilidad condicionada de un hecho con respecto a otro.

La distancia es otra técnica de Minería de Datos empleada cuando se disponen de variables continuas y numéricas, mediante la raíz cuadrada de la suma al cuadrado de las distancias en cada eje. Una medida que complementa a la precisión es el Ratio de error, que mide el porcentaje de casos en los que el resultado no coincide con la realidad.

B. Indicadores de relevancia del resultado

Los indicadores más representativos en este grupo son el Coeficiente de cobertura, el Coeficiente de apoyo y el Coeficiente de significación. Estos indicadores tienen que ver directamente con la importancia que tiene el resultado arrojado por las técnicas de minería y miden la aportación a la situación actual y la frecuencia de utilidad del resultado, cuando la presentación de éstos se hace en forma de reglas.

El Coeficiente de cobertura mide el porcentaje de registros en los cuales se puede aplicar la regla. Por otro lado, el Coeficiente de apoyo permite mostrar el porcentaje de ocasiones en que globalmente aparece la relación descrita por la regla, se recomienda representar el resultado en porcentaje. Por último, el Coeficiente de significación sirve para medir el grado de importancia de la regla a través de la aportación que supone respecto a la pura probabilidad.

C. Indicadores de novedad del resultado

Cuando la información es excesivamente abundante y obvia, puede presentarse el problema al generar reglas. Para ello, existe el Coeficiente de novedad, creado para

indicar si una regla es interesante o no en función del número de reglas ya generadas, para un área de conocimiento concreta. Su objetivo es ayudar a evitar las redundancias en su obtención. Autores como Buchner et al. (1999), entre otros, abogan por la inclusión del conocimiento previo del negocio, e intuición que detentan las decisiones para de esta manera: restringir el espacio de búsqueda, obtener conocimiento más preciso y eliminar aquél que resulte no interesante.

D. Indicadores de aplicabilidad del resultado

La dinámica de las organizaciones actuales demanda cada vez más, tiempos de respuesta más rápidos, por lo cual es necesario que tanto la creación o generación de modelos como los resultados del mismo, deben estar disponibles en el menor tiempo posible. Para lograr esto, hay que buscar la simplicidad de los modelos y de la forma de representar la salida o resultados del análisis, para transformar el conocimiento obtenido y poder aplicarlo al negocio; para lograr esto, se cuenta con el Coeficiente de Simplicidad, la Tasa Interna de Retorno y el Valor Actual Neto.

La aplicación de técnicas de Minería de Datos mediante métodos estadísticos avanzados y la ayuda de softwares permite la extracción de conocimiento en grandes base de datos, ayudando a determinar las características contables de las empresas más rentables, al igual que el perfil de sus clientes. Se hace imprescindible, por un lado, un análisis exploratorio profundo de la base de datos y el empleo de métodos robustos, que hagan que dichos componentes sean menos sensibles a los amplios casos estadísticos. Por otro lado, es aconsejable diseñar con base a opiniones de expertos, si no hay información adecuada, o utilizar algún sistema de aprendizaje, por ejemplo, la utilización de redes neuronales, para el descubrimiento de patrones y extraer la información de la base de datos disponible.

En fin, estos métodos y procedimientos se han convertido en retos tecnológicos para procesar los datos y convertirlos en conocimiento útil para la toma de decisiones. Este camino se presenta como una opción para las organizaciones que quieran ser competitivas, valiéndose de la experiencia acumulada, la cual sin duda alguna constituye el principal activo del que se dispone para la creación de valor. De esta manera, una organización que reflexiona, documenta y aprende, está en condiciones de innovar y obtener ventajas competitivas, por lo tanto la minería de datos se

comporta como una herramienta en pro de facilitar este proceso y obtener óptimos resultados en tiempos más cortos y con menor inversión de capital.

Como evidencia del trabajo realizado, se puede aportar entre otras las siguientes investigaciones las cuales sirve como guía y soporte teórico – práctico para desarrollar la investigación mediante el software SSPS, luego de una búsqueda exhaustiva en internet se encontraron varios proyectos de grado de distintas universidades extranjeras en el área de sistemas, las cuales ilustran ejemplos de aplicación de la minería de datos en empresas y su respectiva contribución a la toma de decisiones, a continuación se ilustran los más relevantes:

- *Formulación de Minería de Datos para la Empresa Distribuidora de Productos Espinoza Aguilar S.A.* de la Universidad Tecnológica del Perú, el cual plantea un estudio de la viabilidad, adaptación y beneficios que puede ofrecer la metodología de la minería de datos aplicado a la pequeña empresa, que no cuenta con plan de proyección estructurado, de los análisis internos y externos, que van cambiando durante el ciclo de vida de la empresa.

- *Modelo de Minería de datos para la identificación de patrones que influyen en el aprovechamiento académico* del Instituto Tecnológico de la Paz, donde se realiza un análisis de las aplicación de técnicas de minería de datos para identificar patrones de comportamiento con el fin de predecir el fracaso escolar y abandono, se llevó a cabo en una institución de nivel medio privada de México.

- *Minería de datos una herramienta para la toma de decisiones*, universidad de San Carlos de Guatemala, donde se evaluó el uso de la minería de datos como una herramienta que sirve para la toma de decisiones a nivel gerencial.

- El articulo Minería de Datos como soporte a la toma de decisiones empresariales, cuyo objetivo es examinar y describir las técnicas y herramientas que emergen en la investigación de minería de datos, apoyándose para ello en una reflexión teórica-cualitativa que contribuya a un mayor entendimiento del alcance y limitaciones de la Minería de Datos como

soporte a la toma de decisiones empresariales. Resaltando los beneficios que ofrece la técnica para elevar los niveles de competencia de los negocios, basándose en la rapidez para identificar, procesar y extraer la información que realmente es importante, descubriendo conocimiento y patrones en bases de datos.

Los demás proyectos y artículos que se tuvieron en cuentan presentan el mismo esquema que los expuestos anteriormente. Se basan en el uso de las técnicas de minería de datos con el fin de encontrar algún conocimiento oculto que contribuya en la toma de decisiones. Se debe aclarar que lo importante en este tipo de análisis no es si el resultado es positivo o negativo para los intereses de la empresa, sino la generación de conocimiento como resultado de la interpretación, análisis y validación de los patrones resultantes del proceso de minería de datos.

Teniendo en cuenta lo anterior, a continuación se planteara el marco metodológico que sustenta todo el procedimiento seguido en este estudio.

6 MARCO METODOLOGICO

6.1 TIPO DE ESTUDIO

El presente este proyecto de investigación se inscribe en los estudios de corte exploratorio, ya que se pretende realizar un análisis de la base de datos de un concesionario utilizando las técnicas de minería de datos para identificar la utilidad práctica que estas tienen en este campo. Al entrar en la investigación, no fue posible encontrar investigaciones mediante el uso de la minería de datos en la aplicación específica de un análisis sobre los datos que se toman a diario en un concesionario. De hecho, a raíz de esto se encontraron varias situaciones que sirven como soporte en este proceso.

6.2 MÉTODO

El desarrollo de este proyecto se basa en un método documental donde se observa y reflexiona sistemáticamente sobre algunos ejemplos teórico – prácticos seleccionados tomando como base la similitud que presentan con el presente proyecto y que pueden

ser de utilidad para identificar elementos de utilidad en la aplicación de la minería de datos en el contexto de negocios automotrices.

Como primera estancia se tienen el libro de ***Minería de datos a través de ejemplos***, el cual sirve como guía y soporte teórico – práctico para desarrollar la investigación mediante el software SSPS, luego de una búsqueda exhaustiva en internet se encontraron varios proyectos de grado de distintas universidades extranjeras en el área de sistemas, las cuales ilustran ejemplos de aplicación de la minería de datos en empresas y su respectiva contribución a la toma de decisiones, a continuación se ilustran los más relevantes: *Formulación de Minería de Datos para la Empresa Distribuidora de Productos Espinoza Aguilar S.A.* de la Universidad Tecnológica del Perú, *Modelo de Minería de datos para la identificación de patrones que influyen en el aprovechamiento académico* del Instituto Tecnológico de la Paz, *Minería de datos una herramienta para la toma de decisiones*, universidad de San Carlos de Guatemala y el articulo *Minería de Datos como soporte a la toma de decisiones empresariales.*

Los demás proyectos y artículos que se tuvieron en cuentan presentan el mismo esquema que los expuestos anteriormente. Se basan en el uso de las técnicas de minería de datos con el fin de encontrar algún conocimiento oculto que contribuya en la toma de decisiones. Se debe aclarar que lo importante en este tipo de análisis no es si el resultado es positivo o negativo para los intereses de la empresa, sino la generación de conocimiento como resultado de la interpretación, análisis y validación de los patrones resultantes del proceso de minería de datos.

Como instrumento de estudio se cuenta con la base de datos de un concesionario automotriz y los programas SSPS, para realizar el respectivo análisis.

6.3 UNIDADES DE ANÁLISIS

Para el desarrollo de esta fase se procedió revisar la base de datos y determinar cuáles eran las variables relevantes para el presente estudio, en este caso se desarrolló una caracterización de la base de datos, en el la tabla 1 se registran las variables estudiadas y su caracterización.

Variable	Tipo de variable	Ajuste requerido	Observaciones
1. Mes	Nominal	Cambio de números por nombre de cada mes.	
2. Año Venta	Nominal (si bien es un número de año este es una referencia y no una cuantía)	No se requiere	
3. Concesionario	Nominal	No se requiere	
4. Tipo compra	Nominal	No se requiere	Hace referencia a compras al por mayor y al detal
5. Precios de venta	Cuantitativa	Eliminación de decimales separados por comas	No se usa separador de miles
6. Vendedor	Nominal	No se requiere	
7. Cliente	Nominal	No se requiere	
8. Uso	Nominal	No se requiere	Uso que le dará el cliente al vehículo
9. Grupo	Nominal	No se requiere	Hace referencia a la capacidad de carga o tipo de carga del vehículo
10. Tipo	Nominal	No se requiere	El tipo de uso si es público o particular

11. Marca	Nominal	No se requiere	Marca del fabricante del vehículo
12. Gama	Nominal	No se requiere	Es un código de gama o clase del vehículo
13. Referencia	Nominal	No se requiere	Dato alfa numérico con que el fabricante codifica el tipo de vehículo fabricado
14. Color	Nominal	No se requiere	Color del vehículo sin particularidades de condición de aplicación de la pintura.
15. Modelo	Nominal (si bien es un número no es una variable cuantitativa porque no tiene sentido obtener promedios)	No se requiere	

Tabla 1. *Caracterización de variables Base de datos*

6.4 PROCEDIMIENTO

Como análisis de la base de datos se tomó a modo de ejemplo ilustrativo sobre la minería de datos en el negocio automotriz, se contempló el siguiente procedimiento:

6.4.1 Análisis exploratorio

Se realizó una inspección de los datos donde se identificaron los valores extremos, discontinuidades en los datos y otras peculiaridades, para evitar los errores al ejecutarse el programa ejecutado
Se encontró algunas casillas sin datos y llenas con signos de interrogación, lo cual llevo a suprimir estas, además se escogieron las variables más relevantes y se les realizo con análisis descriptivo con el fin de conocer los datos atípicos.

6.4.2 Análisis de clústeres

Como primera medida se desarrolló una fase descriptiva consiste en la aplicación de la técnica de clustering bietápico en el aplicación del programa SPSS, este método de clúster se utilizó dado que se cuenta con variables cuantitativas y cualitativas, el número de datos es grande en este caso se cuenta con 16958 datos y no se requieres un procedimiento previo de cálculo a priori de los clústeres del modelo.

6.4.3 Minería de Datos

Utilizar técnicas Descriptivas (se realizó un análisis descriptivo general, tanto para variables cualitativas como cuantitativas)

Utilizar técnicas predictivas (No se utilizaron)

7 RESULTADOS Y ANÁLISIS

7.1 ANÁLISIS DESCRIPTIVO

Inicialmente se realizó un análisis exploratorio para las variables cualitativas, utilizando el software SSPS, con la opción *Analizar, estadísticos descriptivos, explorar*, lo cual arrojo los resultados de la tabla 2 se observa que no hay valores perdidos, en este caso solo se realiza el proceso para variables cuantitativas

La exploración de los datos muestra que no existen valores inusuales, valores extremos, discontinuidades de los datos u otras peculiaridades.

En cuanto a los estadísticos se obtiene media, mediana, media recortada al 5% de error típico, varianza, desviación típica, minimo, máximo, amplitud, amplitud intercuartil, asimetría y curtosis y sus errores típicos, intervalo de confianza para la media (y el nivel de confianza especifico), percentiles, estimador –M de Huber (*Estimador M de posición Las puntuaciones típicas que sean menores que una constante, reciben un peso de 1. Los casos que tienen los mayores valores absolutos tienen pesos tanto más pequeños cuanto mayor es su distancia respecto a cero*), estimador en onda de Andrews, estimador –M, redescedente de Hampel, estimador biponderado de Tukey, los cinco valores mayores y los cinco valores menores, estadisco de Kolmogovor – Smirnov con el nivel de significación de Lillierfors para contrastar la normalidad y estadístico de Shapiro-Wilk. La prueba Kolmogorov-Smirnov se aplica para contrastar la hipótesis de normalidad de la población, el estadístico de prueba es la máxima diferencia:

$$D = máx\,|F_n(x) - F_0(x)|$$

Siendo $Fn(x)$ la función de distribución muestral y $Fo(x)$ la función teórica o correspondiente a la población normal especificada en la hipótesis nula.

La distribución del estadístico de Kolmogorov-Smirnov es independiente de la distribución poblacional especificada en la hipótesis nula y los valores críticos de este estadístico están tabulados. Si la distribución postulada es la normal y se estiman sus parámetros, los valores críticos se obtienen aplicando la corrección de significación propuesta por Lilliefors.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
FECHA FACT	16958	100,0%	0	0,0%	16958	100,0%
DIA	16958	100,0%	0	0,0%	16958	100,0%
AÑO	16958	100,0%	0	0,0%	16958	100,0%
TOTAL VENTA	16958	100,0%	0	0,0%	16958	100,0%
AÑO MODELO	16958	100,0%	0	0,0%	16958	100,0%

Tabla 2. Resumen del procesamiento de los casos

Descriptivos

			Estadístico	Error típ.
FECHA FACT	Media		18-JUN-2011 19:45:26,736	5
	Intervalo de confianza para la media al 95%	Límite inferior	07-JUN-2011	
		Límite superior	29-JUN-2011	
	Media recortada al 5%		26-JUL-2011	
	Mediana		13-FEB-2012	
	Varianza		429268898 4096590,500	

			758	
	Desv. típ.		07:36:55,554	
	Mínimo		06-DEC-2006	
	Máximo		31-DEC-2013	
	Rango		2582 00:00:00	
	Amplitud intercuartil		1007 00:00:00	
	Asimetría		-,846	,019
	Curtosis		-,576	,038
	Media		22,40	,060
	Intervalo de confianza para la media al 95%	Límite inferior	22,28	
		Límite superior	22,51	
	Media recortada al 5%		22,84	
	Mediana		25,00	
DIA	Varianza		61,602	
	Desv. típ.		7,849	
	Mínimo		1	
	Máximo		31	
	Rango		30	
	Amplitud intercuartil		14	
	Asimetría		-,632	,019
	Curtosis		-,816	,038
	Media		2010,93	,016
AÑO	Intervalo de confianza para la media al 95%	Límite inferior	2010,90	
		Límite superior	2010,96	

	Media recortada al 5%		2011,04	
	Mediana		2012,00	
	Varianza		4,250	
	Desv. típ.		2,062	
	Mínimo		2006	
	Máximo		2013	
	Rango		7	
	Amplitud intercuartil		3	
	Asimetría		-,844	,019
	Curtosis		-,597	,038
	Media		42816236,26	141546,574
	Intervalo de confianza para la media al 95%	Límite inferior	42538790,27	
		Límite superior	43093682,25	
	Media recortada al 5%		41024641,50	
	Mediana		39900000,00	
TOTAL VENTA	Varianza		339760867265628,200	
	Desv. típ.		18432603,377	
	Mínimo		0	
	Máximo		225000000	
	Rango		225000000	
	Amplitud intercuartil		20516143	
	Asimetría		2,332	,019
	Curtosis		11,614	,038
	Media		2011,53	,017
AÑO MODELO		Límite inferior	2011,50	

	Intervalo de confianza para la media al 95%	Límite superior	2011,56	
	Media recortada al 5%		2011,65	
	Mediana		2012,00	
	Varianza		4,711	
	Desv. típ.		2,171	
	Mínimo		2006	
	Máximo		2014	
	Rango		8	
	Amplitud intercuartil		3	
	Asimetría		-,850	,019
	Curtosis		-,541	,038

Tabla 3. Descriptivos

Estimadores-M

	Estimador-M de Huber[a]	Biponderado de Tukey[b]	Estimador-M de Hampel[c]	Onda de Andrews[d]
FECHA FACT	06-JAN-2012	15-MAR-2012	12-DEC-2011	20-MAR-2012
DIA	23,93	24,14	23,42	24,16
AÑO	2011,62	2011,91	2011,60	2011,92
TOTAL VENTA	40557442,05	39568888,40	40178242,30	39563886,73
AÑO MODELO	2012,08	2012,36	2012,10	2012,37

a. La constante de ponderación es 1,339.

b. La constante de ponderación es 4,685.

c. Las constantes de ponderación son 1,700, 3,400 y 8,500.

d. La constante de ponderación es 1,340*pi.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado(definición 1)	FECHA FACT	17-MAY-2007	27-SEP-2007	30-APR-2010	13-FEB-2012	31-JAN-2013	14-AUG-2013	21-OCT-2013
	DIA	8,00	10,00	16,00	25,00	30,00	31,00	31,00
	AÑO	2007,00	2007,00	2010,00	2012,00	2013,00	2013,00	2013,00
	TOTAL VENTA	21488431,05	24000000,00	30267857,00	39900000,00	50784000,00	59000000,00	70964912,00
	AÑO MODELO	2007,00	2008,00	2010,00	2012,00	2013,00	2014,00	2014,00
Bisagras de Tukey	FECHA FACT			30-APR-2010	13-FEB-2012	31-JAN-2013		
	DIA			16,00	25,00	30,00		
	AÑO			2010,00	2012,00	2013,00		
	TOTAL VENTA			30267857,00	39900000,00	50784000,00		
	AÑO MODELO			2010,00	2012,00	2013,00		

Tabla 4. Percentiles

Valores extremos

			Número del caso	Valor
FECHA FACT	Mayores	1	1904	31-DEC-2013
		2	2987	31-DEC-2013
		3	3206	31-DEC-2013
		4	3412	31-DEC-2013
		5	3634	31-DEC-2013[a]
	Menores	1	6673	06-DEC-2006
		2	6672	06-DEC-2006
		3	6445	06-DEC-2006
		4	5771	06-DEC-2006
		5	5545	06-DEC-2006[b]
DIA	Mayores	1	8	31
		2	11	31
		3	15	31
		4	20	31
		5	25	31[c]
	Menores	1	15940	1
		2	2109	1
		3	16344	2
		4	14188	2

		5	11438	2[d]
		1	1	2013
		2	2	2013
	Mayores	3	1813	2013
		4	1814	2013
		5	1815	2013[e]
AÑO		1	6910	2006
		2	6909	2006
	Menores	3	6906	2006
		4	6904	2006
		5	6903	2006[f]
		1	15938	225000000
		2	15941	225000000
	Mayores	3	15942	225000000
		4	15940	215000000
TOTAL		5	14990	210000000
VENTA		1	2441	0
		2	11350	15500000
	Menores	3	11518	15658722
		4	11619	15900000
		5	14243	16500000
		1	1836	2014
		2	1837	2014
	Mayores	3	1838	2014
		4	1839	2014
AÑO		5	1840	2014[g]
MODELO		1	15407	2006
		2	15396	2006
	Menores	3	15439	2007
		4	15354	2007
		5	9985	2007[h]

a. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 31-Dec-2013.

b. En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 06-Dec-2006.

c. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 31.

d. En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 2.

e. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 2013.

f. En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 2006.

g. En la tabla de valores extremos mayores sólo se muestra una lista parcial de los casos con el valor 2014.

h. En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 2007.

Tabla 5. Valores Extremos

Pruebas de normalidad

	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
FECHA FACT	,144	16958	,000
DIA	,143	16958	,000
AÑO	,222	16958	,000
TOTAL VENTA	,120	16958	,000
AÑO MODELO	,247	16958	,000

a. Corrección de la significación de Lilliefors

Tabla 6. Prueba de normalidad

El estadístico del contraste Kolmogorov-Smirnov para la variable fecha toma el valor 0,144 que permite rechazar la hipótesis nula de normalidad para niveles de significación inferiores a 0,2, al igual que en las otras variables mediante la utilización

del programa SPSS se aplicó el comando clúster bietápico, el cual se caracteriza por ser clúster en dos etapas está pensado para minería de datos, es decir para estudios con un número de individuos grande que pueden tener problemas de clasificación con los otros procedimientos. Se puede utilizar tanto cuando el número de clúster es conocido a priori y cuando es desconocido. Permite trabajar conjuntamente con variables de tipo mixto (cualitativas y cuantitativas) para la base de datos a las 15 variables definidas en el cuadro 1 y 16958 datos por variable, registrados en la base de datos, de este procesamiento se obtuvo como primer resultado la figura 2 se resumen el modelo conglomerado bietápico base de datos del presente trabajo de grado.

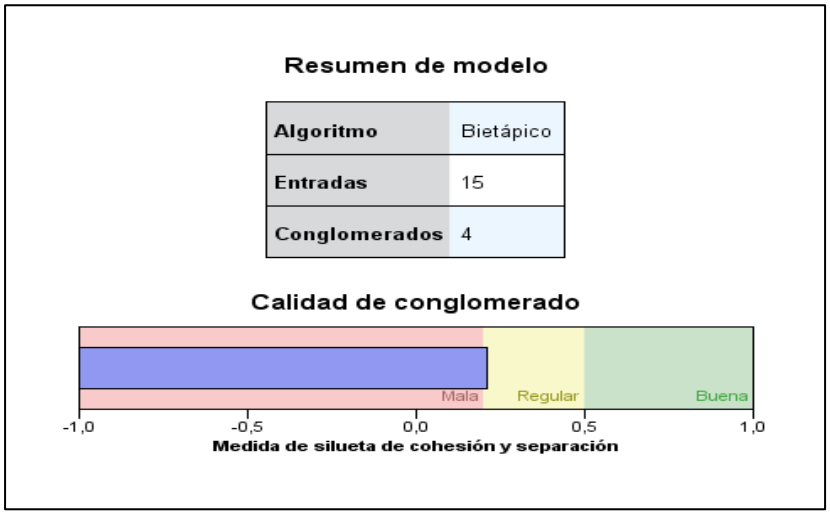


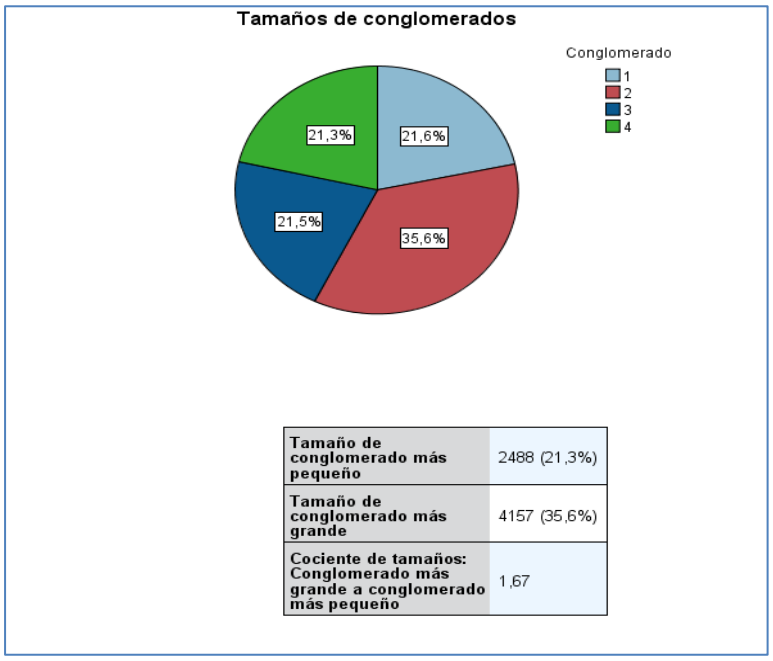
Figura 2. Resumen del modelo análisis de clúster

7.1.1 Caracterización de clústeres

Se realiza un resumen detallado de la relación interna de variables a nivel de cada uno de los conglomerados, en la figura 2 se ilustra el tamaño de los conglomerados obtenidos. Se genera cuatro conglomerados, la calidad de la técnica es aceptable.

Los datos obtenidos en los conglomerados uno y dos difieren en un 14%, mientras que los conglomerados tres y cuatros son muy similares en su tamaño y tan solo difieren en un 0.2 %, los datos expresados en porcentaje son la proporción de las observaciones que acumula cada uno de los clúster, en relación al total de casos registrados *(16958).*

Al explorar cada uno de los conglomerados se determina las variables tienen más relevancia al momento de definir el número de unidades vendidas y las características de las mismas en los marcos del negocio de la empresa que es la venta de vehículos desde los automóviles hasta camiones de gran capacidad de carga.



Tamaño de conglomerado más pequeño	2488 (21,3%)
Tamaño de conglomerado más grande	4157 (35,6%)
Cociente de tamaños: Conglomerado más grande a conglomerado más pequeño	1,67

Figura 3. Tamaño de clúster generados

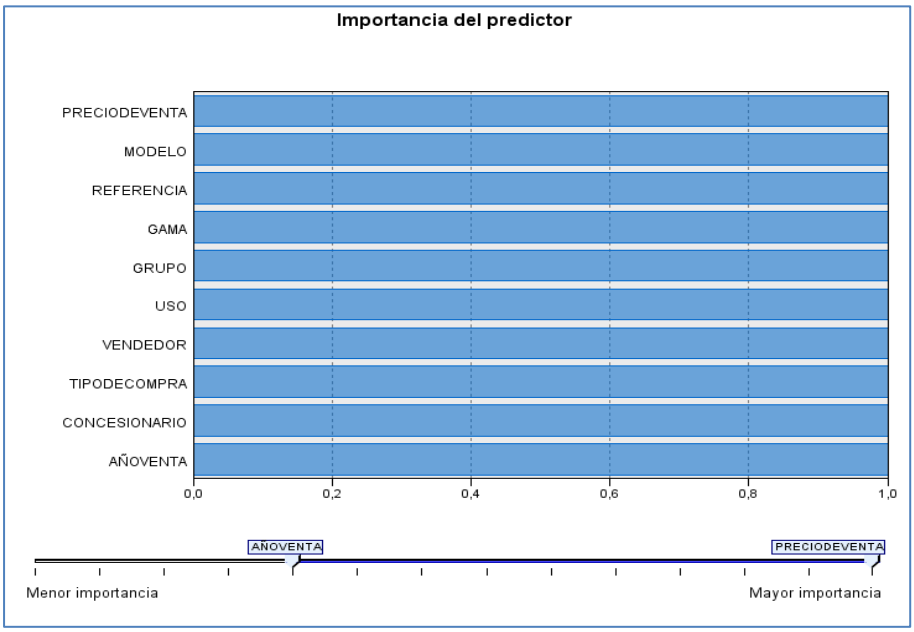
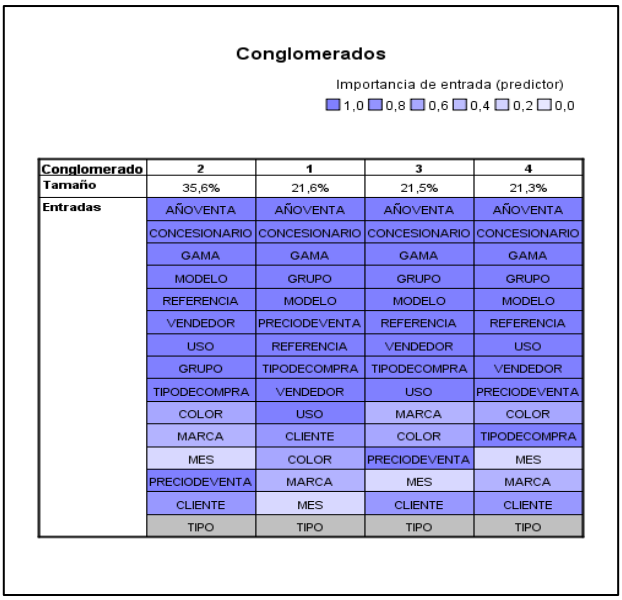


Figura 4. Aporte de las diferentes variables estudiadas

Acorde a lo observado en el gráfico de importancia del predictor, una de las variables que más puede influir en la decisión de compra es el precio de venta, que tiene una estrecha relación el modelo, la gama y el color.

7.2 COMPORTAMIENTO DE LAS VARIABLES EN LOS CLÚSTERES GENERADOS



Conglomerado	2	1	3	4
Tamaño	35,6%	21,6%	21,5%	21,3%
Entradas	AÑOVENTA	AÑOVENTA	AÑOVENTA	AÑOVENTA
	CONCESIONARIO	CONCESIONARIO	CONCESIONARIO	CONCESIONARIO
	GAMA	GAMA	GAMA	GAMA
	MODELO	GRUPO	GRUPO	GRUPO
	REFERENCIA	MODELO	MODELO	MODELO
	VENDEDOR	PRECIODEVENTA	REFERENCIA	REFERENCIA
	USO	REFERENCIA	VENDEDOR	USO
	GRUPO	TIPODECOMPRA	TIPODECOMPRA	VENDEDOR
	TIPODECOMPRA	VENDEDOR	USO	PRECIODEVENTA
	COLOR	USO	MARCA	COLOR
	MARCA	CLIENTE	COLOR	TIPODECOMPRA
	MES	COLOR	PRECIODEVENTA	MES
	PRECIODEVENTA	MARCA	MES	MARCA
	CLIENTE	MES	CLIENTE	CLIENTE
	TIPO	TIPO	TIPO	TIPO

Figura 5. Clasificación de Clústeres por tamaños y variables acorde a su importancia en cada clúster

Se observa en el figura 5 que en todos los clúster las variables da mayor importancia al año de venta, el concesionario y la gama, luego de estas variables se destacan el grupo y modelo, por ultimo encontramos la referencia, estas son la variables que mejor pueden predecir el comportamiento de los conglomerados, que no son más que un registro de las condiciones del mercado de vehículos, lo que el cliente demanda con más frecuencia, consistente con la importancia de estas variables se procedió a desarrollar un estudio del comportamiento al interior de los clústeres.

7.2.1 Variable año de venta

Esta variable nos muestra el comportamiento de las ventas en cada año en cada uno de los conglomerados consolidados.

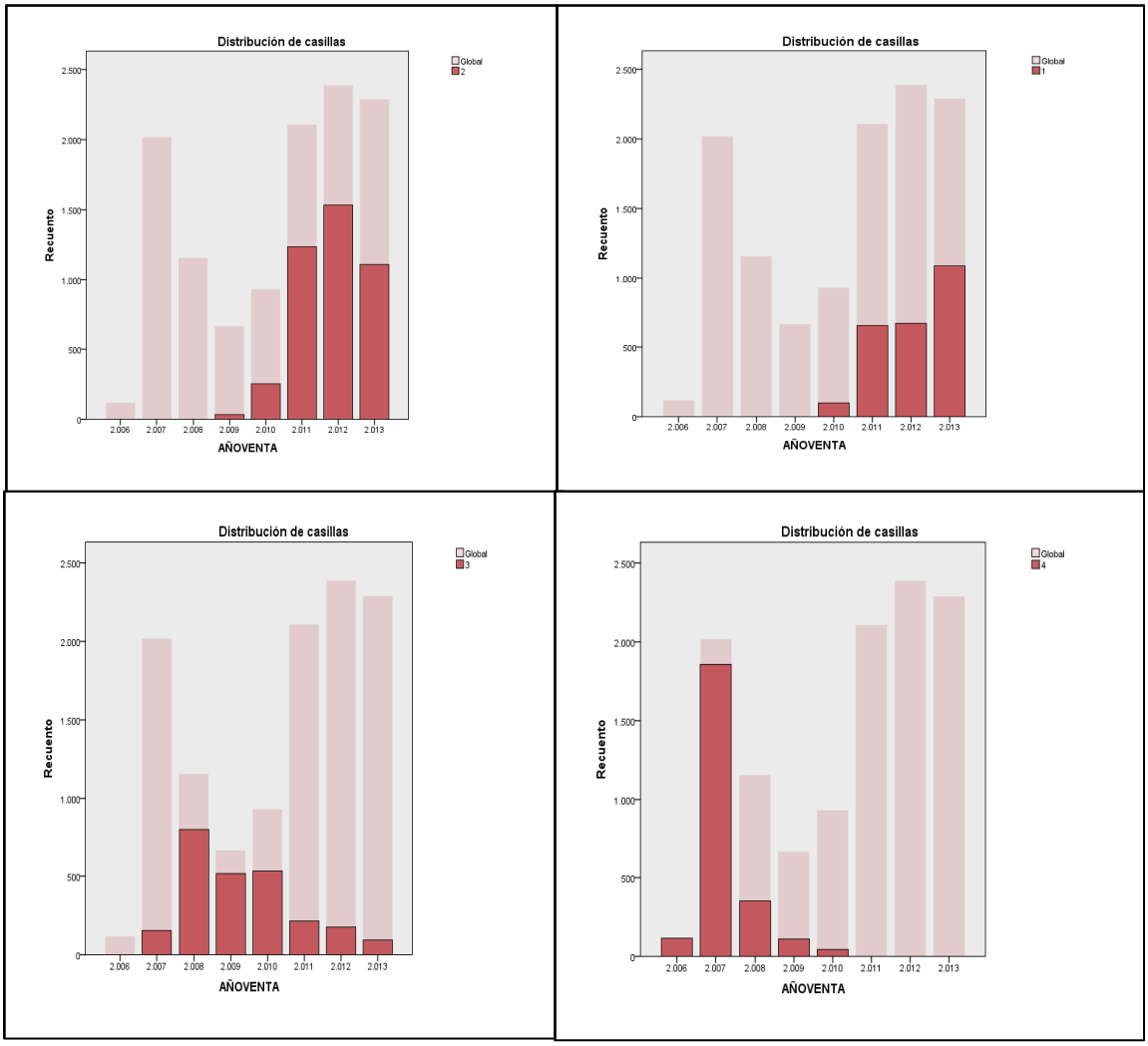


Figura 6. Detalle de ventas por año en cada clúster

En ventas por años los clústeres 1 y 2, presentan un reporte significativo de ventas en los últimos tres años, mientras que los clúster 3 y 4 presentan un decrecimiento acentuado de sus ventas en los últimos tres años, en el caso concreto del clúster número cuatro las ventas tienen una mayor dispersión, lo que hace que estos vendedores y concesionarios tienen una tendencia a salir de la empresa. Ver figura 6.

7.2.2 Variable concesionario

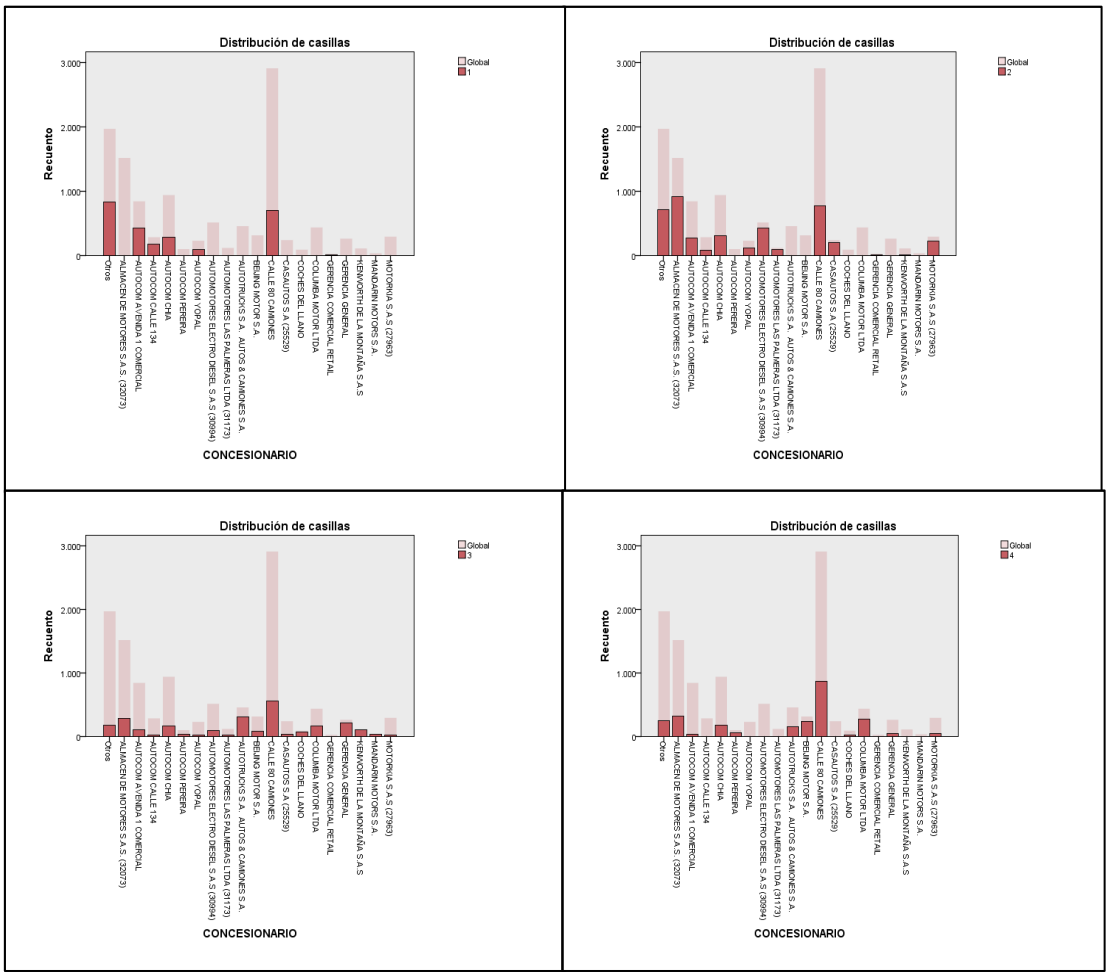


Figura 7. Comportamiento de los concesionarios en cada uno de los Clúster generados

Al detallar los gráficos de figura 7, el concesionario que más rota unidades es el concesionario calle 80 camiones y el almacén de motores SAS, los demás concesionarios se comportan con unos niveles bajos de ventas, en estos es de vital importancia detectar e indagar en el portafolio de vehículos ofertados, cuales son los más demandados y si la zona tiene clientes que geográficamente ven estos sitios como una gran opción y sitio de gran demanda para sus transacciones en materia de adquisición de automotores.

8.2.1 Variable gama

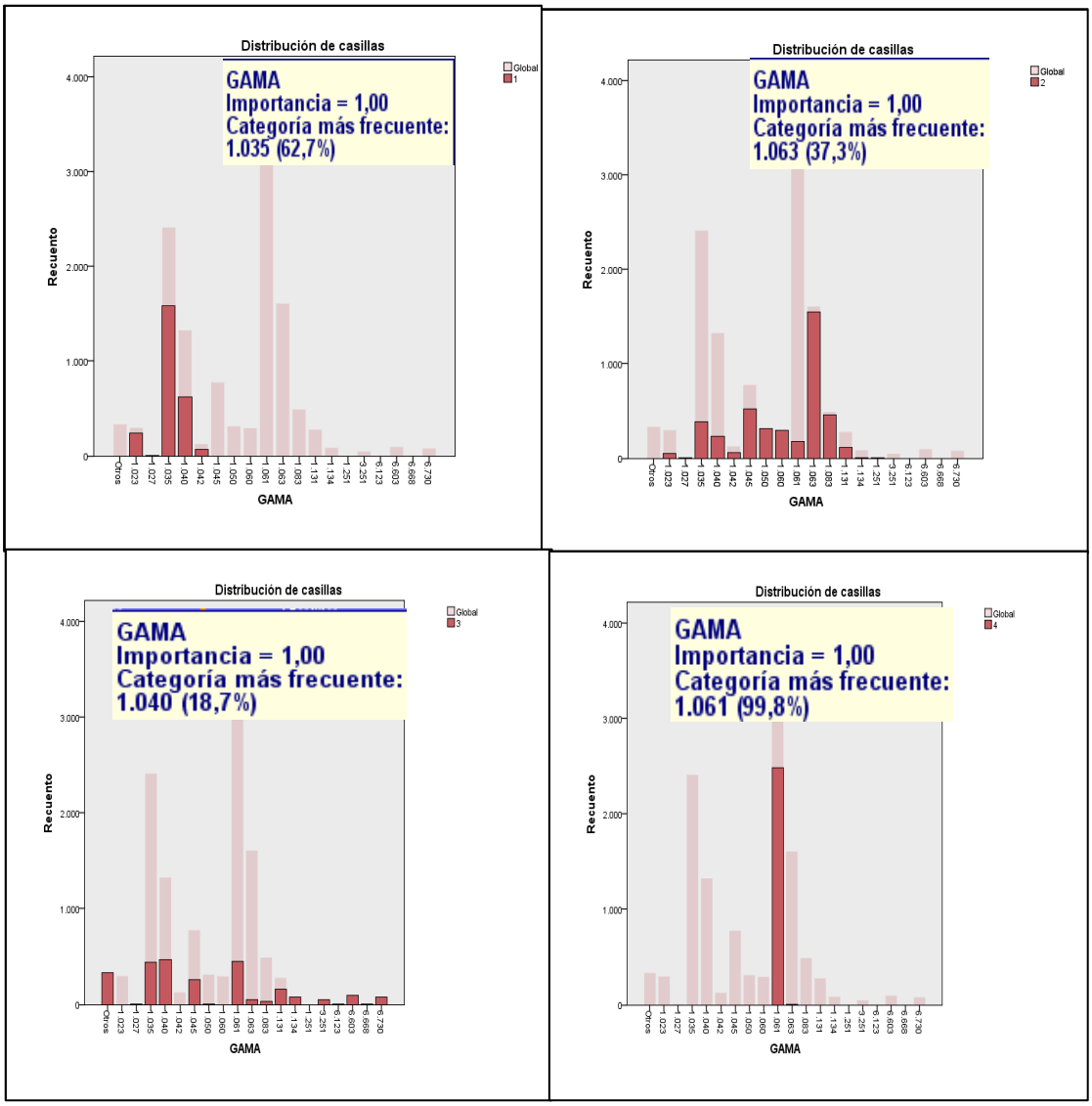


Figura 8. Comportamiento por gamas en los clústeres generados para el presente estudio

Al desarrollar el análisis de la figura 8 Hay una relativa preferencia por unas gamas de cada uno de los clúster, en el clúster 2 la gama 1063, en el clúster 1 la gama 1035, en el clúster 3 la gama 1040 y en el clúster 4 la gama 1061, en tres de los clúster sobre sale la gama 1035, y en caso del cuarto clúster hay una marcada demanda al 1061, de hecho este un clúster que está muy ligado a esta gama, lo que tiene serios inconvenientes por no hay otras gamas como opción de elección en este clúster.

7.2.3 Variable grupo

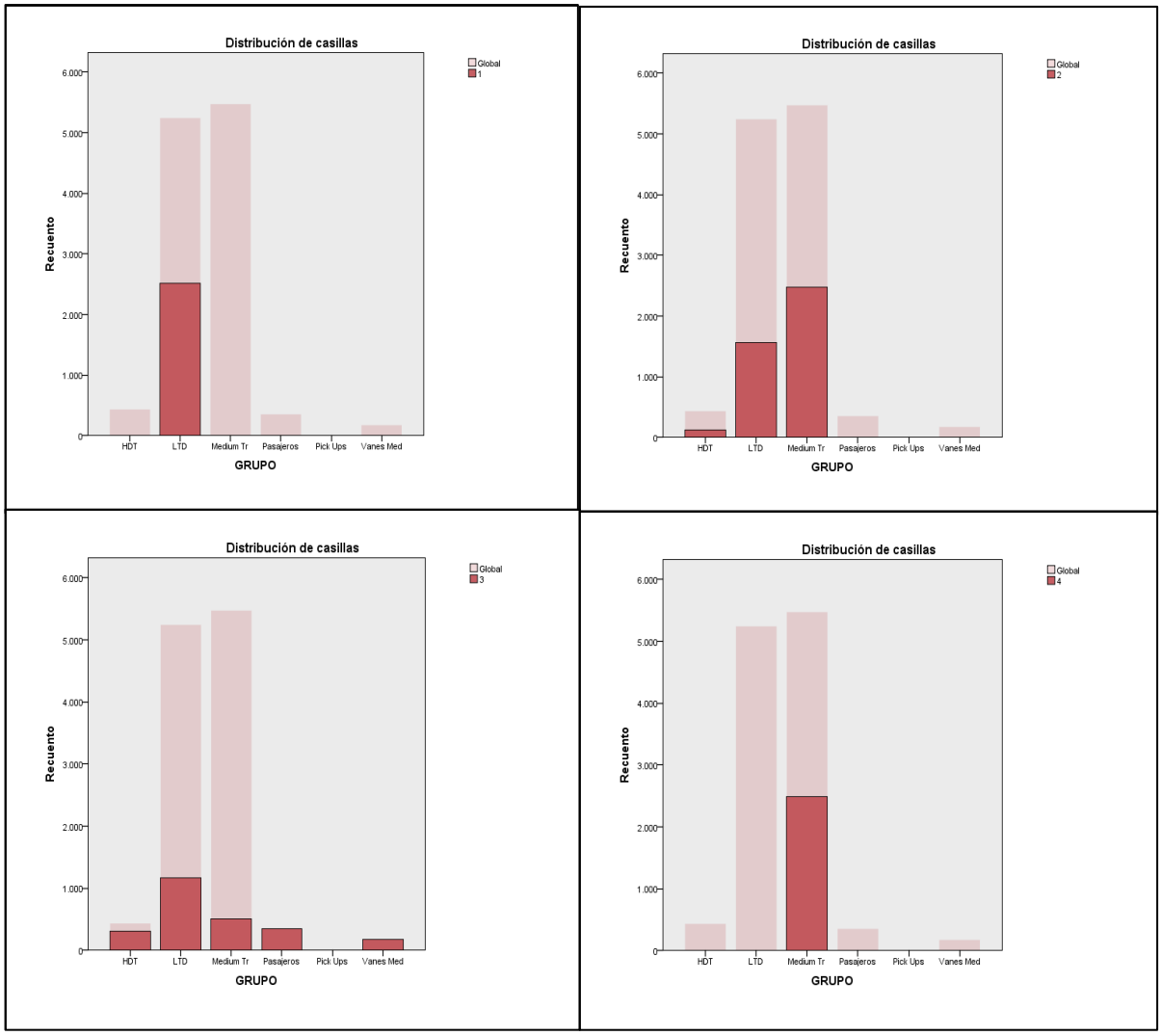


Figura 9 Comportamiento de los grupos de vehículos en cada uno de clústeres

Analizando el comportamiento de variable grupo de vehículo, el grupo más aceptado es el médium Tr, seguido de LTD. Médium al ser el grupo más aceptado demarca el sendero a seguir, en la selección en los vehículos a ofrecer a los clientes, pues se demuestra una preferencia indiscutible, por un par de grupos y este deben ser blanco de estrategias para mantenerse en el mercado y buscar opciones en estos grupo de vehículos, con valores agregados de diseño, como una opción de confort para un grupo de vehículos ya posesionado en el mercado.

7.2.4 Variable modelo

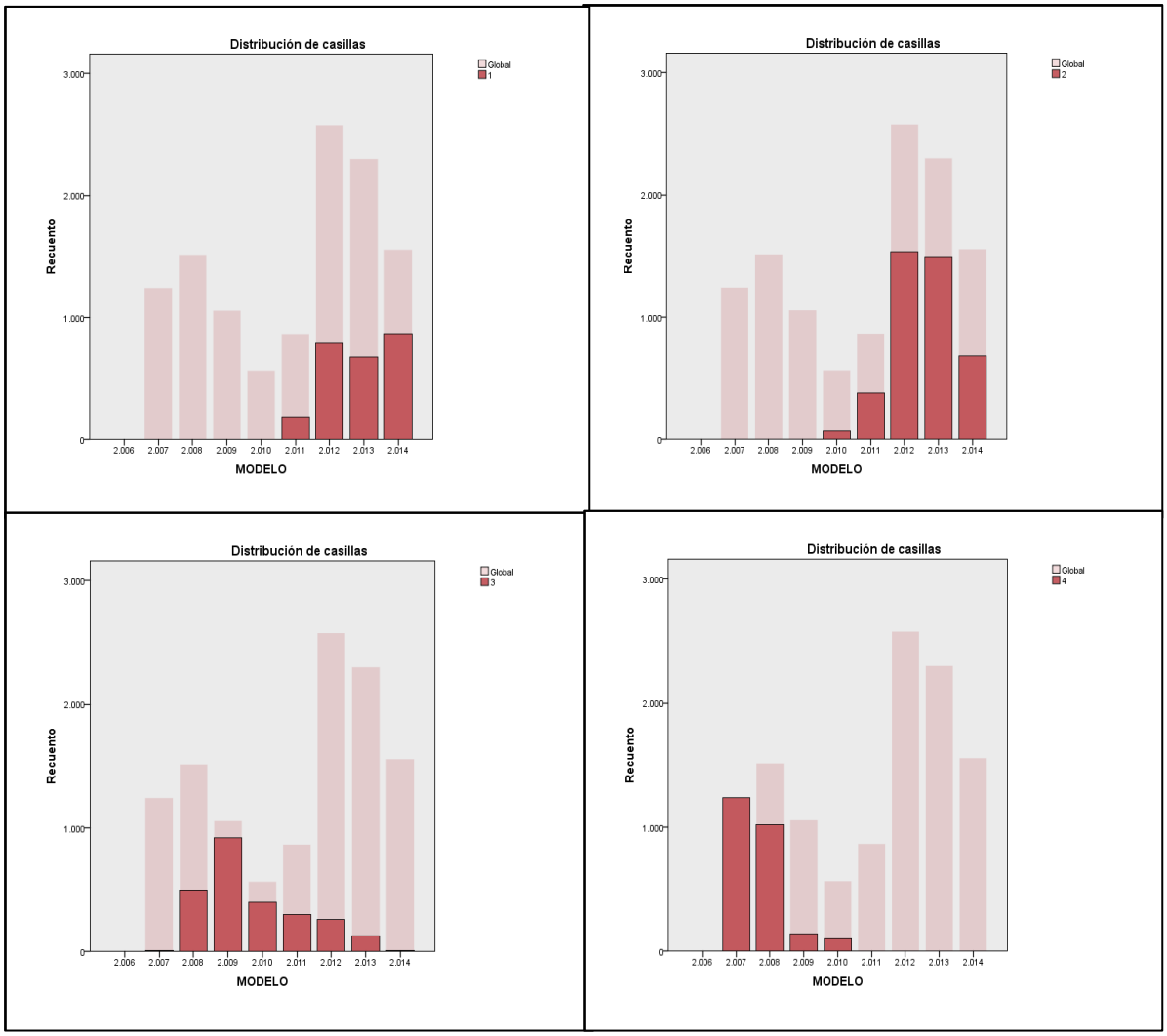


Figura 10. Modelos demandados por clúster

En la Figura 9 se determina que los clúster 1 y 2 son demandantes de vehículos recientes, concentrando su intención de compra en los vehículos de los modelos 2012 y 2013, hay un decrecimiento de la demanda para vehículos del 2014, en caso de los clústeres 3 y 4, hay una inclinación a la compra de vehículos de más de dos años de uso, concentrándose en el años 2008.

7.2.5 Variable referencia

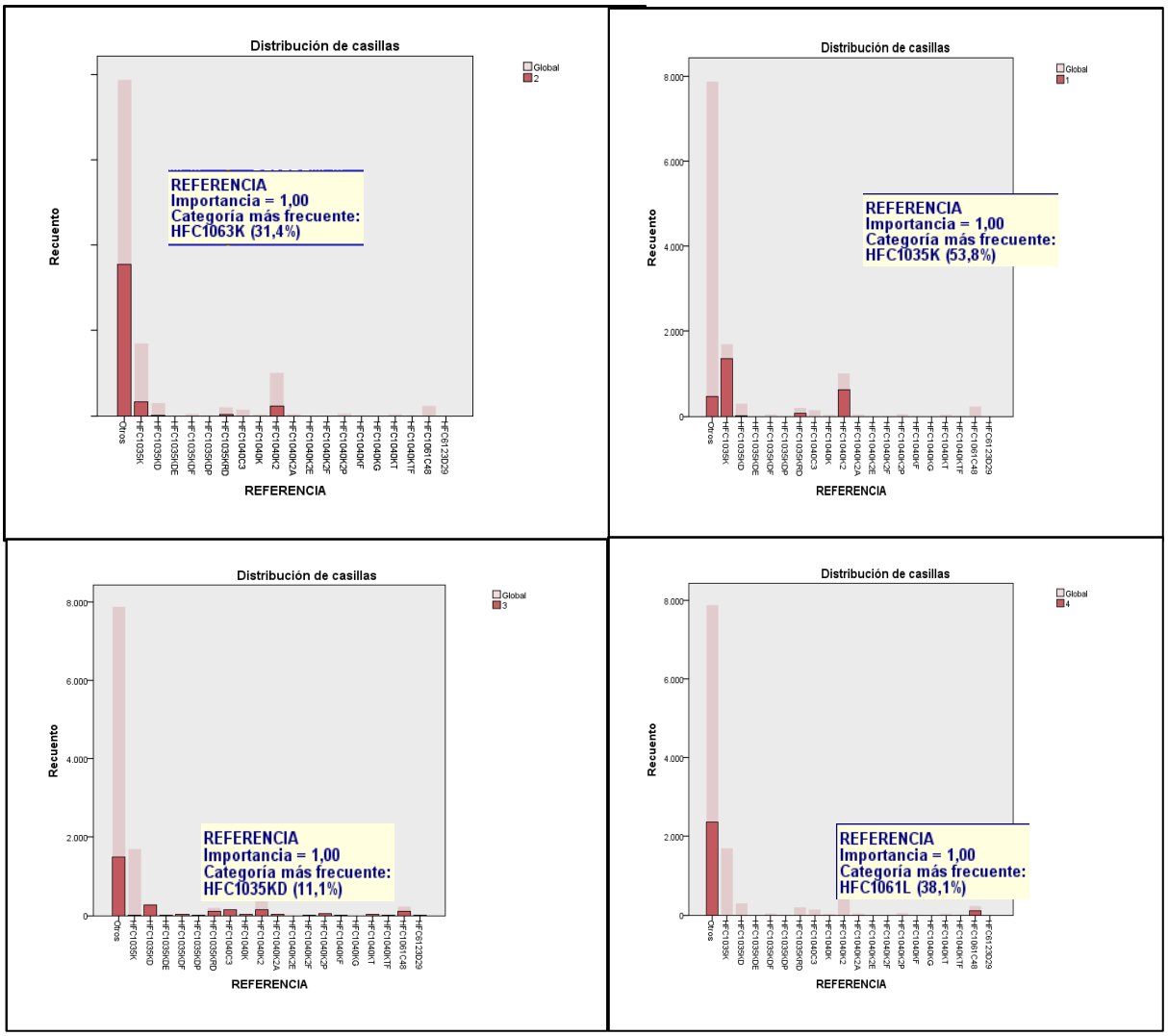


Figura 11. Referencias de más aceptación en cada uno de los clústeres generados

En materia de referencias en los datos analizados de los vehículos comercializados sobresalen las referencias HFC1063K, HFC1035K, HFC1035KD y HFC1061L, se puede observar que línea de vehículos de más venta es la HFC y es estos se observa la aceptación de los vehículos 1035K y su variante el 1935KD, es importante que la compañía, trabaje en determinar cuál es la razón de que estas referencias sean las más demandadas.

7.2.6 Variable marcas

Esta variable se incluye pues para el comercializados es vital determinar, si hay una preferencia por alguna de sus marcas de vehículos comercializadas.

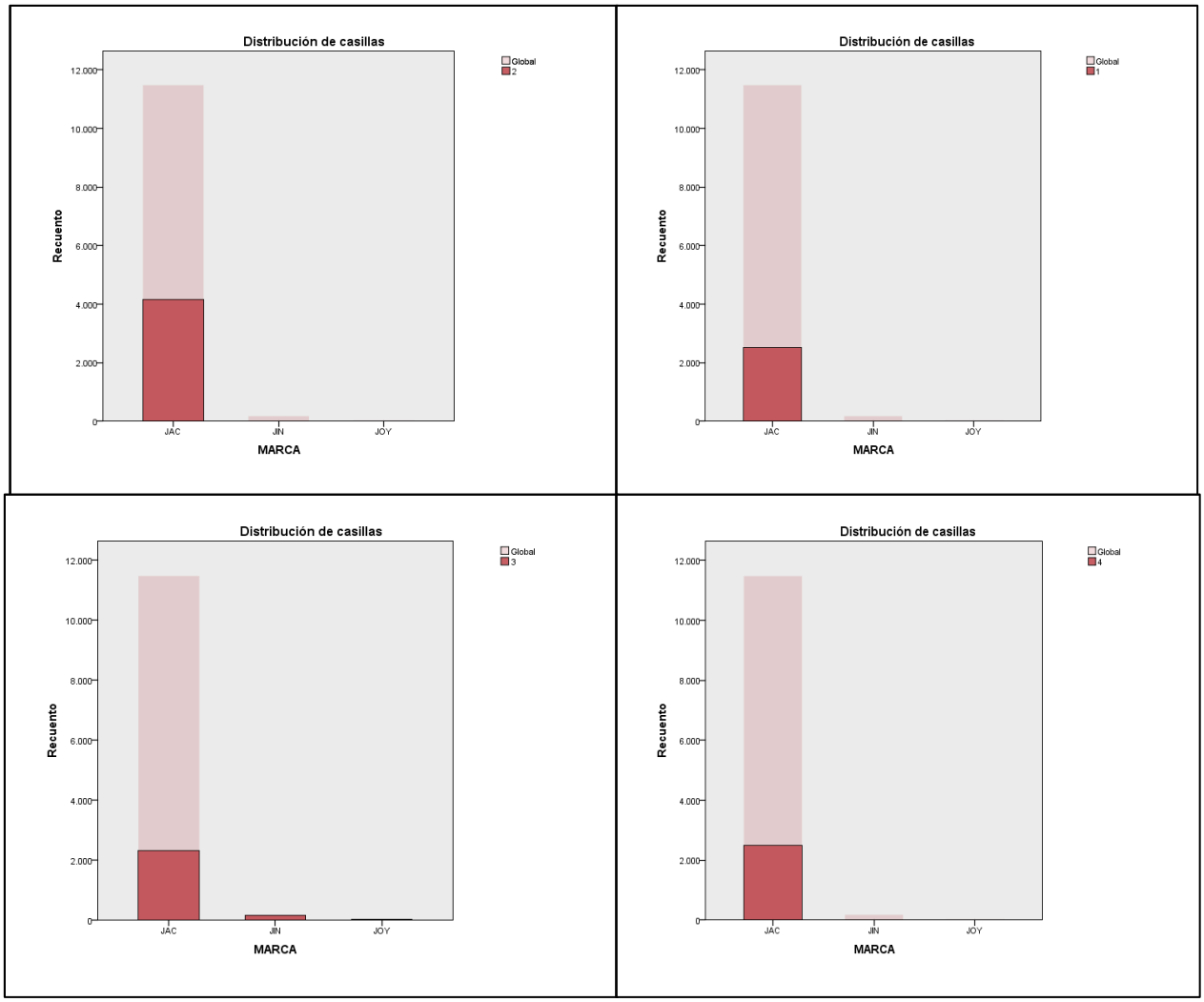


Figura 12. Marcas comercializadas en cada uno de los clúster generados

Se puede observar que la marca más posicionada en el mercado es JAC, las demás marcas JIN y JOY no presentan un balance de unidades comercializadas, competitivo con JAC, en este orden de ideas la compañía debe identificar nuevos versiones de vehículos en esta marca, dado que los clientes confían en ella para sus adquisiciones y pueden aceptar nuevos productos de la marca, ya que la han usado y les representa respaldo, de venta y soporte en materia de eventuales refacciones.

8 CONCLUSIONES Y DISCUSIÓN

- La minería de datos es una valiosa herramienta que nos permite extraer información de una base de datos, permitiendo un acceso más minucioso a todos los datos y variables, generando conglomerados con características homogéneas como se ilustro en los resultados obtenidos.

- Se determinó que las herramientas y técnicas de análisis en minería de datos permiten crear escenarios, de los cuales se puede obtener información útil para la toma de decisiones a nivel gerencial.

- Las técnicas que utiliza la minería de datos para la exploración consisten en la identificación de patrones y agrupar datos homogéneos creando clústeres, y a la vez permite realizar el respectivo análisis estadístico por cada clúster dependiendo de las características que tengan los datos.

- El proceso de la minería de datos genera conocimiento por medio de la depuración, enriquecimiento y transformación de datos que sirve para la creación de un modelo en el que se evalúa un conjunto de casos.

- El proceso de limpieza de datos nos ayudó a encontrar los datos erróneos, atípicos y las casillas sin datos, luego se exploro cada uno de los conglomerados para determinar las variables tienen más relevancia al momento de definir el número de unidades vendidas y las características de las mismas en los marcos del negocio de la empresa que es la venta de vehículos desde los automóviles hasta camiones de gran capacidad de carga.

- Con respecto a las ventas sobresalen las referencias HFC1063K, HFC1035K, HFC1035KD y HFC1061L, se puede observar que línea de vehículos de mayor venta es la HFC, al igual que la aceptación de los vehículos 1035K y su variante el 1935KD, por tanto se sugiere a la compañía determinar las características de

estas referencias, con el fin de ampliar las ventas de las referencias menos ofertadas.

- Los concesionarios con mayores ventas son “Calle 80 camiones” y “Motores SAS”, los demás concesionarios tienen niveles bajos de ventas, en estos es de vital importancia detectar e indagar el portafolio de vehículos ofertados y realizar un estudio para determinar la demanda de vehículos según el sector donde se encuentran los concesionarios y así implementar estrategias para aumentar las ventas.

9 REFERENCIA

Anderson, R. Sweeney, D. Williams, *T. Estadística para administración y economía*. México, D.F. CENAGE Learning. 10ª. Edición

Berreiro Fernandez, J. M., Diez de Castro, J. A., Ruzo, S. E., & Losada, P. F. (2003). *Gestión Científica Empresarial Temas de Investigación Actuales.* Coruña: NETBIBLO.

Canavos, G. (1998). *Probabilidad y Estadística Aplicaciones y métodos*. Mexico. The McGRAW – HILL

Estadística General. Apuntes. Elaborado por: Arturo Rubio Donet

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (ed.) (1996).*Advances in knowledge and data mining*. Cambridge (Massachussets): AAAI/MIT Press

Marcano Aular, Y. J., & Talavera Pereira, R. (2007). Mineria de Datos como soporte a la toma de decisiones empresariales. *Opción*, 104-1587.

Pérez, M. (2015). *Mineria de Datos A traves de Ejemplos,* Madrid España, Alfaomega

Pérez, C. (2004). *Técnicas de Ánalisis Multivariante de Datos Aplicaciones con SPSS,* Madrid España, Pearson Educación S.A

Pérez, C,. Santín D. (2004). *Minería de Datos Técnica y Herramientas,* Madrid España, Thomson Ediciones Paraninfo, S.A

Navidi, W. (2006). *Estadística para Ingenieros y Científicos*. Mexico, D.F. The McGRAW – HILL.

Uriel E., Aldas J.(2006). *Análisis Multivariante Aplicado*. Editorial Thompson

Webster, A. (2001). *Estadística aplicada a los negocios y la economía.* Tercera Edición. Colombia. Irwin McGraw – Hill