**Revista Global de Negocios**
Vol. 8, No. 1, 2020, pp. 1-21
ISSN: 2328-4641 (print)
ISSN: 2328-4668 (online)

IBFR
www.theIBFR.com

# CLASIFICACIÓN, CONGLOMERACIÓN Y PROBABILIDAD CONDICIONAL PARA DEFINIR EL PERFIL DEL CONSUMIDOR DE TARJETA DE CRÉDITO COLOMBIANO: UNA METODOLOGÍA DE ANÁLISIS SUBYACENTE EN LA MINERÍA DE DATOS

Santiago García Carvajal, Universidad Militar Nueva Granada

**RESUMEN**

*El articulo propone, una metodología de investigación sin perjuicio de normas especiales que disponen la confidencialidad o reserva de información registrada en los bancos, de naturaleza pública, para definir el perfil del consumidor de tarjetas de crédito en Colombia a partir de La ley 1266 de 2008, conocida como "LEY DE HABEAS DATA", particularmente en relación con la información financiera, crediticia, comercial, de servicios y la proveniente de terceros países. La metodología identifica perfiles genéricos de consumidores de tarjeta de crédito a partir del análisis de conglomerados según la cercanía entre el comportamiento de las variables. Los perfiles genéricos se identifican por medio del teorema bayesiano de probabilidades condicionales, donde a partir de la intersección entre variables de clasificación y conglomeración, subyace el comportamiento de un tarjetahabiente.*

**PALABRAS CLAVE:** Minería de Datos, Investigación de Mercados, Metodología de Investigación

# CLASSIFICATION, CONGLOMERATION AND CONDITIONAL PROBABILITY TO DETERMINE CREDIT CARD CONSUMER PROFILES IN COLOMBIA: METHODOLOGY USING DATA MINING

**ABSTRACT**

*The article proposes a research methodology. The goals is to, without infringing on laws or rules governing bank -customer data confidentiality, to create consumer profile for credit cards in Colombia under Law 1266 of 2008, known as "HABEAS DATA LAW". We wish to specifically examine financial information, credit level, commercial, service and third-country services. The methodology identifies generic profiles of credit card consumers based on clusters analysis related to the proximity between the behavior of the variables. Generic profiles are identified by means of the Bayesian theorem of conditional probabilities, where the intersection between classification and conglomeration variables determines potential cardholders.*

**JEL:** M3

**KEYWORDS:** Data Mining in Marketing, Marketing Research, Research Methods

## INTRODUCCIÓN

En estos tiempos se recopilan grandes volúmenes de información tanto escrita como audiovisual en las redes sociales, internet y plataformas públicas de datos abiertos. El concepto "Minería de Datos" tiene una connotación de intriga y curiosidad entre los profesionales de hoy, ante la facultad que ésta tiene de identificar patrones subyacentes de conducta en los consumidores. Qué comportamientos de consulta presenta un internauta, de qué forma se comporta una ama de casa atreves de los pasillos de un supermercado, qué tipos de video disfruta un individuo según su historial de vistos recientemente, o hasta qué cupo se le puede otorgar a su próxima tarjeta de crédito según su historial crediticio.

La metodología CRISP-DM se describe en términos de un modelo de proceso jerárquico, que consta de conjuntos de tareas descritas en cuatro niveles de abstracción (de general a específico): fase, tarea genérica, tarea especializada e instancia de proceso. La minería de datos se ubica dentro del núcleo de todo un proceso llamado, Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases – KDD). Las etapas del proceso de descubrimiento son: (1) el entendimiento del negocio, en esta etapa se interpreta el contexto en el cual se recopilan los datos. (2) El entendimiento de los datos, proceso en el cual se establecen relaciones entre los datos en términos de variables relacionadas con el contexto del negocio y se identifican patrones o segmentos de mercado. (3) La presentación del modelo. En ésta etapa se elige el paradigma con el cual se hará el análisis subyacente de los datos, para el caso de estudio del presente artículo, se construyó una matriz de perfiles de consumidores a partir de los datos publicados por la superintendencia financiera sobre tarjeta de crédito y el análisis de conglomerados de las variables, con el fin de identificar las probabilidades condicionales para cada perfil de consumidores. La siguiente etapa es: (4) la evaluación del modelo, o etapa de meta-aprendizaje, es decir, el modelo podría ser interpretado de varias formas, y es el criterio del investigador quien se encarga de tomar la decisión de trabajar con los datos de una u otra forma y decidir qué hacer con los resultados, hasta saber qué resultados son más coherentes con las propiedades de los datos. La última etapa es (5) la presentación de resultados, en este punto se evidencia el impacto o la consecuencia del hallazgo subyacente en la minería de datos, a la luz de la información existente. Los resultados se relacionan con cursos de acción y estrategias de mercadeo en la industria de tarjetas de crédito; por otra parte; se evidencia el beneficio de utilizar todo el proceso de descubrimiento a pesar de las limitaciones que tiene la industria en razón a las leyes de protección de datos.

La naturaleza del producto, se refleja en la configuración del mercado. Es decir, se identifican tres patrones esenciales que segmentan y proveen la selección natural de la clientela al interior de cada segmento. Estos patrones son: La facilidad que tiene el producto de generar apertura de crédito a nivel internacional, la ambivalencia de la característica de la tarjeta como línea de crédito y como medio de pago, y el poder adquisitivo del mercado por el total de montos realizados y el saldo en tarjeta de crédito. El mercadeo directo es uno de los métodos más efectivos orientados a maximizar el ciclo de vida de un consumidor. Zicari, A (2008) Se han propuesto muchos métodos de aprendizaje sensibles al costo, que identifican consumidores valiosos con la expectativa de maximizar ganancias; sin embargo, es frecuente observar que estos métodos, a pesar de estar orientados a la maximización de ganancias; no identifican la probabilidad de deserción en el ciclo de vida del consumidor.

Desafortunadamente, las campañas de mercadeo que aparentemente son muy exitosas maximizando ganancias, fracasan en la minimización de probabilidades de deserción debido a un conflicto de intereses entre estos dos objetivos. W. Fan, S.J. Stolfo, J. Zhang, P.K. Chan (1999) El artículo está organizado de la siguiente manera: La revisión literaria, presenta un problema de mercado derivado de un acto legislativo en Colombia, como lo es la ley de habeas data. Se abordaron estudios realizados en otros países, donde por medio de minería de datos, se ha logrado caracterizar segmentos de mercado, que cumplan con las condiciones de libertad, autoridad, mensurabilidad y sustancia para que reaccionen y sean sujetos de acciones subsecuentes y proactivas de mercadeo, paralelamente, se articula otro grupo estudios orientados a incorporar el uso de la inteligencia artificial en la toma de decisiones de mercadeo con bases de datos;

por otra parte se identifica una miopía de mercadeo citando autores que inviten a desarrollar el juicio del observador con miras a la construcción de un modelo para la toma de decisiones, adaptado al entorno Colombiano, en virtud del análisis subyacente de los datos, y presenta la importancia de algunas técnicas estadísticas preliminares que sirvan para entender los datos y diseñar adecuadas estrategias de análisis. En la sección de metodología se desarrolla la adaptación del modelo CRISP-DM a la industria de tarjeta de crédito en Colombia, a partir de la construcción manual de una matriz de datos que cubre los periodos de junio de 2014 a junio de 2019. La información es pública y fue tomada del sitio web de la superintendencia financiera de Colombia. En la sección de presentación de resultados, se propone una interpretación del ciclo de vida de la industria, capaz de observar el comportamiento subyacente de los segmentos de mercado y las consecuencias que las estrategias de los bancos han generado en la segmentación natural de la clientela en función del cupo de crédito otorgado por la entidad. Las características observadas en los datos hicieron evidente la necesidad de segmentar el mercado para identificar tarjetahabientes homogéneos entre sí; pero heterogéneos entre grupos de ellos, razón por la cual la matriz de datos construida era anormal, es decir, se observó que una porción de las cifras, registró valores totalmente por fuera de los parámetros de normalidad de la base de datos, al identificar esos segmentos de mercado por medio del análisis de clúster. Se realizaron cálculos de probabilidades posteriores y condicionales que permitieran observar el efecto no solo del consumidor financiero; sino de la estrategia bancaria para acomodarse al comportamiento del consumidor, las cifras fueron luego ordenadas, en la secuencia lógica en la que ocurren dentro del ciclo de vida de la categoría del producto y graficadas de manera longitudinal. Este proceso en resumen, permitió vislumbrar el comportamiento de la industria que evoluciona, tanto por los efectos del comportamiento del consumidor; como por las acciones de los bancos que buscan segmentar el mercado con múltiples ofertas. En la sección de conclusiones se hace una síntesis de la metodología y los resultados y se resalta en general como el presente trabajo le aporta a la literatura existente sobre tarjetas de crédito en Colombia, porque esta invita a los analistas de marca de la industria, a emplear herramientas concretas de minería de datos e inteligencia artificial, aplicadas a la toma de decisiones de mercados.

## REVISIÓN LITERARIA

Según el artículo 15 de la Constitución Política de Colombia, modificado por el Acto Legislativo 02 de 2003, declarado inexequible por la Corte Constitucional mediante Sentencia C-816 de 2004, por el vicio de procedimiento ocurrido en el sexto debate de la segunda vuelta: "*Todas las personas tienen derecho a su intimidad personal y familiar y a su buen nombre, y el Estado debe respetarlos y hacerlos respetar. De igual modo, tienen derecho a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en los bancos de datos y en archivos de entidades públicas y privadas... para efectos tributarios judiciales y para los casos de inspección, vigilancia e intervención del Estado, podrá exigirse la presentación de libros de contabilidad y demás documentos privados, en los términos que señale la ley." expedientes D-5121 y D-5122.* En ese ese sentido la ley estatutaria 1266 de 2008, confiere, *las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial, de servicios y la proveniente de terceros países y se dictan otras disposiciones.* Bautista A, (2015).

Los datos personales se refieren a toda aquella información que permita la identificación singular de un individuo. Su documento de identidad, información demográfica y más aún, información sensible sobre su estado de salud, comportamientos sociales, estilos de vida o ideología política. Las entidades financieras, exigen a sus clientes diligenciar formularios que les permitan endilgar obligaciones financieras para asegurar sus rendimientos, acaparar un mercado con futuras ofertas, tasas de intereses, y así diversificar su portafolio de banca de consumo. García S & Alemán F (2011) Igualmente, existen políticas de protección de datos según la naturaleza de su contenido: el dato público es de libre acceso, abierto al debate y especulación, necesario para la creación de políticas públicas, es el que la constitución política determina como tal; por otra parte, está el dato semiprivado, cuyo contenido es de importancia para un determinado grupo de interés, más no a un solo individuo, es por ende no reservado. También existe el dato privado.

Este debe ser concerniente solamente para el titular de la información por su contenido íntimo para la toma de decisiones a nivel individual; en contraste, está el dato sensible. Este dato contiene información que, en manos ajenas, puede causar daños al titular, sin este ser consciente de su uso. Riquelme Santos, J.C., Ruíz, R. & Gilbert, K. (2006). Los procesos de investigación sindicada, que se realizan con información secundaria, o datos abiertos, tienen la ambivalente responsabilidad de construir información valiosa para la toma de decisiones a partir de datos cuya naturaleza sensible, al no poder ser divulgada, presenta un obstáculo en la construcción de modelos científicos. Blazquez Ochando, M. (2010) En el caso de las tarjetas de crédito, los bancos en razón a las políticas de protección de datos, no divulgan información sobre sus tarjetahabientes; sin embargo, están en la obligación de reportar el volumen y la naturaleza de las transacciones a la superintendencia financiera. Roagna, I. (2012).

La minería de datos se conoce como el proceso de crear mayor interacción entre el humano y las bases de datos a partir de su contexto y objetivos y hacer menor énfasis en la automatización en general. Provost F y Fawcet T (2013) La minería de tiene como objetivo descubrir patrones en grandes volúmenes de datos, haciendo una analogía, es como como la mano oculta que sujeta un magneto bajo una hoja de papel. (Chopra, 1996, p. 176). A simple vista, por encima de la hoja, parece tener vida propia aquella limadura de hierro que se mueve a lo largo y ancho del papel y se alinea según los campos magnéticos, pero es la intención de aquel quien mueve la mano lo que interesa. La minería de datos puede interpretarse como un sistema informático para la toma de decisiones, sobre la base del descubrimiento y la trasferencia de los hallazgos a gran escala, donde como herramienta se utilicen los métodos de aprendizaje automático supervisado, no supervisado, pruebas paramétricas de interdependencia, cálculo de probabilidades, medidas de asociación y prueba de hipótesis. (Berndt, D., y Clifford, J. 1996).

El análisis de conglomerados es un grupo de técnicas de reducción de datos diseñados para generar un contraste entre observaciones similares en un set de datos, de modo tal que unas observaciones se agrupen dentro de un mismo espacio común; mientras que otro grupo observaciones agrupadas de manera similar, pero en otro espacio común, sean tan diferentes como sea posible. Si se compara esta técnica con otras técnicas de reducción de datos tales como el análisis factorial se diferencia en que ésta última agrupa en cargas factoriales las correlaciones entre las variables de una base de datos en columnas, Comrey, A. L. (1973) generando así un análisis estadístico muy profundo; en contraste, el análisis de conglomerados agrupa las observaciones según su proximidad y distancia, de tal forma que abre la puerta a un análisis mucho más subjetivo. Aldenderfer MS & Blashfield RK (1984)

K-medias es un método de análisis de conglomerados que agrupa las observaciones por medio de la minimización de las distancias euclidianas entre ellas. Las distancias euclidianas se interpretan de manera análoga a la hipotenusa de un triángulo. Las diferencias entre observaciones de dos variables (x, y) son reemplazadas en una ecuación Pythagorea para hallar la distancia más corta entre dos puntos (la distancia de la hipotenusa). Lorr M (1983). Las distancias euclidianas pueden extenderse en n- dimensiones y las distancias hacen referencia a diferencias numéricas sobre una variable continua, no solamente distancias espaciales o geométricas. Esta definición de distancia Euclidiana, requiere que todas las variables de insumo sean continuas Everitt BS, Landau S, Leese M, Stahl D (2011).

El teorema de bayes relaciona la probabilidad de ocurrencia de un evento con la ocurrencia o no ocurrencia de un evento asociado. Por ejemplo, la probabilidad de sacar un az de poker es 0.077 (4 ÷ 52). Si dos cartas son sacadas al azar, la probabilidad que tiene la secunda carta escogida de ser un az, depende de si la primera fue un az o no: de ser así, entonces la probabilidad de que la segunda carta escogida sea un az es 0.058 (3 ÷ 52); de lo contrario, la probabilidad continúa siendo 0.077. El teorema de Bayes provee una estimación o predicción a la luz de la experiencia u observación. Este difiere de otros métodos de testeo de hipótesis en cuanto a que asigna “después del hecho” (posterior), probabilidades a las hipótesis en lugar de solamente aceptarlas o rechazarlas. Stone, J. V (2013). La tarjeta de crédito es uno de los enfoques de pago electrónico más populares en el comercio electrónico en línea actual. Para consolidar clientes valiosos, los emisores de

tarjetas invierten mucho dinero para mantener una buena relación con sus clientes. A pesar de haber realizado varios esfuerzos para estudiar la motivación del uso de la tarjeta, pocas investigaciones enfatizan el análisis del comportamiento del uso de la tarjeta de crédito cuando los períodos de tiempo cambian de t a t + 1.

Tsai C. (2007) desarrollo un significativo caso de estudio utilizando una base de datos proporcionada por un importante emisor de tarjetas de crédito en Taiwán. En la base de datos de dicho estudio, había 314.339 usuarios de tarjetas activas que generaron 2.153.062 transacciones en el año 2001 (período de tiempo t) y 2.561.202 transacciones en el año 2002 (período de tiempo t + 1). Los gerentes de marketing querían concentrarse en el comportamiento del cliente de sus clientes VIP. Los criterios de selección VIP se basaban en las regulaciones corporativas, las políticas de evaluación de crédito y las evaluaciones del valor de vida del cliente. Los criterios típicos establecidos fueron: "No se realiza ningún pago retrasado en los últimos nueve meses" y "el monto límite más bajo se paga en los últimos dos meses". Una serie de programas COBOL (lenguaje orientado al negocio común) y JCL (lenguaje de control del trabajo) estaban codificados para recuperar perfiles de clientes y datos de comportamiento del cliente de los archivos VSAM (método de acceso de almacenamiento virtual) en el sistema operativo OS / 390 de una computadora con marco principal IBM 9121. Como resultado, se identificaron 9.086 clientes VIP quienes realizaron 354,063 transacciones en el año 2001 y 440,010 transacciones en el año 2002.

Continuando con el caso de estudio en Taiwán, el magnífico aumento en los mercados de tarjetas de crédito para el comercio electrónico, llevó a los emisores de tarjetas a hacer más esfuerzos para comprender su comportamiento de uso. En realidad, el comportamiento del cliente generalmente cambió con el tiempo. Algunos patrones frecuentes en un período de tiempo pueden no ser válidos para otro período de tiempo. Para satisfacer esta necesidad, la investigación de Tsai propuso un enfoque de minería de datos integrado para el análisis del comportamiento de uso de tarjetas de crédito. El marco de análisis de comportamiento de uso de tarjeta de crédito propuesto, constaba de cuatro etapas principales. La primera etapa fue la extracción de datos y el preprocesamiento. En esta etapa, el perfil del cliente y sus datos de transacción en el período de tiempo t se recuperaron de las bases de datos. En la segunda etapa, se llevó a cabo la segmentación de clientes utilizando la red neuronal etiquetada autorganizada, por sus siglas en inglés, Self Organizing Maps (SOM) LabelSOM. LabelSOM agrupó de manera adaptativa a los clientes en grupos e identificó automáticamente las características demográficas críticas para cada grupo. En la tercera etapa, el comportamiento de uso de los clientes en el grupo de interés, se generó utilizando el algoritmo de árbol de decisión difuso (FDT) que representaba el comportamiento de uso como un conjunto de reglas IF-THEN, una vez se obtuvieron los patrones de uso del grupo de clientes de interés en el período de tiempo t, fue posible rastrear los cambios de comportamiento de estos clientes desde el período de tiempo t hasta t + 1. Al recuperar sus datos correspondientes en el período de tiempo t + 1. El modelo propuesto se implementó con éxito utilizando datos reales de tarjetas de crédito proporcionados por un banco comercial en Taiwán. El procedimiento de análisis proporcionado debería proporcionar a los emisores de tarjetas un enfoque sistemático para establecer estrategias de marketing para grupos de clientes de interés; sin embargo, en esa investigación todavía se identificaron algunos espacios para mejorar en el futuro, en relación a los algoritmos de ajuste automático de funciones de membresía que mejor se adapten al marco propuesto. Además, identificaron la oportunidad de investigar qué estrategias de marketing pueden afectar el comportamiento de diferentes grupos de clientes.

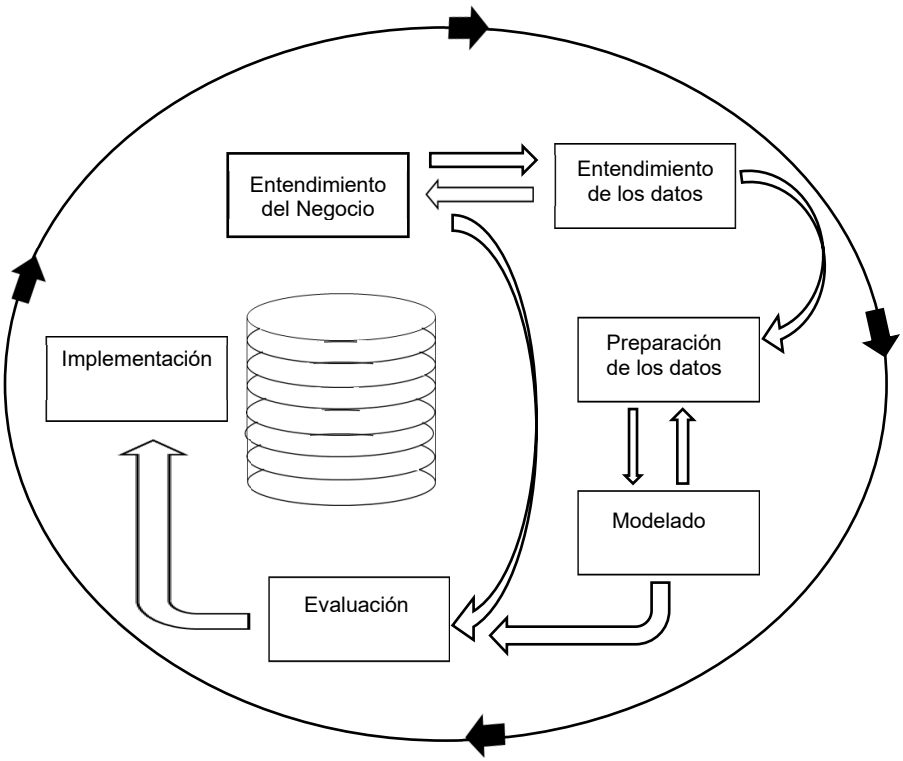
La estandarización es el paso central de preprocesamiento en la minería de datos, para estandarizar valores de características o atributos de diferentes rangos dinámicos en un rango específico. Bin Mohamad, I & Usman, D (2013) realizaron una comparación sobre el rendimiento de los tres métodos de estandarización en el algoritmo convencional de K-medias. Al comparar los resultados en los conjuntos de datos de enfermedades infecciosas, se descubrió que el resultado obtenido por el método de estandarización de puntaje z es más efectivo y eficiente que los métodos de estandarización de escala mínima-máxima y decimal. K-medias. Lilliefors, H. (1967) perfeccionó la prueba de Kolmogorov-Smirnov agregando un

mayor nivel de sensibilidad a la prueba de hipótesis que compara las distribuciones normales con las distribuciones anormales, a partir de la discrepancia entre función de distribución acumulativa y la función de distribución empírica, de esta forma se logra entender con mayor precisión la naturaleza de los datos y proceder con una mejor de preparación de estos en el proceso de minería de datos.

## METODOLOGÍA

El modelo de referencia CRISP-DM propone una revisión de las fases del ciclo de vida de un proyecto de minería de datos incluyendo sus respectivas tareas y relaciones mutuas; sin embargo, las relaciones y tareas implicadas pueden variar dependiendo del interés del usuario, sus objetivos, sus antecedentes, y lo más importante, los datos. El ciclo de vida de un proyecto de minería de datos consiste de un proceso de seis faces exhibidas en la Figura 1. No es una secuencia rígida, el proceso debe fluir de atrás para adelante y el resultado de cada fase determina qué tarea dentro de cada fase debería ser la siguiente en llevarse a cabo. Los vectores de la gráfica indican las instancias o dependencias más frecuentes sobre las cuales fluye el proceso. El circulo exterior circundante simboliza la naturaleza cíclica de la minería de datos en sí misma; de otra parte, el proceso como tal no culmina una vez se implementa una solución. Las lecciones aprendidas durante el proceso y los efectos de la implementación subsecuente sobre los datos, deben orientar el enfoque a futuras preguntas a realizar en el negocio. A continuación, se presenta una propuesta de minería de datos sobre la información pública de tarjetas de crédito en Colombia. Se organizó inicialmente una matriz de datos en Excel de 23 columnas por 1096 filas, con las cifras mes a mes entre junio de 2014 y junio de 2019, tomadas del sitio web de la superintendencia financiera de Colombia, posteriormente de exportó la matriz de datos al Sofware SPSS para el análisis desarrollar el análisis descriptivo e inferencial.

Figura 1: Modelo de Referencia Para las Fases del Modelo CRISP-DM




*CRISP-DM (del inglés Cross Industry Standard Process for Data Mining) se trata de un modelo estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos. La fortaleza que identifica el investigador en este proceso está en la profundidad con la que se puede hacer el análisis del insumo inicial, con el propósito de formular una estrategia de minado eficiente y efectivo, el resultado de este proceso definitivo puede aplicado sistemáticamente a matices de datos posteriores, por medio de la programación de datos*

Entendiendo el Negocio: La tecnología es el factor externo de mayor impacto en la industria del dinero plástico, en virtud de su capacidad de interacción entre sus constituyentes, es decir, las marcas de aceptación (Visa, Master Card, American Express y Diners Club), los bancos emisores y comercios afiliados. La tecnología facilita la penetración de mercados, disminuye costos de manejo y articulación entre economías de alcance y economías de escala. Para los tarjetahabientes, se hace cada vez más atractivo poseer una tarjeta de crédito gracias a la cada vez mayor cantidad de comercios afiliados que facilitan y disminuyen el costo de las transacciones. Las franquicias internacionales como actores principales del canal de distribución, celebran relaciones jurídicas complejas entre los constituyentes del negocio, con el objetivo de facilitar la conectividad y transferencias electrónicas eficientes. Por ejemplo: entre franquicia internacional y bancos emisores, se celebra un contrato de licencia, con el fin de autorizar a los bancos emitir plásticos, los bancos emisores a su vez celebran un contrato de prestación de servicios con los comercios afiliados, para facilitar las transacciones y las comisiones de adquirencia; por su parte, entre bancos y tarjetahabientes, se celebra un contrato de apertura de crédito, con el fin de otorgar cupos de crédito y bifurcar un cobro derivado de dos momentos de la vida del contrato: una comisión de apertura que será cobrada aun en el caso de no utilizar la tarjeta y otra comisión por la administración de la tarjeta aparte del interés corriente sobre su utilización.

En el momento de la transacción con tarjeta de crédito, se materializa un contrato de compraventa o de prestación de servicios, pero se extingue la obligación entre el consumidor y el comercio una vez el segundo provee bienes o servicios y el primero los obtiene o los compra y paga por ellos haciendo uso del plástico. La tarjeta de crédito permite al individuo la realización de pagos y el otorgamiento de crédito, a cambio de acordar una fecha y número de cuotas de pago, tasa de interés anual, y según sea el caso una cuota de manejo. Existe sin embargo, un elemento de crédito gratis otorgado a los tarjetahabientes, en razón a que los recibos abarcan las compras realizadas a partir del último recibo y existe un lapso de 25 días para cancelar el balance. Si una compra se realiza un día después de llegar el último recibo, entonces este plazo en total podría alcanzar hasta 55 días de crédito gratis. Ciertamente, está en el mejor interés de las compañías, el motivar a los tarjetahabientes, no solo a incrementar uso de sus tarjetas; sino de disponer de su facilidad de crédito, puesto que esa es la base fundamental de ingresos para los bancos.

Entendiendo los datos e identificación de patrones**:** Todas las variables continuas fueron exploradas previamente sobre la prueba de medias para determinar la naturaleza de la base de datos que se ingresaría al análisis. Las pruebas de Kolmogorov & Smirnov arrojaron **0.005 entre todas las variables, con lo cual se determinó que la base de datos es anormal y el tipo de análisis estadístico a realizar debería hacerse con pruebas no paramétricas. La prueba de bondad de ajuste de Kolmogorov-Smirnov (prueba K-S) compara la matriz de datos contra una distribución conocida de forma tal que permita saber si tienen la misma distribución. Aunque la prueba no es paramétrica, no asume ninguna distribución subyacente en particular, se usa comúnmente como una prueba de normalidad para determinar si la muestra se distribuye normalmente. También se usa para verificar la suposición de normalidad en el Análisis de varianza. Más específicamente, la prueba compara una distribución de probabilidad hipotética conocida, por ejemplo, contra la distribución generada por los datos: la función de distribución empírica.

$$F_n\,(x) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1\ si\ \ y_i\ \leq\ x_i \\ 0\ alternativa \end{cases} \tag{1}$$

Para dos colas el estadístico viene dado por:

$D_n^+ = \max(\,F_n\,(x) - F(x))$

$D_n^- = \max(\,F\,(x) - F_n(x))$

Donde *F(x)* es la distribución presentada como hipótesis

La prueba de Lilliefors es una versión corregida de la prueba K-S para la normalidad, ésta generalmente proporciona una aproximación más precisa de la distribución de la estadística de prueba. De hecho, muchos paquetes estadísticos (como SPSS) combinan las dos pruebas como una prueba K-S "corregida por Lilliefors". La prueba procede de la siguiente manera: A partir de los datos, se estima la media y la varianza de la población, se encuentra la discrepancia máxima al comparar tanto la función de distribución empírica como la función de distribución acumulativa. Luego se determina si la discrepancia es lo suficientemente amplia como para ser estadísticamente significativa. En tal caso, se rechaza la hipótesis nula. Las hipótesis para la prueba son:

*Hipótesis nula* ($H_o$): los datos provienen de la distribución especificada.
*Hipótesis alternativa* ($H_1$)*:* al menos un valor no coincide con la distribución especificada.

Es decir,

$H_0$: $p = p_0$; $H_0$: $p \neq p_0$.

Donde: $p$ es la distribución de su muestra y $p_0$ es una distribución especificada

Tabla 1: Prueba de Kolmogorov and Smirnov Sobre el Ajuste de Lilliefors

| | N | Media | Desviación Estándar | Diferencia Absoluta | Positivo | Negativo | Estadístico de Prueba | Sig. Asintótica (Bilateral) |
|---|---|---|---|---|---|---|---|---|
| Vigentes a la fecha de corte | 1096 | 624536.98 | 684929.049 | 0.193 | 0.193 | -0.181 | 0.193 | 0.000 |
| Vigentes durante el mes | 1096 | 12583.16 | 14480.613 | 0.193 | 0.143 | -0.193 | 0.193 | 0.000 |
| Canceladas | 1095 | 10825.5 | 13664.317 | 0.214 | 0.194 | -0.214 | 0.214 | 0.000 |
| Bloqueadas temporalmente | 1073 | 76537.35 | 85816.563 | 0.205 | 0.205 | -0.186 | 0.205 | 0.000 |
| Compras nacionales | 1096 | 809786.54 | 1029061.207 | 0.216 | 0.204 | -0.216 | 0.216 | 0.000 |
| Monto compras nacionales | 1096 | 1.70656E+11 | 2.1237E+11 | 0.211 | 0.195 | -0.211 | 0.211 | 0.000 |
| Avances nacionales | 1096 | 117611.94 | 186490.387 | 0.265 | 0.21 | -0.265 | 0.265 | 0.000 |
| Monto avances nacionales | 1096 | 58295890818 | 82295687893 | 0.24 | 0.211 | -0.24 | 0.24 | 0.000 |
| Compras en el exterior | 1097 | 236651.15 | 382896.917 | 0.268 | 0.192 | -0.268 | 0.268 | 0,000 |
| Monto compras en el exterior | 1096 | 35756448566 | 57060372423 | 0.265 | 0.223 | -0.265 | 0.265 | 0.000 |
| Avances en el exterior | 979 | 905.37 | 2107.198 | 0.334 | 0.33 | -0.334 | 0.334 | 0.000 |
| Monto avances en el exterior | 981 | 469999897.1 | 1106536518 | 0.336 | 0.326 | -0.336 | 0.336 | 0.000 |
| Saldo de tarjeta de crédito | 1100 | 1.22662E+12 | 1.42228E+12 | 0.196 | 0.196 | -0.194 | 0.196 | 0.000 |
| Cupo de crédito no utilizado | 1096 | 2.06783E+12 | 2.3917E+12 | 0.194 | 0.183 | -0.194 | 0.194 | 0.000 |
| Intereses por compra y avances corrientes | 1096 | 19984437022 | 23637240030 | 0.199 | 0.182 | -0.199 | 0.199 | 0.000 |
| Intereses por compra y avances de mora | 1096 | 1196420927 | 2373630916 | 0.307 | 0.249 | -0.307 | 0.307 | 0.000 |
| Castigos de cartera capital | 868 | 8363129816 | 9159066934 | 0.182 | 0.182 | -0.181 | 0.182 | 0.000 |
| Castigos de cartera diferente a capital | 880 | 749310287.1 | 1236216783 | 0.272 | 0.247 | -0.272 | 0.272 | 0.000 |
| Total montos | 1096 | 2.65224E+11 | 3.41064E+11 | 0.219 | 0.168 | -0.219 | 0.219 | 0.0000 |
| Montos nacionales | 1096 | 2.28983E+11 | 2.87082E+11 | 0.213 | 0.167 | -0.213 | 0.213 | 0.0000 |
| Montos internacionales | 1096 | 36273785684 | 58383682569 | 0.267 | 0.225 | -0.267 | 0.267 | 0,0000 |

*La prueba de Kolmogorov –Smirnov facilita el entendimiento de los datos, por que el estadístico de prueba se obtiene al realizar pruebas de hipótesis. Si la hipótesis nula se rechaza, la matriz de insumo, no está correlacionada con la población; es decir existen parámetros en el mercado que están por fuera de la normalidad y esto advierte la necesidad de segmentar los datos.*

En la Tabla 1 se observa, la media y la desviación estándar para cada una de las variables, según los parámetros de distribución normal; en contraste se compara contra las diferencias máximas extremas absolutas, positivas y negativas. Como resultado, el valor de *p* asintótico bilateral (para un test de dos colas) con la prueba de ajuste de Lilliefors ** 0.005 rechaza la hipótesis nula. Esto demuestra que los datos tienen observaciones que no provienen de la distribución especificada, con un nivel de significación del 5%. Se hace entonces pertinente, profundizar sobre el análisis de clústeres, hasta identificar cuál de estos presenta comportamientos diferenciados en la categoría de tarjetas de crédito. Preparación de los datos: Con el fin de identificar patrones en la información, era necesario vislumbrar el comportamiento del individuo, pero la información registra frecuencias y cantidades de transacciones. En ese sentido el análisis de conglomerados se eligió para inferir comportamientos de consumidores detrás de estas transacciones; sin embargo, una cosa es la cantidad en la que se realiza la transacción en cada conglomerado; y otra cosa es el tamaño del conglomerado. Para entender a fondo la naturaleza del contenido es relevante establecer como condición, la probabilidad de que un comportamiento aislado en una variable se realice, teniendo en cuenta el comportamiento del conglomerado al que esta pertenece.

Una de las técnicas generalmente utilizadas para crear agrupaciones mediante la optimización de la función de criterio de calificación, definida globalmente (diseño total) o local (en el subconjunto de los diseños), es la técnica K-medias .La agrupación de K-medias es una de las *n* observaciones predictivas más antiguas en el espacio *d* dimensional (se da un número entero *d*) y el problema es determinar un conjunto de puntos *c* para minimizar la distancia cuadrática media desde cada punto de los datos hasta su centro más cercano, al cual cada observación pertenece. No se conocen algoritmos exactos de tiempo polinómico para este problema. El problema se puede configurar como un problema de programación de enteros, pero debido a que resolver programas de enteros con una gran cantidad de variables lleva mucho tiempo, los clústeres a menudo se calculan utilizando un método heurístico rápido que generalmente produce buenas soluciones (pero no necesariamente óptimas). El algoritmo K-medias es uno de esos métodos donde la agrupación requiere menos esfuerzo. Al principio, se determina el número del grupo *c* y se determina un centro aleatorio paras estos grupos. Se puede tomar cualquiera de los primeros *k* objetos en que secuencia pueden servir como centroide inicial. Sin embargo; si hay algunas características, con un gran tamaño o una gran variabilidad, este tipo de características afectará fuertemente el resultado de la agrupación. En este caso, la estandarización de datos sería una tarea importante de preprocesamiento, que se hace necesario para escalar o controlar la variabilidad de los conjuntos de datos.

El algoritmo K-medias realizará los tres pasos a continuación hasta lograr convergencia e iterar hasta que sea estable (sin grupo de movimiento de objetos): Determinar la coordenada centroide, determinar la distancia de cada objeto a los centroides agrupar el objeto según la distancia mínima. El objetivo de la agrupación sería descubrir las similitudes y diseños de los grandes conjuntos de datos dividiendo los datos en grupos. Dada la suposición que los conjuntos de datos no están etiquetados, la agrupación se considera con frecuencia como el problema de aprendizaje no supervisado más valioso. Una aplicación principal de medidas geométricas (distancias) a entidades con rangos grandes, asignará implícitamente mayores esfuerzos en las métricas en comparación con la aplicación realizada a entidades que contengan rangos más pequeños. Además, las características deben ser adimensionales ya que los valores numéricos de los rangos de las características dimensionales dependen de las unidades de medida y, por lo tanto, una selección de las unidades de medida puede alterar significativamente los resultados de la agrupación. Por lo tanto, no se deberían emplear medidas de distancia como la distancia euclidiana sin tener normalización de los conjuntos de datos. Una etapa de preprocesamiento es realmente esencial antes de usar cualquier algoritmo de exploración de datos para mejorar el rendimiento de los resultados. La normalización del conjunto de datos se encuentra entre los procesos de preprocesamiento en la exploración de datos, en los que los datos de los atributos se escalan a todos en un pequeño rango específico. La normalización antes de la agrupación se necesita específicamente para la métrica de distancia, como la distancia euclidiana que es sensible a las variaciones dentro de la magnitud o escalas de los atributos. En aplicaciones reales, debido a las variaciones en la selección del valor del atributo, un atributo puede dominar a otro. La normalización evita que las

características de mayor peso tengan un gran número sobre las características con números más pequeños. El objetivo sería igualar las dimensiones o la magnitud y también la variabilidad de esas características. Las técnicas de preprocesamiento de datos se aplican a datos sin procesar para hacer que los datos sean limpios, libres de ruido y consistentes.

La normalización de datos estandariza los datos sin procesar al convertirlos en un rango específico mediante una transformación lineal que puede generar agrupaciones de buena calidad y mejorar la precisión de los algoritmos de agrupación. No existe una regla universalmente definida para normalizar los conjuntos de datos y, por lo tanto, la elección de una regla de normalización particular se deja a discreción del usuario. Por lo tanto, los métodos de normalización de datos incluyen el puntaje *Z*, Min-Max y escala decimal. En la puntuación *Z*, los valores para un atributo *X* están estandarizados en función de la media y la desviación estándar de *X*, este método es útil cuando se desconocen el mínimo y el máximo reales del atributo *X*; por otra parte, en la escala decimal estandarizada, al mover el punto decimal de los valores del atributo *X*, el número de puntos decimales movidos depende del valor absoluto máximo de *X*; en contraste e procedimiento *Min-Max* transforma el conjunto de datos entre *0.0* y *1.0* restando el valor mínimo de cada valor dividido por el rango de valores para cada valor individual.

La función $Y = \{X_1, X_2, \cdots, X_n\}$ denota el conjunto de datos brutos *d*-dimensiones. Entonces la matriz de datos es una matriz $n \times d$ dada por:

$$X_1, X_2, X_3, X_n = \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nd} \end{pmatrix} \quad (2)$$

Puntaje Z: El puntaje Z es una forma de estandarización utilizada para transformar variantes normales en forma de puntaje estándar. Dado un conjunto de datos sin procesar *Y*, la fórmula de estandarización del puntaje Z se define como:

$$X_{ij} = \mathrm{Z}\left(X_{ij}\right) = \frac{X_{ij} - \overline{X}_j}{\sigma_j} \quad (3)$$

Donde, $\bar{X}_j$ xj y $\vartheta_j$son la media muestral y la desviación estándar del atributo $j_{th}$ , respectivamente. La variable transformada tendrá una media de *0* y una varianza de *1*. La información de ubicación y escala de la variable original se ha perdido. Una restricción importante de la estandarización de la puntuación Z.

$$\underset{s}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{i_j \in S_i} \left\| X_j - \mu_i \right\|^2 \quad (4)$$

Agrupación de K-medias: dado un conjunto de observaciones $(X_1, X_2, \cdots, X_n)$, donde cada observación es un vector real d-dimensional, la agrupación de K-medias tiene como objetivo dividir las *n* observaciones en *k* conjuntos $(K \leq n)$ , S = $\{S_1, S_2, \cdots, S_k\}$ para minimizar la suma de cuadrados dentro del clúster. El puntaje Z es el método de estandarización más apropiado para producir agrupaciones de calidad para la técnica de clústeres.

*Presentación del Modelo:* Se realizaron pruebas para modelos de dos, tres y cuatro conglomerados, el modelo de segmentación de tres conglomerados arrojo la separación más clara entre los segmentos. La Tabla 2 exhibe una solución de tres clústeres con la respectiva media en su solución final (todas significativas en **0.01). El conglomerado 1 arroja una frecuencia de 102, el conglomerado 2 una frecuencia de 647, y el conglomerado 3 una frecuencia de 357; como resultado el porcentaje valido es 9.2%, 58.5% y 32.3% respectivamente. Sin embargo; al comparar estos resultados con la Tabla 2, las medias

individuales para las variables que conforman cada clúster, se obtiene que el clúster 1 de menor frecuencia, es el que mayor cantidad transacciones aporta a la categoría. El clúster 1 se caracteriza por ser el de mayor cantidad de montos, totales realizados tanto en compras y avances nacionales e internacionales; contraste el clúster 2 es el que menores montos realizan, maneja los saldos en tarjeta más bajos, los más bajos niveles por compras y avances y los niveles más bajos de cupo de crédito no utilizado; por otra parte el clúster 3, tiene niveles sostenibles de tarjetas vigentes a la fecha de corte y tarjetas bloqueadas, y por consiguiente sostenibles de intereses por compra y avances corrientes.

Tabla 2: Preparación de Conglomerados (Acumulado Junio 2014-Junio 2019)

| | **Clúster 1 (9.2%)** | **Clúster 2 (58.5%)** | **Clúster 3 (32.2%)** |
|---|---|---|---|
| Monto compras en el exterior | 2.58601 | -0.51107 | 0.18737 |
| Montos internacionales | 2.58549 | -0.50737 | 0.18080 |
| Compras en el exterior | 2.54430 | -0.51414 | 0.20484 |
| Monto avances nacionales | 2.52864 | -0.53172 | 0.24118 |
| Total montos | 2.50360 | -0.61377 | 0.39703 |
| Montos nacionales | 2.44971 | -0.62611 | 0.43480 |
| Avances nacionales | 2,36154 | -0.46578 | 0.16942 |
| Monto avances en el exterior | 2.34614 | -0.31319 | -0.10272 |
| Monto compras nacionales | 2.33018 | -0.64018 | 0.49445 |
| Avances en el exterior | 2.29708 | -0.30830 | -0.09757 |
| Compras nacionales | 2.28539 | -0.64008 | 0.50707 |
| Cupo de crédito no utilizado | 2.15095 | -0.66014 | 0.58182 |
| Intereses por compra y avances corrientes | 2.14514 | -0.66303 | 0.58873 |
| Canceladas | 2.12885 | -0.58406 | 0.45027 |
| Saldo de tarjeta de crédito | 2.08264 | -0.67786 | 0.63346 |
| Vigentes durante el mes | 1.99134 | -0.60223 | 0.52248 |
| Intereses por compra y avances de mora | 1.71172 | -0.44051 | 0.30929 |
| Vigentes a la fecha de corte | 1.69913 | -0.67191 | 0.73225 |
| Castigos de cartera diferente a capital | 1.69608 | -0.34858 | 0.14715 |
| Castigos de cartera capital | 1.61987 | -0.50741 | 0.45677 |
| Bloqueadas temporalmente | 1.48158 | -0.59847 | 0.66132 |

*El Análisis Clúster, conocido como Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. La mejor analogía para explicar el análisis de conglomerados es por ejemplo, cuando se necesita saber cuál es el punto medio de diez disparos realizados a una diana. La dispersión es la distancia más cercana entre todos ellos, hasta obtener lo que se conoce como un centroide, si luego se realizan otros diez disparos y se repite el proceso, se toma ahora la distancia entre los dos nuevos centroides, y así sucesivamente, el proceso de repetir cada grupo de disparos se le conoce como iteraciones.*

Para llevar a cabo el conglomerado K-medias, el algoritmo inicialmente asigna k centros iniciales (k es especificado por el usuario), ya sea por medio de selección aleatoria dentro del espacio euclidiano definido por *n* variables o por la selección muestral de k puntos en todas las observaciones, con el fin de ubicar los centroides iniciales. Posteriormente asigna de manera iterativa cada observación al centroides más cercano. Luego, calcula nuevos centros para cada conglomerado a medida que la media de los centroides pertenecientes a las variables de cada conglomerado va agregando nuevas observaciones. K-medias reitera este proceso, asignando nuevas observaciones al centro más cercano (algunas observaciones cambiaran el conglomerado). Este proceso se repite hasta cuando una nueva iteración no reasigne nuevas observaciones al nuevo conglomerado. En este punto, el algoritmo se considera haber hecho convergencia y las asignaciones siguientes constituyen la solución final del conglomerado. Posteriormente se filtró la variable clúster para cada uno de los valores en la matriz inicial de datos, y se totalizó el aporte de cada variable a cada clúster, hasta obtener el total de cada variable como se muestra en la Tabla 2.

Tabla 3: Tabla Cruzada de Resultados de Aporte Total de Cada Variable a Cada Clúster

| | **Clúster 1** | **Clúster 2** | **Cluster3** | **Total** |
|---|---|---|---|---|
| Vigentes a la fecha de corte | 1783045.57 | 166414.28 | 1123801.71 | 3073261.56 |
| Vigentes durante el mes | 41288.19 | 3902.06 | 20114.69 | 65304.94 |
| Canceladas | 39769.57 | 2884.53 | 16947.4 | 59601.5 |
| Bloqueadas temporalmente | 201768.83 | 25951.2 | 132435.48 | 360155.51 |
| Compras nacionales | 3150926.84 | 154090.94 | 1329222.78 | 4634240.56 |
| Monto compras nacionales | 6,6327e+11 | 3,5317e+10 | 2,7519e+11 | 9,7377e+11 |
| Avances nacionales | 556019.38 | 31141.78 | 149064.7 | 736225.86 |
| Monto avances nacionales | 2,65448e+11 | 1.4736e+10 | 7,8054e+10 | 3.5824e+11 |
| Compras en el exterior | 1206880.08 | 40593 | 314763.96 | 1562237.04 |
| Monto compras en el exterior | 1,82646e+11 | 6726771100 | 4.6399e+10 | 2.3577e+11 |
| Avances en el exterior | 5459.12 | 294.2 | 711.95 | 6465.27 |
| Monto avances en el exterior | 2914843722 | 143628998 | 362963206 | 3421435926 |
| Saldo de tarjeta de crédito | 4.18066e+12 | 2.6514e+11 | 2.1251e+12 | 6.5709e+12 |
| Cupo de crédito no utilizado | 7.18893e+12 | 4.9614e+11 | 3.4531e+12 | 1.1138e+13 |
| Intereses por compra y avances corrientes | 70459557859 | 4383331280 | 3.3837e+10 | 1.0868e+11 |
| Intereses por compra y avances de mora | 5240991543 | 155552907 | 1927220383 | 7323764833 |
| Catigos de cartera capital | 21505081074 | 4246553039 | 1.2069e+10 | 3.782e+10 |
| Catigos de cartera diferente a capital | 2619368101 | 364972864 | 911552348 | 3895893313 |
| Total montos | 1.11524e+12 | 5.684e+10 | 4.0002e+11 | 1.5721e+12 |
| Montos nacionales | 9.29062e+11 | 5.0054e+10 | 3.5324e+11 | 1.3324e+12 |
| Montos internacionales | 1.8654e+11 | 6786280603 | 4.6782e+10 | 2.4011e+11 |

*La Tabla explica el paso de obtener el porcentaje total que cada variable le aporta a cada clúster, para despejar el camino de establecer probabilidades posteriores y condicionales. La tabulación cruzada clúster x variable permite identificar de forma cuantitativa el valor y el aporte de cada variable a cada clúster.*

El total de cada variable obtenido para cada clúster, se divide por el valor individual de la variable para cada fila de la Tabla 3 hasta obtener el porcentaje total que cada variable le aporta a cada clúster, con la condición, de que cada uno aporte un 100%, al sumar los tres clústeres como se observa en la Tabla 4. El clúster 1, clúster 3 y clúster 3, se reemplazan en las formulas del teorema de bayes como $C_1, C_2, C_3$ hasta obtener las probabilidades posteriores; es decir si se tiene en cuenta que, a juzgar por el tamaño de los clústeres, el mercado de tarjetas de crédito tiene una baja penetración en el mercado Colombiano; sin embargo el clúster más pequeño es el que guía la industria por su nivel transaccional y montos efectuados por franquiciamiento internacional, ¿ qué ocurriría entonces en el mercado una vez el clúster 1, ha desempeñado su función? o ¿Qué características remanentes no observa la industria para los comportamientos que manifiesta el clúster 2 y clúster 3?

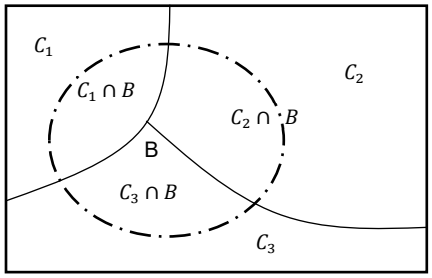
Tabla 4: Porcentajes de Aporte de Cada Variable a Cada Clúster

| | **Clúster 1** | **Clúster 2** | **Clúster 3** | **Total** |
|---|---|---|---|---|
| **Tamaño del Clúster** | **9.22%** | **58.49%** | **32.27%** | 100.00% |
| Vigentes a la fecha de corte | 58.02% | 5.41% | 36.57% | 100.00% |
| Vigentes durante el mes | 63.22% | 5.98% | 30.80% | 100.00% |
| Canceladas | 66.73% | 4.84% | 28.43% | 100.00% |
| Bloqueadas temporalmente | 56.02% | 7.21% | 36.77% | 100.00% |
| Compras nacionales | 67.99% | 3.33% | 28.68% | 100.00% |
| Monto compras nacionales | 68.11% | 3.63% | 28.26% | 100.00% |
| Avances nacionales | 75.52% | 4.23% | 20.25% | 100.00% |
| Monto avances nacionales | 74.10% | 4.11% | 21.79% | 100.00% |
| Compras en el exterior | 77.25% | 2.60% | 20.15% | 100.00% |
| Monto compras en el exterior | 77.47% | 2.85% | 19.68% | 100.00% |
| Avances en el exterior | 84.44% | 4.55% | 11.01% | 100.00% |
| Monto avances en el exterior | 85.19% | 4.20% | 10.61% | 100.00% |
| Saldo de tarjeta de crédito | 63.62% | 4.03% | 32.34% | 100.00% |
| Cupo de crédito no utilizado | 64.54% | 4.45% | 31.00% | 100.00% |
| Intereses por compra y avances corrientes | 64.83% | 4.03% | 31.13% | 100.00% |
| Intereses por compra y avances de mora | 71.,56% | 2.12% | 26.31% | 100.00% |
| Castigos de cartera capital | 56.86% | 11.23% | 31.91% | 100.00% |
| Castigos de cartera diferente a capital | 67.23% | 9.37% | 23.40% | 100.00% |
| Total montos | 70.94% | 3.62% | 25.45% | 100.00% |
| Montos nacionales | 69.73% | 3.76% | 26.51% | 100.00% |
| Montos internacionales | 77.69% | 2.83% | 19.48% | 100.00% |
| Total castigos de cartera | 57.83% | 11.05% | 31.12% | 100.00% |

*La tabla presenta las probabilidades en iniciales de ponderación de cada clúster en la muestra de datos, la suma de los tres cada clúster debe arrojar el 100%, de tal forma que se cumpla la condición de ser exhaustivos y mutuamente excluyentes, es decir, un clúster debe estar los más separado posible del otro antes de formular la probabilidad del espacio común entre ellos.*

El teorema de bayes va mucho más allá de un evento de cara y sello en una moneda o el giro de unos dados, el concepto de probabilidad es muy importante cuando se trata de análisis de datos. Las probabilidades conllevan a distribuciones y a su vez, las distribuciones son como un mapa de lectura de los datos, como se puede apreciar en la Tabla 4, el mapa de distribución de comportamientos subyacentes es como la ruta de viaje para interpretar los tarjetahabientes. El teorema de bayes es la aplicación de la probabilidad condicional. Así como el análisis de conglomerados provee todos los resultados posibles de un experimento $C_1, C_2, C_3$ ; existe un espacio mutuo $B$ en el que un comportamiento se realiza, una vez ocurre el conglomerado al que pertenece. La Figura 2, representa la variación en el comportamiento de una misma variable, sujeta a la condición del conglomerado al que pertenezca, por ejemplo, el comportamiento de la variable tarjetas bloqueadas, está sujeto o condicionado al comportamiento del segmento de mercado o conglomerado según ecuación (1) y (2) que representa la intersección entre cada conglomerado y la variable común.

Figura 2: Gráfico de Intersección Conglomerado vs Variable Constituyente




*La gráfica explica de una forma más didáctica el proceso de aplicar el teorema de bayes y la definición de las variables. Se observa que cada clúster C, posee su espacio propio; sin embargo existe la probabilidad de coexistir un espacio común B en medio de todos ellos. La probabilidad posterior se hace relevante si dado que existe un $C_1$ menor en tamaño mucho mayor en volumen transaccional, se necesita saber, como este segmento modifica el mercado una vez este se ha manifestado.*

$$\mathrm{P}\,(\mathrm{B}) = \mathrm{P}(\mathrm{C}_1 \cap \mathrm{B}) + \mathrm{P}(\mathrm{c}_2 \cap \mathrm{B}) + \mathrm{P}(\mathrm{c}_3 \cap \mathrm{B}) \quad (5)$$

$$P\,(B) = P(B|c_1) * P(c_1) + P(B|c_2) * P(c_2) + P(B|c_3) * P(c_3) \quad (6)$$

$$\mathrm{P}(B) = \sum_{i=1}^{n} P(B|c_i) * P(c_i) \quad (7)$$

$$\mathrm{P}(c_i|B) = \frac{P(B|c_i)*P(c_i)}{P(B)} \quad (8)$$

$$\mathrm{P}(B) = \sum_{i=1}^{n} P\,(B|c_i) * P(c_i) \quad (9)$$

$$\mathrm{P}(c_i|B) = \frac{P(B|c_i)*P(c_i)}{\sum_{i=1}^{n} P(B|c_i)*P(c_i)} \quad (10)$$

La Figura 2 representa la probabilidad de que ocurra si $C_2$ ha ocurrido $C_1$con anterioridad. Se necesita conocer el espacio común en el que ocurre $C_1$, pero además puede ocurrir $C_2$ y $C_3$. Se necesita conocer la disponibilidad de $C_2$ en $C_1$, osea P $(C_2|C_1)$ y esa es la probabilidad entre ( $C_1 \cap C_2$) dividido, la probabilidad en $C_1$, osea P ($C_1$); por otra el espacio común *B* es la suma de la intersección entre ($C_1 \cap B$), ($C_2 \cap B$), ($C_3 \cap B$), hasta llega a la ecuación (1) que defina la probabilidad del espacio común *B*.

*Evaluación del Modelo:* Basado en la información histórica a junio de 2014 a junio de 2019, sobre transacciones con tarjeta de crédito, se desea conocer cómo se diferencian los segmentos de mercado según el monto de transacciones internacionales. Los bancos tienen catalogado el clúster 1 (9,22%) como, alto monto de transacciones internacionales, clúster 2 (48,49%) como mediano monto de transacciones internacionales y clúster 3 (32,27%) como bajo monto de transacciones internacionales. Según la Tabla 1, el monto de transacciones internacionales para el clúster 1 fue de 77,69%, para el clúster # 2 fue de 2,83% y para el clúster 3 fue de 19,48%.

Tabla 5: Mapa de Distribución de Probabilidades Para Interpretación de Comportamientos Subyacentes (Acumulado Junio 2014- Junio 2019)

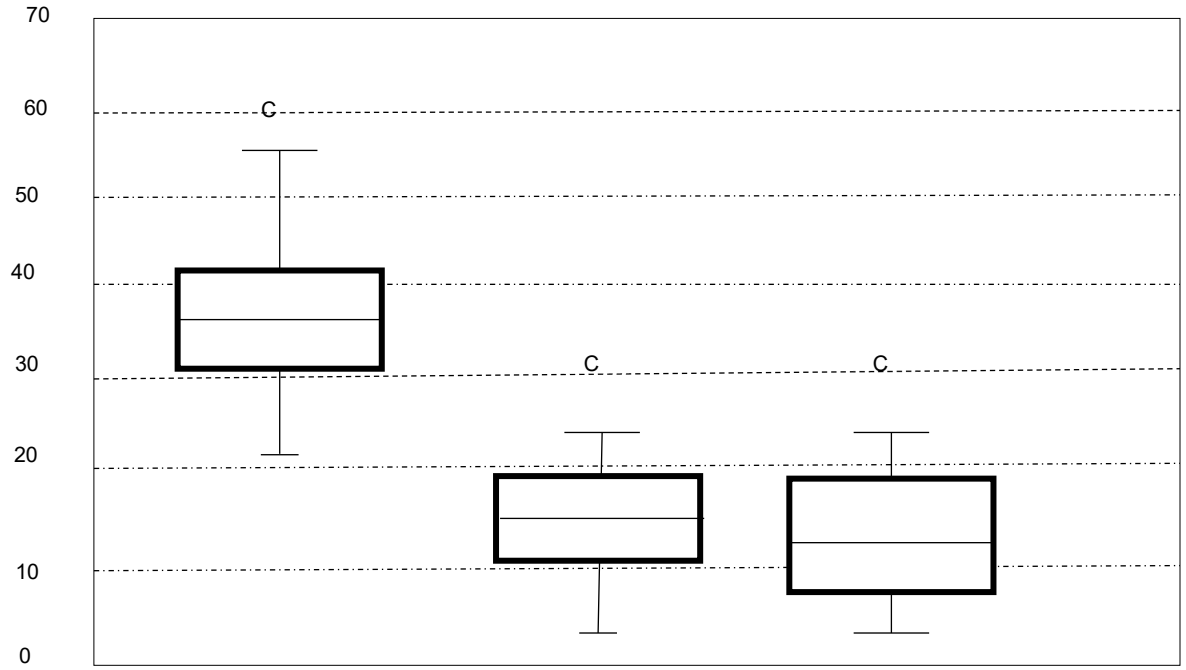
| | **Clúster 1** | **Clúster 2** | **Clúster 3** |
|---|---|---|---|
| Monto compras nacionales | 26.33% | 15.59% | 58.08% |
| Compras nacionales | 30.26% | 18.14% | 23.33% |
| Bloqueadas temporalmente | 33.88% | 15.59% | 17.09% |
| Montos nacionales | 24.31% | 19.84% | 21.70% |
| Total montos | 35.88% | 11.13% | 21.58% |
| Monto avances nacionales | 35.84% | 12.11% | 14.86% |
| Intereses por compra y avances de mora | 43.60% | 15.49% | 9.98% |
| Cupo de crédito no utilizado | 41.99% | 14.79% | 7.29% |
| Vigentes a la fecha de corte | 47.03% | 10.04% | 7.11% |
| Total castigos de cartera | 47.11% | 11.01% | 6.21% |
| Compras en el exterior | 55.61% | 19.01% | 3.20% |
| Avances nacionales | 57.19% | 17.88% | 1.59% |
| Montos internacionales | 31.43% | 12.65% | 5.94% |
| Castigos de cartera diferente a capital | 32.06% | 14.04% | 19.56% |
| Vigentes durante el mes | 32.52% | 12.83% | 18.68% |
| Intereses por compra y avances corrientes | 40.40% | 7.61% | 14.99% |
| Canceladas | 23.71% | 29.71% | 17.02% |
| Avances en el exterior | 32.24% | 28.50% | 15.97% |
| Monto avances en el exterior | 38.78% | 12.54% | 11.03% |
| Catigos de cartera capital | 37.42% | 12.79% | 11.97% |
| Monto compras en el exterior | 47.43% | 10.95% | 8.69% |
| Saldo de tarjeta de crédito | 24.41% | 29.61% | 11.82% |

*Descubrir patrones en grandes volúmenes de datos es la tarea de traer al presente la información histórica acumulada. Este proceso se denomina minería de datos. La articulación de análisis inferencial estadístico, administración de bases de datos, inteligencia artificial y visualización en tiempo real, atreves de interfaces gráficas de interpretación, son las herramientas que utiliza el Gracias a las facultades del "Knowledge Discovery in Databases" se puede abordar la pregunta de investigación: .¿Cómo desarrollar una metodología que ayude a saber quién está detrás de la hoja de papel que sostiene la limadura de hierro, que a los ojos de un observador externo, se mueve por si sola?*

Al aplicar la ecuación (6), se desea conocer, por ejemplo, cuál es la probabilidad de obtener un tarjetahabiente que pertenezca al clúster 1, si aleatoriamente se elige alguno de la variable "montos internacionales". La Tabla 3 representa el mapa de distribuciones para interpretar cada perfil del consumidor de tarjeta de crédito, sujeto a la condición de que cada variable se comporta de una forma diferente dependiendo del conglomerado o clúster al que pertenezca, es decir, el objetivo del análisis no es el de conocer el comportamiento de una variable; sino el de inferir el comportamiento futuro del mercado. A partir de la estructura del negocio, subyace la urgencia de los bancos por aumentar la penetración de su servicio, al emitir tarjetas nuevas mes a mes y asignar un cupo de crédito a nuevos tarjetahabientes dentro del portafolio de sus tarjetas. Como el individuo entiende la función de la tarjeta como medio de pago, como línea de crédito y como una mezcla de ambos servicios; existe una selección natural de la clientela en virtud de su capacidad de pago, por ende, cupos de crédito asignados y ejercicio de la línea de apertura de crédito. Existe el consumidor ($C_1$) al que se le otorgan altos cupos de crédito, hace uso de éste para compras nacionales y compras internacionales, eventualmente compras por internet y posiblemente hace uso de una tarjeta corporativa en el exterior para gastos de representación. Este consumidor tiene mayor probabilidad a pagar altos intereses corrientes por mantener altos balances en su tarjeta de crédito, por su monto de compras y número de veces que la usa, por eso, para este consumidor es imperativo mantener su tarjeta vigente durante el mes, pero por los altos montos y avances en mora, con frecuencia encuentra su tarjeta bloqueada temporalmente.

Otro tipo de consumidor ($C_2$) es el que no posee un cupo de crédito muy alto, pero hace uso de este casi hasta el límite, con frecuencia encuentra su tarjeta bloqueada por no poseerla vigente durante el mes, y paga intereses de mora con mucha frecuencia, tiene un alto monto de compras nacionales. El ultimo tipo de consumidor ($C_3$) es el que entiende la tarjeta de crédito estrictamente como medio de pago por que su cupo de endeudamiento es muy bajo, tiene un bajo cupo de crédito asignado, una tarjeta vigente durante el mes, paga bajos intereses corrientes, bajo monto y número de compras a nivel nacional y es el tipo de consumidor que los bancos buscan para penetrar mercados con tarjetas nuevas mes a mes. En la Tabla 5, se observa a la izquierda, el clúster 1. Este representa el 9,22% del tamaño del mercado, pero el 35,88% del total de los montos realizados. Esto infiere que el producto de tarjeta de crédito tiene unos comportamientos de Pareto muy marcados dentro del comportamiento del mercado. Cámara, D. (1994); en contraste el clúster 2, representa el 58, 49% del mercado, pero un total de montos de solo el 11,3%. Es determinante para el producto, el papel de las franquicias internacionales, porque estas son las que dividen la estructura y el comportamiento del mercado en virtud del enrutamiento internacional que ofrecen. El clúster 3 a la derecha de la gráfica, representa el 32, 27% con un valor extremo del 58,08% sobre el monto de compras nacionales, en relación al comportamiento del resto de variables dentro del clúster.

Figura 3: Gráfico de Cajas y Bigotes Para Probabilidades Posteriores




*También conocido como diagrama de caja y bigote, box plot, box-plot o boxplot. Es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, el diagrama de caja muestra a simple vista la mediana y los cuartiles de los datos,1 de forma tal que permita además, representar los valores atípicos de estos.*

## RESULTADOS

El ciclo de vida de un producto se interpreta sobre la variable tiempo. Lo primero que ocurre dentro de la vida de la tarjeta es la otorgación del plástico al usuario. En ese momento la tarjeta se encuentra vigente, el usuario hace uso de un cupo de crédito, para compras nacionales, compras internacionales, avances nacionales, avances en el exterior y recibe a fin de mes un extracto en el que se le liquidan sus intereses corrientes sobre el número de cuotas que elija pagar. El saldo en tarjeta de crédito es el balance remanente que se acumula mes a mes, sobre la diferencia entre el saldo mínimo cancelado y el total de motos realizados. En el evento en el que un tarjetahabiente no logre cancelar su cuota mínima de pago, la tarjeta

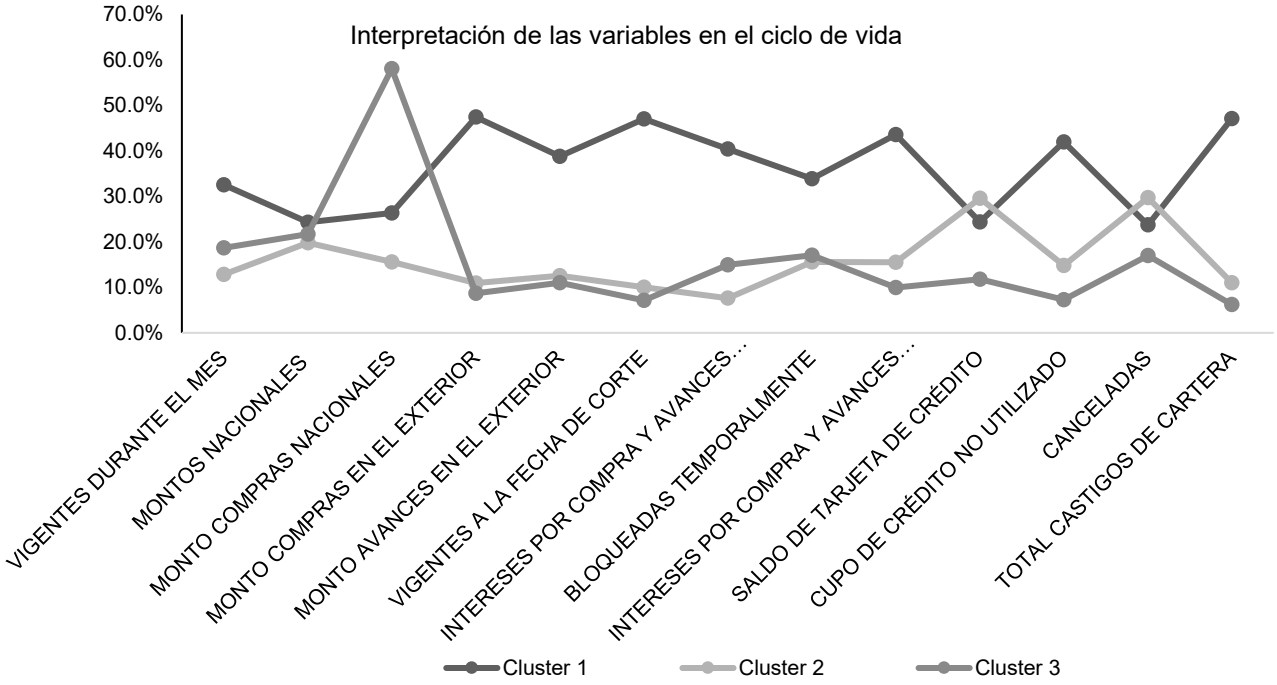
puede aparecer bloqueada temporalmente, causando intereses por compra y avances de mora, hasta llegar a tener la cancelación definitiva de su tarjeta. Después de generarse todo este ejercicio en el que interviene el tarjetahabiente, el banco emisor, el establecimiento y el banco adquiriente, la entidad bancaria obtiene una cartera castigada sobre el crédito vencido.

La Figura 4 interpreta las probabilidades posteriores independientes para cada clúster a lo largo del ciclo de vida propuesto para tarjeta de crédito, es decir el comportamiento de cada variable, en función del clúster o segmento al que pertenece. Los resultados de la minería de datos, permiten observar las estrategias de posicionamiento y colocación que implementan los bancos en el mercado. El posicionamiento de las tarjetas se logra con la red de franquicias que ofrecen enrutamiento internacional a tarjetahabientes que viajan con mucha frecuencia al exterior, utilizan tarjetas corporativas y realizan avances en efectivo en el exterior. Se observa que el clúster 1 es el tipo de tarjetahabiente, sobre el cual tiene mayor efecto el empaquetamiento de beneficios, tales como, acumulación de millas viajeras, y fidelidad a franquicias internacionales; de otro lado los clústeres 2 y 3 reciben las mayores estrategias de colocación de tarjetas nuevas, reciben descuentos de temporada en establecimientos, y obtienen cupos de crédito más limitados, donde se les hace mayor énfasis en la tarjeta como medio de pago para transacciones nacionales.

La Figura 4 es básicamente una interpretación gráfica de las cifras que arroja la Tabla 5, con la diferencia de que está vez, las variables están ordenadas en el orden en el que ocurren las transacciones en el ciclo de vida del producto según se ha explicado anteriormente. La razón principal por la cual se ha incorporado el teorema de bayes en el análisis de clústeres es, por que el "driver" o líder de la categoría es el clúster 1; a pesar de este ser el clúster más pequeño, genera la mayor cantidad de transacciones. En ese sentido, se hace necesario entender que pasa con el mercado bajo la influencia de un segmento de mayor ponderación sobre la categoría. Los montos nacionales son similares para los tres clústeres, son tarjetas emitidas en Colombia, que realizan compras en establecimientos con una frecuencia equivalente; sin embargo, se resalta una diferencia importante del clúster 2 en el monto de compras nacionales, donde se resalta el uso de la tarjeta como medio de pago alternativo para compras, a partir de ese punto se evidencia el efecto de la probabilidad posterior, como efecto del comportamiento del clúster 1 sobre el clúster 2 y 3.

El clúster 1 presenta un el monto de compras y avances en el exterior, aproximadamente tres veces mayor que el clúster 2 y 3, del mismo modo las tarjetas vigentes la fecha de corte, indican la frecuencia de uso de la tarjeta; sin embargo por los altos balances en tarjeta de crédito y el número de responsabilidades asociadas a la tarjeta, se presentan bloqueos frecuentes por los altos intereses corrientes y en mora. El saldo en tarjeta de crédito vuelve a ser equivalente para los tres clústeres, como muestra de la alta capacidad de pago del segmento; sin embargo en el clúster 1, se dispara la cartera castigada al nuevamente al final del ciclo de vida del producto. Se evidencia la estrategia de los bancos en la otorgación de cupos de crédito ilimitado a estos segmentos y la importancia de ejercitar convenios con franquicias internacionales, para fortalecer el clúster 1.

Figura 4: Estrategias Genéricas de Tarjeta de Crédito Para Clústeres




*La gráfica presenta los comportamientos de los tarjetahabientes al interior de cada segmento y la interpretación que los bancos le dan a la industria, el $C_1$ es un segmento de tarjetas emitidas en Colombia peor de uso primordialmente internacional, el $C_2$ es una tarjeta utilizada en Colombia con altos montos, sujeta a cantidad de beneficios y convenios y el $C_3$ es una tarjeta utilizada básicamente como medio de pago a nivel nacional. La cartera castigada tan alta del $C_1$ infiere que el ciclo de vida de esta clúster en mucho menos que el de los otros dos.*

## CONCLUSIONES

La minería de datos requiere de un depurado proceso de preparación de los datos, para que el efecto de realizarla arroje resultados diferenciados como resultado del análisis subyacente. El análisis de clústeres permite identificar contextos singulares para las variables de datos y establecer condiciones sobre las cuales las variables evolucionan a lo largo del tiempo. El teorema de bayes aplicado a las bases de datos permite identificar comportamientos subyacentes al interior de los segmentos de mercado para diferenciar el comportamiento y la evolución de estos en el ciclo de vida de un producto. Las leyes de protección de datos infieren que el mercadeo de servicios financieros y en particular la generación de estudios sindicados de mercados para el sector, deban realizarse, con tecnologías cada vez más avanzadas, porque el reto de los tiempos y el recelo de los bancos por divulgar información privilegiada sobre sus usuarios, hace que los investigadores deban perfeccionar sus metodologías tanto para mejorar el servicio como para reducir el fraude. Existe una ambivalencia entre la necesidad de mejorar la tecnología de análisis de los datos y la limitación jurídica que imponen los entes locales e internacionales en relación a la información de usuarios a nivel bancario, como activo empresarial y como mecanismo de anticipación al comportamiento de la industria. La revisión del marco teórico alrededor del problema de investigación presenta la importancia de incorporar mecanismos de inteligencia artificial en la información de usuarios bancarios, la necesidad de interpretar el cliente en diferentes etapas y periodos de tiempo, la subsecuente toma de decisiones alrededor de mapas de distribución de la probabilidad de los comportamientos y la importancia de articular a todos los constituyentes de la industria en un mismo constructo tecnológico que maximice el servicio al usuario. Los segmentos de mercado o clústeres arrojados, interpretados a lo largo del ciclo de vida del producto, explican la importancia de las franquicias internacionales en el posicionamiento de las marcas independientemente de quien sea el emisor de la tarjeta de crédito. Se vislumbran oportunidades de investigación muy interesantes alrededor de este sujeto de estudio.

Por ejemplo: ¿Cómo impactaría el comportamiento de las variables públicas sobre en la industria, el hecho de que una tarjeta sea emitida por un banco, un establecimiento comercial o una compañía de financiamiento? ¿La penetración de las franquicias en el mercado nacional, tiene un efecto en el incremento en las transacciones internacionales de tarjeta de crédito? En conclusión, la pregunta de investigación inicial: ¿Cómo desarrollar una metodología de investigación que ayude a saber quién mueve la limadura de hierro sobre de la hoja de papel? a simple vista esta limadura mueve por si sola. Esta misma analogía sirve para reiterar el objetivo del estudio: Desarrollar una adaptación del proceso convencional de minería de datos, para la industria de tarjeta de crédito en Colombia, que permita inferir quien es el usuario que registra las cifras sobre manejo de tarjetas de crédito, y sus intenciones futuras a partir del análisis de las cifras que son publicadas acerca de la industria. Las leyes de protección de datos, las características anormales de las bases de datos públicas sobre tarjetas de crédito, la dificultad para construir matrices de datos para un correcto análisis, son barreras que dificultan conocer quién está detrás de las transacciones. La motivación principal del investigador era la de encontrar una manera educada de adivinar mediante un cálculo probabilístico, la mejor manera de matizar el comportamiento de las variables que describen la industria de tarjetas de crédito, condicionadas por la influencia que el sujeto de estudio tiene sobre ellas. El proceso de preparación de los datos y la normalización de las variables, advierte la necesidad de construir segmentos de mercado por medio del análisis de clústeres.

La metodología del estudio se resume de la siguiente manera: Construcción de una matriz de datos obtenidos del sitio web de la superintendencia financiera, exportación de la matriz al software SPSS, realización de la prueba kolmogorov- smirnov para someter a prueba de hipótesis la homogeneidad de la muestra con la población, elegir una estrategia de segmentación a partir de la prueba de hipótesis, hasta identificar el segmento que causa la anormalidad de los datos, estandarizar las variables con el puntaje Z, realizar pruebas de conglomerados y elegir el número de conglomerados que conformarán el análisis, agrupar los datos por clúster en columnas y magnitud de las variables en las filas, obtener los porcentajes de cada clúster a partir del total de cada variable, aplicar el teorema de bayes a la tabla de porcentajes iniciales, obtener la probabilidad posterior y condicional, graficar las variables en el orden secuencial en el que ocurren en el ciclo de vida de la categoría tarjetas de crédito. Los resultados arrojan 1 clústerde menor tamaño pero de mayor volumen de transacciones internacionales y cupos de crédito aprobados, contra dos clústeres similares dentro de la distribución normal de la industria, esté fenómeno a su vez es aprovechado por los bancos de mayor penetración y cubrimiento en el mercado para orientar sus estrategias de mercadeo en función del comportamiento del consumidor. Las limitaciones de esta investigación están dadas por la incapacidad de entrevistar directamente al sujeto de estudio y la calidad de las cifras que reporte la superintendencia financiera que son reportadas con dos meses de atraso, este hecho hace imperativo desarrollar modelos de inferencia para interpretar el comportamiento. Futuras avenidas de investigación se vislumbran a partir de este estudio si, se logra tomar nuevos modelos econométricos orientados a la investigación del comportamiento y ser adaptados a un marco de referencia para el análisis del consumidor, también se vislumbran oportunidades de enseñanza sobre contenidos y procedimientos de la minería de datos para estudiantes de programas gerenciales y que tengan que ver con la gestión de la innovación y la tecnología.

## REFERENCIAS

Aldenderfer MS and Blashfield RK (1984). “Cluster Analysis. Sage University Paper series on Quantitative Applications in the Social Sciences”, series no. 07-044. Newbury Park, California: Sage Publications. The cluster analysis “green book”.

Bautista A, (2015). “El derecho a la intimidad y su disponibilidad publica”, recuperado de http://catalogoenlinea.bibliotecanacional.gov.co/client/es_ES/search/asset/107191/0

Berndt, D., and Clifford, J. 1996. "Finding Patterns in Time Series: A Dynamic Programming Approach. In Advances in Knowledge Discovery and Data Mining, eds". U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 229–248. Menlo Park, Calif.: AAAI Press.

Bin Mohamad, I & Usman, D (2013) "Standarization and Its Effects on K-means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 16(7), 3033-3299

Blazquez Ochando, M. 2010. [Tesis Doctoral] "Aplicaciones de la sindicación para la gestión de catálogos bibliográficos". Universidad Complutense de Madrid.

Cámara, D. (1994) "Cuándo y cómo utilizar el marketing de relaciones", en Harvard Deusto, Marketing & Ventas., Mayo-Junio, Págs. 12-13

Chopra, Deepak M.D. (1991). "La curación cuántica". Plaza &Janes Editores S.A ISBN: 84-01-45115-9 Enric Granados ,86-88.08008 Barcelona

Comrey, A. L. (1973) "A first course in factor analysis". Nueva York: Academic Press.

Everitt BS, Landau S, Leese M, Stahl D (2011). "Cluster Analysis", 5th ed. Wiley Series.

Garcia S, Alemán F (2011) "La Gerencia de portafolio de Tarjetas de Crédito en Colombia", *Revista Internacional de Administración & Finanzas* Vol 4 p. 103.120

Lilliefors, H. (June 1967), "On the Kolmogorov–Smirnov test for normality with mean and variance unknown", *Journal of the American Statistical Association*, Vol. 62. pp. 399–402.

Lorr, M (1983). "Cluster Analysis for Social Scientists". Jossey-Bass Social and Behavioral Science Series

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). "CRISP-DM 1.0. Step-by-step data mining guide".

Provost F y Fawcet T (2013) "Data Science for Business, What you need to know about data mining and Data analytic Thinking" O'Reilly Media, Inc. 2013 p. 50 ISBN 1449361323 9781449361327

Riquelme Santos, J.C., Ruíz, R. y Gilbert, K. (2006). "Minería de Datos: Conceptos y Tendencias". Inteligencia Artificial: *Revista Iberoamericana de Inteligencia Artificial*, 10 (29), 11-18.

Roagna, I. (2012) "Protecting the right to respect for private and family life under the European Convention on Human Rights", Council of Europe Human Rights handbook, Strasbourg, 2012.

Stone, J. V (2013). "Bayes' Rule: A Tutorial Introduction to Bayesian Analysis Author" ISBN 978-0-9563728-4-0

Tsai C. (2007). "A Credit Card usage behaviour Analysis framework- a Data Mining approach". Proceedings of the Second International Conference on e-Business, pages 219-226 DOI: 10.5220/0002108102190226

W. Fan, S.J. Stolfo, J. Zhang, P.K. Chan (1999) "AdaCost: misclassification cost-sensitiveboosting" , Proceedings of the. Sixteenth International Conference onMachine Learning (ICML'99), 1999, pp. 97–105

Zicari, A (2008) "Finanzas personales y ciclo de vida: Un desafío actual", *Invenio*, Vol 11 No.020 p 63-71

## RECONOCIMIENTOS

Agradezco a la Universidad Militar Nueva Granada por su apoyo al desarrollo de la investigación en Colombia.

## BIOGRAFIA

Docente asociado Universidad Militar Nueva Granada, Facultad de Estudios a Distancia. Experiencia en periodismo, Mercadeo con base de datos, Minería de Datos, Investigación de Mercados, Docencia Universitaria. Magister en Administración de Empresas de Southern New Hampshire University.