Universidad Cenfotec

Maestría en Tecnología de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

Desarrollo de un Cubo y Aplicación de Minería de Datos para el Análisis de Datos y Soporte a la Toma de Decisiones

Cordero Sancho, Edgar

Urdaneta Pulgar, Heber

Enero, 2014

DECLARATORIA DE DERECHOS DE AUTOR

DEDICATORIA

Dedicado a mi familia, y a aquella gente que lucha por sus sueños.

Edgar Cordero.

Dedicado a todas las personas que han sido importantes en mi vida y en mi formación académica, profesional y personal, que me han ayudado a crecer y alcanzar las metas planteadas, a mis amigos, mi novia, mis padres y familiares.

Heber Urdaneta.

AGRADECIMIENTO

Gracias a Dios por darnos la fuerza para lograr este gran objetivo. Le agradecemos profundamente, a todos nuestros amigos, compañeros de estudio y de trabajo, con quiénes hemos tenido experiencias y generado conocimientos importantes. A todos los profesores que han contribuido a nuestra formación académica. A nuestra tutora, que nos guió y ayudó en el transcurso de la investigación. Finalmente, a nuestros padres, de quiénes no tenemos palabras suficientes para expresar nuestros agradecimientos. Muchas gracias por su apoyo.

Los autores.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Hoy en día, gracias a los avances informáticos, las organizaciones disponen de grandes volúmenes de información. Esto en conjunto con la necesidad del personal gerencial de las organizaciones de tomar decisiones estratégicas, ha generado una creciente demanda de hacer un mejor uso de la información almacenada. Los datos de una organización representan un activo valioso, estos son la materia prima del conocimiento para el soporte a la toma de decisiones a nivel estratégico y gerencial.

Tomando esto en cuenta, es importante, estudiar las mejores opciones y tecnologías para aprovechar los datos almacenados, de manera que contribuyan con el bienestar de la organización, ayudando a cumplir los objetivos planteados.

La presente investigación tiene como finalidad el desarrollo de un cubo y la posterior aplicación de técnicas de minería de datos para el soporte a la toma de decisiones. El cubo fue desarrollado para una empresa de comercio electrónico que se dedica al procesamiento de pagos en línea, y por facilidad de ahora en adelante la denominaremos EPPL. Basado en el proceso de negocio principal de EPPL, se tomaron los requerimientos necesarios para realizar un análisis de la información que posee la organización para colaborar en el proceso de toma de decisiones.

La investigación, se encuentra enmarcada bajo la metodología “Business Dimensional Lifecycle” expuesta por Kimball, para el desarrollo del cubo, y bajo la metodología “Cross Industry Standard Process for Data Mining” para la aplicación de técnicas de minería de datos. La primera representa una alternativa efectiva y eficiente para abarcar todo el proceso de planificación, creación y mantenimiento de un cubo, y la segunda permite estructurar adecuadamente, un proyecto de minería de datos.

El estudio actual, se encuentra estructurado y organizado en cuatro (4) capítulos. En el primero se describen los planteamientos del problema y los objetivos; se justifica la investigación y se definen los alcances. El capítulo II es conformado por las bases teóricas, donde se desglosan las definiciones de interés en el marco de la investigación.

Seguidamente, en el capítulo III, se describen algunas técnicas e instrumentos de recolección de datos, se explican las metodologías seleccionadas y cada una de las fases aplicadas. Finalmente, el capítulo IV, donde se detalla el resultado de la investigación, se realiza el análisis de datos, describiendo cada una de las fases desarrolladas; y por último se discuten los resultados obtenidos, señalando las conclusiones y sus respectivas recomendaciones.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1. ANTECEDENTES

En los últimos años el comercio electrónico ha tomado un gran auge en lo referente a productos y servicios, EPPL se dedica al procesamiento de pagos en línea y ofrece servicios a través de un sistema, en el cual, los comercios se suscriben al servicio de pago; y de esta forma brindan a sus clientes la posibilidad de realizar sus transacciones a través de Internet, de forma segura.

Por tal motivo, EPPL tiene muy claro que día a día el comercio electrónico se incrementa a nivel mundial, sin embargo, en la región, las empresas o negocios pequeños no siempre están listos para aventurarse o arriesgarse en este ámbito; inclusive los bancos o agentes de pago no siempre brindan las facilidades para que estos negocios se incorporen a este escenario.

Es en este segmento de mercado donde EPPL colabora para garantizar un proceso ágil, seguro, confiable y accesible para que los comercios coticen sus servicios vía internet, garantizando alta disponibilidad, seguridad y escalabilidad.

El principal problema de EPPL recae en la diversidad de organizaciones y de mercados a los cuales brinda servicios, esto no permite tener una visión global clara; se tiene información de sistemas transaccionales, pero no a un nivel de toma de decisiones.

En el sistema actual que posee la empresa, se realizan reportes que consultan directamente la base de datos transaccional; acción que afecta el desempeño y rendimiento de la plataforma. Además, los reportes inicialmente, fueron construidos para el trabajo diario de las áreas operativas y no con el fin

de soportar la toma de decisiones de la gerencia organizacional.

Estos reportes también, son el centro de atención de los comercios que utilizan los servicios. Por esto se genera la necesidad de separar estos dos ambientes, el primero para garantizar un rendimiento óptimo en la plataforma transaccional; y el segundo, que colabore con el soporte a la toma de decisiones de EPPL.

2. JUSTIFICACIÓN

La necesidad de EPPL de tener una herramienta eficaz que le permita tomar decisiones oportunas en un nicho de mercado muy competitivo, se acrecienta a medida que pasa el tiempo y el mercado se vuelve más agresivo.

Por esto se plantea un sistema que permita posicionarse en el mercado, para analizar la información de sus operaciones y a su vez tomar decisiones inteligentes basadas en conocimiento, que los haga ser más competitivos y alcancen las metas de la organización.

En el último año EPPL ha estado teniendo dificultades a la hora de consultar la actividad transaccional. Los reportes creados funcionaron adecuadamente, por un tiempo, sin embargo, en las horas en las que el volumen transaccional es muy elevado, se optaba por no utilizar dichos reportes con el fin de no afectar el rendimiento de las aplicaciones, además, estos no cumplían las necesidades del entorno gerencial que los consultaba.

Las soluciones de inteligencia de negocios aumentan considerablemente, la agilidad empresarial, al ofrecer a grupos estratégicos importantes dentro y fuera de la organización puntos de vista específicos de la información corporativa, que se requiere para obtener el éxito. En este caso se optimiza la visualización de los datos y generación de conocimiento sin comprometer el desempeño de la plataforma.

3. Planteamiento del Problema

En esta sección se plantean los problemas generales y específicos que se desarrollan durante la investigación.

a. Problema General

¿Cómo desarrollar un cubo y aplicar técnicas de minería de datos para el análisis de datos y soporte a la toma de decisiones?.

b. Subproblemas

- ¿Cómo determinar cuál metodología se utilizará para construir el cubo?
- ¿Cómo establecer cuáles son los requerimientos del negocio?
- ¿Cómo seleccionar las tecnologías que serán utilizadas en la creación de la solución?
- ¿Cómo diseñar el cubo con base en requerimientos obtenidos?
- ¿Cómo desarrollar los reportes que servirán para el soporte a la toma de decisiones?
- ¿Cómo identificar los problemas específicos para el análisis con minería de datos?
- ¿Cómo encontrar patrones ocultos con técnicas de minería de datos con base en los problemas analizados?

c. Objetivo General

Desarrollar un cubo y aplicar técnicas de minería de datos para el análisis de datos y soporte a la toma de decisiones.

d. Objetivos Específicos

- Determinar una metodología para la construcción del cubo.
- Establecer los requerimientos del negocio.
- Identificar las tecnologías que serán utilizadas en la creación de la solución.
- Diseñar el cubo con base en los requerimientos obtenidos.
- Desarrollar los reportes que servirán para el soporte a la toma de decisiones.
- Identificar los problemas específicos para el análisis con minería de datos.
- Encontrar patrones ocultos a través de técnicas de minería de datos con base en los problemas analizados.

4. ALCANCES

Se diseña la solución mediante un cubo donde se integra la información de la base de datos relacional con la que cuenta EPPL. Ésta es la única fuente que es evaluada para determinar los datos que se cargan al cubo. No se aplica un proceso de limpieza de datos.

Con base en los requerimientos del negocio se diseñan y desarrollan algunos reportes, estos brindan la posibilidad al usuario de analizar por demanda dinámicamente los datos, según sea el interés para la toma de decisiones. Los indicadores de desempeño a implementar son los de mayor importancia en proceso de negocio que se está evaluando.

La arquitectura física no será implementada, es decir, EPPL debe de adquirir e implementar el hardware y software necesarios. Por ende, la solución no será puesta en una plataforma de producción.

No se definirá una estrategia de índices, particionamiento o tareas relacionadas a la administración del almacén de datos, estas especificaciones quedarán a cargo de EPPL en la plataforma que sea implementada la solución.

Con respecto a la utilización de la solución no se considera el entrenamiento al personal operativo y gerencial que utilizará las herramientas de consultas y reportes.

El soporte a la toma de decisiones está delimitado al área de negocio operativa, y no a otras áreas, limitando así que la toma de decisiones abarque la totalidad de la organización.

En lo que se refiere a la minería de datos, se va a realizar un proceso de una iteración, debido a que es un proceso exhaustivo que requiere de múltiples fases de aprendizaje, entrenamiento y evaluación. Básicamente, lo que se desea es efectuar una exploración que siente las bases, para explotar en un mediano plazo la información que pueda ser aprovechada por la organización.

A nivel de minería, el enfoque se dará con el fin de examinar un problema del negocio, e intentar explotar la información recolectada que gira en torno a dicho problema para buscar soluciones alternas al mismo.

5. LIMITACIONES

El cubo a desarrollar brindará la información generada a partir de lo que se encuentre presente en el sistema transaccional, estos datos pasarán por ciertas transformaciones, pero, no se cuenta con el tiempo para implementar un proceso de limpieza o mecanismo de integración. Por lo tanto, los resultados obtenidos dependerán de los datos encontrados en las fuentes.

Los datos utilizados en todo el desarrollo de la investigación presentan limitaciones en el ámbito de la confidencialidad con los comercios, es debido a esto que no se presentan datos muy detallados sobre los productos, servicios y clientes de cada comercio.

Los datos que EPPL brinda para esta investigación, están limitados por contratos de confidencialidad, por lo que, algunos datos mostrados han sido modificados con el fin de preservar la integridad de dicha confidencialidad.

CAPÍTULO II

MARCO TEÓRICO

A continuación, se presentan una serie de elementos teóricos que permiten la sustentación de la investigación, específicamente, de los conceptos principales de estudio “Inteligencia de Negocios”, “Data Warehouse”, “Minería de Datos” y “Comercio Electrónico”; la definición de términos básicos y la funcionalidad de cada variable aplicada y relacionada a la investigación.

1. INTELIGENCIA DE NEGOCIOS.

En cuanto al aprovechamiento de la inteligencia de negocios, esta misma se ha convertido en una herramienta importante que puede ayudar a las organizaciones a obtener conocimiento, es así como (Rahman El Sheikh, 2011) la describen:

“La inteligencia de negocios puede ayudar a las organizaciones a administrar, desarrollar y comunicar su conocimiento e información. Por lo tanto, puede ser considerado como un enfoque imperativo en esta era del conocimiento. Estas aplicaciones son principalmente, definidas por la flexibilidad y adaptabilidad, en el que las aplicaciones tradicionales casi siempre fallan, el proceso tradicional en general involucra reportes estáticos y mucha documentación, siendo el enfoque tradicional incapaz de cumplir con los requerimientos dinámicos en un ambiente cambiante”.

El autor, (Kimball, The Data Warehouse Toolkit, 2002) La describe brevemente, indicando que es un término genérico para referirse al aprovechamiento de la información interna y externa de una organización para realizar mejores decisiones de negocio.

En este caso, para el proyecto propuesto la idea es estudiar los datos provenientes del sistema transaccional, en el que se cuenta con reportes que no son capaces de cumplir con las necesidades de la gerencia, y de esta manera complementarlos con aplicaciones de inteligencia de negocios para aprovechar al máximo la información almacenada en el mismo, brindando así a la organización información y conocimiento que puedan diferenciarlos en el mercado.

2. DATA WAREHOUSE

Al respecto, (Inmon, 2005) Define los “Data Warehouses” (bodegas de datos) como: “Una colección de datos orientados a temas, integrados, no volátiles y variante en el tiempo, organizados para soportar decisiones empresariales”. Nuestro proyecto se enfoca específicamente, al área de procesamiento transaccional de la EPPL.

Por otro lado, el enfoque de (Kimball, The Data Warehouse Toolkit, 2002) expresa que un Data Warehouse es la conglomeración de los almacenes de datos, áreas de "staging" y áreas de presentación de una organización, donde los datos operacionales son específicamente, estructurados para consultas y análisis de rendimiento.

2.1. SISTEMAS FUENTE

(Kimball, The Data Warehouse Toolkit, 2002) Define los sistemas fuente como sistemas operacionales cuyas funciones son realizar transacciones o cualquier otra métrica del rendimiento de un proceso del negocio. Los sistemas fuente pueden ser externos a la organización, pero, pueden tener datos necesarios en el "data warehouse".

2.2. DATA MART

Es un segmento lógico y físico de un "data warehouse". En un principio, se define como datos agregados, regularmente orientados a responder una pregunta específica. Ahora generalmente, se refiere a un conjunto de datos flexible, siendo lo más atómico posible y presentado con frecuencia, en un modelo dimensional para soportar consultas de usuarios de forma más eficiente. Un "data mart" representa datos de un proceso de negocios específico. (Kimball, The Data Warehouse Toolkit, 2002)

Para (Inmon, 2005), un "data mart" es una estructura de datos orientada a un departamento, donde los datos están desnormalizados basándose en las necesidades de información del departamento. Por ejemplo, en nuestro caso un área del negocio que puede ser vista desde esta perspectiva corresponde al procesamiento transaccional, otra perspectiva es el área financiera contable, ambos departamentos distintos en EPPL.

2.3. CUBO

Es el nombre que recibe una estructura multidimensional o una plataforma de base de datos para el procesamiento analítico en línea conocido como OLAP, originalmente, se refería a un típico y simple caso de tres dimensiones en el que se analizaban producto, tienda y tiempo. (Kimball, The Data Warehouse Toolkit, 2002)

2.3.1. DIMENSIONES

En relación a este tema el autor, (Kimball, The Data Warehouse Toolkit, 2002) argumenta que las dimensiones son entidades independientes de un modelo dimensional que sirven como un punto de entrada o como un mecanismo de filtrado para las medidas localizadas en la tabla de hechos del modelo dimensional.

2.3.2. TABLA DE HECHOS

En un esquema de estrellas, la tabla de hechos es la tabla central con las medidas, en general, numéricas, y compuesta por llaves foráneas de las tablas de dimensiones. (Kimball, The Data Warehouse Toolkit, 2002)

2.3.3. MEDIDAS

Las medidas son típicamente, números o campos aditivos que están almacenados en las tablas de hecho, estos permiten describir un comportamiento o campo de un proceso de negocio. (Kimball, The Data Warehouse Toolkit, 2002)

2.4. ETL

ETL es definido por (Kimball, The Data Warehouse Toolkit, 2002) como un conjunto de procesos en el que la fuente operacional de datos es preparada para el "data warehouse". Son los principales procesos en la preparación de un "data warehouse", antes que cualquier presentación de datos o consultas. Consiste en extraer los datos de una fuente, transformarlos, cargarlos e indexarlos, asegurando la calidad e integridad de los mismos.

También, el investigador, (Inmon, 2005) Lo define brevemente, como un proceso en el que se encuentran los datos, luego de varias fuentes se integra o se transforman, y luego se ubican dentro del "data warehouse".

3. MINERÍA DE DATOS

Tal como señalan (Fayyad U. M., 1997), la gran cantidad de datos que se almacenan en las organizaciones hace imposible la utilización de métodos manuales para su análisis. Por ello son necesarias técnicas y herramientas informáticas capaces de ayudar de forma inteligente y automática en el análisis de grandes cantidades de datos.

Estos mismos autores, definen minería de la siguiente forma:

"Minería de datos es un proceso no trivial de identificación de patrones de datos válidos, nuevos, potencialmente usables y comprensibles", es decir, se trata de un proceso concreto, específico, con un objetivo, que busca identificar repeticiones y/o tendencias en un conjunto de datos que resulten útiles y sean veraces.

Para el autor, (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004), en esta definición se resumen las propiedades del conocimiento extraído: válido, novedoso, potencialmente útil y comprensible.

Así pues, la minería de datos es un campo que se compone de otras disciplinas como la estadística, ciencias de la computación y la inteligencia artificial. De este modo, la minería de datos supone una reivindicación del valor de las fuentes de datos estadísticas para la gestión de la información.

3.1. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de minería de datos permiten plantear los problemas de la forma más adecuada, hay que recordar que cuantas más variables entran en el problema, más difícil resulta encontrar hipótesis de partida interesantes. O, aun cuando pudiera, el tiempo necesario no justificará la inversión.

Según, indica (Agrawal, 1993):

"La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente, patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada"

De esta forma, las técnicas se basan en algoritmos estadísticos que se clasifican en dos grandes categorías:

- Supervisados o predictivos.
- No supervisados o de descubrimiento del conocimiento.

Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otras series de atributos.

Cuando una aplicación no es lo suficientemente madura, es decir, no tiene el potencial necesario para ser una solución predictiva, se debe recurrir a los métodos no supervisados o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales.

3.1.1. REGLAS DE ASOCIACIÓN

Las reglas de asociación son el producto del descubrimiento de relaciones de asociación o correlación en un conjunto de datos. Las asociaciones se expresan como condiciones atributo – valor y deben estar presentes varias veces en los datos.

Indican la fuerza de la asociación de dos o más atributos de datos. El interés en las reglas de asociación es que ellas entregan la promesa (o ilusión) de causalidad, o al menos de relaciones predictivas. Sin embargo, las reglas de asociación sólo calculan la frecuencia de ocurrencias de uniones de dos o más atributos de datos; ellas no expresan una relación causal. Para (Meyer, 1998):

"Derivan de un tipo de análisis que extraen información por coincidencias. Este análisis a veces llamado "cesta de la compra" permite descubrir correlaciones o co-ocurrencias en los sucesos de la base de datos a analizar y se formaliza en la obtención de reglas".

3.1.2. ÁRBOLES DE DECISIÓN

Para el autor, (Bramer, 2007), los árboles de decisión conforman un método para construir un modelo a partir de un conjunto de datos en forma de un árbol, que contiene reglas de decisión. Es a menudo considerado como una representación de los datos con cierta ventaja sobre otras técnicas, ya que, se considera significante y fácil de interpretar.

El árbol es creado a partir de un proceso conocido como “splitting on attributes”, que consiste en probar el valor de un atributo y crear una rama para cada uno de los posibles valores. En el caso de los valores continuos se prueba normalmente, si el valor es mayor o igual a un valor conocido de referencia. El proceso continúa hasta que cada rama sea etiquetada con una clasificación.

3.1.3. BAYES INOCENTE

Para (Bramer, 2007), la técnica Bayes inocente de minería de datos, es una técnica que no utiliza reglas, árboles de decisiones u otras representaciones explícitas. En lugar de esto se vale de la rama de la matemática conocida como, probabilidad teórica, para encontrar lo más probable de las posibles clasificaciones.

Es necesario, señalar, que la palabra inocente, hace referencia a la premisa que el método hace que el efecto del valor de un atributo en la probabilidad de una clasificación es independiente de los valores de otros atributos.

Este algoritmo proporciona una manera de combinar probabilidad y condiciones posibles en una simple fórmula, que puede usarse para calcular la probabilidad de cada posible clasificación. Al hacer esto se escoge la clasificación con el valor mayor.

3.1.4. CLUSTERS

Para este mismo autor, (Bramer, 2007), “Clustering” es una técnica que se concentra en el agrupamiento de objetos que comparten características similares a otros objetos y características no similares a objetos que pertenecen a otros “clusters”.

Asimismo, existen varios algoritmos para realizar “clusters”. Estos se basan en la similitud entre los objetos basándose en la distancia entre ellos. Usualmente, es fácil de visualizar los “clusters” en gráficos de dos dimensiones y generalmente, lo importante, es analizar las características que hacen que los puntos se agrupen. Antes de decidir cuál algoritmo se utilizará para agrupar los objetos es significativo decidir primero cómo se medirá la distancia entre los puntos.

4. E-COMMERCE Y SISTEMAS DE PAGOS.

Para (Turban, Rainer, & E. Potter, 2005) el comercio electrónico describe la compra, venta e intercambio de productos, servicios e información a través de redes computacionales, principalmente, la Internet. Algunas personas ven el término de comercio como un proceso de transacciones entre dos socios de negocios.

Además, para (Khosrow-Pour & Khosrowpour, 2006), se refiere a la transacción de bienes y servicios por medio de comunicaciones electrónicas. Existen dos tipos principales de comercio electrónico, siendo estos "business to business" y "business to consumer".

Según, (Chan, Lee, Dillon, & Chang, 2001) el comercio electrónico tiene como meta maximizar el valor de los clientes redefiniendo el modelo de negocios utilizando tecnologías de la información. (Gary & James, 2000) ven el comercio electrónico como una vía efectiva para mejor la cadena de valor utilizada en varias actividades funcionales de una compañía, facilitando los flujos de información para reducir los costos asociados a las transacciones. Para esto, es preciso, tener en el punto de vista tres relaciones: La relación del negocio con los clientes, con los socios comerciales y con los empleados. El "e-commerce" provee una herramienta efectiva para construir, administrar y mejorar estas relaciones.

4.1. TIPOS DE E-COMMERCE

Los autores generalmente, coinciden en dividir el comercio electrónico en dos principales categorías:

4.1.1. NEGOCIO-CONSUMIDOR

Para (Chan, Lee, Dillon, & Chang, 2001), en este tipo de comercio electrónico el vendedor es una organización o negocio y el comprador es el

consumidor, esto emula una venta física corriente y es llamado comúnmente, venta electrónica. Típicamente, los negocios montan en internet tiendas electrónicas para vender sus bienes a los consumidores. Para (Khosrow-Pour & Khosrowpour, 2006) es el tipo de comercio en que las compañías venden productos y servicios directamente a los clientes, siendo este abierto al público. "business to business" es la forma que actualmente, domina el mercado del "e-commerce" en términos de ganancias.

4.1.2. NEGOCIO-NEGOCIO

Para (Khosrow-Pour & Khosrowpour, 2006) es el tipo en que las compañías venden productos y servicios a otros socios de negocios, distribuidores, u otros. Mientras que, para (Turban, Rainer, & E. Potter, 2005) el "business to business" es el tipo en que dos o más negocios hacen transacciones o colaboran electrónicamente, este es el que más ganancias genera globalmente.

4.2. SISTEMAS DE PAGOS

Para (Nakajima, 2011) un sistema de pagos es un mecanismo que permite que los fondos se transfieran sin problemas entre el comprador, el vendedor y los bancos. Este consiste de instrumentos, procedimientos bancarios y típicamente, sistemas de transferencias interbancarias que aseguran la circulación del dinero.

También,(Montague, 2011), explica que la mayoría de las personas piensa que los pagos en línea son únicamente, pagos directos con tarjetas de crédito, es cierto que representan la mayoría de los tipos de pagos, pero, existen otras opciones disponibles.

4.2.1. ENTIDADES INVOLUCRADAS

A continuación, se describen algunas de las entidades involucradas en un sistema de pagos.

4.2.1.1. TARJETA DE CRÉDITO

Una tarjeta de crédito es una parte de un sistema de pagos que permite que el poseedor de la tarjeta compre bienes y servicios basados en la premisa de que el cliente pagará por ellos. El banco emisor de la tarjeta le da una línea de crédito al consumidor por medio del cual puede utilizar dinero a préstamo para pagarle a un comercio. Generalmente, las compañías de tarjetas de crédito obtienen una comisión que varía dependiendo del país.

4.2.1.2. COMERCIOS

Además, (Montague, 2011) define a los comercios o "merchants" como los que poseen los bienes y servicios y están buscando vendérselos a los consumidores o negocios. Los consumidores se ven motivados a seleccionar un comercio en particular basándose en varios aspectos como precios, servicios, o preferencias. El propósito principal de los comercios consiste en hacer dinero, vendiendo sus bienes y servicios, es importante, que tengan varias formas para realizar los pagos como cheques, dinero efectivo, tarjetas de débito o crédito. Los comercios y los consumidores han aceptado las tarjetas de crédito como la forma de pago dominante por su facilidad de uso. Para (Bradley, 2007) los comercios son compañías que aceptan pagos de tarjetas de crédito, a cambio de bienes o servicios.

4.2.1.3. BANCOS

Los bancos, son divididos de dos formas, bancos adquirientes y bancos emisores.

Para (Bradley, 2007) el banco adquiriente es el banco que procesa las transacciones de un comercio, puede ser una marca de tarjetas. (Montague, 2011) los identifica como los que representan al comercio, procesando todos los pagos por tarjetas de crédito que el comercio autorice. Estos almacenan el dinero de cada transacción procesada.

Asimismo, (Bradley, 2007) expone que los bancos emisores son las instituciones que emiten las tarjetas de crédito utilizadas por los clientes. Del mismo modo, (Montague, 2011) lo define como el banco que proporciona al cliente la tarjeta de crédito. Estas son instituciones que trabajan detrás de estas tarjetas de crédito y administran los créditos. El propósito de estos bancos es garantizar el crédito directo al consumidor, definiendo los límites basándose en el historial crediticio y las deudas actuales.

4.3. MÉTODOS DE PAGOS

Para definir algunas de las formas de pago por tarjeta de crédito alternas a la comunicación directa entre el banco y el comercio, se dejan definiciones de (Montague, 2011).

4.3.1. AGREGADORES DE PAGO

Los agregadores de pago o comercios maestros son proveedores de servicios para comercios electrónicos que procesan sus transacciones. En algunos casos estos pueden incluso tener la cuenta de los comercios ante los procesadores y bancos.

Estas compañías tienen la información de tarjetas de crédito para permitir realizar compras más rápidas o tener dinero en una cuenta para futuras compras. Estos son atractivos para los comercios que tienen dificultades en abrir cuentas con un banco, ya que, al tener un comercio maestro ese será el que administre las cuentas de los sub comercios correspondientes. Son de los métodos de pago más aceptados entre los comercios electrónicos.

4.3.2. PROCESADORES DE PAGO

No hay nada que impida que los comercios se conecten directamente, con los bancos adquirientes, pero, existen varias razones por las que los comercios no quieren o pueden. Existen varios requerimientos técnicos y de negocio para conducir el procesamiento del pago de tarjetas de crédito y la mayoría de los comercios no quieren preocuparse por estos detalles. En lugar de esto escogen un servicio de terceros entre ellos y los bancos llamados procesadores de pagos. Estos ofrecen una arquitectura física para que los comercios se comuniquen con los bancos adquirientes.

Además, son los encargados de conectar todas las entidades, estos permiten que algunos bancos pequeños ofrezcan servicios a los comercios que normalmente, no hubiesen podido.

Los procesadores de pago hacen dinero cobrando una comisión por transacción a los comercios, a cambio de proporcionar la infraestructura y el servicio de redirigir las transacciones a los bancos para que sea procesadas.

4.4. OTROS TIPOS DE TRANSACCIONES

4.4.1. CRÉDITO

El crédito es un tipo de transacción que ocurre cuando el consumidor devuelve el bien o servicio adquirido. Cuando se procesó la transacción, se

realizó una autorización, se hizo el asentamiento y el dinero fue transferido, debe ejecutarse una devolución del dinero al consumidor. Esto es realizado procesando un crédito. Cuando esto se efectúa el banco adquiriente pasará el dinero de vuelta a la tarjeta de crédito del consumidor.

4.4.2. CHARGEBACK

Los "chargebacks" ocurren cuando los consumidores contactan a sus bancos emisores para indicar que ellos no realizan las órdenes o no reciben lo ordenado. Hay dos categorías de "chargeback", la primera es fraudulenta y consiste en que los consumidores explican que no hicieron la orden o no recibieron los bienes o servicios, y la otra es cuando los consumidores admiten que hicieron la orden, pero, entran en disputa por cargos u otras razones. Estos ocasionan problemas para los comercios, ya que, cada vez que los bancos emisores se contactan con ellos para notificarles el problema ocurrido, existen comisiones que los comercios deben pagar.

CAPÍTULO III

MARCO METODOLÓGICO

En este capítulo se describen los aspectos metodológicos de esta investigación. Se detalla el tipo de investigación, la técnica de recolección de datos y las metodologías aplicadas, desglosando cada una de las fases utilizadas.

1. TIPO DE INVESTIGACIÓN

De acuerdo con el autor (Hernández Sampieri, 1997), la metodología en el proceso de investigación son los pasos o etapas que expresan a profundidad los datos y los contextualizan en el ambiente, en este caso conocer el ambiente es muy importante, pues, es la base para lograr que las necesidades de toma de decisiones de cada usuario se cumplan. Esta sección describe la forma en que se va a llevar a cabo la investigación.

En este estudio, se establece bajo un enfoque cualitativo porque, los datos que se reúnen son recolectados en forma de texto basados en características o cualidades del procesamiento transaccional de la empresa EPPL, y se realiza un período de análisis para obtener los datos relevantes que sean útiles en el proceso de toma de decisiones.

Además, se define que el diseño de esta investigación es evaluativo, debido a que está orientado a examinar los procesos de toma de decisiones a través de una solución de inteligencia de negocios, así, se buscan nuevas ideas en la toma de decisiones teniendo en cuenta información relevante.

Para profundizar más en las categorías de investigación en el ámbito cualitativo, el diseño evaluativo presenta las características que se desean en el desarrollo del proyecto, según, el estudioso (Sandín, 2003) la investigación

evaluativa es decisiva para la toma de decisiones y está orientada a determinar la eficacia de organizaciones. Si se retrocede y verifica los objetivos de esta investigación se encuentra, que la toma de decisiones es parte fundamental de la misma, y los métodos evaluativos complementan estos objetivos así como se expresa (Cabrera, 1987):

"La investigación evaluativa se trata de una forma de investigación pedagógica aplicada que tiene por objetivo valorar la eficacia o éxito de un programa de acuerdo a unos criterios y todo ello en orden a tomar decisiones presumiblemente optimizantes la situación".

Como puntualiza Cabrera, cuando la evaluación tiene por objeto valorar la eficacia ya sea de algún elemento, del proceso, o de un programa en su totalidad, tiene el significado de investigación evaluativa. Es decir, este término es utilizado con el propósito de precisar que determinar el valor de los fenómenos exige un proceso sistemático y riguroso que aporte evidencias basadas en dicho proceso y no debidas meramente a la intuición (Cabrera, 1987).

2. RECOLECCIÓN DE DATOS

A continuación se analizan algunos conceptos relacionados a la recolección de datos.

2.1. POBLACIÓN Y MUESTREO

En la realidad hay que tener en cuenta que nunca es posible estudiar la totalidad de la población, por ello el muestreo es una parte importante que se realiza en el proceso de investigación.

El muestreo como lo define (Piergiorgio, 2007), es el procedimiento, por

el cual, de un conjunto de unidades que forman el objeto de estudio (población), se elige un número reducido de unidades (muestra) aplicando criterios tales que permitan generalizar los resultados obtenidos del estudio de la muestra de la población.

Para nuestro caso específico, la muestra la integran los usuarios que consultan los reportes y toman decisiones en la compañía que son tres personas, al ser una población pequeña, la muestra incluye a toda la población, siendo esta realmente significativa.

2.2. Recopilación de datos

Según, (Piergiorgio, 2007), para la información y datos que se van a recolectar se ingresa al paradigma interpretativo, es decir, desde el punto de vista conceptual las técnicas de análisis cualitativo no difieren mucho entre sí, por ejemplo, la investigación de campo, estudios empresariales, observación participante, entrevistas libres, entrevistas no estructuradas, etc. son todas técnicas de recopilación similares.

Se caracterizan por utilizarse simultáneamente, en varias etapas de la investigación, por último es difícil dividir el procedimiento de la investigación cualitativa en fases separadas y bien diferenciadas entre sí, las fases de recopilación y análisis se alternan. Por lo tanto y según lo confirma (Bryman, 1994), en la investigación cualitativa hay que señalar más de técnicas o de fases que de proceso de investigación, "la investigación cualitativa no puede reducirse a técnicas específicas, ni a una sucesión de estados, sino que consiste más bien en un proceso dinámico que une problemas, teorías y métodos".

Del mismo modo, menciona (Piergiorgio, 2007), una de las diferencias principales entre el enfoque cualitativo y cuantitativo está relacionado con el diseño de la investigación, es decir, dónde cómo y cuándo se recopilan los

datos, los instrumentos usados para la investigación, entrevistas u observación participante, cuestionarios o experimentos, etc., la localización, cuáles y cuántos sujetos se entrevistarán, cuáles y cuántas organizaciones se estudiarán, etc. En la investigación cualitativa, el diseño no tiene una estructura fija, es abierto, de modo que permita captar lo imprevisto en el curso del proceso.

2.3. INSTRUMENTOS PARA RECOLECCIÓN DE DATOS

Las técnicas de recopilación de datos de la investigación cualitativa se agrupar en tres categorías basadas en la observación directa, las entrevistas y el uso de documentos, en síntesis: observar, preguntar y leer. Dada esta premisa se concluye que para este proyecto se requiere de este conjunto de técnicas.

2.3.1. OBSERVACIÓN PARTICIPANTE

Según (Piergiorgio, 2007) se caracteriza en que existe la intervención directa del investigador en el objeto estudiado, puesto que implica mirar y escuchar, pero, al mismo tiempo conlleva contacto personal e intenso entre el sujeto que estudia y el sujeto estudiado, básicamente, el investigador participa en el día a día de los sujetos estudiados mientras dure el estudio. Esta es la principal característica, ni en las entrevistas estructuradas, el análisis de las fuentes estadísticas ni en el experimento, ni en el análisis de documentos o en entrevistas a profundidad, el entrevistador participa en el fenómeno estudiado, en la observación participante se "baja al campo", se adentra en el contexto que quiere estudiar, vive como y con las personas objeto de estudio, comparte con ellas la cotidianidad, les pregunta, descubre preocupaciones y desarrolla una visión "desde adentro".

2.3.2. ENTREVISTA A LOS USUARIOS DE REPORTES Y GERENTES DE LA COMPAÑÍA

Según, (Piergiorgio, 2007) la entrevista cualitativa se considera equivalente a la observación participante en cuanto a los objetos del investigador, claramente la inmersión en el mundo del objeto no es profunda, ni se pretende que sea así. La entrevista cualitativa se define como una conversación provocada por el entrevistador, realizada a sujetos seleccionados a partir de un plan de investigación en un número considerable; que tiene una finalidad cognitiva; guiada por el entrevistador y con un esquema de preguntas flexible y no estandarizado.

2.4. LECTURA DE INFORMACIÓN

En general, el mundo produce gran cantidad de documentos y en este caso se observa cómo se utiliza para complementar una investigación que se basa en la poca experiencia que tiene EPPL en materia del tema que se quiere abordar, en este caso existen varias técnicas que se abordan, pero primero hay que definir que es un documento, según (Piergiorgio, 2007), un documento es un material informativo sobre un determinado fenómeno que existe con independencia de la acción del investigador, estos documentos son generados con fines distintos a los de la investigación, entre ellos se encuentran, las cartas, periódicos, balances de empresas, manuales, tutoriales, estos documentos presentan una gran ventaja pues, son independientes de cualquier investigación y en general sirven como un marco para enfocarse en áreas de interés bajo visiones que se pueden considerar “particulares”.

Fundamentalmente, esta investigación se basa en documentos empresariales manuales generados por EPPL.

2.5. ANÁLISIS DE LA INFORMACIÓN QUE SE RECOLECTARÁ

El análisis de la información es probablemente, una de las partes más difíciles de la investigación, ya que, se debe de transformar los productos de las técnicas utilizadas que se encuentran en el lenguaje de los sujetos analizados en categorías conceptuales.

El análisis cualitativo de los datos como explica (Piergiorgio, 2007) se centra en los sujetos y no en las variables, es decir, el individuo es observado y estudiado en su totalidad, con la convicción de que cada ser humano es algo más que la suma de sus partes, en este caso el objetivo es comprender a las personas, lo que hacen y lo que esperan de ello, más que analizar las relaciones entre las variables. Por lo que se expresan sus sentidos en forma de requerimientos o deseos para que su experiencia de análisis a través del sistema de toma de decisiones sea el deseado.

Inicialmente, la presentación de estos resultados se realizara en forma de narración mediante la descripción de casos que presentan cada uno de los entrevistados.

3. METODOLOGÍA SELECCIONADA

En esta investigación, se toma la metodología expuesta por Kimball en su libro "The Data Warehouse Toolkit", como una opción válida y eficiente para desarrollar un proyecto de inteligencia de negocios. El enfoque dimensional permite que el desarrollo de un "data warehouse" sea orientado a procesos de negocios, optimizando los tiempos de entregas y promoviendo así mejores resultados, enfocados en la comunicación entre el grupo de trabajo y el cliente.

Para el análisis de los datos mediante técnicas de minería, se utiliza la metodología "Cross Industry Standard Process for Data Mining", que a través de sus fases permite identificar y estructurar los procesos necesarios para

obtener resultados óptimos en las aplicaciones de las técnicas.

A continuación, se definen las fases que componen las metodologías.

3.1. BUSINESS DIMENSIONAL LIFECYCLE

La metodología propuesta por Kimball, se compone de varios procesos, se muestran en el diagrama de la figura 1. Posteriormente, las fases son explicadas.

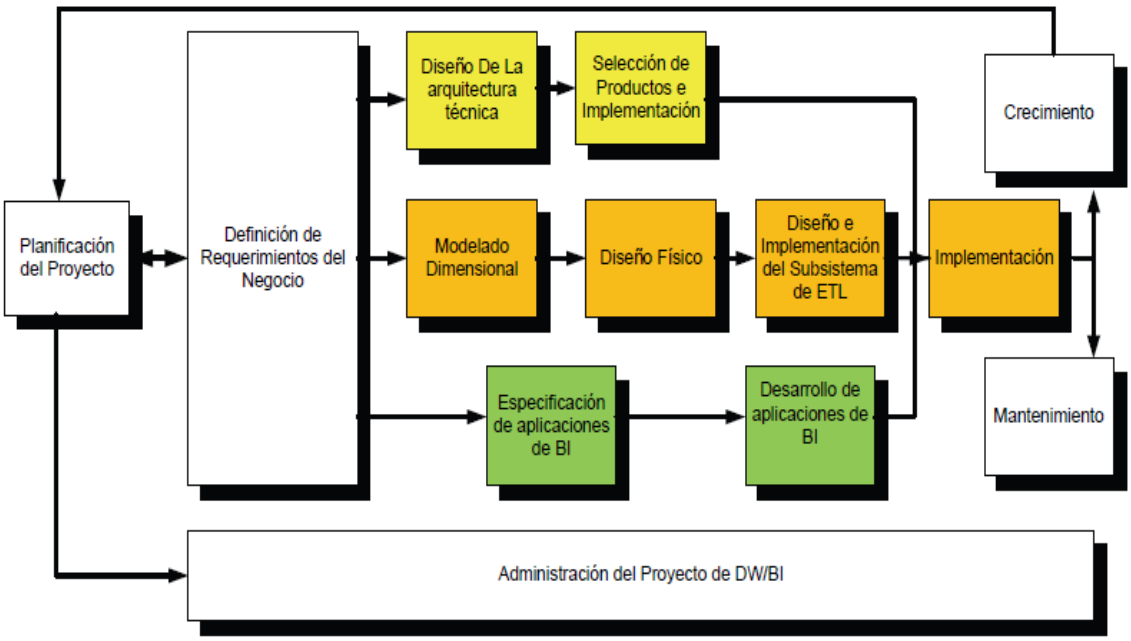


FIGURA 1: BUSINESS DIMENSIONAL LIFECYCLE.

a. PLANIFICACIÓN DEL PROYECTO

El ciclo comienza con la planificación del proyecto, en esta fase se busca identificar la justificación y el alcance que tiene el proyecto. Esta etapa se concentra sobre la definición del proyecto, obtención de recursos y lanzamiento del proyecto. También, se incluyen actividades como identificar y programar tareas, uso de los recursos y plan de proyecto.

b. DEFINICIÓN DE LOS REQUERIMIENTOS DEL NEGOCIO

La segunda etapa se concentra en la definición de los requerimientos, se tiene una flecha de dos flujos entre la planificación del proyecto y los requerimientos del negocio porque debe existir mucha interacción entre estos dos procesos. Este es un proceso crucial, ya que, el "data warehouse" está alineado con los requerimientos del negocio. Los diseñadores deben entender muy bien las necesidades para plasmarlas en el diseño. Este proceso tiene impacto en cómo se diseña y desarrolla el proyecto. Es recomendable, leer todos los informes de la organización, entrevistar a los empleados e interesados del negocio.

c. DISEÑO DE LA ARQUITECTURA TÉCNICA

Esta fase establece la infraestructura para soportar la integración de las tecnologías. Comprende tres factores: los requerimientos de negocio, los actuales entornos técnicos, las directrices técnicas y estratégicas futuras planificadas por la compañía, lo que permitirá establecer el diseño de la arquitectura técnica. En la arquitectura se identifican los componentes necesarios y se evitan problemas futuros a la hora de implementar el proyecto.

d. SELECCIÓN DE PRODUCTOS E INSTALACIÓN

En esta fase se evalúa y selecciona cuáles son los componentes necesarios específicos de la arquitectura. Algunas recomendaciones son comprender el proceso necesario de EPPL para realizar compras, desarrollar una matriz de evaluación de productos, efectuar una investigación de los productos en el mercado, y negociar el producto elegido.

e. MODELADO DIMENSIONAL

Luego de definir los requerimientos del negocio, se elabora la matriz del "data warehouse" que permite ayudar a definir procesos de negocios y sus respectivas dimensiones. Se recomienda seguir cuatro pasos; definir los procesos del negocio involucrados, determinar la granularidad que tendrá el diseño del cubo, luego de esto, se identifican las dimensiones relacionadas a estos procesos de negocio, y finalmente, se determinan las tablas de hechos y sus medidas.

f. DISEÑO FÍSICO

El diseño lógico creado anteriormente, es traspasado a un diseño físico. El diseño físico se parecerá mucho. Un elemento principal de este proceso es la definición de estándares del entorno de la base de datos. Los índices, agregaciones y las estrategias de particionamiento se determinan en esta etapa.

g. ESPECIFICACIÓN DE Y DESARROLLO DE LA PRESENTACIÓN DE DATOS

Esta fase se refiere al diseño y desarrollo del área de "staging" o procesos de ETL. Esto consiste en tomar los datos de los sistemas transaccionales y prepararlos para el modelo dimensional. Estos procesos influirán en la calidad de datos del proyecto, particularmente, el proceso de transformación debe ser el adecuado para combinar datos, resolver problemas de calidad, identificar datos ya actualizados y manejar errores de la manera más adecuada.

h. Especificación de Aplicaciones analíticas

En esta última ruta se diseñan aplicaciones que cumplen con los requerimientos de los usuarios analíticos. En esta fase, es necesario, revisar algunos reportes que utilicen los usuarios y comenzar a diseñar múltiples reportes. Es recomendable, seguir un estándar para las aplicaciones, así como también, facilitar accesos estructurados y portales para satisfacer las necesidades de los usuarios. De la misma manera, se identifican los roles o perfiles de usuarios para los diferentes tipos de aplicaciones necesarias con base en el alcance de los perfiles detectados.

i. Desarrollo de aplicaciones analíticas

Esta fase se refiere al desarrollo de las aplicaciones analíticas diseñadas, se recomienda nuevamente concentrarse en los estándares. El desarrollo comienza una vez el diseño de los datos, los accesos y metadatos estén realizados. Al existir muchas herramientas para diseñar estas aplicaciones, la recomendación es concentrarse en una sola y capacitar a todo el equipo de trabajo. Es importante, verificar si existe alguna falla probable que pueda ser corregida en el diseño paralelamente.

j. Despliegue.

Al finalizar las tres rutas paralelas, se continúa con la fase de despliegue. Esta fase representa la convergencia entre las rutas de datos y de aplicaciones analíticas, asegurando el correcto funcionamiento de lo realizado en el proyecto. Desplegar el producto final requiere mucho planeamiento y cautela. Además, del despliegue físico del "data warehouse" y entregables de las aplicaciones analíticas, es necesario, proveer además, educación y soporte para el despliegue. Al igual que el desarrollo de software debe pasar por una

serie de fases para que se encuentre finalmente, disponible.

k. MANTENIMIENTO Y CRECIMIENTO

La fase posterior al despliegue consiste en seguir invirtiendo recursos para darle soporte y mantenimiento al "data warehouse", en particular en las áreas de soporte al usuario, esto para asegurar que se resuelvan sus inquietudes y que puedan utilizar sin problemas las herramientas analíticas y de acceso a datos. La educación al usuario tiene que ser continua. El soporte técnico al "warehouse" también, es necesario, hay que verificar el rendimiento y las capacidades del mismo.

3.2. METODOLOGÍA CRISP

La metodología CRISP-DM según, (Chapman, Clinton, & Kerber, 2000) describe en términos de un modelo de proceso jerárquico, que consiste en conjuntos de tareas que describen a los cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, y la instancia de proceso, como se observa en la figura 2.

En el nivel superior, el proceso de minería de datos se organiza en una serie de fases, cada fase se compone de otras de segundo nivel o tareas genéricas. Este segundo nivel se llama genérico, ya que, está destinado a ser suficientemente, general para abarcar todas las posibles situaciones de minería y es estable, que significa que el modelo es válido para nuevas técnicas de modelado.

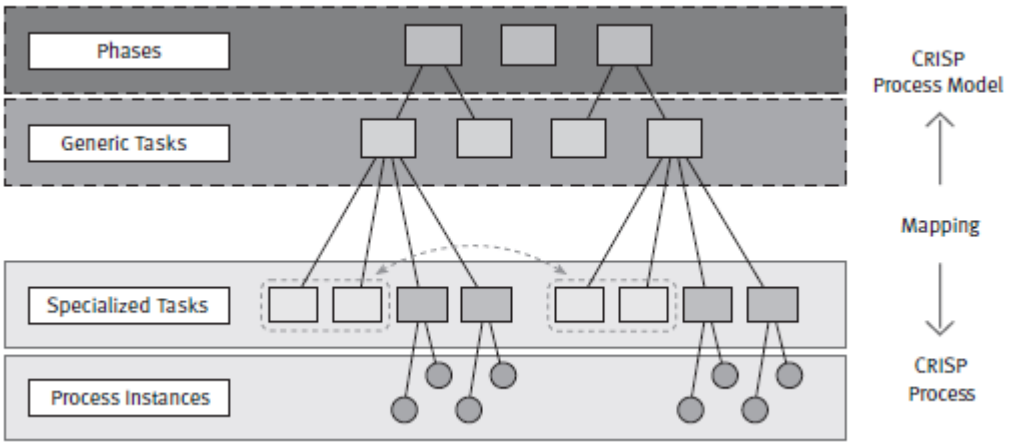


FIGURA 2: COMPOSICIÓN DEL MODELO DE CUATRO NIVELES DE LA METODOLOGÍA CRISP.

El tercer nivel, el nivel de tarea especializada, es el lugar para describir cómo deberían llevarse a cabo acciones en las tareas genéricas en ciertas situaciones específicas. La descripción de las fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de las tareas se pueden realizar en un orden diferente y, a menudo será necesario dar marcha atrás. Nuestro modelo de proceso no intenta capturar todas las rutas posibles a través del proceso de minería de datos, pues, esto requeriría un modelo de proceso demasiado complejo.

El cuarto nivel, la instancia de proceso, es un registro de las acciones, decisiones y resultados de una participación real de la minería de datos. Una instancia de proceso está organizada de acuerdo a las tareas definidas en los planos superiores, pero, representa lo que realmente, ocurrió en el proceso.

En la metodología CRISP-DM el ciclo de vida se resume en el cuadro de la figura 3.

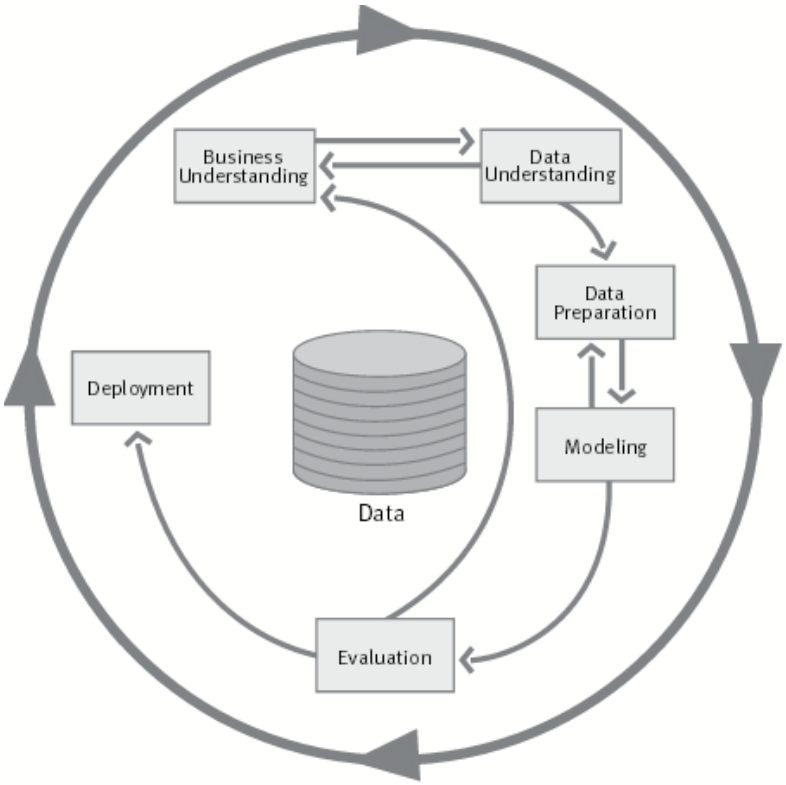


FIGURA 3: CICLO DE VIDA DE METODOLOGÍA CRISP DM.

A continuación, se observa, a que se refiere cada una de las fases identificadas por CRISP-DM.

En la fase de comprensión del negocio se entiende los objetivos del proyecto y los requerimientos desde una perspectiva del negocio y luego convertir este conocimiento en un concepto de un problema de explotación, de información y diseñar un plan preliminar para lograr dichos objetivos.

El conocimiento de los datos comienza con su recolección inicial y procede con las acciones para familiarizarse con ellos, identificar problemas de calidad, identificar primeras pautas en los datos o detectar subconjuntos interesantes de las hipótesis de información oculta.

La fase de preparación de los datos cubre todas las actividades para construir el conjunto de datos final desde los datos iniciales, las tareas de esta

fase son realizadas muchas veces y sin un orden preestablecido, incluye tanto la selección de tablas, registros, atributos como transformación y limpieza de datos para herramientas de modelado.

El modelado incluye la selección de técnicas de modelado y la calibración de sus parámetros a los valores óptimos, suelen existir distintas técnicas para un mismo problema de explotación de información y cada una de ellas suele tener ciertos requisitos sobre los datos, muchas veces, es necesario volver a la fase de desarrollo de los datos.

La evaluación requiere la construcción de uno o varios modelos que aparentan tener la mayor calidad desde una perspectiva de análisis, requiere la evaluación del modelo y estudio de los pasos ejecutados para la construcción del modelo para asegurarse de lograr los objetivos de negocio, al final de esta fase se debe haber alcanzado una decisión en el uso de los resultados.

Por último, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso de explotación de información repetible a través de EPPL.

CAPÍTULO IV

RESULTADOS DE LA INVESTIGACIÓN

A continuación, se procede a detallar cada una de las fases de dichas metodologías previamente, elegidas. El proceso en general se enfoca en desarrollar cada una de las actividades correspondientes al desarrollo del cubo y la aplicación de las técnicas de minería de datos según, las metodologías planteadas.

1. DESARROLLO DE LA METODOLOGÍA

1.1. BUSINESS DIMENSIONAL LIFECYCLE

1.1.1. PLANIFICACIÓN DEL PROYECTO

En esta sección se desglosan las tareas que se realizan para cumplir con los objetivos de la investigación. A continuación se presenta el cronograma del proyecto donde las etapas de la metodología mostrada anteriormente, se dividen y se les asignan sus respectivos tiempos y predecesores. También, se definen los recursos que trabajarán en el proyecto.

Una vez listo el cronograma, se procede a realizar el lanzamiento del proyecto para dar inicio al desarrollo.

El cronograma está definido en la figura 4.

Nombre de tarea	Duration	Predecessors	Resource Names
Desarrollo Proyecto de Investigación	**50 days**		
Inicio	**0 days**		
Kickoff	0 days		
Analisis y Diseño	**16 days**		
Definicion de requerimientos del negocio	4 days	3	Edgar,Heber
Diseño de la arquitectura técnica	4 days	5	Edgar,Heber
Diseño del ETL	3 days	6	Edgar,Heber
Selección de producto e implementación	1 day	7	Edgar,Heber
Modelado Dimensional	2 days	8	Edgar,Heber
Modelado fisico	2 days	9	Edgar,Heber
Desarrollo	**26 days**		
Preparacion de ambiente	**6 days**		
Instalación de ambiente de desarrollo	2 days	10	Heber,Edgar
Instalación de la base de datos	1 day	13	Heber
Recuperación de base de datos	3 days	14	Heber
ETL	**6 days**		
Implementación	**6 days**		
Extracción	1 day	13	Edgar
Transformacion	3 days	14	Edgar
Carga	1 day	15	Edgar
Pruebas	1 day	20	Edgar
Creación de analíticos	**13 days**		
Creación de Jerarquias	2 days	15	Heber
Creación de Roles	1 day	23	Heber
Creación de perspectivas	2 days	24	Heber
Creación de KPI	2 days	25	Heber
Reportes	6 days	26	Heber
Minería de datos	**18 days**		
Comprender el negocio	1 day	21	Edgar
Comprender datos	4 days	29	Edgar
Preparación de datos	4 days	30	Edgar
Modelado	2 days	31	Edgar
Evaluación	3 days	32	Edgar
Desarrollo	4 days	33	Edgar
Pruebas	**4 days**		
Pruebas integrales	4 days	34,27	Edgar,Heber
Transición	0 days	36	Edgar,Heber
Cierre	**0 days**		
Reunión de cierre	0 days	37	Edgar,Heber
Administración del proyecto	4 days	39	Edgar,Heber

FIGURA 4: CRONOGRAMA DEL PROYECTO.

1.1.2. DEFINICIÓN DE LOS REQUERIMIENTOS DEL NEGOCIO

El desarrollo de la toma de requerimientos del negocio, según la metodología que se está utilizando, se realiza mediante entrevistas. Para ello se define a quien se va a entrevistar, en este caso se especifica los entrevistados, según, la siguiente tabla de personas y responsabilidades.

Por cuestiones de confidencialidad de EPPL se reserva los nombres de las personas.

Puesto	Descripción
Gerente general	El gerente general vela por todas las funciones de mercado y ventas de una empresa, así como las operaciones del día a día. Es el responsable de liderar y coordinar las funciones de la planificación estratégica.
Gerente de operaciones	Es la persona, que debe de, configurar los sistemas basados en los reportes del sistema actual, es la que en EPPL debe tener la mayor visibilidad para transmitirla al gerente general, para que, en conjunto tomen decisiones sobre el rumbo del negocio.
Gerente financiero	Es responsable de la planificación, ejecución e información financieras, cataliza las nuevas actuaciones financieras que se van a llevar a cabo y debe implementar estrategias que aseguren un eficiente aprovechamiento de los recursos financieros de EPPL, para sacar el máximo partido de los mismos.
Desarrollador del sistema	Es la persona que ha plasmado durante 5 años los requerimientos y datos que maneja el sistema.

TABLA 1: RESPONSABLES DEL NEGOCIO.

Luego de identificar a las personas, a las cuales, se va a entrevistar en este proceso, se procede a realizar la toma de los requerimientos bajo el alcance del proyecto, se determinan la siguiente lista de requerimientos.

- El sistema debe permitir la utilización de filtros dinámicos.
- El sistema debe filtrar por comercio, banco, canal de pago, tiempo, estado de transacción, conexión de la transacción.
- Los datos que se quieren analizar son monto de la transacción en dólares, montos de créditos y montos de "chargebacks", y los respectivos cargos de cada tipo de transacción.
- Se debe de tener visibilidad con respecto a la ubicación geográfica de los clientes que procesan, tanto globalmente, como por comercio.
- Se debe de tener visibilidad a un nivel de estado y ciudad, solo para Estados Unidos, pues, los datos de otros países son insuficientes para lograr el requerimiento.
- Se debe de tener visibilidad de la cantidad de transacciones aprobadas, denegadas y fallidas por comercio, canal de pago y banco.
- Se debe de tener transparencia de los créditos y "chargebacks" por canales de pago.
- Deben existir reportes de procesamiento a lo largo del tiempo, por mes, semestre, trimestre, mes, semana y día.
- Debe haber una visualización de procesamiento, comisiones por comercio y tipo de transacción.

1.1.3. DISEÑO DE LA ARQUITECTURA TÉCNICA

De acuerdo con los requerimientos obtenidos en la fase anterior se determina que la arquitectura expuesta en la Figura 5 es la adecuada porque es la más sencilla de implementar y cumpliendo con las necesidades solicitadas.

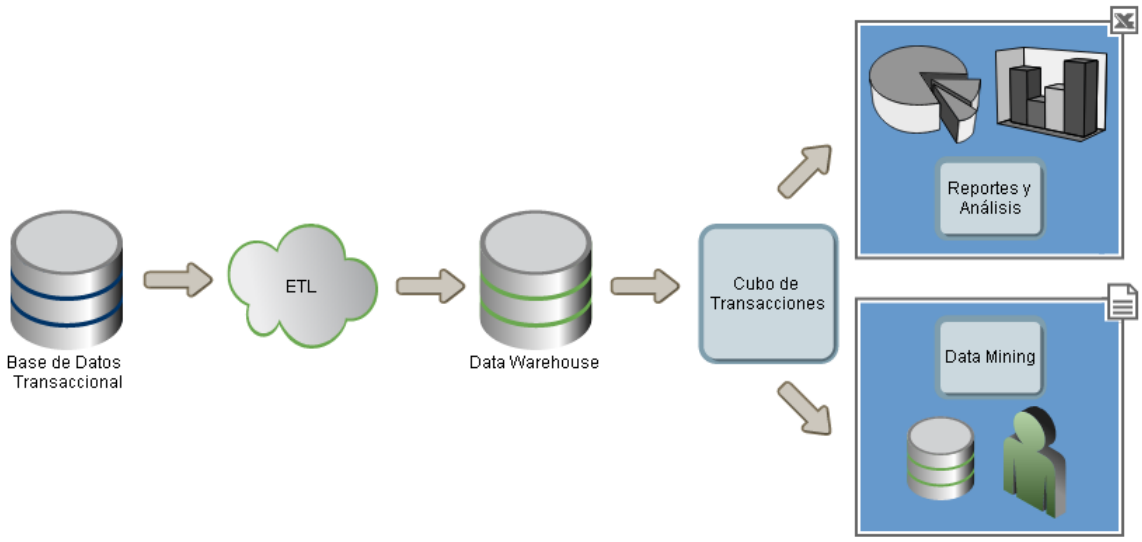


FIGURA 5: ARQUITECTURA TÉCNICA.

En esta fase se establece que los datos se obtienen únicamente, de la base de datos transaccional principal de EPPL, la cual, cuenta con toda la información de los procesos de negocios que son tomados en cuenta para esta investigación.

Además, como sólo se cuenta con una fuente de datos, el proceso de ETL se torna más simple y no se considera necesario un área de "staging", por lo que las operaciones de transformación se realizarán directamente, en un "script" y este es cargado directamente, al almacén de datos. A partir del mismo se genera el cubo OLAP.

Una vez implementado el cubo se procede a aplicar las herramientas de consulta y análisis de datos, y luego desarrollar las técnicas de explotación de minería.

La arquitectura a nivel de configuración de hardware y servidores no fue tomada en cuenta por el alcance del proyecto.

1.1.4. SELECCIÓN DE PRODUCTOS E INSTALACIÓN

A continuación, siguiendo lo expuesto en la fase de diseño de arquitectura, es preciso identificar los componentes necesarios para implementarla.

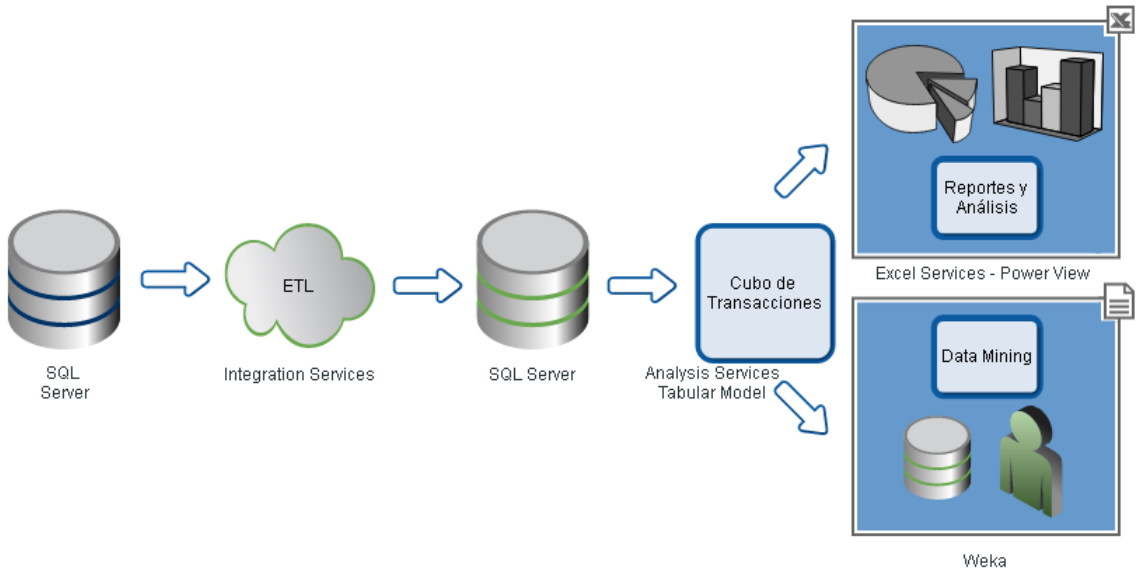


FIGURA 6: PRODUCTOS E INSTALACIÓN.

La base de datos transaccional actual está implementada en un servidor con el software de "Microsoft SQL Server 2008 R2", que cuenta en este momento, con las licencias para implementar el resto de componentes en tecnologías, se decide que se continúe en esta línea de productos.

El proceso de ETL es realizado directamente, a través de la base de datos transaccional al "data warehouse", no se utiliza un área de "staging". La extracción, transformación y carga se crea con paquetes de "SQL Server

Integration Services", estos mueven los datos y aplican las transformaciones necesarias para cargar la información en el "data warehouse".

El "data warehouse" se implementa en un servidor de bases de datos diferente al que tiene la base de datos transaccional. Este servidor cuenta con "SQL Server Analysis Services", y se configura como un modelo tabular. Este permite a través de un diseño multidimensional implementar un cubo, permitiendo identificar las medidas, jerarquías, KPIs y otros aspectos físicos de forma sencilla.

Finalmente, los datos del cubo son consultados mediante reportes físicos o “ad hoc” mediante "Excel Services". Se utilizará principalmente, "Power View" para los reportes estadísticos y gráficos.

1.1.5. MODELADO DIMENSIONAL

La metodología recomienda realizar el diseño del modelo dimensional siguiendo cuatro pasos:

1.1.5.1. IDENTIFICAR PROCESOS DE NEGOCIO

El primer paso a seguir, es seleccionar el proceso de negocio a modelar. Este proceso es una actividad de la organización que típicamente, está soportada por sistemas fuente. Es importante, no referirse a un departamento en particular, sino a un proceso de negocio que abarque a toda la organización.

Entonces, si se enfoca, en un proceso de negocio en lugar de un departamento (ventas, atención al cliente, recursos humanos) se obtiene información consistente que sirva y a la vez sea reutilizada por toda la organización. Asimismo, tener múltiples modelos que contengan información parecida puede llevar a una inconsistencia en los datos. La mejor manera de

asegurar la consistencia es publicar los datos una sola vez, de esta manera también, se facilitan los esfuerzos de extracción, transformación y carga de datos y se economiza en cierta medida recursos tecnológicos.

Para identificar los procesos de negocios de EPPL que deben ser modelados en forma dimensional, se recomienda siempre pensar primero en el proceso principal, o que sea el de mayor importancia. Éste es el proceso responsable de generar mayores ganancias, esto se considera porque, es el que tendrá la información que más interesa a la alta gerencia y contiene los datos clave para la toma de decisiones vitales en la organización.

Luego de las reuniones respectivas para la toma de requerimientos se determina que los procesos de negocio que quieren analizarse son:

- Procesamiento de transacciones, créditos y "chargebacks".

En la siguiente imagen, se observa un esquema del procesamiento de las transacciones.

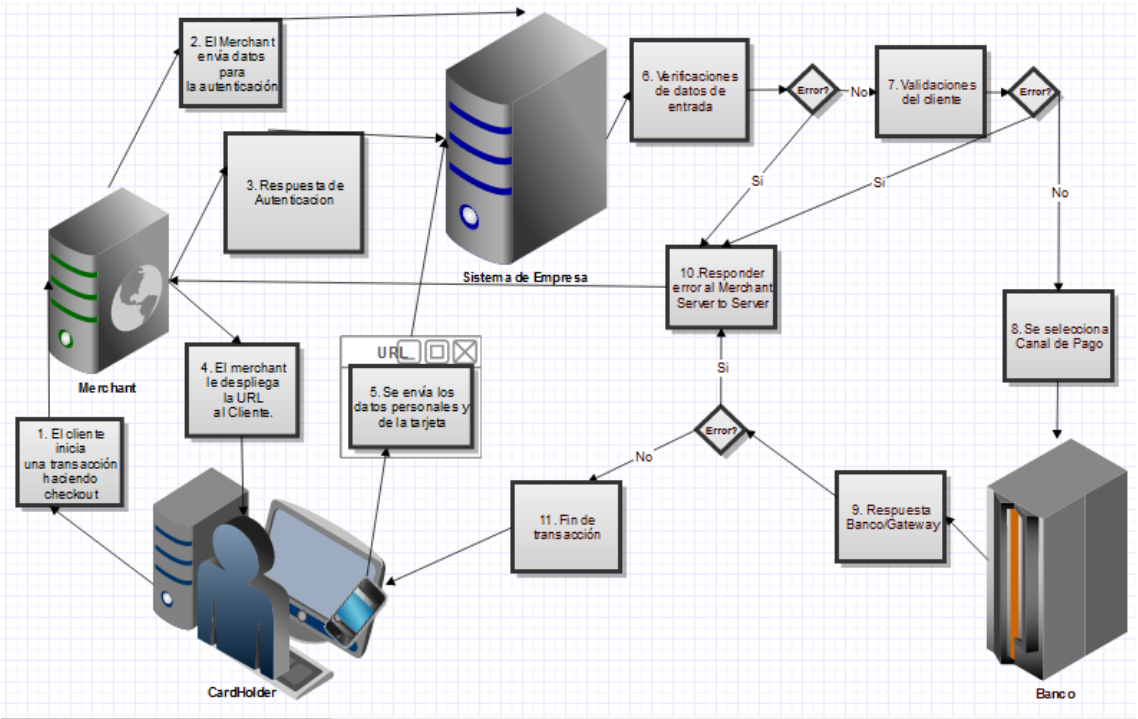


FIGURA 7: PROCESAMIENTO DE TRANSACCIONES.

Este proceso de negocio, propio del procesamiento de tarjetas de crédito y del "e-commerce", permite a los clientes realizar transacciones vía web. El proceso inicia cuando una persona física realiza un "checkout" para comprar algún producto en una tienda electrónica, posteriormente, esta tienda se autentica contra el sistema de EPPL.

El sistema de EPPL valida la información de la tienda y luego se procede a la autorización de los datos de la persona y de la tarjeta. Una vez verificada la información del usuario y la tarjeta, se procede a seleccionar un canal de pago. Estos canales de pago se comunican con su banco respectivo y se procede a finalizar la transacción mostrando el mensaje de resultado al cliente.

Por otro lado, en el mismo ámbito del procesamiento de transacciones se quiere también, tomar en cuenta el manejo de créditos o "refunds". En este proceso el cliente, al estar insatisfecho con el producto recibido, procede a comunicarse con el comercio respectivo para exigir la devolución de su dinero.

Posteriormente, el comercio comunica las devoluciones realizadas semanalmente, para tener el registro en el sistema y contabilizar también las comisiones generadas, ya que, este proceso involucra costos adicionales por lo que siempre se intenta evitar.

Finalmente el último proceso del negocio que se requiere analizar consiste en los "chargebacks", que consiste en la devolución del dinero a un cliente y es iniciada de manera forzada por el interesado, en comunicación directa con el banco emisor argumentando transferencias no autorizadas. Esto puede incurrir en fraudes, pues, el cliente posiblemente, si autorizó la transacción, no obstante, luego lo negó fraudulentamente, para obtener la devolución del dinero. Esto al igual que las devoluciones conlleva diversos costos y comisiones, por lo que debe siempre evitarse.

Ahora bien, si se conocen los procesos de negocio más importantes o interesantes para modelarlos en una base de datos de consulta, se procede a analizarlo con el departamento de contabilidad, la gerencia, y los clientes de

EPPL, y se identifica que los procesos principales del negocio son:

- Generación de transacciones.
- Generación de créditos.
- Generacion de "chargebacks".

Se aprecia, en el diagrama que el entorno de estos procesos de negocio gira alrededor de las siguientes entidades o fuentes:

- Cliente
- Comercio
- Autenticación del comercio
- Canal de Pago
- Banco
- Tarjeta de crédito
- Tiempo o fecha de transacción

1.1.5.2. DEFINIR LA GRANULARIDAD.

Una de las principales necesidades de consulta a los datos se presenta cuando el personal administrativo necesita revisar o buscar transacciones específicas de un cliente, esto en la actualidad, lo realizan a través de consultas sobre la base de datos transaccional que generalmente, contienen múltiples filtros que afectan al rendimiento del sistema.

Entre los reportes que se tienen, en este momento, uno de los más utilizados es el de buscar transacciones específicas para un día y de un cliente especifico. Esto es requerido por el personal para dar soporte a solicitudes y preguntas de los clientes. Entre los filtros habitualmente usados, se tiene la IP utilizada por el cliente en la transacción, también, el código de la tienda o comercio donde se realizó la compra y el nombre del canal de pago donde se cumplió la transacción. Asimismo, es necesario, verificar con los clientes los

números de tarjeta que procesaron la transacción.

Por otro lado, el personal gerencial utiliza actualmente, reportes que muestran las transacciones resumidas por comercio, sin ver el detalle del cliente que realiza cada transacción, aunque les gustaría tener la posibilidad de llegar hasta la información del cliente. Igualmente, es posible realizar agregaciones por canal de pago, y subir hasta los bancos que los contienen y a su vez también, ver las regiones en donde se encuentran dichos canales de pago.

Es de interés de EPPL conocer más de su clientela y aprovechar al máximo la información que se obtiene de ellos. Por esto se requiere que en las consultas a realizar puedan ver datos más específicos de los clientes, por ejemplo, clasificar por la edad o por la nacionalidad, datos que actualmente, se guardan en el sistema pero, no se aprovechan en los reportes actuales.

Con base en lo expuesto anteriormente, la granularidad requerida para que tanto los requerimientos contables como gerenciales debe de ser a nivel transaccional.

1.1.5.3. DEFINIR LAS DIMENSIONES.

Es importante, que las dimensiones que contenga el cubo permitan describir los datos resultantes de estos procesos de negocio descritos. Es necesario, que las dimensiones elegidas consientan en representar todas las posibles respuestas a las preguntas del negocio. Si se conoce ahora los procesos que se requieren modelar, se han identificado una serie de entidades que son modelados como dimensiones. Todas ellas intervienen de manera directa en las medidas que se tendrán en la tabla de hechos y permiten describir los datos resultantes del proceso de negocio analizado. Es preciso permitir filtrar los hechos de manera representativa y que cumpla con las exigencias del negocio, admitiendo responder las preguntas clave de la

organización.

Las dimensiones planteadas son las siguientes:

Nombre de la dimensión	Descripción
Dim_Conexion	Guarda los atributos de la conexión realizada previa a la transacción. En esta se almacenan atributos como el navegador utilizado, la IP del cliente, etc.
Dim_Comercio	Guarda los atributos de los comercios que realizan transacciones. Es una dimensión muy importante para EPPL ya que permitirá filtrar los hechos por comercio, pudiendo analizar las ventas de cada uno de ellos y facilitarles información.
Dim_Tiempo	Guarda los atributos del momento en el que se efectúa una transacción.
Dim_Estado_Transaccion	Guarda los atributos del estado de transacción, este indica si la transacción pudo realizarse correctamente o no. Indicando el estado final de la transacción.
Dim_Tarjeta	Guarda los atributos de la tarjeta de crédito utilizada en la transacción.
Dim_Cliente	Guarda los atributos que caracterizan a la entidad del cliente en el negocio.
Dim_Canal_Pago	Guarda los atributos del canal de pago utilizado para el procesamiento de una transacción

TABLA 2: TABLAS DE DIMENSIÓN.

En el Apéndice A se expone con mayor detalle la información de las dimensiones.

1.1.5.4. DEFINIR LAS MEDIDAS Y HECHOS.

Luego de definir las dimensiones con las que se va a trabajar, es necesario, determinar cuáles son las medidas numéricas que se van a incluir en cada tabla de hecho. Generalmente, para definirlas se busca que es lo que se está tratando de observar. En el caso de esta tesis, son los procesos de negocio que incluyen varias medidas que abarcan cantidades y montos, cuyas granularidades ya han sido definidas.

En la definición del proceso de negocio se identificaron tres tablas de hechos, la tabla de hechos de "Fac_Transacción", la tabla de hechos de "Fac_Credito" y la tabla de hechos "Fac_Chargeback".

Cada una de estas tablas de hechos presentan relación entre las dimensiones que se observan en la sección anterior de este documento, estas medidas se crearon en distintas tablas de hechos debido a que por su significado a nivel de negocio cada una de estas medidas tienen un concepto distinto, y cada una de estas medidas se equipara mediante identificadores únicos.

A continuación, se detallan las medidas para cada tabla de hecho identificada.

Nombre del hecho	Fac_Transaccion	
Atributos	Monto_Rolling_Reserve	Es el monto correspondiente a un porcentaje de la transacción que se mantiene como reserva.
	Monto_Dolares	Es el monto de la transacción en dólares
	Monto_Service_Comission	Monto de comisión por servicio.
	Monto_Transaction_Fee	Es el monto, por comisión por realizar una transacción.
	Cant_Transaccion	Representa la cantidad de transacciones.

TABLA 3: TABLA DE HECHOS DE TRANSACCIONES.

Nombre del hecho	Fac_Credito	
Atributos	Mon_Refund	Es el monto total de un Refund
	Monto_Refund_Fee	Es el monto por comisión por realizar un Refund.
	Cant_Refund	Representa la cantidad de Refund

TABLA 4: TABLA DE HECHOS DE CRÉDITOS.

Nombre del hecho	Fac_Chargeback	
Atributos	Monto_Chargeback	Es el monto total de un Chargeback
	Monto_Chargeback_Fee	Es el monto por comisión por realizar un Chargeback.
	Cant_Chargeback	Representa la cantidad de Chargeback.

TABLA 5: TABLA DE HECHOS DE CHARGEBACKS.

1.1.6. DISEÑO FÍSICO

En cuanto al diseño físico, se parece mucho al diseño lógico presentado anteriormente. Los nombres utilizados para las columnas de las dimensiones y los hechos se mantienen. Los índices creados son los índices únicos en las llaves primarias de cada tabla. Por el alcance de proyecto no se define una estrategia de agregaciones ni de tablas o medidas "pre calculadas", éstas quedarán desplegadas en el cubo. Adicionalmente, se crearon columnas calculadas que se convirtieron en indicadores de desempeño, a continuación se explican con detalle.

En la figura 8 se aprecia el diseño del cubo. Se observan las medidas, las dimensiones, las tablas de hecho y las jerarquías.

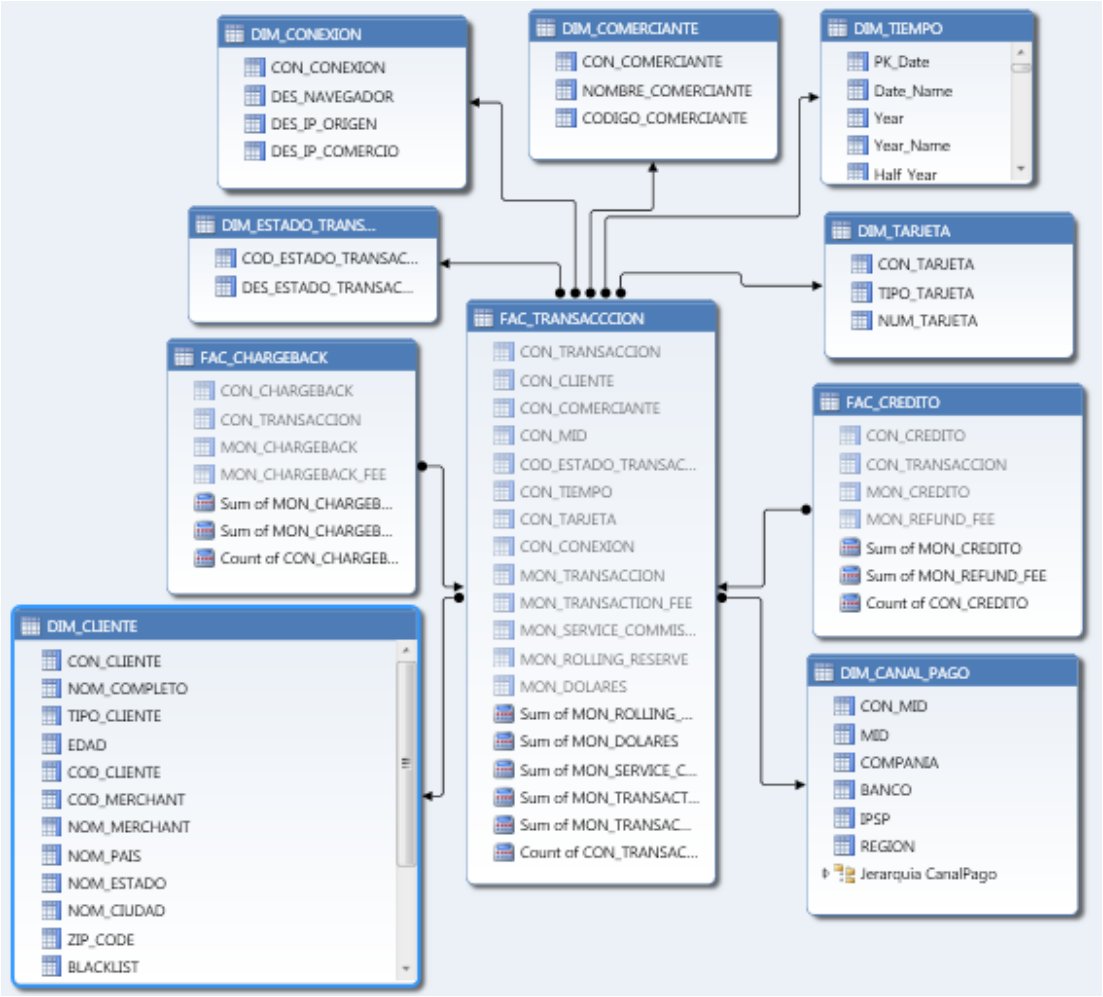


FIGURA 8: DISEÑO DEL CUBO.

1.1.6.1. CREACIÓN DE KPIS

Para el seguimiento de indicadores esenciales en los procesos del negocio anteriormente, expuestos, se tomaron en consideración en los requerimientos la creación de "Key Performance Indicators" o indicadores de desempeño para monitorear algunos aspectos relevantes.

Los valores de referencia para la definición de los indicadores de desempeños fueron suministrados por la empresa EPPL.

A. KPI DE CRÉDITOS

Con este KPI se pretende evaluar la cantidad de créditos en un determinado tiempo, comercio, o canal de pago. La idea sobre todo es analizar que el porcentaje de créditos se encuentre en un valor normal, ya que, es esencial para ver la calidad del servicio o producto que está brindando el comercio.

El KPI será calculado con la siguiente forma:

PorcentajeRefunds = (CantidadRefunds / CantidadTransacciones) x 100

El valor meta es de 0.5%. El rango ideal es entre 0 y 1%. El rango intermedio o de advertencia es de 1% a 1.99%, y finalmente, el rango de alerta sería de 2% en adelante.

B. KPI DE CHARGEBACKS

Con este KPI, se pretende evaluar la cantidad de "chargebacks" en un determinado tiempo, comercio, o canal de pago. La idea sobre todo es analizar que el porcentaje de "chargebacks" se encuentre en un valor normal, porque es esencial para detectar comportamientos fraudulentos en un comercio.

El KPI, es calculado, con la siguiente forma:

PorcentajeChargebacks = (CantidadChargebacks / CantidadTransacciones) x 100

El valor meta es de 0.05%. El rango ideal es entre 0 y 0.10%. El rango intermedio o de advertencia es de 0.11% a 0.99%, y finalmente, el rango de alerta sería de 1% en adelante.

C. KPI DE TRANSACCIONES APROBADAS

Con este KPI, se pretende examinar, la cantidad de transacciones que están siendo aprobadas en un determinado tiempo, comercio, o canal de pago. La idea sobre todo es analizar que el porcentaje de transacciones aprobadas para detectar anomalías en el procesamiento de las transacciones, que pueden ser indicios de problemas de validación de clientes o problemas puntuales con algún comercio, o banco.

El KPI, será calculado con la siguiente forma:

PorcentajeAprobadas = (CantidadTransAprobadas / CantidadTransacciones) x 100

El valor meta es de 80%. El rango ideal es entre 100% y 70%. El rango intermedio o de advertencia es de 70% a 60%, y finalmente, el rango de alerta sería de 60% o menor.

D. KPI DE TRANSACCIONES DENEGADAS

Con este KPI, se pretende valorar la cantidad de transacciones que están siendo denegadas en un determinado tiempo, comercio, o canal de pago. La idea sobre todo es estudiar que el porcentaje de transacciones denegadas para detectar anomalías en el procesamiento de las transacciones, que son

indicios de problemas de validación de clientes o problemas puntuales con algún comercio, o banco.

El KPI, será calculado con la siguiente forma:

PorcentajeDenegadas = (CantidadTransDenegadas/ CantidadTransacciones) x 100

El valor meta es de 15%. El rango ideal es entre 100% y 70%. El rango intermedio o de advertencia es de 70% a 60%, y finalmente, el rango de alerta sería de 60% o menor.

E. KPI DE TRANSACCIONES FALLIDAS

Con este KPI, se pretende evaluar, la cantidad de transacciones que están siendo fallidas en un determinado tiempo, comercio, o canal de pago. La idea sobre todo es examinar que el porcentaje de transacciones fallidas para detectar anomalías en el procesamiento de las transacciones, que pueden ser indicios de problemas de validación de clientes o problemas puntuales con algún comercio, o banco.

El KPI será calculado con la siguiente forma:

PorcentajeFallidas = (CantidadTransFallidas/ CantidadTransacciones) x 100

El valor meta es de 15%. El rango ideal es entre 100% y 70%. El rango intermedio o de advertencia es de 70% a 60%, y posteriormente, el rango de alerta sería de 60% o menor.

1.1.7. DISEÑO Y DESARROLLO DE DATA STAGING

EPPL cuenta únicamente, con un sistema transaccional y se pretende consolidar los datos de ese sistema fuente. Debido a esto no se consideró necesaria la implementación de un "staging area", y las transformaciones se hacen directamente, con las consultas en lenguaje SQL que permiten extraer los datos.

Cómo fue explicado en la fase de diseño de arquitectura, las operaciones de transformación se realizan directamente, en un "script" y son cargadas al almacén de datos. Los paquetes de "SQL Server Integration Services" mueven los datos, aplican las transformaciones y cargan en el "data warehouse".

En la figura 9, se muestran los paquetes diseñados, se creó, uno para cada dimensión y tabla de hechos que fue diseñada.

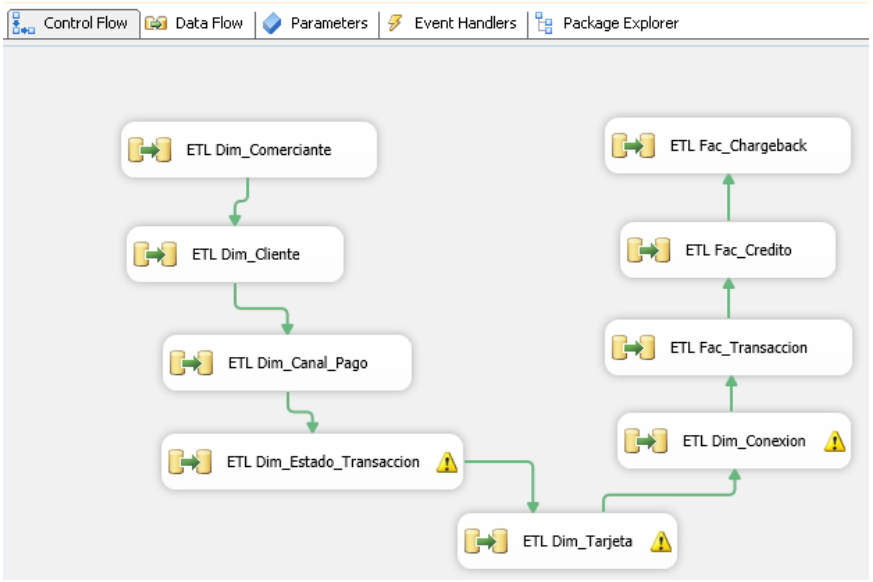


FIGURA 9: PAQUETES SSIS PARA ETL.

También, se deja un ejemplo de las transformaciones realizadas en la figura 10. En este caso el script para extraer los datos de la tabla de clientes, aplicar las transformaciones y cargarlas en la dimensión Dim_Cliente.

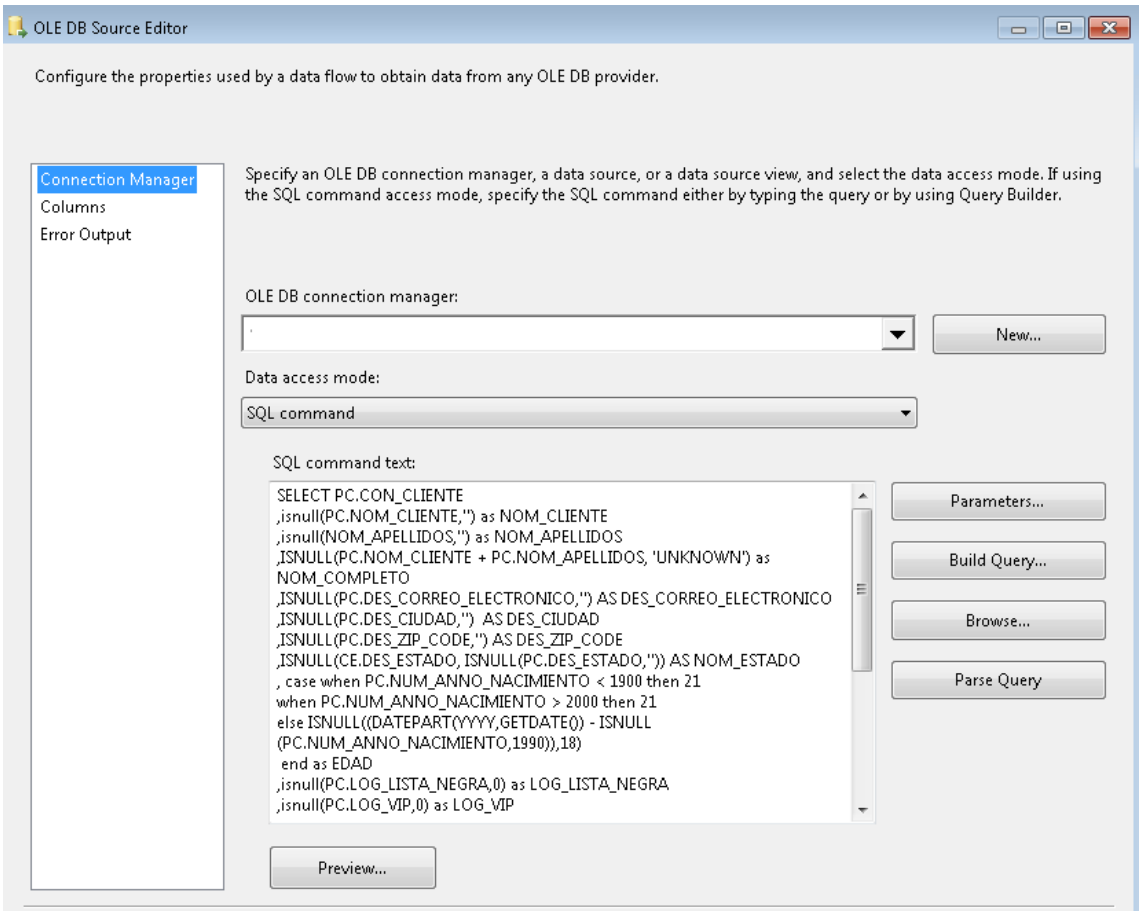


FIGURA 10: SCRIPT DE ETL DIMENSION CLIENTE.

1.1.8. ESPECIFICACIÓN DE APLICACIONES ANALÍTICAS

Las aplicaciones analíticas contempladas en esta investigación involucran la creación de reportes y la posibilidad de crear reportes dinámicos que contengan gráficos y tablas desplegando la información de manera clara y concisa permitiendo al personal gerencial consultarlo de forma simple y fácil.

Para la generación de reportes se utilizan las herramientas de "Excel Services", que son ampliamente, utilizadas por los usuarios gerenciales. Una de ellas es "Power View", que permite consumir el cubo y crear reportes "ad-hoc", permite plantear, diseñar y crear sus propios reportes, aplicando los filtros requeridos y se analizando los datos e indicadores de rendimiento creados.

Con base en los requerimientos obtenidos en la entrevista, se pensó primero en que los filtros sean aplicados dinámicamente, de manera que puedan agregar a sus reportes varios filtros y ver reflejados inmediatamente los cambios en las medidas.

En cuanto al requerimiento de crear perfiles para los comercios, se creó una perspectiva que funciona como rol, el cual, restringe algunos datos y dimensiones por seguridad, esto fue configurado para cada comercio para filtrar los valores que acceden. Esto está pensado para ofrecerle a los comercios la posibilidad de consultar sus datos de procesamiento.

De lo anterior, resulta en dos tipos de usuarios que consultarán los datos, los usuarios gerenciales de EPPL, y los usuarios de los comercios.

Para los usuarios gerenciales de EPPL, se plantearon varios reportes predefinidos, se diseñaron reportes que incluyan a los comercios, canales de pagos, fechas, estados de transacción y que tengan las medidas requeridas como el monto de transacciones, créditos y "chargebacks", con sus respectivos montos de comisiones. Se diseñaron reportes también, que incluyen la ubicación geográfica de los clientes.

1.1.9. DESARROLLO DE APLICACIONES ANALÍTICAS

En referencia a lo expuesto en la especificación de los reportes, se exponen ejemplos, de los reportes, siendo visualizados para el análisis de los datos.

Reporte de Transacciones globales por País, Estado de los Clientes.

FIGURA 11: IMAGEN DEL REPORTE GENERADO DE TRANSACCIONES GLOBALES POR PAÍS DE LOS CLIENTES.

En el reporte de la figura 11, se observa la distribución de montos de transacciones realizadas por país, el tamaño de los círculos varía en función de la cantidad de transacciones realizadas

Desde el reporte anterior, se accede, al detalle del país. En este caso se ve igualmente, la distribución de montos por estados del país USA. Figura 12.

FIGURA 12: IMAGEN DEL REPORTE GENERADO DE MONTOS POR TRANSACCIONES Y ESTADOS.

Reporte de Estado de Transacciones en porcentajes por Comercio.

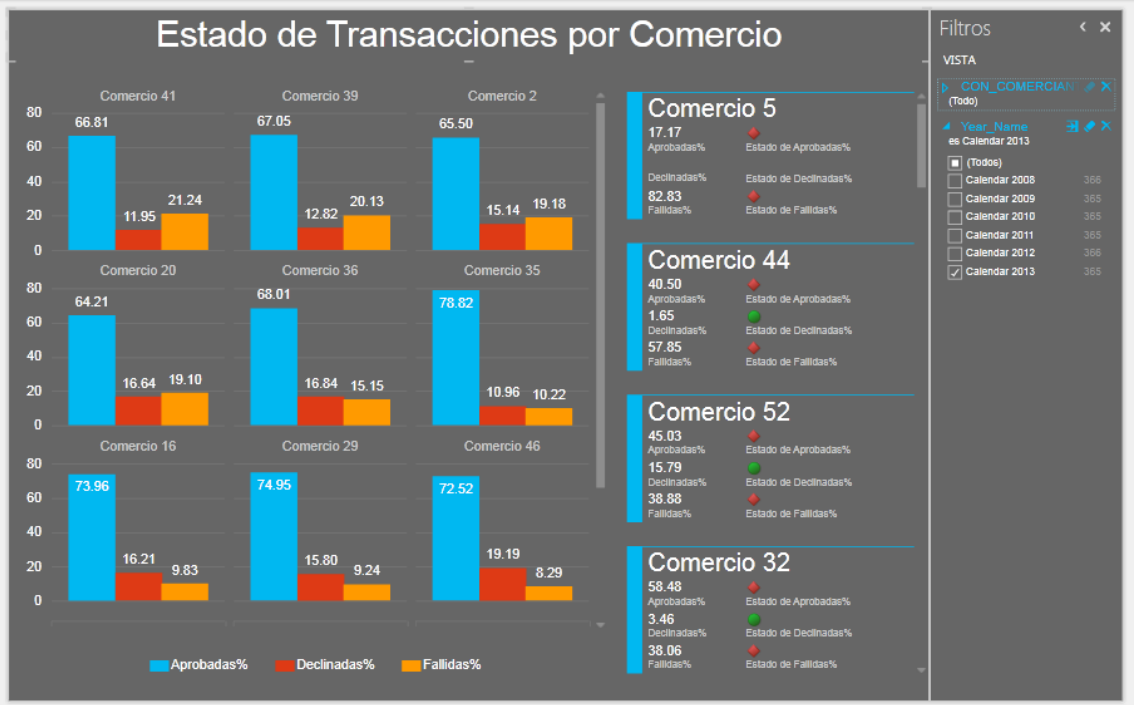


FIGURA 13: IMAGEN DEL REPORTE DE ESTADOS DE TRANSACCIONES POR COMERCIO.

En la figura 13 se aprecia la información del estado de las transacciones por comercio. De esta se forma visualiza el estado de las transacciones de los comercios. Se muestran también, los KPIs "estado de aprobadas", "estado de declinadas" y "estado de fallidas" para analizar si un comercio está infringiendo los valores normales del procesamiento. En el ejemplo mostrado, se distingue que varios comercios tienen valores en estado de error, por ejemplo, el Comercio 52 tiene un porcentaje de aprobadas de 45.84, muy por debajo de los valores esperados.

Reporte de Porcentaje de Créditos, por Comercio, Canal de Pago, y Banco.

La figura 14 presenta un reporte con los datos de créditos y son analizados por comercio, por canal de pago, y por banco por separado. De esta manera se quiere visualizar el porcentaje de créditos principalmente, de los comercios. Se muestran también, el KPI de "porcentaje de créditos" para analizar si un comercio está infringiendo de los valores normales de créditos. En este caso se valora, por ejemplo, que el Canal 361 se encuentra en estado de alerta.

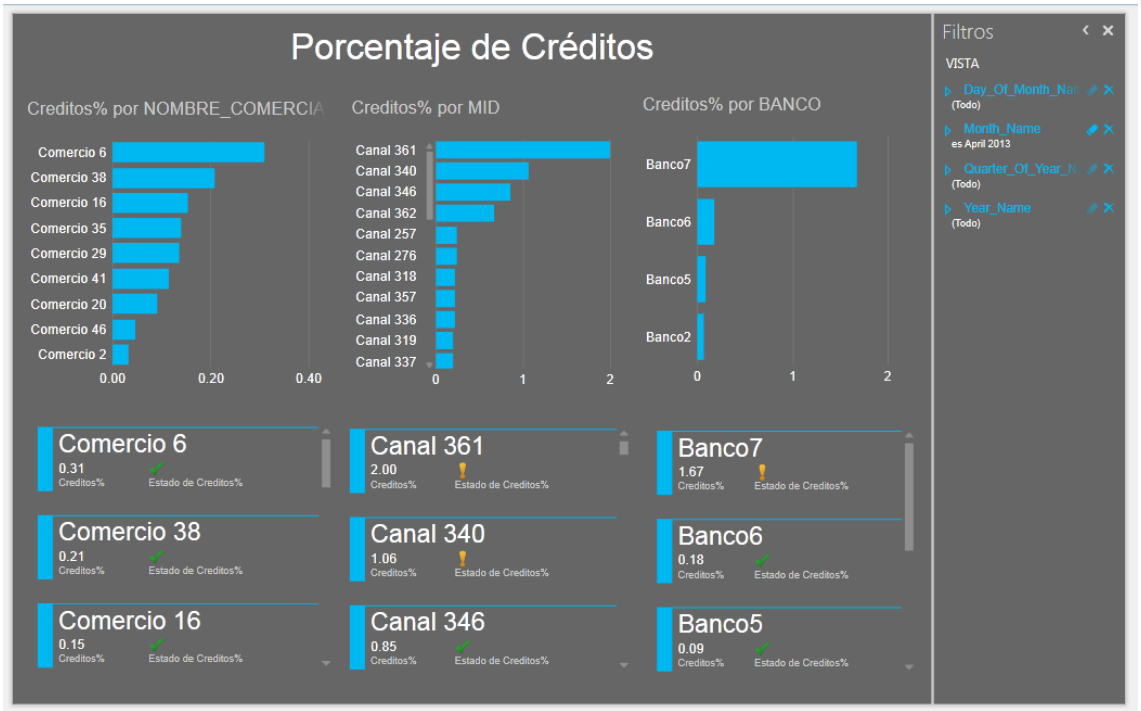


FIGURA 14: IMAGEN DE PORCENTAJE DE CREDITOS POR COMERCIOS.

Si se hace clic sobre el costo en este reporte el resto de valores mostrados se filtra por el que fue escogido, en este caso al seleccionar el Comercio 6, se muestra el "porcentaje de créditos" de los canales de pago y bancos que utilizó el Comercio 6 y se compara con el valor promedio. Esto se observa en la figura 15.

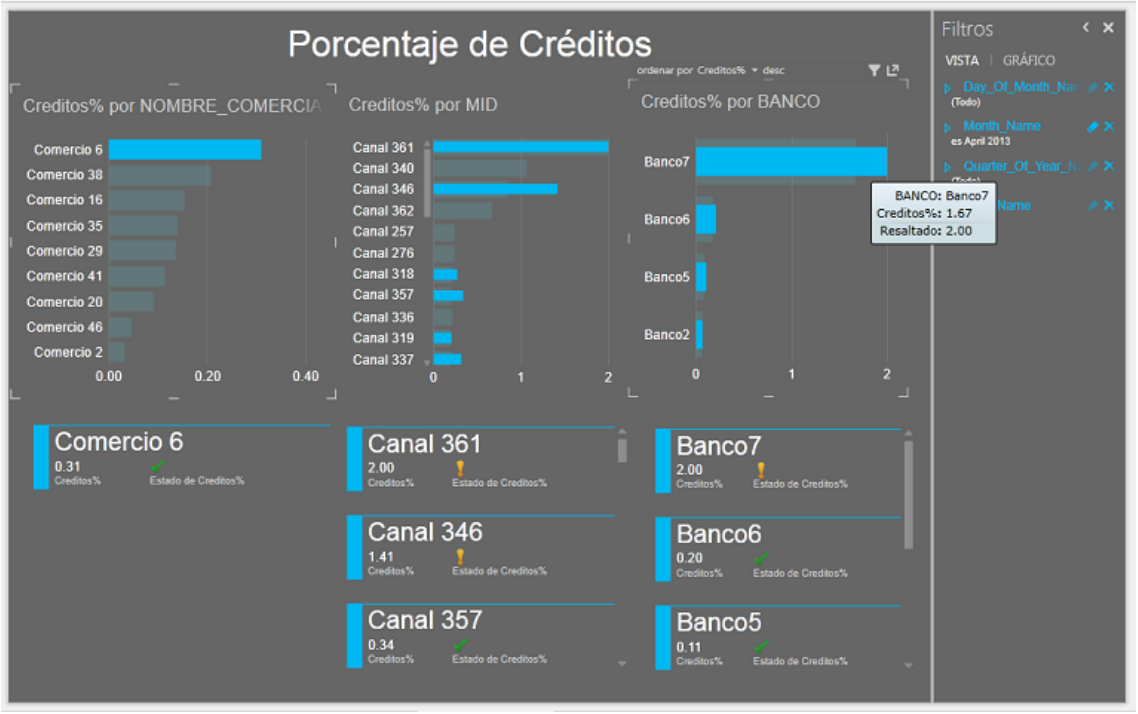


FIGURA 15: IMAGEN DEL REPORTE DE PROCENTAJES DE CRÉDITOS POR CANALES DE PAGO, COMERCIOS Y BANCOS.

Reporte de Porcentaje de Chargebacks.

El reporte de la figura 16 presenta los datos de "chargebacks" y son analizados por comercio, por canal de pago, y por banco por separado. Se muestra asimismo, el KPI de "porcentaje de chargebacks" para analizar si un comercio, canal de pago o banco está infringiendo los valores normales de "porcentaje de chargebacks".

Si se hace clic sobre un valor en este reporte el resto de valores mostrados se filtra por el que fue escogido, en este caso al seleccionar el Banco 7, se muestra el "porcentaje de chargebacks" de los canales de pago y bancos que utilizó el Banco 7 y se compara con el valor promedio. Esto se aprecia en la figura 17.

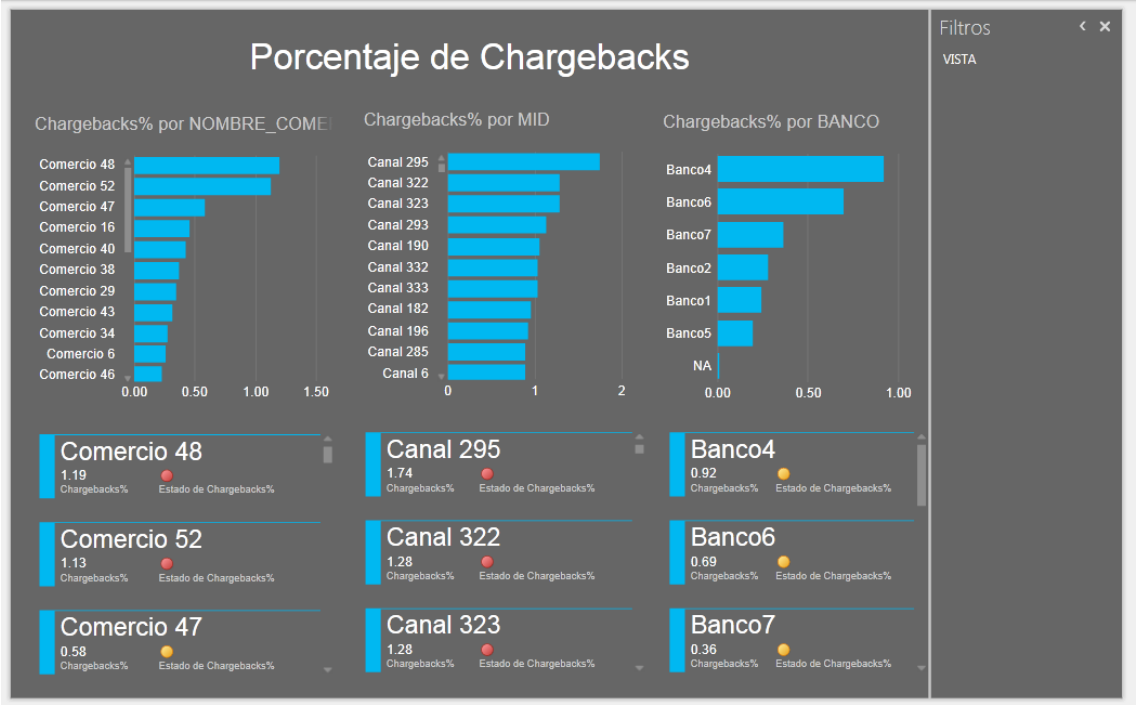


FIGURA 16: IMAGEN DEL REPORTE DEL PORCENTAJE DE CHARGEBACS POR CANAL DE PAGO.

De la misma manera, se observa, en la figura 16 al analizar el reporte, que el Comercio 48 tiene un "porcentaje de chargeback" que está en estado de alerta. En la figura 17 se muestra el "porcentaje de chargebacks" que tuvieron los canales de pago y comercios con el Banco 7.

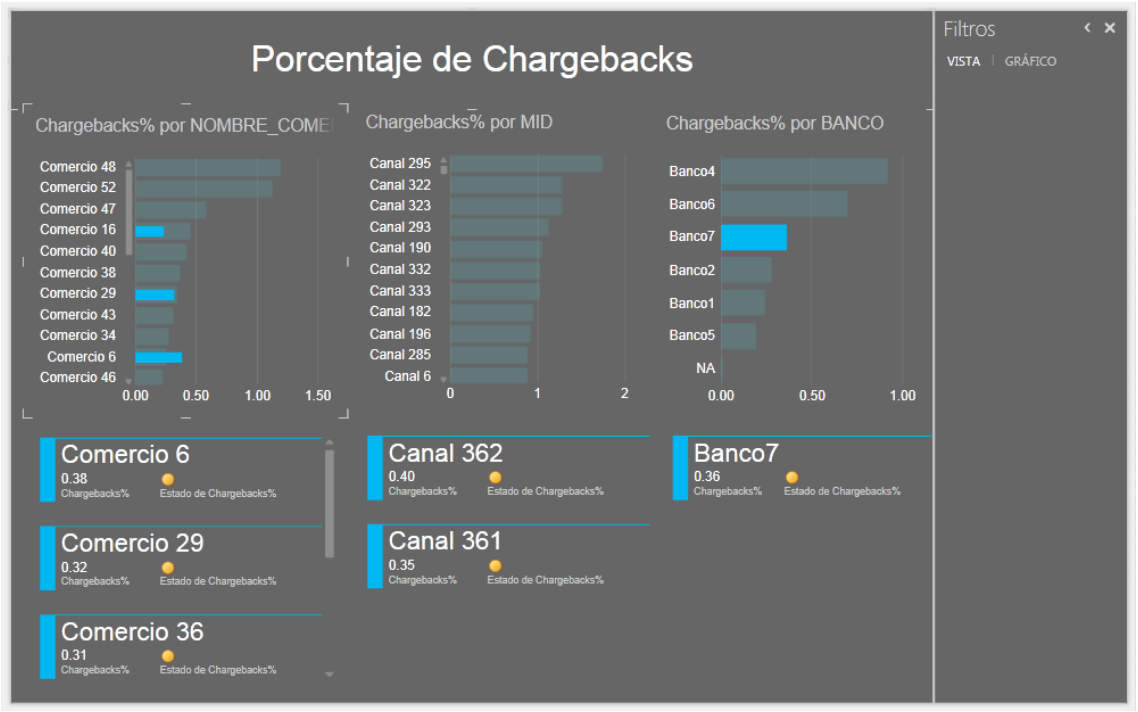


FIGURA 17: IMAGEN DEL REPORTE DE CHARGEBACKS POR CANAL DE PAGO, COMERCIO, BANCOS.

Reporte de procesamiento por año.

La figura 18 permite visualizar la información esencial acerca del procesamiento de EPPL. Se divide en dos gráficas, una para el procesamiento en montos y la otra gráfica para el procesamiento en cantidad de transacciones. Estos gráficos permiten ver en una línea los cambios que van ocurriendo en el transcurso del tiempo.

Al seleccionar un año es posible ver ordenado por cuatrimestres la información del procesamiento.

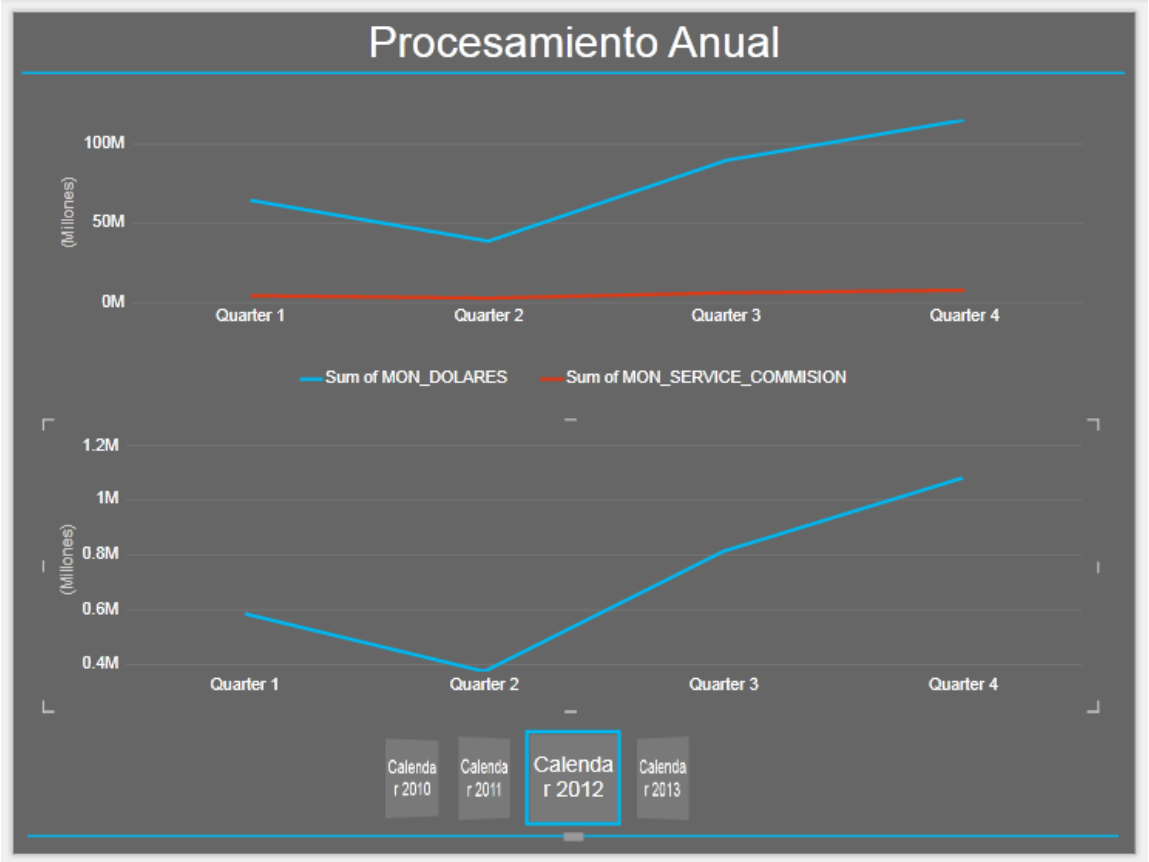


FIGURA 18: IMAGEN DEL REPORTE DEL PROCESAMIENTO POR TIEMPO.

Si se realiza "drill down" en un cuatrimestre específico, es posible, ver la información del procesamiento agrupada por los meses que conforman el cuatrimestre, y de esta manera analizar la información a nivel de mes.

Esto se ve en la figura 19, se distingue el caso del cuatrimestre 3, del año 2011 y se aprecia la información del procesamiento de los meses 7, 8 y 9.

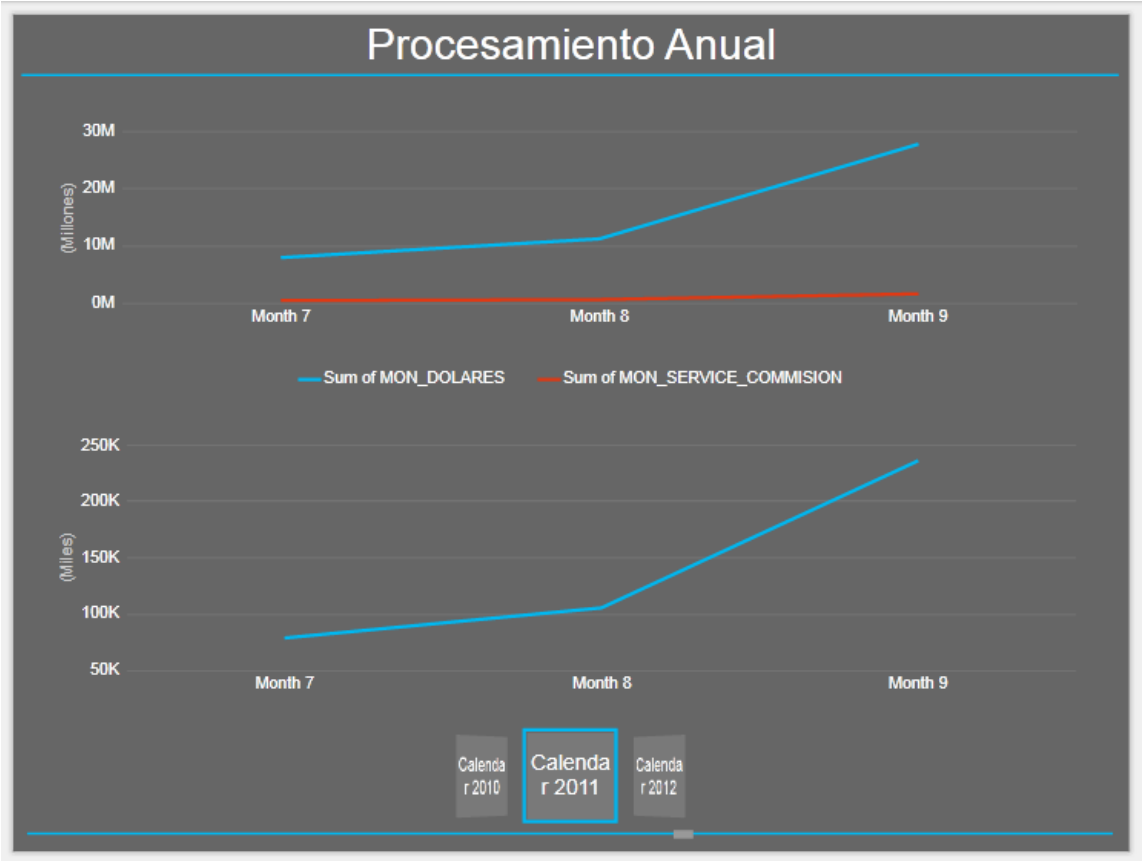


FIGURA 19: IMAGEN DEL REPORTE DEL PROCESAMIENTO POR UN LAPSO DE TIEMPO.

Procesamiento y Comisiones por Comercio.

Este reporte permite analizar y representar la información del procesamiento de transacciones y comisiones organizadas por comercio.

Se divide en dos gráficas, una para el procesamiento en montos de las transacciones y la otra gráfica para el procesamiento en montos de comisiones.

Al seleccionar un comercio se despliega la información de los montos procesados en cada tipo de transacción, que son débitos, créditos y "chargebacks", organizados por año.

En la figura 20, se muestra que es posible realizar "drill down" en cada año y ver la información de procesamiento por mes. En el primer ejemplo, se ve la información del Comercio 6, organizada por años.

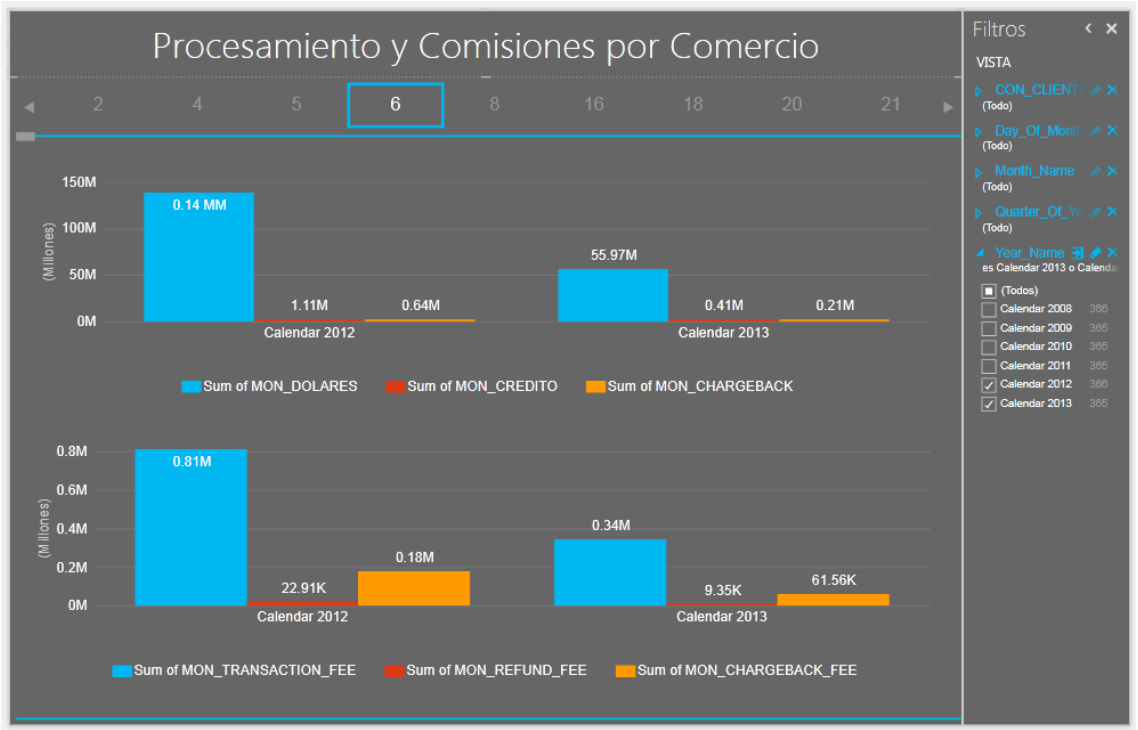


FIGURA 20: IMAGEN DEL REPORTE DEL PROCESAMIENTO Y COMISIONES POR COMERCIO.

Para el segundo ejemplo, se ingresó en el año 2013, y es posible analizar la información por mes del año seleccionado. Es la figura 21.

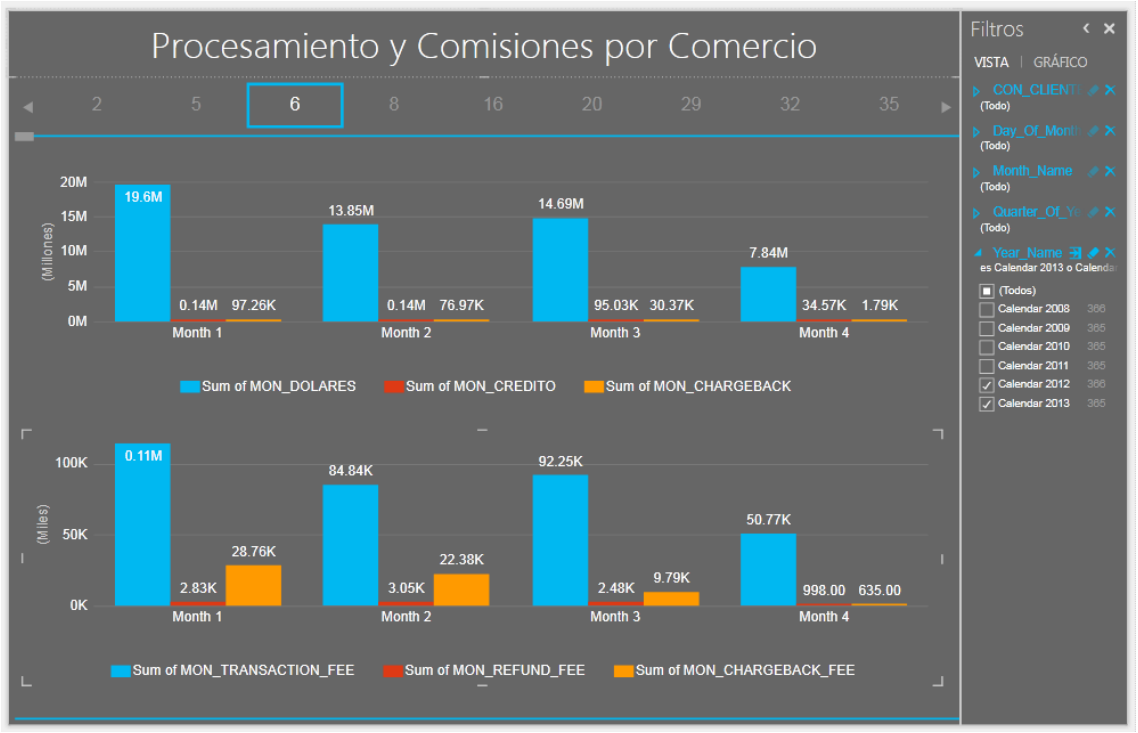


FIGURA 21: IMAGEN DEL REPORTE DEL PROCESAMIENTO Y COMISIONES.

1.1.10. DESPLIEGUE

Esta etapa de la metodología, no forma parte del presente proyecto, EPPL será responsable de instalación de los componentes y versiones finales.

1.1.11. MANTENIMIENTO Y CRECIMIENTO

No está planificado en los alcances del proyecto abordar esta fase, ya que, ésta se tornaría importante una vez el proyecto haya sido desplegado y sea utilizado constantemente, requiriendo así contar con el mantenimiento adecuado y la planificación del crecimiento para tomar las medidas que aseguren el correcto funcionamiento del mismo.

1.2. DESARROLLO DE LA METODOLOGÍA CRISP PARA MINERÍA DE DATOS

El proceso en general, se enfocará en desarrollar cada una de las actividades correspondientes al desarrollo de la aplicación de las técnicas de minería de datos según, la metodología CRISP.

1.2.1. PLANTEAMIENTO DEL PROBLEMA

1.2.1.1. DESCRIPCIÓN DEL PROBLEMA

El principal problema para las entidades procesadoras de pago es el manejo de los créditos y "chargebacks", debido al comportamiento por parte de los tarjetahabientes que incurren en este tipo de transacciones. Por la naturaleza de negocio, EPPL tiene la responsabilidad de velar por minimizar la probabilidad de que ocurra este tipo de transacciones, pues, incurre en gastos adicionales acarreando fuertes costos financieros que reducen significativamente, la utilidad obtenida del procesador, perjudican al comercio y ponen en riesgo la confianza del banco.

Esto hace que el negocio piense en utilizar la información de clientes - suministrada por cada tienda - para replantear y a buscar estrategias que minimicen los comportamientos explicados anteriormente.

Entre las estrategias, la más común, por adaptar es la de reducción del riesgo, a través, del análisis del comportamiento de los clientes de los comercios.

Actualmente, el sistema transaccional cuenta con una estrategia preventiva basada en una serie de herramientas de mitigación de riesgo, las cuales, tienen dos puntos.

- Asegurarse que la persona que está procesando la tarjeta, sea la

dueña de la tarjeta.

- Evaluar comportamientos sospechosos. Por ejemplo, que las tarjetas emitidas en una región repentinamente, procesen transacciones a una distancia geográfica considerable de donde se emitieron, o de lugares sospechosos, como cárceles.

Estas herramientas consisten en servicios dados por empresas de terceros que analizan ciertos datos de ubicaciones y clientes, sin embargo, no analizan comportamientos de las personas que incurren en estos tipos de transacciones.

El problema radica en que no se tiene, hoy por hoy, un método para detectar cómo está relacionado el comportamiento de los clientes con respecto a la generación de créditos o "chargebacks". Además, no se cuenta con una forma de realizar de manera fácil y dinámica este tipo de análisis.

El conocer de alguna tendencia en los datos detrás del negocio que permita encontrar una población más riesgosa o comportamientos sospechosos, es muy importante, para evitar costos adicionales a la organización.

1.2.2. FORMULACIÓN

Problema General

¿Cómo descubrir información que ayude a mitigar el riesgo de transacciones de tipo créditos?

Problemas Específicos

- ¿Cuál es la relación entre las edades de los clientes y los créditos?
- ¿Cuál es la relación entre la cantidad de transacciones de los clientes y los créditos?
- ¿Cuál es la relación entre los montos que los clientes procesan y los

créditos?

- ¿Cuál es la relación entre la cantidad de transacciones aprobadas, denegadas y fallidas de los clientes y los créditos?
- ¿Cuál es la relación del país de origen de los clientes y los créditos?

1.2.3. DELIMITACIÓN DEL PROCESO DE MINERÍA

1.2.3.1. ESPACIAL

El período que se tiene de captura de datos viene desde el 01 de diciembre de 2008, fecha en que se iniciaron actividades de ventas. Se trabajará con los datos almacenados hasta el mes de mayo del 2013, es decir, se tomará en cuenta, prácticamente, toda la operación de ventas de la empresa.

1.2.3.2. CONCEPTUAL

Esta metodología se enmarca en los parámetros de generación de conocimiento establecidos por la KDD (Knowledge Discovery in Databases). Se iniciará con una recopilación de los datos necesarios, de éstos se seleccionará, limpiará y transformará únicamente, la información necesaria, para continuar con la aplicación de técnicas de minería de datos y así encontrar posibles patrones de comportamiento que den indicios sobre la relación de las variables analizadas. Posteriormente, estos patrones serán evaluados e interpretados, y sus resultados luego de ser validados, serán finalmente, difundidos.

1.2.3.3. TECNOLÓGICA

Se trabaja bajo una plataforma Windows, con una base de datos almacenada en "SQL Server 2012", "Analysis Services 2012" y "Visual Studio 2010".

1.2.4. DESARROLLO DEL PROBLEMA

1.2.4.1. ENTENDIMIENTO DEL NEGOCIO

A continuación, se desglosan las actividades para esquematizar el conocimiento a un nivel de minería de datos, y de esta forma iniciar el plan que ayude a llegar a los objetivos.

A. DETERMINAR LOS OBJETIVOS DEL NEGOCIO

Descubrir información que ayude a mitigar el riesgo de transacciones de tipo créditos.

B. EVALUAR SITUACIÓN

Es pertinente, comprender el concepto de los créditos, los cuales, se realizan en dos esquemas: una devolución simple, que se da entre el comercio y el comprador, una devolución compleja que es la que involucra al comercio, el comprador y el banco donde se emitió la tarjeta. Ambos tipos de devolución se consideran un problema grave, pues, afecta los costos para EPPL, el comercio involucrado y también, la confianza de los bancos.

Para esta investigación, nos enfocaremos en las devoluciones simples o créditos, esto debido a que son las más comunes y tienden a estar relacionadas con las devoluciones complejas.

Además, una devolución, se da por razones fraudulentas o por simple voluntad del comprador, ambas formas para EPPL son consideradas un problema. Debido a esto, en la actualidad, EPPL cuenta con un sistema que utiliza una serie de herramientas de mitigación de riesgo, pero, su aplicación es global, es decir, se realiza sobre todos los clientes que procesan transacciones.

Aplicar estas herramientas también, implica un costo de comisión por la utilización de las mismas ante las empresas que brindan el servicio.

La idea planteada es utilizar los datos que se tienen para sugerir tendencias que ayuden a considerar qué comportamientos son peligrosos y de esta forma aplicar las herramientas de riesgo a poblaciones específicas.

C. DETERMINAR OBJETIVOS DE MINERÍA DE DATOS

- Determinar la relación entre las edades de los clientes y los créditos.
- Establecer la relación entre la cantidad de transacciones de los clientes y los créditos
- Determinar la relación entre los montos que los clientes procesan y los créditos
- Determinar la relación entre la cantidad de transacciones aprobadas, denegadas y fallidas de los clientes y los créditos
- Determinar la relación entre los países de los clientes y los de un crédito.

1.2.4.2. COMPRENSIÓN DE DATOS

A. RECOGER DATOS INICIALES

Los datos se han tomado de la base de datos relacional, la cual, es un respaldo a mayo del año 2013.

B. DESCRIBIR LOS DATOS

En esta fase se va a tomar una colección inicial de datos con el objetivo de familiarizarse con los mismos. Se identifican problemas de calidad y se detectan conjuntos interesantes que permitan formar hipótesis. En el apéndice B, se detalla la tabla con la información de los datos analizados.

Luego de hacer la descripción de los datos definimos los rangos de valores con el fin de discretizarlos.

Dato	Rangos
Rango de edades	• Adolecente (menor a 18 años) • Joven temprano (entre 18 y 24) • Joven tardío (entre 24 y 29) • Adulto (entre 30 y 37) • Adulto medio (entre 38 y 49) • Adulto tardío (mayor a 50 años)
Rango de montos, esto se va a realizar por transacciones aprobadas, denegadas y fallidas.	• $0-$50 • $51-$100 • $101-$150 • $151-$200 • $201-$300 • $301-$400 • $400-$500 • mayor a $500

TABLA 6: RANGOS DE DATOS.

Con respecto a los datos que se van a utilizar, se realiza el siguiente análisis.

Dato	Origen	Comentario
Rango de edades		Campo calculado de una vista
Nombre del país	NOM_PAIS	Nombre del país
Código del tienda	COD_MERCHANT	
Total de transacciones por cliente	TOTAL_TRANSACTIONS_BY_CLIENT	
Cantidad de transacciones aprobadas por cliente.	TOTAL_APPROVED_TRANSACTIONS_BY_CLIENT	Campo calculado de una vista
Cantidad de transacciones declinadas por cliente.	TOTAL_DECLINED_TRANSACTIONS_BY_CLIENT	Campo calculado de una vista
Cantidad de transacciones fallidas por cliente.	TOTAL_FAILED_TRANSACTIONS_BY_CLIENT	Campo calculado de una vista
Monto total de transacciones aprobadas por cliente	TOTAL_AMOUNT_APPROVED_TRANSACTIONS_BY_CLIENT	Campo calculado de una vista
Monto total de transacciones denegadas por cliente	TOTAL_AMOUNT_DECLINED_TRANSACTIONS_BY_CLIENT	Campo calculado de una vista
Monto total de transacciones fallidas por cliente		Campo calculado de una vista
Monto promedio de transacciones aprobadas por cliente		Campo calculado de una vista
Monto promedio de transacciones denegadas por cliente		Campo calculado de una vista

Monto promedio de transacciones fallidas por cliente	Campo calculado de una vista
Rango de montos aprobados	Campo calculado de una vista
Rango de montos fallidos	Campo calculado de una vista
Rango de montos denegados	Campo calculado de una vista

TABLA 7: ANÁLISIS DE CAMPOS CALCULADOS.

C. EXPLORAR LOS DATOS

En esta sección, se analizan los datos y se observan sus características en la vista seleccionada. Para efectos del documento se muestra como ejemplo, 10 filas, en ese caso cada uno de los campos no es nulo y tienen los datos de los clientes con respecto a su comportamiento transaccional.

CON_CLIENTE	RANGO_EDADES	NOM_PAIS	COD_MERCHANT	TOTAL_TRANSACCIONES
642300	Adulto medio	USA	NCL5001-2008	4
924581	Adulto tardío	USA	AS5005-2009	1
1000991	Adulto tardío	USA	WEL25021-2010	1
1233660	Adulto tardío	USA	AS5005-2009	2
1351192	Joven temprano	USA	AS5005-2009	1
117770	Adulto	USA	AS5005-2009	1
642323	Adulto	USA	NCL5001-2008	7
1053543	Joven tardío	USA	AS5005-2009	1
1440621	Joven tardío	USA	AS5005-2009	1
147554	Adulto	USA	AS5005-2009	1

TABLA 8: DATOS DE MUESTRA DE CLIENTES.

En esta tabla se observa, la clasificación a nivel de rango de edades, país, comercio y el total de transacciones realizadas por el cliente en el sistema.

CON_CLIENTE	TOTAL_APROBADAS	MONTO_TOTAL_APROBADAS	PROMEDIO_MONTO_APROBADAS
642300	4	500	125
924581	1	50	50
1000991	1	245.93	245.93
1233660	2	500	250
1351192	1	100	100
117770	1	28	28
642323	7	550	78.571428
1053543	1	40.23	40.23
1440621	1	150	150
147554	1	25	25

TABLA 9: DATOS DE MUESTRA DE TRANSACCIONES APROBADAS.

En esta tabla se distingue, el total de transacciones aprobadas, el monto total aprobado y el promedio del monto aprobado de cada cliente, es decir, el monto total entre la cantidad de transacciones aprobadas.

CON_CLIENTE	TOTAL_DENEGADAS	MONTO_TOTAL_DENEGADAS	PROMEDIO_MONTO_DENEGADO
642300	0	0	0
924581	0	0	0
1000991	0	0	0
1233660	0	0	0
1351192	0	0	0
117770	0	0	0
642323	0	0	0
1053543	0	0	0
1440621	0	0	0
147554	0	0	0

TABLA 10: DATOS DE MUESTRA DE TRANSACCIONES DENEGADAS.

En esta tabla se muestra el total de transacciones declinadas por cliente, el monto total declinado, y su respectivo promedio.

CON_CLIENTE	TOTAL_FALLIDAS	MONTO_TOTAL_FALLIDO	PROMEDIO_MONTO_FALLIDO
642300	0	0	0
924581	0	0	0
1000991	0	0	0
1233660	0	0	0
1351192	0	0	0
117770	0	0	0
642323	0	0	0
1053543	0	0	0
1440621	0	0	0
147554	0	0	0

TABLA 11: DATOS DE MUESTRA DE TRANSACCIONES FALLIDAS.

De la misma forma, en la tabla anterior se muestra el total de transacciones fallidas, el monto total fallido y el promedio respectivo.

CON_CLIENTE	**ES_CREDITO**	**RANGO_MONTOS_APROBADOS**
642300	Yes	101-150
924581	Yes	0-50
1000991	Yes	201-300
1233660	Yes	201-300
1351192	Yes	51-100
117770	Yes	0-50
642323	Yes	51-100
1053543	Yes	0-50
1440621	Yes	101-150
147554	Yes	0-50

TABLA 12: DATOS DE MUESTRA DE RANGOS DE MONTOS APROBADOS.

En la tabla anterior, se ve si este cliente ha hecho devoluciones o créditos, y si ha realizado "chargebacks", además, el rango de montos aprobados, al que el cliente pertenece.

CON_CLIENTE	**RANGO_MONTOS_DENEGADOS**	**RANGO_MONTO_FALLIDOS**
642300	0-50	0-50
924581	0-50	0-50
1000991	0-50	0-50
1233660	0-50	0-50
1351192	0-50	0-50
117770	0-50	0-50
642323	0-50	0-50
1053543	0-50	0-50
1440621	0-50	0-50
147554	0-50	0-50

TABLA 13: DATOS DE MUESTRA DE RANGOS DE MONTOS DENEGADOS.

D. COMPRUEBE LA CALIDAD DE LOS DATOS

En esta etapa, se revisan los datos, descartando valores nulos e inconsistencias de cualquier tipo.

1.2.4.3. PREPARACIÓN DE DATOS

A. SELECCIÓN DE LOS DATOS

La selección de los datos se basa en la exploración y el análisis realizado en los puntos anteriores.

B. LIMPIAR DATOS

El proceso de limpieza de datos consiste en eliminar los valores nulos.

C. CONSTRUIR DATOS

La construcción de los datos se basa en la vista expuesta en el apéndice C.

D. INTEGRAR LOS DATOS

Las integraciones de los datos son las siguientes:

Columna	**Integración**
CON_CLIENTE	No tiene integración de datos.
RANGO_EDADES	Rango de edades, que es la clasificación de las edades de los clientes.
NOM_PAIS	No tiene integración de datos.
COD_MERCHANT	No tiene integración de datos.

TOTAL_TRANSACCIONES	No tiene integración de datos.
TOTAL_APROBADAS	No tiene integración de datos.
MONTO_TOTAL_APROBADAS	No tiene integración de datos.
PROMEDIO_MONTO_APROBADAS	MONTO_TOTAL_APROBADAS / TOTAL_APROBADAS
TOTAL_DENEGADAS	No tiene integración de datos.
MONTO_TOTAL_DENEGADAS	No tiene integración de datos.
PROMEDIO_MONTO_DENEGADO	MONTO_TOTAL_DENEGADAS / TOTAL_DENEGADAS
TOTAL_FALLIDAS	No tiene integración de datos.
MONTO_TOTAL_FALLIDO	No tiene integración de datos.
PROMEDIO_MONTO_FALLIDO	MONTO_TOTAL_FALLIDO / TOTAL_FALLIDO
ES_CREDITO	Si existe al menos un crédito el valor es "YES"
ES_CHARGEBACK	Si existe al menos un crédito el valor es "NO"
RANGO_MONTOS_APROBADOS	Rango de montos, que es la clasificación de la columna MONTO_TOTAL_APROBADO
RANGO_MONTOS_DENEGADOS	Rango de montos, que es la clasificación de la columna MONTO_TOTAL_DENEGADO
RANGO_MONTO_FALLIDOS	Rango de montos, que es la clasificación de la columna MONTO_TOTAL_FALLIDO

TABLA 14: TABLA DE INTEGRACIONES DE DATOS.

E. FORMATO DE DATOS

Columna	**Formato**
CON_CLIENTE	Entero, [0-9]
RANGO_EDADES	String [A-Z,a-z], largo máximo 15
NOM_PAIS	String [A-Z,a-z], largo 3
COD_MERCHANT	String [A-Z,a-z], largo máximo 15
TOTAL_TRANSACCIONES	Entero, [0-9]
TOTAL_APROBADAS	Entero, [0-9]
MONTO_TOTAL_APROBADAS	Entero, [0-9]
PROMEDIO_MONTO_APROBADAS	Entero, [0-9]
TOTAL_DENEGADAS	Entero, [0-9]
MONTO_TOTAL_DENEGADAS	Entero, [0-9]
PROMEDIO_MONTO_DENEGADO	Entero, [0-9]
TOTAL_FALLIDAS	Entero, [0-9]
MONTO_TOTAL_FALLIDO	Entero, [0-9]
PROMEDIO_MONTO_FALLIDO	Entero, [0-9]
ES_CREDITO	String, [Yes, No]
RANGO_MONTOS_APROBADOS	String [A-Z,a-z], largo máximo 15
RANGO_MONTOS_DENEGADOS	String [A-Z,a-z], largo máximo 15

RANGO_MONTO_FALLIDOS	String [A-Z,a-z], largo máximo 15

TABLA 15: TABLA DE FORMATO.

1.2.4.4. MODELADO

A. SELECCIONE TÉCNICA DE MODELADO

En esta sección, se definen la(s) técnicas de modelado que se van a utilizar. Las técnicas seleccionadas son:

- Reglas de asociación.
- Árboles de decisión.
- "Clúster"
- "Neural Network"
- "Naive Bayes"

B. CONSTRUIR EL MODELO

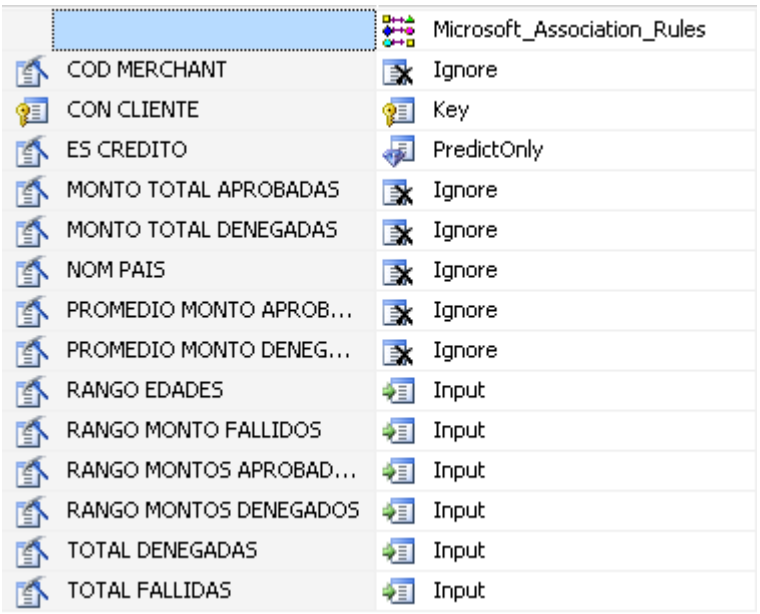
En la herramienta de Microsoft Visual Studio 2010, se crea el modelo de datos, el cual, se basa en la información a nivel de datos definida en las fases anteriores.

En la siguiente figura se observa el modelo general, a continuación se aprecia de forma más detallada, dividido por cada uno de los modelos definidos para cada una de las técnicas a utilizar.

	Microsoft_Association_Rules	Microsoft_Neural_...	Microsoft_Naive_...	Microsoft_Decisio...	Microsoft_Clustering
COD MERCHANT	Ignore	Input	Input	Input	Ignore
CON CLIENTE	Key	Key	Key	Key	Key
ES CREDITO	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly
MONTO TOTAL APROBADAS	Ignore	Input	Input	Ignore	Ignore
MONTO TOTAL DENEGADAS	Ignore	Input	Ignore	Ignore	Ignore
NOM PAIS	Ignore	Input	Input	Input	Ignore
PROMEDIO MONTO APROB...	Ignore	Input	Ignore	Input	Ignore
PROMEDIO MONTO DENEG...	Ignore	Ignore	Ignore	Input	Ignore
RANGO EDADES	Input	Input	Input	Input	Input
RANGO MONTO FALLIDOS	Input	Ignore	Ignore	Ignore	Ignore
RANGO MONTOS APROBAD...	Input	Input	Input	Input	Input
RANGO MONTOS DENEGADOS	Input	Ignore	Input	Input	Input
TOTAL DENEGADAS	Input	Ignore	Input	Ignore	Ignore
TOTAL FALLIDAS	Input	Ignore	Ignore	Ignore	Ignore

FIGURA 22: MODELO GLOBAL.

El siguiente modelo corresponde a las reglas de asociación, en este caso se va a predecir para la variable "es crédito" y se van a ignorar las variables "cod_merchant", "monto total aprobadas", "monto total denegadas", "nom_pais", "promedio monto aprobadas", "promedio monto denegadas", esto porque los datos aportan poco a este modelo.



	Microsoft_Association_Rules
COD MERCHANT	Ignore
CON CLIENTE	Key
ES CREDITO	PredictOnly
MONTO TOTAL APROBADAS	Ignore
MONTO TOTAL DENEGADAS	Ignore
NOM PAIS	Ignore
PROMEDIO MONTO APROB...	Ignore
PROMEDIO MONTO DENEG...	Ignore
RANGO EDADES	Input
RANGO MONTO FALLIDOS	Input
RANGO MONTOS APROBAD...	Input
RANGO MONTOS DENEGADOS	Input
TOTAL DENEGADAS	Input
TOTAL FALLIDAS	Input

FIGURA 23: MODELO ASSOCIATION RULES.

La técnica de redes neuronales al igual que el algoritmo anterior va a predecir para la variable "es crédito", las variables de entrada son "cod_merchant", "monto total aprobadas", "monto total denegadas, "nom_pais", "promedio monto aprobadas", "rango edades" y "rango montos aprobados"

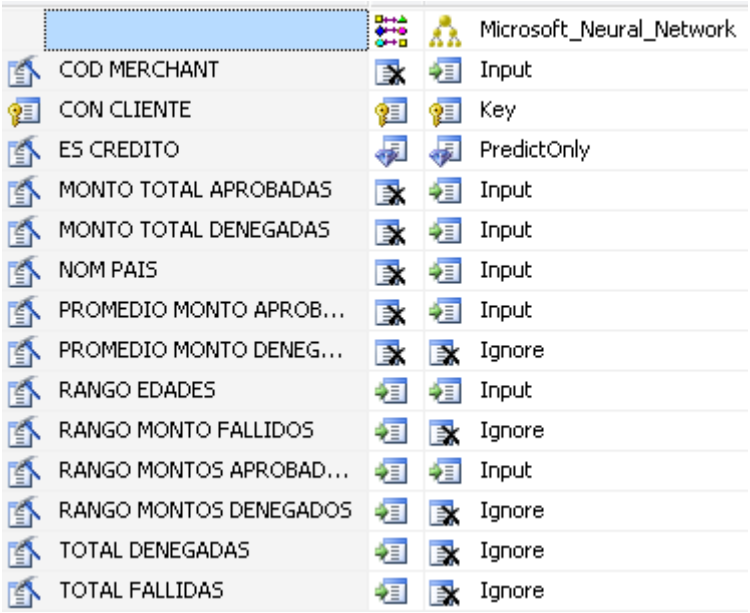


FIGURA 24: MODELO DE REDES NEURONALES.

El modelo de "naive bayes", predice para "es crédito" e ignora las variables de "estado fallido", "rango de montos fallidos", "promedio de montos denegados", "promedio de montos aprobados" y "monto total denegadas".

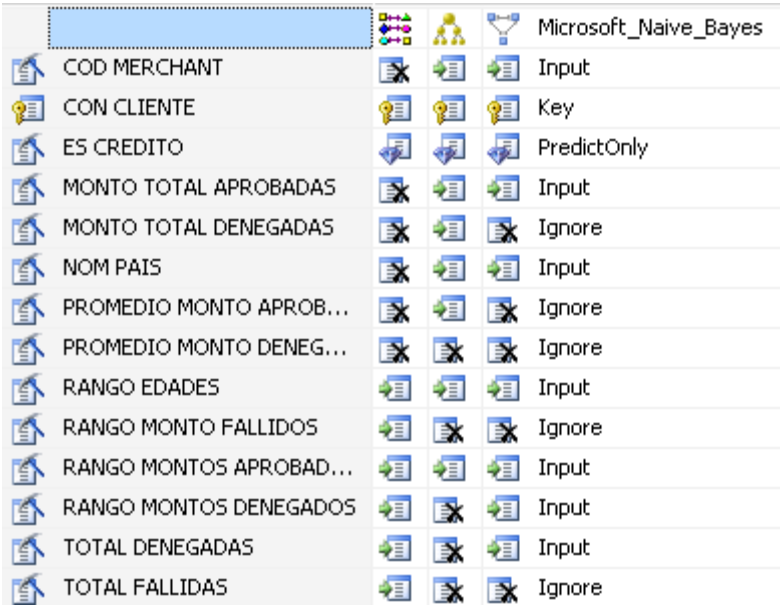
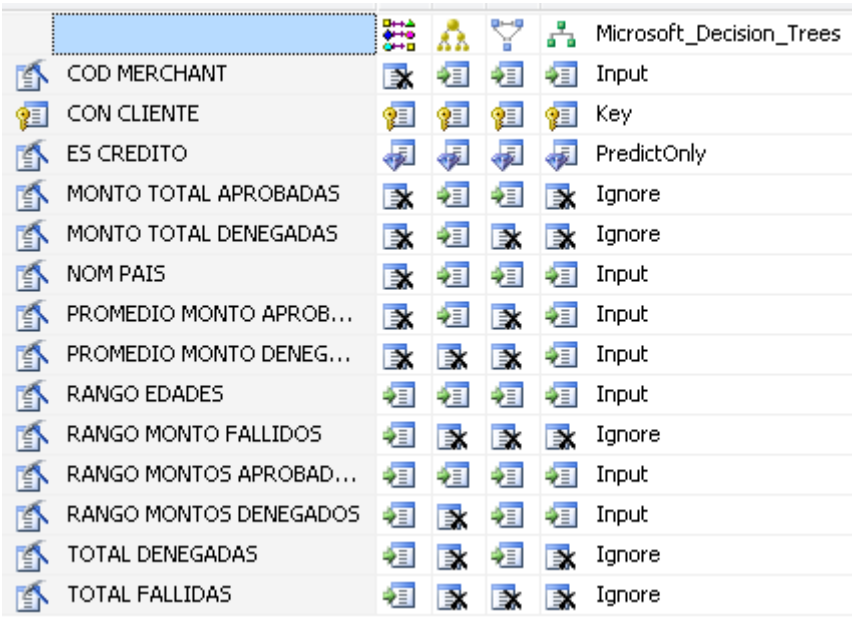


FIGURA 25: MODELO DE NAIVE BAYES.

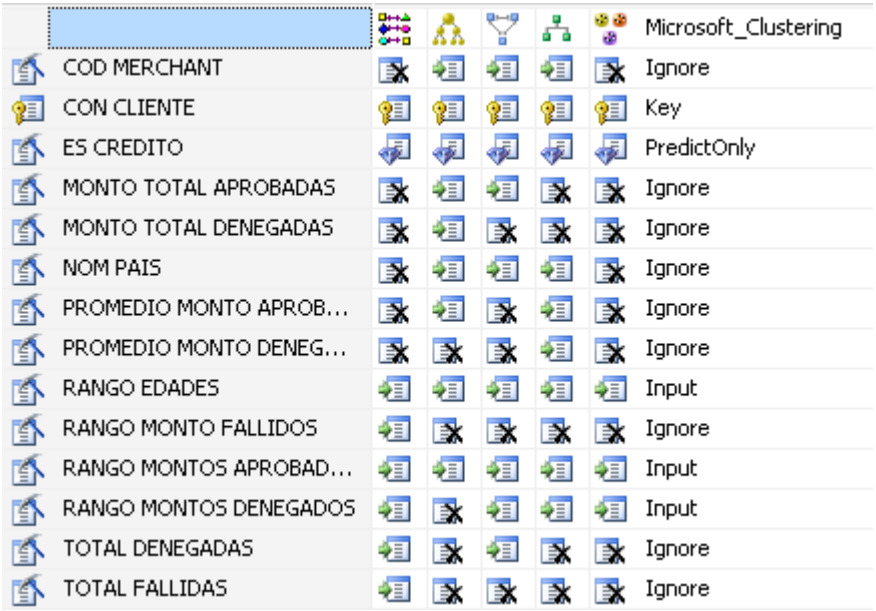
El modelo de árboles de decisión predice para la variable "es crédito" e ignora, las variables de "monto total aprobadas", "monto total denegadas", el "rango de montos fallidos", "total denegadas" y el "total fallidas".



	Microsoft_Decision_Trees
COD MERCHANT	Input
CON CLIENTE	Key
ES CREDITO	PredictOnly
MONTO TOTAL APROBADAS	Ignore
MONTO TOTAL DENEGADAS	Ignore
NOM PAIS	Input
PROMEDIO MONTO APROB...	Input
PROMEDIO MONTO DENEG...	Input
RANGO EDADES	Input
RANGO MONTO FALLIDOS	Ignore
RANGO MONTOS APROBAD...	Input
RANGO MONTOS DENEGADOS	Input
TOTAL DENEGADAS	Ignore
TOTAL FALLIDAS	Ignore

FIGURA 26: MODELO DE ÁRBOLES DE DECISIÓN.

Por último, el modelo de cluster, se va a basar en el "rango de edades", el "rango de montos aprobados", el "rango de montos denegados" y la variable "es crédito".



	Microsoft_Clustering
COD MERCHANT	Ignore
CON CLIENTE	Key
ES CREDITO	PredictOnly
MONTO TOTAL APROBADAS	Ignore
MONTO TOTAL DENEGADAS	Ignore
NOM PAIS	Ignore
PROMEDIO MONTO APROB...	Ignore
PROMEDIO MONTO DENEG...	Ignore
RANGO EDADES	Input
RANGO MONTO FALLIDOS	Ignore
RANGO MONTOS APROBAD...	Input
RANGO MONTOS DENEGADOS	Input
TOTAL DENEGADAS	Ignore
TOTAL FALLIDAS	Ignore

FIGURA 27: MODELO CLUSTERING.

C. EVALUAR MODELO

Para el proceso de evaluación se utiliza lo que se conoce como un gráfico de mejora ("lift chart"). Respecto al modelo predictivo representa gráficamente, la mejora que proporciona un modelo de minería de datos en comparación con una estimación aleatoria, y mide el cambio en términos de puntuación de la mejora respecto al modelo predictivo. Al comparar las puntuaciones de mejora respecto al modelo predictivo para las distintas partes del conjunto de datos y para los diferentes modelos, permite determinar cuál es el mejor modelo y qué porcentaje de casos del conjunto de datos se beneficiaría de aplicar las predicciones del modelo.

Con un gráfico de mejora respecto al modelo predictivo, se compara la precisión de las predicciones para varios modelos que tienen el mismo atributo de predicción. También, se evalúa la exactitud de la predicción para un único resultado (un único valor del atributo de predicción) o para todos los resultados (todos los valores del atributo especificado). Para nuestro caso, en específico, se evaluará el gráfico de mejora para la variable de "es crédito" para los valores de "Yes" y el valor de "No".

Para el valor de "Yes" se tiene:

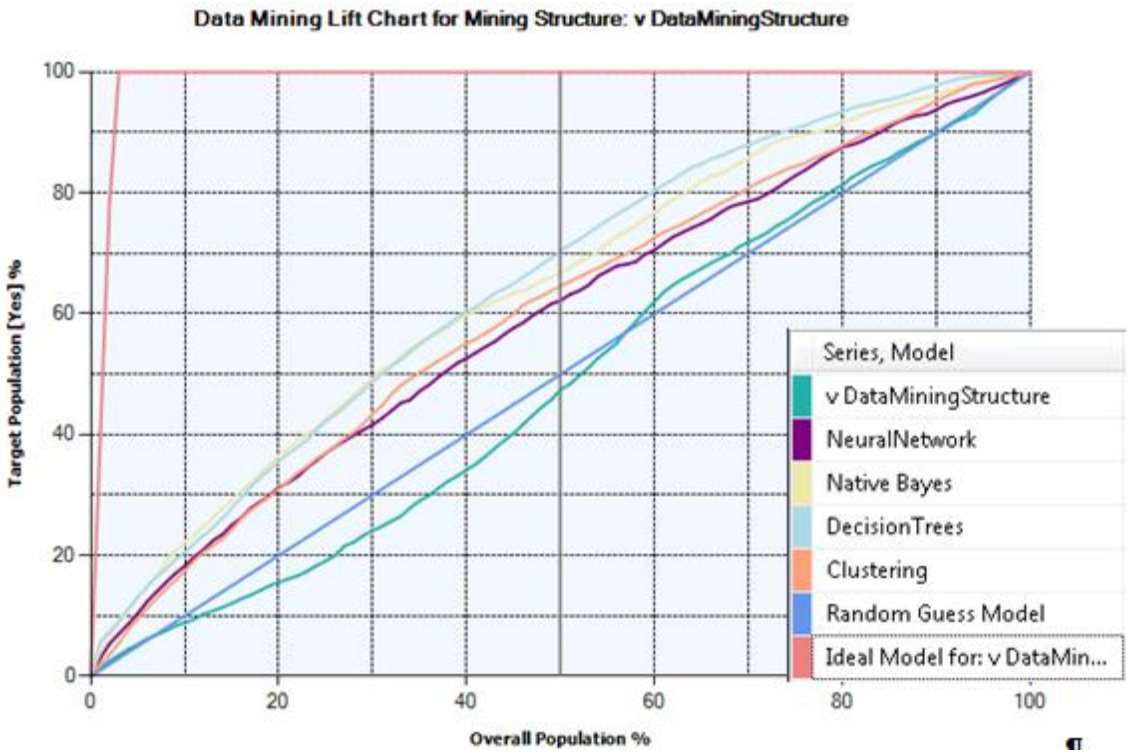


FIGURA 28: GRÁFICO DE MEJORA PARA "YES".

El eje X del gráfico representa el porcentaje del conjunto de datos de prueba que se usa para comparar las predicciones. El eje Y del gráfico representa el porcentaje de valores de predicción.

La línea recta diagonal, mostrada aquí en azul, que aparece en el gráfico, representa los resultados de la estimación aleatoria y es la línea base con la que se evalúa la mejora respecto al modelo predictivo. Con cada modelo que se agrega, se obtienen dos líneas adicionales: que muestra los resultados ideales para los conjuntos de datos de entrenamiento establecidos y la segunda línea, que muestra la mejora respecto al modelo predictivo real, o progreso en los resultados, para el modelo, que en este caso es la línea anaranjada que se encuentra más arriba, ésta para todos los modelos, lo que significa que el porcentaje de predicción para este modelo es bajo.

Además, en el gráfico, se observa, una línea vertical gris llamada leyenda de minería de datos, la cual, contiene los valores reales de cualquier punto de las curvas. En el gráfico, la línea gris se ha movido al 49.5%, porque se trata del punto donde tanto el modelo filtrado como el modelo sin filtrar parecen ser más eficientes, y después de este punto la cantidad de mejora respecto al modelo predictivo decae. La leyenda de minería de datos también, contiene puntuaciones y estadísticas que ayudan a interpretar el gráfico. Estos resultados representan la exactitud del modelo en la línea gris como se distingue en la siguiente figura:

Series, Model	Score	Target population	Predict probability
v DataMiningStructure	0.50	47.46%	2.55%
NeuralNetwork	0.60	62.20%	2.05%
Native Bayes	0.65	66.80%	1.82%
DecisionTrees	0.66	70.39%	2.66%
Clustering	0.61	64.58%	2.04%
Random Guess Model		50.00%	
Ideal Model for: v DataMin...		100.00%	

FIGURA 29: LEYENDA DEL GRÁFICO DE MEJORA PARA "YES".

Como se ve en la figura anterior, hay tres datos que se están valorando en el gráfico con base en la línea de minería, la puntuación ("score"), la población objetivo y la probabilidad de predicción, la puntuación es el indicador que compara un modelo con respecto a los otros.

En estos resultados, se observa, que, cuando se mide en el 49.5% de la población objetivo, los modelos predicen el comportamiento de créditos en un 47.46%, 62.20%, 66.8%, 70.39% y 64.58% de la población de destino respectivamente. En otras palabras, si aplicara el modelo de árboles de decisión al 49.5% por ciento de los clientes de la base de datos, este modelo podría ser efectivo a cerca del 70.39% la población analizada.

El valor de probabilidad de predicción representa el umbral necesario para incluir un cliente entre los casos "con probabilidad de crédito". Para cada

caso, el modelo calcula la exactitud de cada predicción y almacena ese valor, que se utiliza para filtrar o elegir clientes.

El valor de puntuación ayuda a comparar los modelos calculando la efectividad del modelo a través de una población normalizada. Una mayor puntuación es mejor, de modo que en este caso podría decidir que seleccionar el modelo de árboles de decisión es la estrategia más eficiente, a pesar tener una probabilidad de predicción baja.

Para el valor de "No" tenemos:

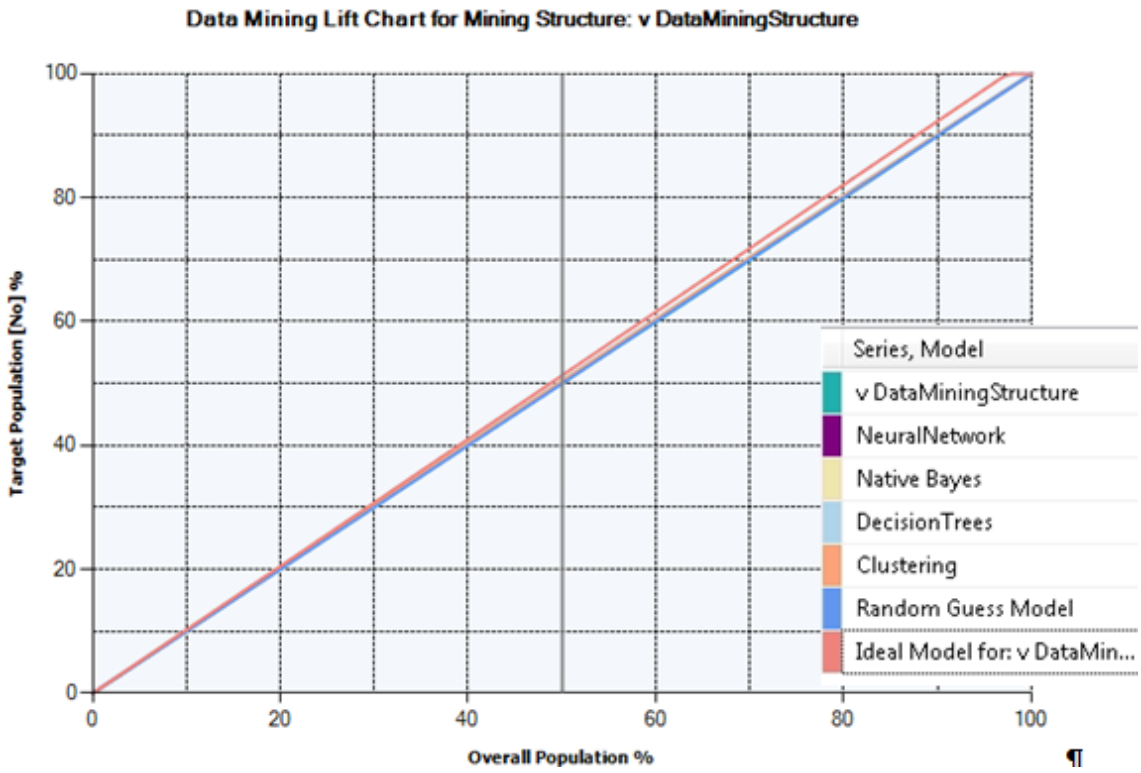


FIGURA 30: GRÁFICO DE MEJORA PARA EL "NO".

En este caso se observa como la población del "No", tiene las líneas bases, óptimas y las líneas de cada uno de los modelos convergiendo en la línea media del gráfico, que generan, puntaciones y predicciones de probabilidad muy altas para cada uno de los algoritmos, todo esto bajo una población cercana al 50% para la mayoría de los modelos utilizados.

Series, Model	Score	Target popu...	Predict probability
v DataMiningStructure	0.98	50.06%	97.45%
NeuralNetwork	0.98	50.31%	97.94%
Native Bayes	0.98	50.43%	98.18%
DecisionTrees	0.98	50.53%	97.34%
Clustering	0.98	50.38%	97.96%
Random Guess Model		50.00%	
Ideal Model for: v DataMin...		51.31%	

FIGURA 31 LEYENDA PARA EL GRÁFICO DE MEJORA DEL "NO"

Con estos resultados se ve que, cuando se mide en el 49.5% de todos los casos, los modelos predicen el comportamiento de créditos del "No" en un aproximado del 50% de la población de destino. Por ejemplo, si aplicara el modelo de árboles de decisión al 49.5% por ciento de los clientes de la base de datos, podría llegar a algo cerca del 50% la población analizada.

1.2.4.5. EVALUACIÓN

A. EVALUAR LOS RESULTADOS

La evaluación de resultados se realiza por modelo, para cada uno de los modelos se analiza con base en los objetivos del negocio cuáles reglas pueden aplicar y el por qué.

- Reglas de asociación.

Por medio del algoritmo de reglas de asociación es posible generar reglas que permiten predecir el resultado de uno o más atributos con base en diversos atributos de entrada.

Los datos de entrada para este modelo son: el nombre del país, el monto total de transacciones denegadas y de aprobadas, rango de montos aprobados y rangos de edad para predecir si hay crédito o no.

A continuación, se muestran las reglas encontradas ordenadas por probabilidad e importancia.

Probability	Importance	Rule
1.000	1...	NOM PAIS = Italy, RANGO MONTOS APROBADOS > 501 -> ES CREDITO = Yes
1.000	1...	NOM PAIS = South Africa, MONTO TOTAL DENEGADAS = 249.6229574656 - 1558.3832354816 -> ES CREDITO = Yes
0.800	1...	NOM PAIS = South Africa, RANGO MONTOS APROBADOS > 501 -> ES CREDITO = Yes
0.833	1...	NOM PAIS = Germany, RANGO MONTOS APROBADOS > 501 -> ES CREDITO = Yes
0.833	1...	NOM PAIS = Italy, MONTO TOTAL APROBADAS = 871.37379584 - 3545.7799606272 -> ES CREDITO = Yes
1.000	1...	NOM PAIS = India, RANGO MONTOS APROBADOS = 401-500 -> ES CREDITO = Yes
1.000	1...	NOM PAIS = India, RANGO MONTOS APROBADOS > 501 -> ES CREDITO = Yes
0.778	1...	NOM PAIS = Germany, MONTO TOTAL APROBADAS = 871.37379584 - 3545.7799606272 -> ES CREDITO = Yes
0.750	1...	NOM PAIS = Great Britain, RANGO MONTOS APROBADOS = 201-300 -> ES CREDITO = Yes
0.750	1...	NOM PAIS = South Africa, MONTO TOTAL APROBADAS = 871.37379584 - 3545.7799606272 -> ES CREDITO = Yes
0.750	1...	NOM PAIS = Germany, RANGO MONTOS APROBADOS = 301-400 -> ES CREDITO = Yes
0.667	1...	NOM PAIS = Australia, RANGO MONTOS APROBADOS = 401-500 -> ES CREDITO = Yes
0.667	1...	NOM PAIS = South Africa, RANGO EDADES = Joven Tardio -> ES CREDITO = Yes
0.667	1...	NOM PAIS = Ireland, MONTO TOTAL APROBADAS = 871.37379584 - 3545.7799606272 -> ES CREDITO = Yes
0.667	1...	NOM PAIS = Great Britain, RANGO MONTOS APROBADOS > 501 -> ES CREDITO = Yes
0.600	1...	NOM PAIS = South Africa, RANGO EDADES = Adulto -> ES CREDITO = Yes
0.571	1...	NOM PAIS = Germany, RANGO MONTOS APROBADOS = 401-500 -> ES CREDITO = Yes
0.538	1...	NOM PAIS = Belgium, RANGO MONTOS APROBADOS > 501 -> ES CREDITO = Yes
0.500	1...	NOM PAIS = South Africa -> ES CREDITO = Yes
0.500	1...	NOM PAIS = India, MONTO TOTAL APROBADAS = 871.37379584 - 3545.7799606272 -> ES CREDITO = Yes

FIGURA 33: REGLAS DE ASOCIACIÓN.

Es posible analizar ciertos aspectos interesantes a partir de las reglas generadas. Se listan algunas que predicen créditos:

- Cuando el cliente es de Italia y su rango de montos aprobados es mayor a $501, existe la posibilidad de incurrir en créditos.
- Se observa, que cuando el cliente es de Sudáfrica y las denegadas se encuentran entre $249 y $1558 es posible que realicen créditos.
- También, para clientes de Sudáfrica, cuando los clientes se encuentran en el rango Adulto o Joven Tardío, existe posibilidad de incurrir en créditos.

Si se analizan globalmente las reglas, se aprecia que cuando los países son diferentes a USA y los montos totales tienden a ser elevados, se incrementa la probabilidad de que el cliente realice créditos.

- Árboles de decisión.

El algoritmo de árboles de decisión presenta el siguiente diagrama de dependencia:

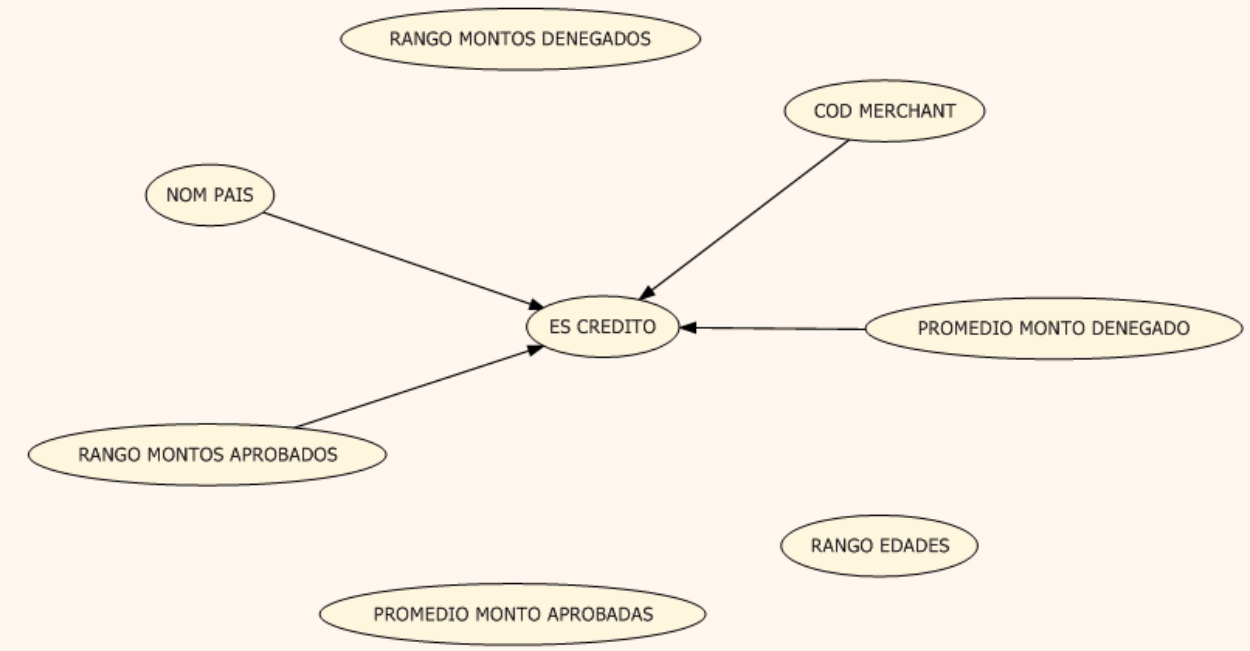


FIGURA 32: DEPENDENCIA DE ARBOLES DE DECISIÓN.

Este diagrama muestra las siguientes asociaciones establecidas por orden de fuerza en su dependencia, es decir, las relaciones más fuertes encabezan la lista.

1. Rango de montos aprobados.
2. Cod_merchant.
3. Promedio monto denegado.
4. Nom_país.
5. Promedio monto aprobadas.
6. Rango edades.
7. Rango montos denegados.

Para este análisis sobre el árbol se observan las secciones que

favorecen la presencia de créditos. En primera instancia, como se observa, en la fuerza de relación de las variables del modelo, el "rango de montos aprobados" es la relación más fuerte, por ende es la raíz del árbol; en este sentido se analizan las derivaciones de los rangos de $300 a $400, de $400 a $500 y de $500 en adelante; esto se observa a continuación.

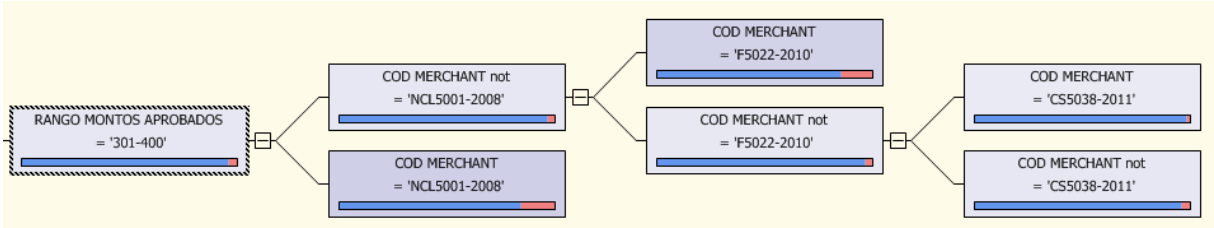


FIGURA 33: RAMA ÁRBOL DE DECISIÓN.

Como se distingue en la figura anterior, que corresponde al rango de montos aprobados entre $300 y $400, según el árbol, hay en el comercio "NCL5001-2008" tiene una probabilidad de créditos de un 16.19%, muy por encima del indicador de negocio menor al 4% que se considera normal, a continuación está el detalle:

Value	Cases	Probability	Histogram
Missing	0	0.00%	
No	233	83.81%	
Yes	45	16.19%	

FIGURA 34: LEYENDA ARBOL DE DECISIÓN 1.

Luego de este nodo del árbol, está el comercio "F5022-2010", este comercio presenta una situación similar al anterior, la probabilidad de crédito está muy por encima de lo que se considera normal, como se ve en el siguiente detalle.

Value	Cases	Probability	Histogram
Missing	0	0.00%	
No	119	84.99%	
Yes	21	15.01%	

FIGURA 35: LEYENDA ARBOL DE DECISIÓN 1.

Ambos comercios en ese rango de procesamiento de sus clientes tienen problemas de créditos, pues, la probabilidad es más alta de lo que se considera normal, para el resto de comercios, por ejemplo, para "CS5038-2011" la probabilidad corresponde al 2.08% y para los demás comercios es del 4.13%, rangos que se consideran aceptables para negocio.

Ahora para el rango de $400 a $500 dólares, está la siguiente figura.

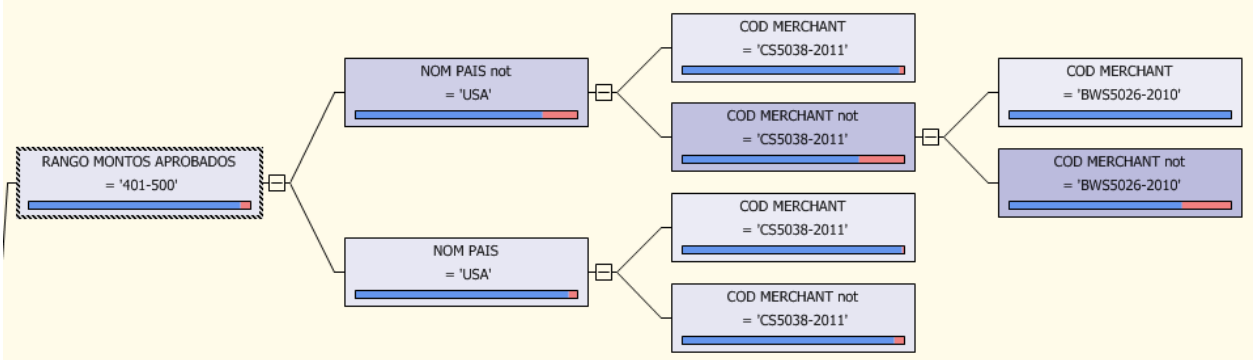


FIGURA 36: RAMA ÁRBOL DE DECISIÓN.

Para esta sección del árbol hay países que no son Estados Unidos y para comercios que no son ni "CS5038-2011" ni "BW55026-2010" (comercios que procesan clientes mayoritariamente de Estados Unidos) la probabilidad de que una transacción sea un crédito es considerable, 22,65%, como se verifica en la siguiente figura:

Value	Cases	Probability	Histogram
Missing	0	0.00%	
No	224	77.35%	
Yes	65	22.65%	

FIGURA 37: LEYENDA PARA ÁRBOL DE DECISIÓN 2.

Se analiza el siguiente rango, que corresponde a montos aprobados por encima de los $500 se presenta un comportamiento similar al anterior para los países que no son Estados Unidos.

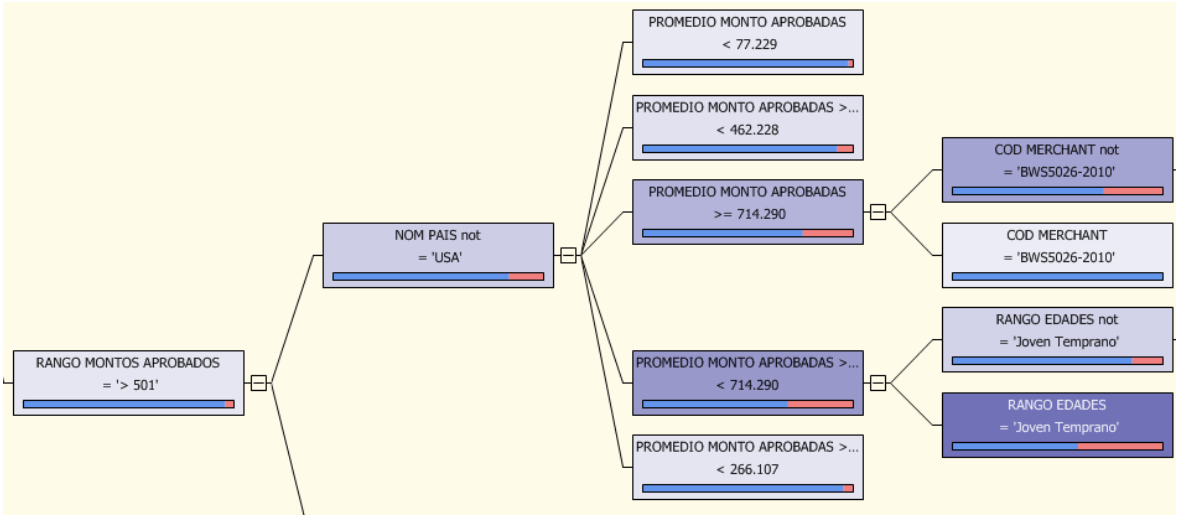


FIGURA 38: RAMA ÁRBOL DE DECISIÓN.

Para el caso en específico, si el país no es Estados Unidos, si el promedio de monto aprobado por cliente es mayor a $714 y el comercio no es "BWS5006-2010" (esto puede explicarse debido a que este comercio procesa clientes de Estados Unidos, casi en su totalidad) la probabilidad de créditos se dispara, a un 28.71%.

Value	Cases	Probability	Histogram
Missing	0	0.00%	
No	107	71.29%	
Yes	43	28.71%	

FIGURA 39: LEYENDA ÁRBOL DE DECISIÓN 3.

De la misma forma, para clientes que no son de Estados Unidos y el promedio de aprobadas es mayor o igual a $462 y menor a $714, es decir, un rango menor al analizado en el caso anterior, la probabilidad de créditos se mantiene alta (31.71%) como se observa en la siguiente figura:

Value	Cases	Probability	Histogram
Missing	0	0.00%	
No	222	68.29%	
Yes	103	31.71%	

FIGURA 40: LEYENDA.

Es particularmente, alta para la población que se refiere a los "jóvenes tempranos", con una probabilidad de que si exista créditos del 40.76%, mientras los demás rangos de edades representan una probabilidad del 14% dentro de su población.

Value	Cases	Probability	Histogram
Missing	0	0.00%	
No	125	59.24%	
Yes	86	40.76%	

FIGURA 41: LEYENDA.

- Clúster

Para este modelo de clúster, no se designa una columna de predicción. El algoritmo de clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

Es decir, agrupa por medio del entrenamiento que realiza el algoritmo y a nivel poblacional se observan los clústeres que se crean y su respectiva densidad y fuerza entre las relaciones de los elementos que lo componen, se distinguen, los clúster 4 y 9 tienen una relación fuerte con el clúster 8, al igual que el 5 y el 9 con el clúster 10, no así los clúster 1, 2 y 3 que no tienen una relación estrecha con el resto de clústeres encontrados

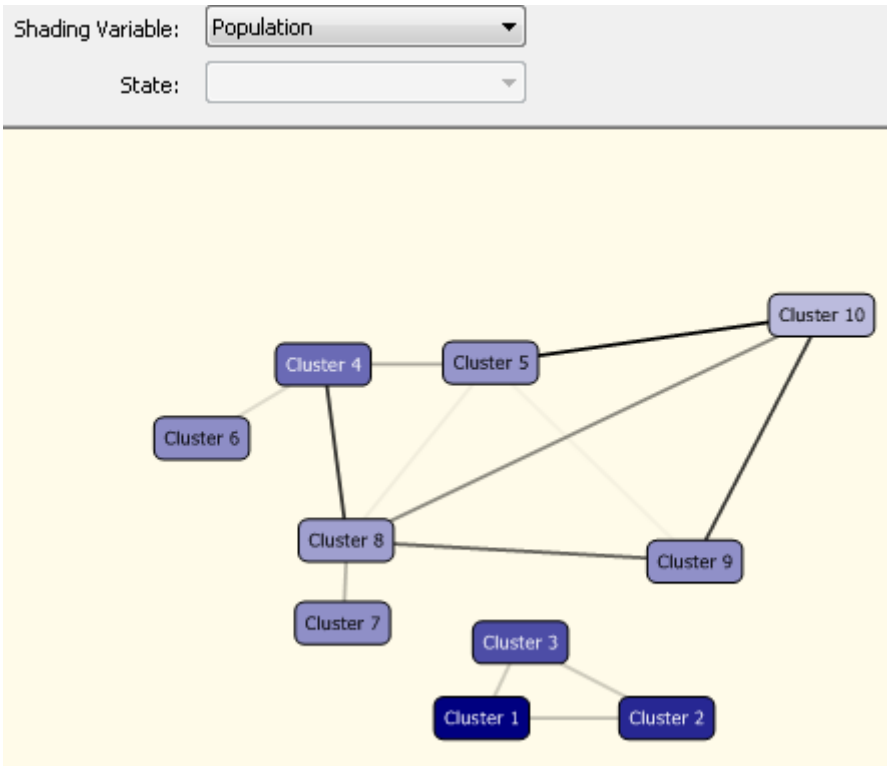


FIGURA 42: DIAGRAMA DE DEPENDENCIA DEL CLÚSTER.

Este es el panorama global, de toda la población, ahora se trata, de caracterizar y dar más énfasis al clúster con la población más densa de créditos, con el fin de conocer sus características.

La siguiente figura muestra el diagrama de densidad con base en la variable "es crédito" y su valor de "Yes".

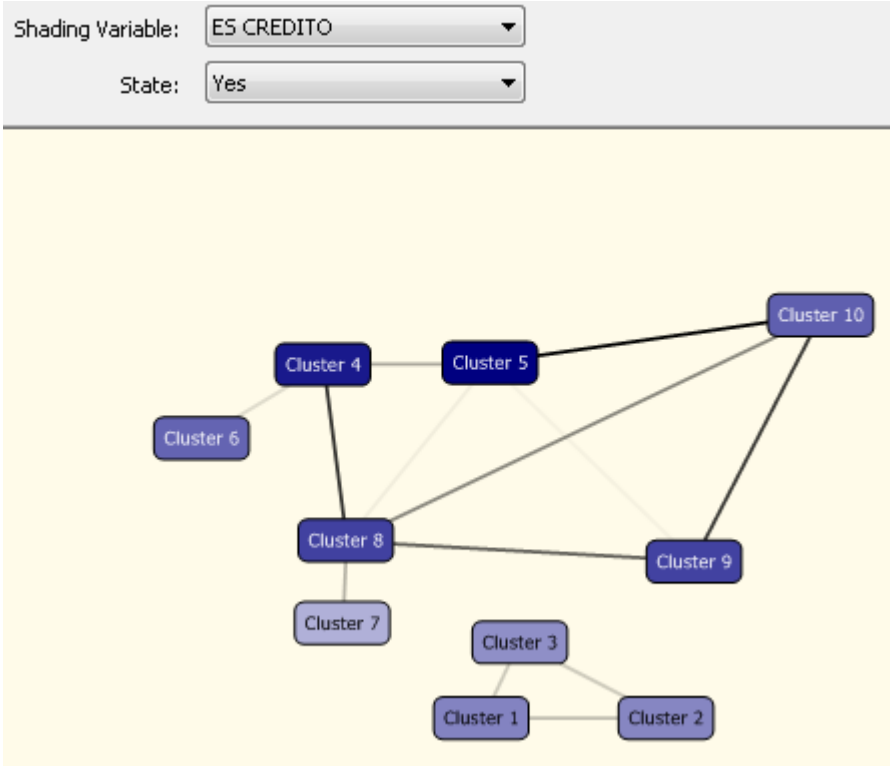


FIGURA 43: DIAGRAMA DE DEPENDENCIA DEL CLÚSTER PARA CREDITOS "YES".

Gráficamente, se observa como los clúster 4 y 5 presentan una mayor densidad con respecto a la existencia de créditos en esa agrupación de poblaciones. A continuación, se observa las características de ambos clústeres de forma más detallada.

FIGURA 44: CARACTERÍSTICAS DEL CLÚSTER 5.

El clúster 5 tiene una gran densidad de valores para los créditos de valor "No", de hecho todos los clúster tienen una alta densidad de este valor, debido a que es el valor más común por lo que la variable no es valiosa para el valor del "No", se percibe, como montos pequeños de $0 a $50 denegados y rangos mayores a $500 aprobados son características importantes para este clúster, al igual que las edades de adulto tardío, adulto medio y adulto y los rangos de montos superiores a los $300.

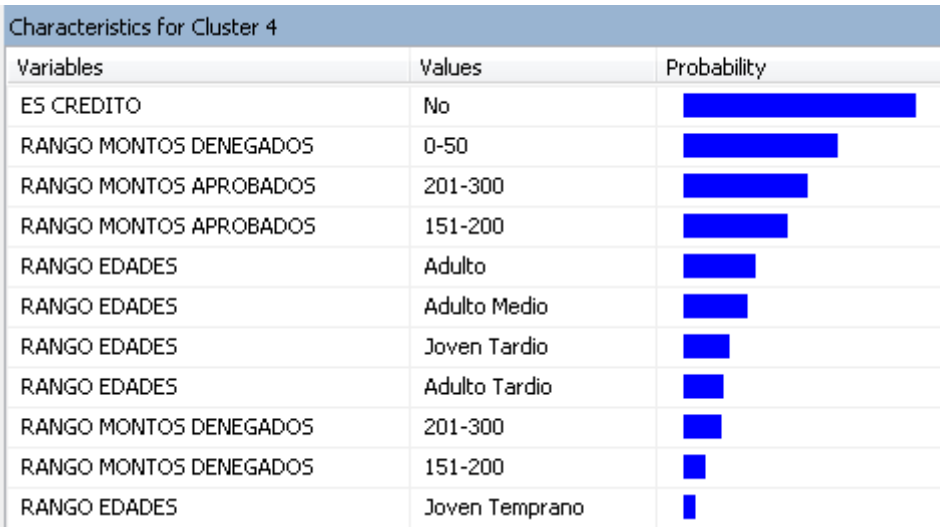


FIGURA 45: CARACTERÍSTICAS DEL CLÚSTER 4.

El clúster 4 por su parte presenta montos aprobados menores, y edades inferiores al clúster 5, montos entre $150 y $300 y rangos de edades entre joven tardío y adulto tardío.

En las siguientes figuras, se ilustran las principales características de estos clústeres con respecto a la densidad de los valores poblacionales característicos.

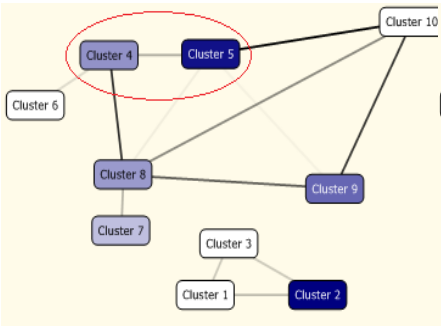
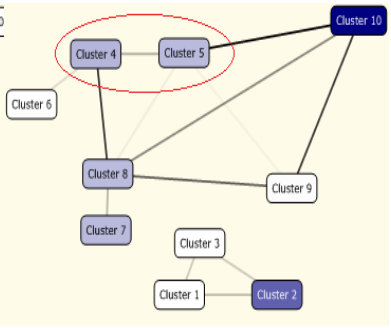
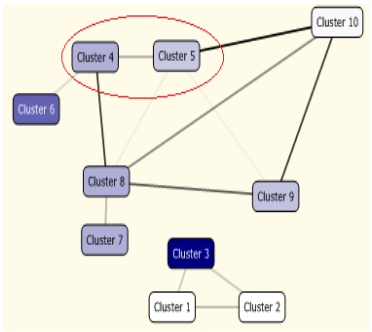
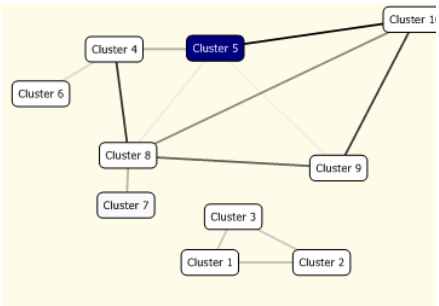
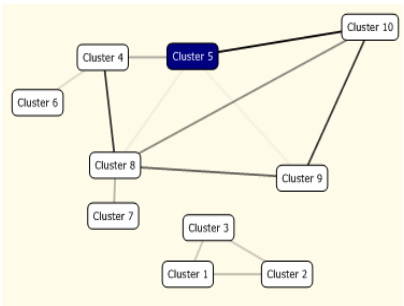
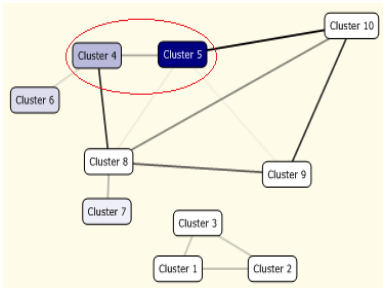
Adulto tardío	Adulto Medio	Adulto
		
Mayor a quinientos dólares	De cuatrocientos a quinientos	Trescientos a cuatrocientos
		

TABLA 16: CLÚSTER DE POBLACIONES DE CARACTERÍSTICAS AFINES.

Básicamente, se ve que la concentración de créditos en ambos clústeres están relacionadas a las edades y montos de rangos transaccionales, entre

mayor es la persona y mayor es el monto es más característico para estos dos clústeres.

Otra forma de ver estos clústeres su gráfico de perfil, que nos permite ver de una forma más clara sus principales características, para el clúster 5 el valor de "No" es predominante, pero, el valor de "Yes" es el más predominante en el resto de clústeres, los rangos de edades por orden tamaño son, adulto tardío, adulto medio, adulto y joven tardío, los montos que procesan los clientes de este clúster superan los $300 y deniegan montos muy pequeños o muy grandes.

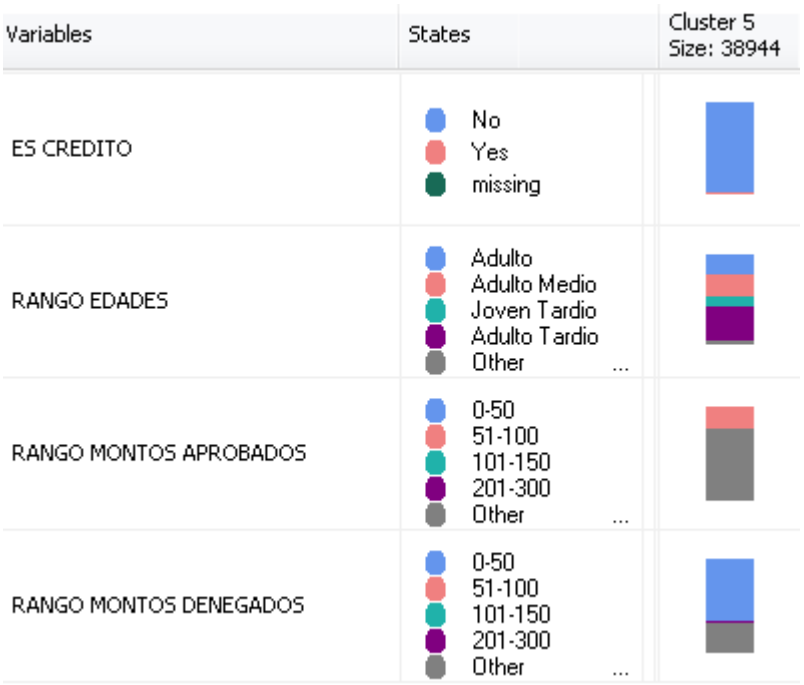


FIGURA 46: PERFIL DE CLÚSTER 5.

Para el clúster 4, hay un comportamiento similar con los valores para la variable de "es crédito", el "No" es mayoría, en el clúster, pero, en general en este clúster presenta los valores más densos; hay una mayor predominancia de los jóvenes tardíos con respecto al clúster 5, sin embargo, se mantienen los

principales rangos de edades; a nivel de tamaño estas poblaciones dentro del clúster mantienen la proporción del clúster 5 y con respecto al rango de montos aprobados procesa montos mayores a los $200 dólares, el clúster 5 se caracterizaba por procesar montos mayores a $300, básicamente, se agrega un rango un poco más permisivo con lo que respecta a un monto alto.

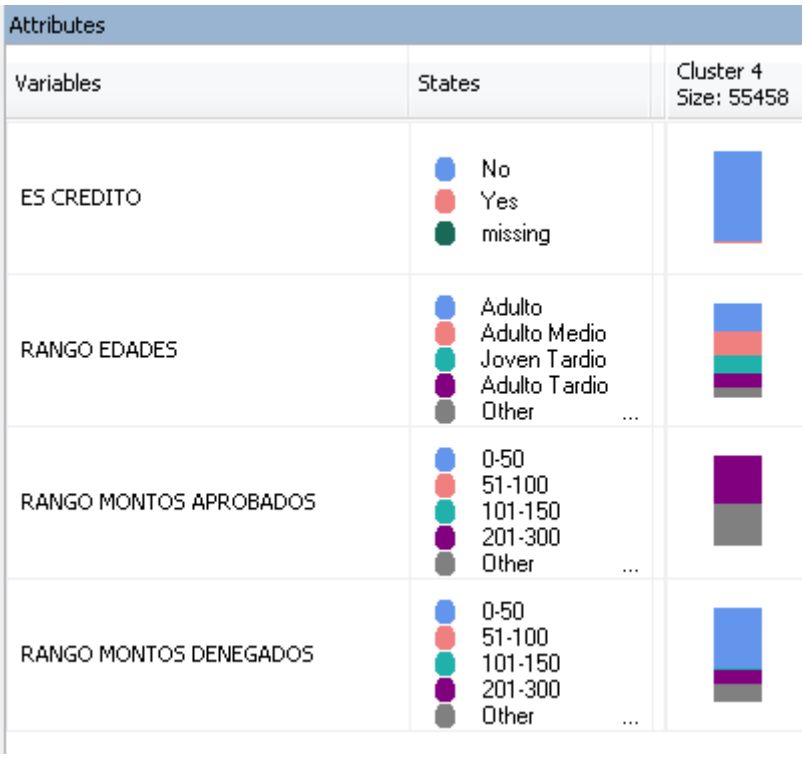


FIGURA 47: PERFIL DE CLÚSTER 4.

- "Neural Network"

El algoritmo “Neural Network” a través de probabilidades y estadística permite modelar predicciones binarias, es decir, salidas de sí o no, entre otras.

En el caso de la investigación fue aplicado para verificar si los valores de entrada predicen si los clientes pueden o no hacer créditos. A continuación se listan varios ejemplos.

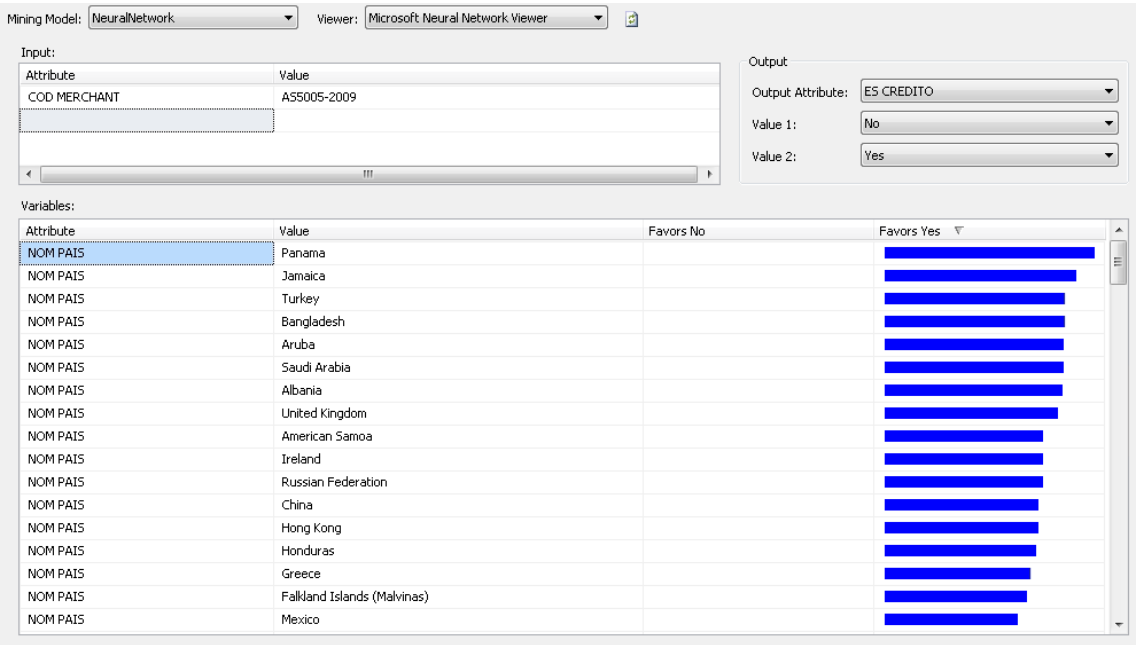


FIGURA 48: REDES NEURONALES, PARA AS5005-2009.

En la figura 49 se observa que al analizar el país del cliente, existen algunos valores que favorece la aparición de créditos. Algunos de los países incluyen a Panamá, Jamaica, Turquía, entre otros.

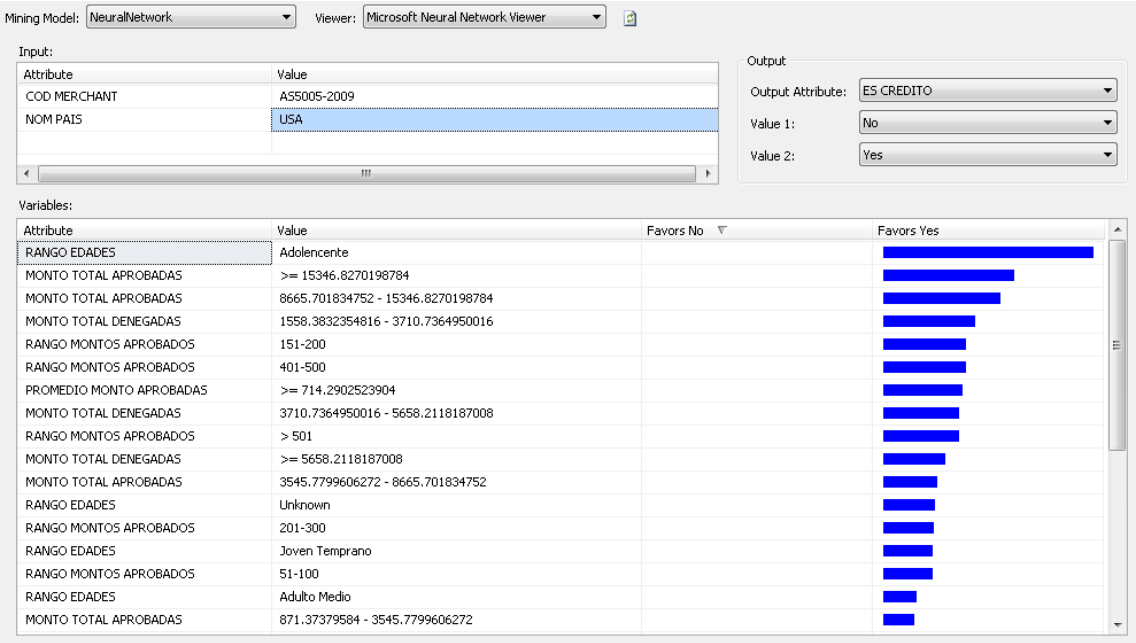


FIGURA 49: REDES NEURONALES, PARA AS5005-2009 Y USA.

En la figura 50 se analizan los valores que favorecen la existencia de créditos para el comercio elegido y para el país USA. A continuación se analizan algunos de los aspectos observados que favorecen la aparición de créditos, ordenadas según, su importancia.

- Es posible analizar que el rango de edades de adolescentes favorece la aparición de créditos con alto grado de importancia.
- Si el monto total de aprobadas se encuentra entre $8600 y $15346.
- Si el rango de montos aprobados varía entre $151 y $200.
- Si el monto total de transacciones denegadas es mayor a 5658.
- Si el rango de edades es Joven temprano, o Adulto medio.

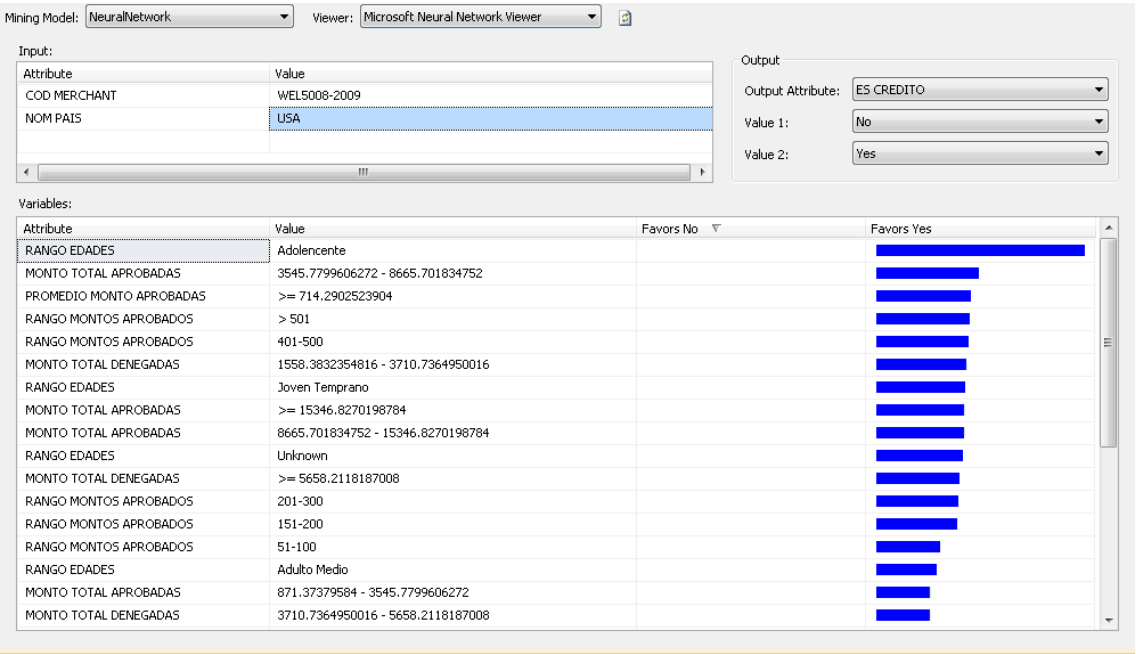


FIGURA 50: REDES NEURONALES, PARA WEL5008-2009.

En la figura 51 se puede apreciar el análisis para un Comercio diferente. Sin embargo, se observan datos similares. A continuación se detallan algunos atributos y valores que favorecen el crédito de los clientes ordenados por su importancia.

- El cliente está en el rango de edades Adolescente.
- Tiene un monto total de aprobadas entre $3,545 y $8,665.
- Tiene rangos de montos aprobados mayores a $501.
- El cliente está en el rango de edades Joven Temprano.
- Tiene montos totales de denegadas mayores a $5,658.

La similitud de los valores que favorecen la existencia de créditos permite identificar de forma más sencilla las características que influyen con mayor probabilidad en la incidencia de un crédito. Como se observa los clientes que procesan en su historial montos mayores a $3,545 y además, promedian por transacción montos mayores a $501, tienden a incidir en créditos.

- "Naive Bayes"

La técnica de clasificación Bayes Inocente, permite identificar ciertos atributos que tienen mayor probabilidad de incidir en un crédito.

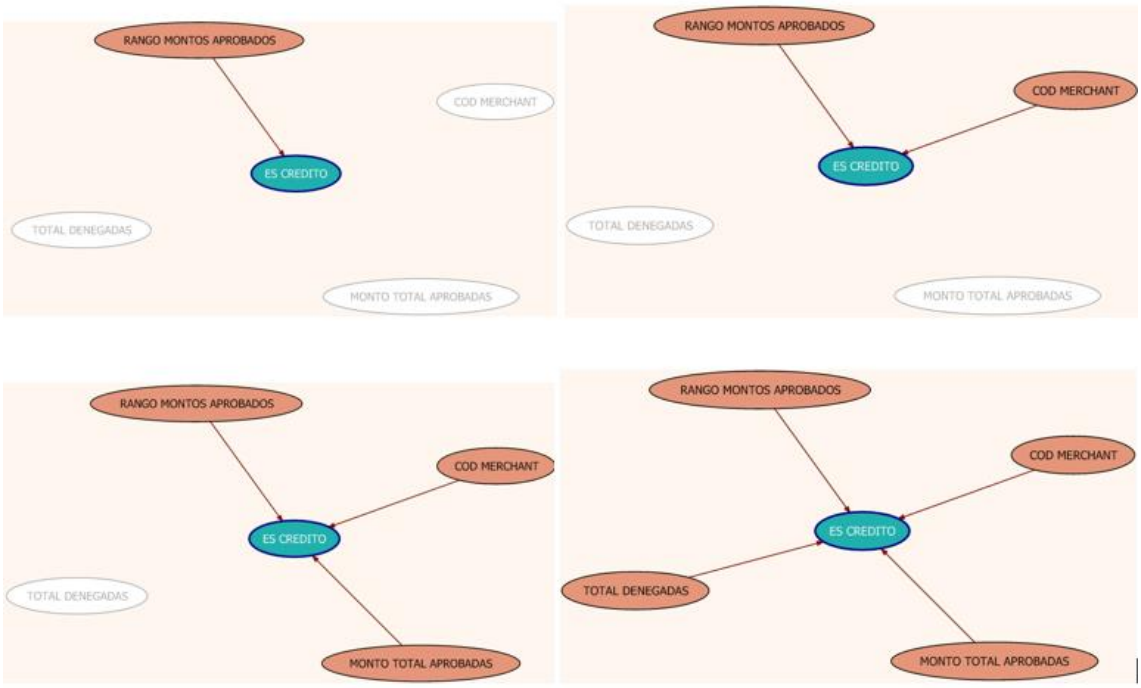


FIGURA 51: NAIVE BAYES DEPENDENCIA.

Fue posible analizar que los atributos que tienen mayores posibilidades de incidir en un crédito son los siguientes, ordenados de mayor a menor importancia; son:

- Los rangos de los montos de transacciones aprobados.
- El comercio.
- Monto total de aprobadas.
- Total de transacciones denegadas.

Posteriormente, con base en estos atributos, se generó un agrupamiento que permite clasificar los atributos que intervienen en un crédito.

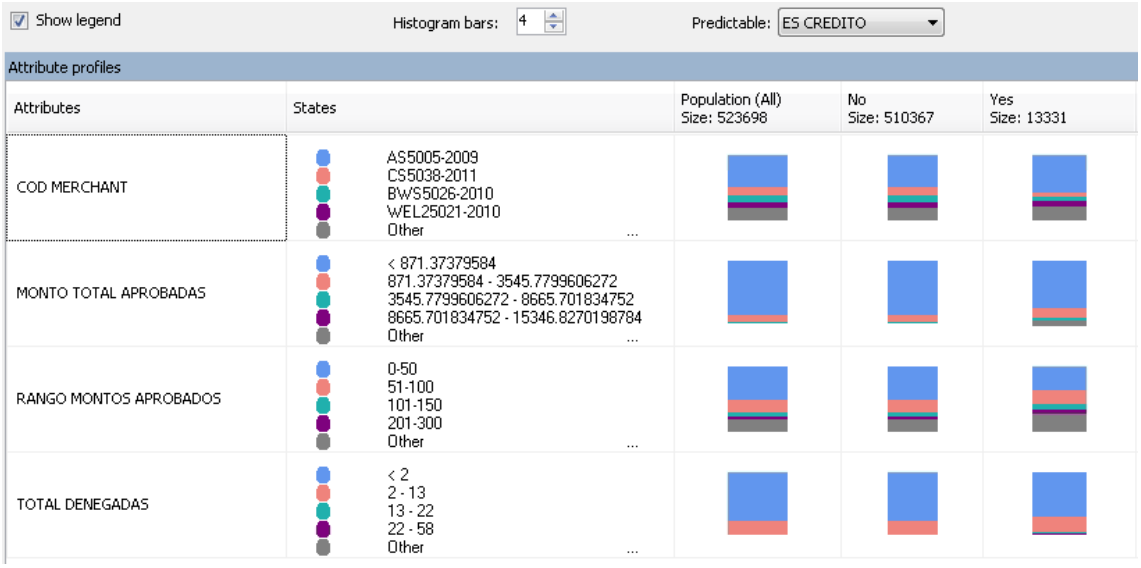


FIGURA 52: NAIVE BAYES CLÚSTER.

En la figura 53 es posible identificar los atributos que intervienen en la incidencia o no de un crédito.

Se analizan que algunos comercios influyen en la generación de créditos, así como también, los montos totales de transacciones aprobadas mayores a $871 hasta $8,600, influyen en un crédito. Los rangos de montos aprobados que favorecen la incidencia de créditos son mayores que los rangos de montos aprobados que no generan créditos. Posteriormente, por el total de denegadas, es posible determinar que los clientes con 22 a 58 transacciones denegadas realizan créditos.

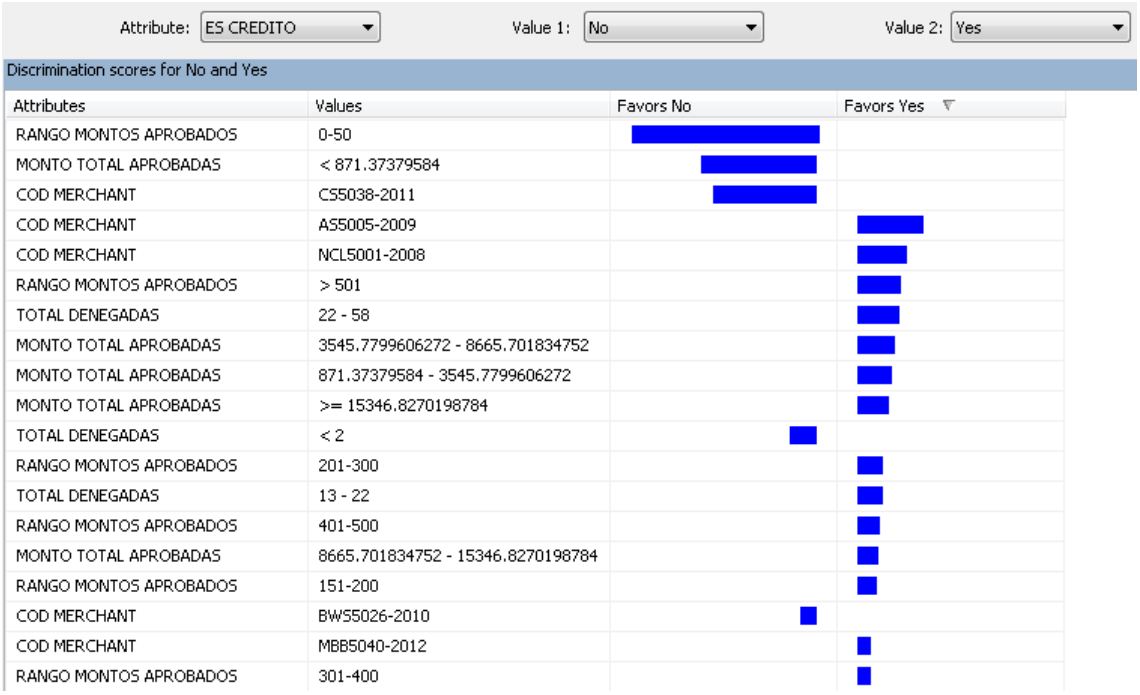


FIGURA 53: NAIVE BAYES, FAVORECIMIENTOS.

Finalmente, en la figura 54 se aprecia qué tanto influye o no en la aparición de créditos el valor de un atributo en específico.

Se destaca que entre los valores que no favorecen créditos, se incluyen montos menores a $871, así como rangos de montos aprobados de 0 a $50. Se destaca el comercio CS500382011 por su tendencia a no incidir en créditos.

Entre los valores que sí favorecen los créditos, se encuentra el comercio AS50052009, así como también, el NCL50012008. Los montos aprobados mayores a $501, el total de denegadas mayor a 22, o rangos de montos elevados; influyen en la incidencia de créditos.

2. DISCUSIÓN DE LOS RESULTADOS

En esta sección de la investigación se hace un repaso de los resultados obtenidos de todos los pasos realizados, basados en la aplicación de dos metodologías establecidas en el estudio, una para el desarrollo del proyecto de "data warehouse" y la otra para la aplicación de técnicas de minería y análisis de los datos.

2.1. BUSINESS DIMENSIONAL LIFECYCLE

Para el desarrollo del "data warehouse" se decidió trabajar con la metodología Business Dimensional Lifecycle, expuesta por Kimball para abarcar todas las etapas del desarrollo de un proyecto de este tipo.

En la primera fase de la investigación, se realiza la planificación del proyecto. En ésta se indican varios aspectos definidos a lo largo de la investigación, entre ellas la delimitación de alcances, la creación del cronograma de actividades con la estimación en tiempo y recursos a utilizar. Esta fase permitió establecer las actividades necesarias para cumplir con los objetivos planteados.

La segunda fase fue una de las más importantes, en la misma se definieron los requerimientos del negocio. La recolección de datos fue realizada a través de entrevistas. Fue clave definir el personal a entrevistar para recolectar datos según, las responsabilidades de los posibles usuarios de la mejor forma posible, posteriormente, se listaron los puntos más importantes en cuánto a requerimientos y solicitudes del personal entrevistado, esta fase sirvió de base para la fase del diseño de la arquitectura y del diseño del cubo.

En la fase de diseño de la arquitectura, se realizó un diagrama que ayudó a identificar los componentes necesarios para implementar la solución planteada en el proyecto. La idea de la arquitectura es contar con la especificación de los componentes donde se estará trabajando para alcanzar

los diferentes objetivos planteados, lo que permite, cubrir las necesidades del sistema. En esta fase se definió que se tendrán los datos de un sistema fuente que pasarán por el proceso de ETL, y luego serán almacenados en el "data warehouse".

En la fase de selección de tecnologías se procede a definir las tecnologías a implementar en la arquitectura planteada. Como la base de datos transaccional fue implementada con "SQL Server 2008 R2", se decidió seguir utilizando las tecnologías de Microsoft, por lo que el proceso de ETLs se desarrolló con "SQL Server Integration Services" y el cubo se implementó con "SQL Server Analysis Services". La selección de estas tecnologías permitió definir cómo implementar el proyecto planteado ajustándose a las especificaciones técnicas de los productos.

Luego de verificar la arquitectura técnica y seleccionar las tecnologías a utilizar, se procede a la ruta de los datos, en ésta primero se trabajó con la fase de modelado dimensional. En esta etapa se identificaron los procesos de negocio que iban a ser modelados, una vez que dichos procesos fueron analizados y verificados se definió la granularidad con la que los usuarios querían analizar la información. Posteriormente, la tarea fue proceder a evaluar las dimensiones que interactúan en los procesos de negocio analizados, y finalmente, elegir las tablas de hecho, que determinan las medidas que los usuarios consultarán. Esta fase permitió diseñar y realizar el modelado lógico del cubo que luego sería implementado físicamente.

En la siguiente fase se analizó cómo sería implementado el diseño físico, para esta sección se utilizó como base el diseño lógico, y se tomaron decisiones en la elección y análisis de la implementación de índices, definición de los indicadores de desempeño, jerarquías y columnas calculadas. Esto permitió determinar cómo sería implementado el cubo físicamente. Los indicadores de desempeño permitieron observar y analizar varios comportamientos muy significativos para la toma de decisiones en el negocio.

La fase siguiente, consistió en diseñar el desarrollo del "data staging".

Básicamente, al tenerse solo un sistema transaccional el proceso se simplificó considerablemente, al punto de no considerar necesario un área de "staging", y en lugar de esto aplicar las transformaciones directamente, en un solo proceso de extracción, transformación y carga realizado en paquetes de SSIS. Esta fase permitió diseñar y desarrollar el proceso en el que se definen los datos que fueron extraídos del sistema fuente y cargados al "data warehouse", permitiendo poblar de datos el mismo.

La fase posterior consistió en la especificación de aplicaciones analíticas. En la misma se detalla la creación de los reportes y se establece cómo serán utilizadas las herramientas dispuestas para dichas tareas. Fue importante, definir que los reportes tenían que aportar un aspecto dinámico y que los usuarios finales pudieran generar sus propios reportes. También, se procedió a plantear ciertos reportes predefinidos. Esta fase sirvió para plasmar los requerimientos recolectados en el diseño de los reportes.

En la siguiente fase de desarrollo de aplicaciones analíticas, se desarrollaron los reportes planteados anteriormente. En esta fase se aprecia más profundamente, el valor del proyecto de investigación, al ver finalmente, los reportes y analizar los datos como usuarios finales. Esta fase fue de mucho valor, ya que, fue la que mostró y presentó los datos, se brinda, el soporte a la toma de decisiones en EPPL.

2.2. CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING.

Para la aplicación de técnicas de minería de datos dónde se decidió trabajar con la metodología “Cross Industry Standard Process for Data Mining”.

En la primera fase, de esta metodología, se realizó el proceso de entendimiento del negocio donde se determinaron los objetivos con respecto a la minería, se evaluó la situación de las necesidades del negocio y las metas a nivel de la exploración de los datos, también, se llegó a delimitar el alcance de esta sección y se asentaron las bases para la siguiente etapa.

La segunda fase, se basó en la recolección, descripción y exploración de los datos a explotar, conociendo formatos, fuentes, agrupaciones, tipos de datos y características generales de la información analizada. Una vez que esta etapa se logró, se pasó a la extracción y selección de los datos, su limpieza, integración y formato respectivo según, se requirió para iniciar con la etapa de modelado.

La etapa de modelado se basó en la selección de técnicas y la generación de los modelos respectivos, cada modelo utilizado representó, una técnica de minería de datos, la cual, bajo sus características, se ajustaba a cumplir el objetivo de encontrar los datos afines a la generación de créditos de los clientes de la organización.

Luego de esto se pasa a la etapa de evaluación, donde cada modelo se puso a prueba y se realizó un análisis de los datos, determinando y evaluando el proceso con el fin de dar recomendaciones fundamentadas al negocio bajo la iteración del proceso de minería realizado, se analizaron, los modelos para determinar la efectividad de los mismos con base en una población característica. Además, se lograron identificar cuáles características están asociadas a la generación de créditos, que era el objetivo fundamental en esta fase de la investigación.

CONCLUSIONES

Atendiendo al objetivo de determinar una metodología para la construcción del cubo, se determina, que la metodología "Business Dimensional Lifecycle", expuesta por Kimball, es válida para el desarrollo de la investigación. La misma se divide en varias fases que facilitaron la construcción del cubo sirviendo como guía para cada etapa del proyecto, y se permite cumplir los objetivos planteados.

Se determinó que el proyecto de investigación permite cumplir los requerimientos del negocio, ya que, los reportes que tenía EPPL no satisfacían al personal gerencial. A través de la entrevista se determinó que el proceso de negocio, más importante, es el procesamiento de transacciones y se plantearon las entidades que intervienen en el mismo.

En relación al objetivo de identificar las tecnologías para la construcción del cubo; por las condiciones de operación del sistema transaccional de EPPL, se identificó continuar con tecnologías de Microsoft, que son las herramientas que mayor grado de compatibilidad ofrecían. El uso de estas herramientas permitió integrar fácilmente, el cubo creado con "SQL Server Analysis Services" con la base de datos transaccional en "SQL Server 2008R2". También, se aprovecha que los colaboradores de la empresa ya están familiarizados con estas herramientas, por lo que, el mantenimiento y capacitación resultan más sencillos. Las herramientas de reportes con Excel, específicamente "Power View", permitieron cumplir con los requerimientos de reportes ad hoc y dinámicos.

Para el diseño del cubo, se expone que los pasos indicados en la metodología permiten identificar fácilmente, los procesos de negocio involucrados, determinando la granularidad que tendrá el cubo e identificando las dimensiones y tablas de hecho que forman parte en los procesos de negocio seleccionados. Se determina que a través de estos pasos se facilita el

diseño del cubo, identificando rápidamente, las dimensiones y hechos que lo componen.

Con respecto a los reportes diseñados, se determina, que el desarrollo de los mismos es la parte que permite plasmar los requerimientos recolectados en información primordial y vital para los usuarios finales, dando un adecuado soporte a la toma de decisiones. Los indicadores de desempeño permiten, a los usuarios gerenciales, estar al tanto de comportamientos clave en varias entidades como canales de pago o bancos, tomando acciones cuando se presenten valores anormales.

También, se establece que la herramienta "Power View" de Excel Services permite diseñar reportes con un aspecto dinámico para los usuarios finales. Es de mucho valor la presentación de los resultados, y permitió corroborar el éxito del proyecto.

Para cumplir con el objetivo de identificar problemas específicos para el análisis con minería de datos, la primera fase de la metodología CRISP ayudó a formular e identificar el problema de la incidencia de créditos con respecto al comportamiento de los clientes. A través de la metodología CRISP es posible desarrollar un proyecto de minería de datos, comenzando por identificar y plantear un problema específico del negocio, como lo fue el caso de generación de créditos por cliente; se establecen objetivos, facilitando el entendimiento del problema y comprensión de los datos a utilizar para construir, evaluar los modelos, y finalmente, descubrir los patrones obtenidos.

Se atiende el objetivo de encontrar patrones ocultos de interés para el negocio, se encuentra que existen varias técnicas de minería de datos que exponen reglas, agrupaciones o clasificaciones que al ser analizadas permiten encontrar y determinar patrones. Es posible, determinar conclusiones de interés, con base en los patrones encontrados al aplicar las técnicas de minería. Entre ellas se listan las siguientes:

- Se identifican comercios que poseen mayor probabilidad de generar créditos.
- Se determina que el país del cliente influye en la incidencia de créditos. Los créditos son comunes en clientes fuera de USA. Esto en combinación con clientes que realizan transacciones por montos elevados, resulta en una alta probabilidad de generar créditos.
- El rango de edad identificado como Adulto Tardío en conjunto con montos elevados, favorece la incidencia de créditos.

RECOMENDACIONES Y TRABAJO FUTURO

Basados en el análisis de los resultados obtenidos, se recomienda a continuación un conjunto de mejoras para que la empresa EPPL ponga en ejecución.

Llevar a cabo un proceso más detallado de limpieza de datos, ya que, es posible que los datos que fueron cargados al cubo presenten cierto grado de impureza. Se recomienda refinar el proceso de extracción, transformación y carga de datos para asegurarse que los datos tengan un mayor grado de calidad.

Se recomienda, la implementación de una arquitectura de hardware más adecuada tomando en cuenta el uso y crecimiento del proyecto. También, se recomienda realizar una optimización en la administración del "data warehouse". Por el alcance del proyecto no se tomó en cuenta las especificaciones y configuraciones para un óptimo rendimiento del cubo, sin embargo, se exhorta, llevar a cabo un proceso de administración de particiones para almacenar los datos de forma más organizada.

Del mismo modo, se aconseja, enriquecer más la parte de análisis de datos. En cuanto a los reportes, es posible diseñar una mayor cantidad de reportes predeterminados. Asimismo, se propone integrar todos los reportes que fueron diseñados en un administrador de sitios o portal web, donde los reportes puedan ser diseñados y compartidos entre el mismo personal gerencial. Por otro lado, se podrían incluir más indicadores de desempeño que no fueron tomados en cuenta y que son de importancia para el negocio.

Con respecto a los modelos de minería se recomienda darle más énfasis a los productos o servicios que dan los comercios para generar más información que sea valiosa para el negocio, es decir, si se quiere realizar una minería más enriquecedora sería de utilidad contar con detalles más precisos de las compras de los clientes.

Igualmente, se recomienda aplicar los modelos de minería de datos para descubrir patrones ocultos en otros procesos, por ejemplo, la incidencia de "chargebacks". Es recomendado además, aplicar el proceso de minería desde otra perspectiva, u otra área de negocio.

Al haber sido analizada sólo una iteración de la metodología CRISP, se recomienda continuar con la siguiente iteración, refinando el proceso, depurando mejor los datos y optimizando los modelos de minería generados.

Finalmente, se recomienda continuar utilizando las metodologías elegidas para el crecimiento del proyecto. Esta investigación ha demostrado ser efectiva para proyectos de este tipo, ya que, al ser orientados a procesos de negocio se realiza en un corto plazo obteniendo pronto los resultados que pueden ser de vital importancia para el negocio.

REFERENCIAS BIBLIOGRÁFICAS

Agrawal, R. (1993). *Fast Algorithms for Mining Association Rules.*

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22, 2 (June 1993). DOI=10.1145/170036.170072 http://doi.acm.org/10.1145/170036.170072* , 207-216.

Berzal, F., Ignacio, B., Daniel, S., & Maria, V. (2002). Measuring the accuracy and interest of. *Department of Computer Science and Artificial Intelligence, University of Granada, E.T.S.I.I., Avda*, 221-235.

Bradley, T. (2007). *PCI Compliance: Implementing Effective PCI Data Security Standard Standards.* Burlington MA: Syngress.

Bramer, M. (2007). *Principles of Data Mining.*

Bryman, A. &. (1994). *Analyzing Qualitative Data.* New York: Routledge.

Cabrera. (1987). *La investigación evaluativa en Educación. Técnicas de Evaluación y Seguimiento de Programas de Formación Profesional.* Barcelona: Largo Caballero.

Chan, H., Lee, R., Dillon, T., & Chang, E. (2001). *E-Commerce, Fundamentals And Applications.* Ontario, Canada: Wiley.

Chapman, P., Clinton, J., & Kerber, R. (2000). CRISP-DM 1.0. *Step-by-step data mining guide*, 1-70.

Fayyad, U. M. (1997). Knowledge Discovery in Databases: An Overview. *ILP*, 3-16.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Advances in knowledge discovery and data mining. *American Association for Artificial Intelligence*, 1-34.

Gary, S., & James, P. (2000). *Electronic Commerce.*

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos.* Madrid: Pearson - Prentice Hall.

Hernández Sampieri, R. (1997). *Metodología de la Investigación.* México: McGraw-Hill.

Inmon, B. (2005). *Building a Data Warehouse. .*

Khosrow-Pour, M., & Khosrowpour, M. (2006). *Encyclopedia of E-commerce, E-government and Mobile Commerce.* Hershey: Idea Group.

Kimball. (2002). *The Data Warehouse Toolkit.*

Kimball. (2004). *The Data Warehouse ETL Toolkit. .*

Kostojohn, P. &. (2011). *CRM Fundamentals.*

Kudybañ, S., & Hoptroff, R. (2001). *Data Mining and Business Intelligence: A Guide to Productivity.* Hershey: Idea Group Publishing.

Margaret, D. (2002). *Data Mining: Introductory And Advanced Topics.* Upper Saddle River, NJ, USA: Prentice Hall.

María N. Moreno García, L. A. (2001). Aplicación de Técnicas de Minería de Datos en la Construcción y Validación de Modelos Predictivos y Asociativos a Partir de Especificaciones de Requisitos De Software.

Meyer, C. (1998). *Building a Better Data Warehouse.* Upper Saddle River, NJ.

Montague, D. A. (2011). *Essentials of Online payment Security and Fraud Prevention.* New Jersey: Wiley.

Nakajima, M. (2011). *Payment System Technologies and Functions: Innovations and Developments.* Japan: Reitaku University.

Piergiorgio, C. (2007). *Metodologia Y Tecnicas De Investigacion Social.* Madrid: McGraw-Hill.

Rahman El Sheikh, A. A. (2011). *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications.* IGI Global 2012.

Sandín, E. M. (2003). *Investigación cualitativa en investigación: Fundamentos y Tradiciones.* Madrid: McGraw-Hill.

Turban, E., Rainer, R. K., & E. Potter, R. (2005). *Turban Introduction to Information Technology.* USA: Wiley.

Uthurusam, U. F. (1996). Data mining and knowledge discovery in databases. *Commun. ACM 39, 11 (November 1996), 24-26. DOI=10.1145/240455.240463*, 24-26.

APÉNDICES

APÉNDICE A

Las tablas a continuación, muestran el detalle de la descripción, justificación de uso y atributos de cada una de las dimensiones.

Nombre de la dimensión	Dim_Conexion	
Descripción	Guarda los atributos de la conexión realizada previa a la transacción. En esta se almacenan atributos como el navegador utilizado, la IP del cliente, etc. Debe ser posible analizar los datos por navegador y también, tener las IP del cliente y del comercio.	
Justificación de uso:	EPPL requiere tener visibilidad con respecto a los atributos propios de las conexiones, de elementos como las direcciones ip de donde se realizaron las conexiones, y los navegadores utilizados, con el fin de prestar atención a tendencias de uso de navegadores y de regiones desde donde se estén realizando conexiones a los sistemas transaccionales.	
Atributos	ID_Conexion	Es el ID auto numérico de la conexión.
	DES_Navegador	Es el navegador utilizado en la transacción.
	DES_IP_Origen	Es la IP del cliente que está realizando la transacción.
	DES_IP_Origen	Es la IP del comercio o tienda electrónica que está procesando la transacción.

TABLA A1: DIMENSIÓN DE CONEXIÓN DE LA TRANSACCIÓN.

Nombre de la dimensión	Dim_Comercio	
Descripción	Guarda los atributos de los comercios que realizan transacciones. Es una dimensión, muy importante, para EPPL, ya que, permitirá filtrar los hechos por comercio, permitiendo analizar las ventas de cada uno de ellos y facilitarles información que sea útil. En esta dimensión se almacenan los atributos de nombre del comercio y el código. Debe ser posible analizar los datos por comercio.	
Justificación de uso	EPPL requiere tener visibilidad sobre los comercios o tiendas que procesan mediante el sistema.	
Atributos	ID_Comercio	Es el ID auto numérico de la conexión.
	Nombre_Comercio	Es el nombre del comercio.
	Codigo_Comercio	Es el código del comercio, que resume el nombre.

TABLA A2: DIMENSIÓN DE COMERCIO

Nombre de la dimensión	Dim_Tiempo	
Descripción	Guarda los atributos del momento en el que se efectúa una transacción.	
Justificación de uso	La dimensión de tiempo es fundamental para que EPPL ubique los hechos y tenga visibilidad a través del tiempo.	
Atributos	Time_key	Llave de la tabla de tiempo que equivale a una fecha en el que se realiza la transacción.
	Year_Name	Año en que se realizó la transacción.
	Half_Year_Name	Semestre en que se efectuó la transacción.
	Quarter_Name	Cuatrimestre en que se cumplió la transacción.
	Trimester_Name	Trimestre en que se ejecutó la transacción.
	Month_Name	Mes en que se formalizó la transacción.
	Week_Name	Semana en que se plasmó la transacción.
	Day_Of_Month	Día en que se realizó la transacción.

TABLA A3: DIMENSIÓN DE TIEMPO.

Nombre de la dimensión	Dim_Estado_Transaccion	
Descripción	Guarda los atributos del estado de transacción, este indica si la transacción se realiza correctamente o no. Indicando el estado final de la transacción. Se debe de filtrar a través de estado de la transacción para ver el estado aprobado, fallido o declinado. Es importante, para analizar el índice de problemas en el procesamiento de las transacciones, permitiendo definir indicadores que evalúen los niveles de regularidad y normalidad en el flujo de transacciones.	
Justificación de uso	Esa dimensión presenta los posibles estados en los que una transacción desea ser vista por EPPL.	
Atributos	Cod_Estado_Transaccion	Es el ID auto numérico del estado de la transacción.
	Des_Estado_Transaccion	Es la descripción del estado de la transacción.

TABLA A4: DIMENSIÓN ESTADO TRANSACCIÓN.

Nombre de la dimensión	Dim_Tarjeta	
Descripción	Guarda los atributos de la tarjeta de crédito utilizada en la transacción. Esta información es utilizada para dar seguimiento a una transacción de algún cliente en particular para verificar la tarjeta utilizada.	
Justificación de uso	EPPL desea realizar clasificaciones por el estado de las transacciones.	
Atributos	Con_Tarjeta	Es el ID auto numérico de la conexión
	Tipo_Tarjeta	Es el tipo de tarjeta de crédito. Puede ser, Visa o MasterCard.
	Num_Tarjeta	Es el número cifrado de la tarjeta.

TABLA A5: DIMENSIÓN TARJETA.

Nombre de la dimensión	Dim_Cliente	
Descripción	Guarda los atributos que caracterizan a la entidad del cliente en el negocio.	
Justificación de uso	EPPL desea observar los datos basados en el comportamiento de los clientes y sus atributos.	
Atributos	Con_Cliente	Llave principal del cliente
	Nom_Completo	Nombre completo del cliente
	Tipo_Cliente	Tipo de cliente en el sistema
	Edad	Edad del cliente
	Cod_Cliente	Código del comercio con el que se identifica al cliente
	Cod_Merchant	Código del comercio al que pertenece el cliente
	Nom_Merchant	Nombre del comercio al que pertenece el cliente
	Nom_Pais	Nombre del país del cliente
	Nom_Estado	Nombre del estado del país del cliente.
	Nom_Ciudad	Nombre de la ciudad del cliente.

	Zip_Code	Código postal del cliente
	BlackList	Si el cliente, es lista negra o no

TABLA A6: DIMENSIÓN CLIENTE.

Dimensión	Dim_Canal_Pago	
Descripción	Guarda los atributos del canal de pago utilizado para el procesamiento de una transacción	
Justificación de uso	EPPL desea observar los datos basados en los canales de pagos utilizados durante el procesamiento.	
Atributos	Con_Mid	Pk del canal de pago
	Mid	Nombre del canal de pago
	Banco	Banco del canal de pago
	Ipsp	Forma de agrupación de los mids para fines contables.
	Región	Formas de agrupación de los mids para fines contables.

TABLA A7: DIMENSIÓN DE CANAL DE PAGO.

APÉNDICE B

La tabla a continuación, muestra el nombre, tipo de dato, descripción, problemas de calidad y el indicador de dicretizable, de los datos utilizados en la metodología CRISP para realizar la minería de datos.

Nombre del dato	Tipo de dato	Descripción	Problemas de calidad	Conjunto encontrado	Se puede discretizar
Identificador **Cliente**	Entero	Es el consecutivo de asignado a cada cliente, no parece significativo para minería	No, todos son enteros consecutivos, no hay nulos.	Ninguno, no parece ser representativo en el comportamiento de clientes	No
Canal de pago	String alfanumérico	Es el canal de pago configurado en el banco, una especie de cuenta bancaria donde el dinero es almacenado.	Los nombres son alfanuméricos. No hay valores vacíos.	Ninguno, no parece ser representativo en el comportamiento de clientes	Si
Compañía	String alfanumérico	Compañia lógica registrada en el banco	No hay valores vacíos.	Ninguno, no parece ser representativo en el comportamiento de clientes	Si
Banco	String alfanumérico	Banco con el que procesa la transacción	No hay valores vacíos.	Ninguno, no parece ser representativo en el comportamiento de clientes	Si
Nombre de cliente	String alfanumérico	Nombre del cliente que procesa	No hay valores vacíos.	Ninguno, no parece ser representativo en el comportamiento de clientes	No
Tipo de cliente	String alfanumérico	Tipo de cliente en el que es categorizado	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente, pues existe una categorización	Si
Edad	Entero	Rango entre 1 y 100 años	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente, pues existe una categorización	Si

Tienda	String alfanumérico	Tienda en la que se compra.	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente, pues ,existe una categorización	Si
País	String alfanumérico	País de origen del cliente que realiza la transacción	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente, pues, existe una categorización	Si
Estado país	String alfanumérico	División geográfica de un país, aplica solo para Estados Unidos.	Hay valores vacios o sucios, para países que no son Estados Unidos.	Si es representativo con respecto a al comportamiento del cliente.	Si
Zipcode	String alfanumérico	Zipcode aplica solo para Estados Unidos	Hay valores vacios o sucios, para países que no son Estados Unidos	Por la cantidad de valores sucios, aunque tiene relación con la ubicación del cliente no va a ser tomado en cuenta	Si
Navegador utilizado	String alfanumérico	Navegador que el cliente utilizó	Hay valores vacios o sucios, es un texto difícil de tratar.		No
Monto de la transacción	Numérico con dos decimales	Monto de transacción por la cual fue procesada.	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente.	Si
Devolución de primer grado	Entero	Número entero que identifica una devolución de primer grado.	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente.	Si
Devolución de segundo grado	Entero	Número entero que identifica una devolución de segundo grado.	No hay valores vacíos.	Si es representativo con respecto a al comportamiento del cliente.	Si

TABLA B1: DESCRIPCIÓN DE DATOS.

APÉNDICE C

El código presentado a continuación, representa la creación de la vista utilizada para la obtención de los datos en la metodología CRISP para la minería de datos.

```
CREATE VIEW [dbo].[vDataMining4](

CON_CLIENTE,RANGO_EDADES,NOM_PAIS,COD_MERCHANT,TOTAL_TRANSACCIONES,TOTAL_APROBADAS
MONTO_TOTAL_APROBADAS,PROMEDIO_MONTO_APROBADAS,TOTAL_DENEGADAS,MONTO_TOTAL_DENE
GADAS,PROMEDIO_MONTO_DENEGADO,TOTAL_FALLIDAS,MONTO_TOTAL_FALLIDO,PROMEDIO_MONTO_FA
LLIDO,ES_CREDITO,ES_CHARGEBACK,RANGO_MONTOS_APROBADOS,RANGO_MONTOS_DENEGADOS,RAN
GO_MONTO_FALLIDOS ) as

select CON_CLIENTE, RangoEdades
NOM_PAIS,COD_MERCHANT,total,aprobadas,montoAprobadas,promedioMontoAprobado,denegadas,montoDenegad
as,promedioDenegado,fallidas,montoFallidas,promedioMontoFallido,

'EsCredito' = Case when creditos = 0 then 'No else 'Yes' end,

'EsChargeback' = Case when chargeback = 0 then 'No' else 'yes' end,

'RangoMontosAprobados'=CASE

WHEN tabla.promedioMontoAprobado between 0  and 50 THEN '0-50'

WHEN tabla.promedioMontoAprobado between 51  and 100 THEN '51-100'

WHEN tabla.promedioMontoAprobado between 101  and 150 THEN '101-150'

WHEN tabla.promedioMontoAprobado between 151  and 200 THEN '151-200'

WHEN tabla.promedioMontoAprobado between 201  and 300 THEN '201-300'

WHEN tabla.promedioMontoAprobado between 301  and 400 THEN '301-400'

WHEN tabla.promedioMontoAprobado between 401  and 500 THEN '401-500'

ELSE '> 501'END,

'RangoMontosDenegados'=CASE

WHEN tabla.promedioDenegado between 0  and 50 THEN '0-50'

WHEN tabla.promedioDenegado between 51  and 100 THEN '51-100'

WHEN tabla.promedioDenegado between 101  and 150 THEN '101-150'

WHEN tabla.promedioDenegado between 151  and 200 THEN '151-200'

WHEN tabla.promedioDenegado between 201  and 300 THEN '201-300'

WHEN tabla.promedioDenegado between 301  and 400 THEN '301-400'

WHEN tabla.promedioDenegado between 401  and 500 THEN '401-500'

ELSE '> 501'END,

'RangoMontosFallidos'=CASE

WHEN tabla.promedioMontoFallido between 0  and 50 THEN '0-50'

WHEN tabla.promedioMontoFallido between 51  and 100 THEN '51-100'

WHEN tabla.promedioMontoFallido between 101  and 150 THEN '101-150'

WHEN tabla.promedioMontoFallido between 151  and 200 THEN '151-200'

WHEN tabla.promedioMontoFallido between 201  and 300 THEN '201-300'

WHEN tabla.promedioMontoFallido between 301  and 400 THEN '301-400'

WHEN tabla.promedioMontoFallido between 401  and 500 THEN '401-500'
```

```
ELSE '> 501'END
from
(
 select  top 10000
 c.CON_CLIENTE,
 'RangoEdades'=CASE
WHEN c.EDAD < 18 THEN 'Adolencente'
WHEN c.EDAD BETWEEN 18 AND 24 THEN 'Joven Temprano'
WHEN c.EDAD BETWEEN 24 AND 29 THEN 'Joven Tardio'
WHEN c.EDAD BETWEEN 30 AND 37 THEN 'Adulto'
WHEN c.EDAD BETWEEN 38 AND 49 THEN 'Adulto Medio'
WHEN c.EDAD > 50 THEN 'Adulto Tardio'
ELSE 'Unknown'END,
 c.NOM_PAIS,
 c.COD_MERCHANT,
   count(*) total,
   sum(case when t.COD_ESTADO_TRANSACCION = 3 then 1 else 0 end) aprobadas,
            sum(case when t.COD_ESTADO_TRANSACCION = 3 then t.MON_DOLARES else 0 end) montoAprobadas,

            sum(case when t.COD_ESTADO_TRANSACCION = 3 then t.MON_DOLARES else 0 end) /
            (case when isnull(sum(case when t.COD_ESTADO_TRANSACCION = 3 then 1 else 0 end), 0) = 0 then 1
else
   (sum(case when t.COD_ESTADO_TRANSACCION = 3 then 1 else 0 end)) end) promedioMontoAprobado,

            sum(case when t.COD_ESTADO_TRANSACCION = 4 then 1 else 0 end) denegadas,
            sum(case when t.COD_ESTADO_TRANSACCION = 4 then t.MON_DOLARES else 0 end) montoDenegadas,

            sum(case when t.COD_ESTADO_TRANSACCION = 4 then t.MON_DOLARES else 0 end) /
            (case when isnull(sum(case when t.COD_ESTADO_TRANSACCION = 4 then 1 else 0 end), 0) = 0 then 1
else
   (sum(case when t.COD_ESTADO_TRANSACCION = 4 then 1 else 0 end)) end) promedioDenegado,
            sum(case when t.COD_ESTADO_TRANSACCION = 10 then 1 else 0 end) fallidas,
            sum(case when t.COD_ESTADO_TRANSACCION = 10 then t.MON_DOLARES else 0 end) montoFallidas,
            sum(case when t.COD_ESTADO_TRANSACCION = 10 then t.MON_DOLARES else 0 end) /
            (case when isnull(sum(case when t.COD_ESTADO_TRANSACCION = 10 then 1 else 0 end), 0) = 0 then 1
else
   (sum(case when t.COD_ESTADO_TRANSACCION = 10 then 1 else 0 end)) end) promedioMontoFallido,
            sum(case when isnull(credit.CON_CREDITO,0) = 0  then 0 else 1 end) creditos,
            sum(case when isnull(ch.CON_CHARGEBACK,0) = 0  then  0 else 1 end) chargeback
from FAC_TRANSACCCION t inner join dim_cliente c
 on t.con_cliente = c.con_cliente
left join dbo.FAC_CREDITO credit on t.con_transaccion = credit.CON_TRANSACCION
 left join dbo.FAC_CHARGEBACK ch on t.CON_TRANSACCION = ch.CON_TRANSACCION
 group by  c.CON_CLIENTE,c.EDAD, c.NOM_PAIS, c.COD_MERCHANT
 ) as tabla
```