

**UNIVERSIDAD NACIONAL AGRARIA LA MOLINA**

**FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**

**DEPARTAMENTO ACADÉMICO DE ESTADÍSTICA E INFORMÁTICA**



**TRABAJO MONOGRÁFICO**

**“IDENTIFICACIÓN DE PERFILES DE CLIENTES CREDITICIOS  
APLICANDO TÉCNICAS DE SEGMENTACIÓN Y REGRESIÓN  
LOGÍSTICA MULTINOMIAL”**

**PRESENTADO PARA OPTAR EL TÍTULO DE INGENIERO ESTADÍSTICO E  
INFORMÁTICO**

**MAGALLY LOIDIT RAMÍREZ SOPLIN**

**Modalidad de Examen Profesional**

**LIMA - PERÚ**

**2014**

*A Dios.*

*A mis padres, Loidit y Winqui.*

*Gracias por todo su cariño y apoyo.*

# ÍNDICE

## RESUMEN

I.	INTRODUCCIÓN .....	1
1.1	Fundamentación del problema de Investigación .....	2
1.2	Formulación del problema.....	2
1.3	Objetivos de la Investigación .....	2
1.4	Justificación de la investigación.....	3
II.	REVISIÓN BIBLIOGRÁFICA.....	3
2.1	Segmentación K-means .....	3
2.1.1	Algoritmo de Segmentación K-means.....	4
2.1.2	Supuestos .....	4
2.1.3	Tratamiento de variables.....	5
2.2	Segmentación Bietápico .....	7
2.2.1	Algoritmo de Segmentación Bietápico .....	8
2.2.2	Supuestos .....	8
2.2.3	Medida de cohesión y silueta.....	9
2.3	Segmentación Kohonen.....	9
2.3.1	Algoritmo de Kohonen .....	11
2.4	Regresión Logística Multinomial.....	12
2.4.1	Calidad del Ajuste.....	12
III.	MATERIALES Y MÉTODOS .....	14
3.1	Procedimiento de análisis de datos.....	14
3.2	Formulación de la hipótesis.....	14

3.3 Definición operacional de variables .....	14
3.4 Diseño de la Investigación .....	16
3.5 Población y muestra .....	16
IV. RESULTADOS Y DISCUSIÓN .....	17
4.1 Aplicación de Análisis Segmentación .....	17
4.1.1 Aplicación de Segmentación K - means .....	18
4.1.2 Aplicación de Segmentación Bietápica .....	19
4.1.3 Aplicación de Segmentación Kohonen.....	21
4.2 Análisis de Regresión Multinomial.....	27
V. CONCLUSIONES Y RECOMENDACIONES .....	30
VI. BIBLIOGRAFÍA .....	32
VII. ANEXOS .....	33

## **RESUMEN**

El presente estudio de investigación se centró en identificar los perfiles más adecuados, en una muestra de 8, 504 clientes que realizaron transacciones crediticias en el primer trimestre del año. Se agruparon los casos mediante las técnicas de segmentación: K-means, Bietápico y Kohonen, utilizando variables cuantitativas y categóricas. De las tres técnicas, la que obtuvo mayor medida de silueta de cohesión y separación, fue K-means, indicando una estructura “buena” en cuanto a la cohesión al interior de los grupos y la separación de los mismos. Por otro lado, también se analizó las proporciones de los conglomerados, siendo la técnica K-means la que presentó las proporciones más adecuadas en función a las variables de historial crediticio y transacciones realizadas. Posterior a la obtención de los conglomerados, se procedió al proceso de obtención de la reglas de clasificación, mediante la técnica de regresión logística multinomial, la cual nos permitirá realizar predicciones futuras. El procedimiento se aplicó a la muestra particionada, es decir, una parte de entrenamiento y otra de comprobación. Finalmente, se obtuvo una adecuada tasa de eficiencia en ambas muestras.

Además, los análisis permitieron identificar a dos conglomerados que muestran una alerta para la empresa, es decir necesitan ser gestionados de forma oportuna, ya que constituyen un futuro comportamiento de no pago de acuerdo a la caracterización obtenida de dichos conglomerados.

## I. INTRODUCCIÓN

En los últimos años el uso de las tarjetas de crédito ha crecido enormemente, así lo dio a conocer La Superintendencia de Banca, Seguros y AFP, dando cuenta el aumento de 6.7 millones en el 2010 a ocho millones en el 2013.

En este contexto, sabemos que cada día se realiza la emisión de tarjetas de crédito por parte de las instituciones bancarias y no bancarias, cuya finalidad es captar nuevos clientes y venderles el servicio del crédito a través de las tarjetas, donde el precio de éste está compuesto por intereses, comisiones y mantención de tarjeta, los cuales están destinados a cubrir los costos y el riesgo crediticio inherente de las tarjetas.

Actualmente, la presente entidad emisora de tarjetas de crédito trabaja principalmente con tres tipos de tarjetas; “Golden”, “Platinum” y “Clásica”, dirigidas a los niveles socioeconómicos A y B. Las transacciones que representan dichas tarjetas son monitoreadas a través de un sistema de gestiones tempranas, este sistema cuenta con ciertos perfiles elaborados empíricamente, es decir por “expertos en el negocio”, sin embargo dichos perfiles no tienen sustento y validez estadística en el procedimiento. Con dichos perfiles empíricos se identifica a los clientes que necesitan ser gestionados de forma anticipada, es decir identifican ciertas características de los clientes que pueden incurrir en un futuro comportamiento de no pago. Ante tal situación, cabría hacerse la siguiente pregunta ¿La entidad emisora está gestionando adecuadamente a sus clientes de acuerdo a los perfiles empíricos? Para responder la interrogante y otras más se propone el uso de las Técnicas de Segmentación: K-means, Bietápico y Kohonen. Posterior a ello con el objetivo de obtener las reglas de clasificación para predicciones futuras, se propone utilizar la técnica Regresión Logística Multinomial.

En esta perspectiva, este estudio tiene como finalidad identificar aquellos perfiles que nos permitan realizar “gestiones tempranas” eficientes y oportunas. Para ello, se trabajará con los clientes activos que realizaron transacciones el primer trimestre del año 2014. Se utilizará variables que se derivan de las transacciones de los clientes y otras que provienen de su historial crediticio. El software a usar será SPSS Clementine 12.

## **1.1 Fundamentación del problema de Investigación**

La gran actividad financiera de hoy en día, propone el constante análisis del comportamiento de clientes. Por un lado, tenemos el criterio y juicio de analistas expertos, quienes utilizan sus conocimientos y experiencias acumuladas a través de años de trabajo para evaluar a los clientes y tomar decisiones al respecto. Sin embargo la falta de validez metodológica y estadística en la construcción de modelos u otros tipos de análisis, es un gran inconveniente en cuanto a la toma de decisiones. En éste contexto sabemos que la entidad emisora de tarjetas cuenta con perfiles construidos con el conocimiento de expertos, dichos perfiles tienen como objetivo identificar clientes que deben ser gestionados anticipadamente, con la finalidad de adelantarse a los eventos de no pago de los mismos.

Desde esta perspectiva se hace imprescindible realizar un nuevo análisis con enfoque estadístico, usando las variables adecuadas y otorgando validez al mismo.

## **1.2 Formulación del problema**

El problema de investigación se formuló a través de las siguientes preguntas:

1. ¿Qué perfiles identifican a los clientes crediticios que realizaron transacciones en el primer trimestre del año? ¿Hay algún perfil, que identifique clientes que deben ser gestionados anticipadamente?

## **1.3 Objetivos de la Investigación**

- Determinar e identificar conglomerados que indiquen la gestión anticipada de los clientes, antes que incurran en un evento de no pago.
- Obtener las reglas de clasificación mediante la técnica de Regresión Logística Multinomial.

## **1.4 Justificación de la investigación**

Para la división de riesgos de la entidad emisora de tarjetas de créditos, es de suma importancia anticiparse a los eventos de no pago de sus clientes, por ello contar con perfiles que identifiquen a sus clientes de forma oportuna, le permitirá realizar una gestión adecuada, cuyo beneficio principal será evitar un deterioro de su portafolio de créditos que vea perjudicada su rentabilidad, evitando el aumento de la proporción de créditos no pagados.

Desde esta perspectiva, ésta investigación se centrará en encontrar los perfiles más adecuados, haciendo uso de tres técnicas de segmentación: Cluster K-means, Bietápico y Kohonen, de tal manera que se pueda comparar y tomar la mejor alternativa de solución.

Al cabo de encontrar los mejores conglomerados de clientes que sean identificables, utilizaremos la técnica de Regresión Logística Multinomial para obtener las reglas de clasificación.

## **II. REVISIÓN BIBLIOGRÁFICA**

### **2.1 Segmentación K-means**

La técnica K-means busca identificar los grupos mediante particiones del conjunto de datos en k grupos definidos al inicio por el investigador.

Esta metodología está basada en la idea de que cada cluster está representado por el centro de cada cluster (MacQueen 1967) o “centroide”, que es la media o mediana de un grupo de puntos. Asimismo esta técnica permite a las observaciones moverse de un cluster a otro.

A diferencia de la segmentación jerárquica, donde se requiere del cálculo de la matriz de similitud para cada par de casos que resulta lento cuando se trabaja con gran cantidad de datos, la segmentación por K-means no requiere el cálculo de todas las posibles distancias.



Esta metodología parte de un determinado número de cluster fijados y, mediante un algoritmo, repetidamente cada caso se asigna a un cluster y mientras el algoritmo avanza, el mismo caso puede moverse de cluster en cluster. Lo cual también es una ventaja en comparación con la segmentación jerárquica, donde la observación una vez añadida a un cluster se queda ahí (Norusis 2011) [8].

### **2.1.1 Algoritmo de Segmentación K-means**

El algoritmo de las K-medias (presentado por MacQueen en 1967) es uno de los algoritmos de aprendizaje no supervisado más simples para resolver el problema de la clusterización. El procedimiento aproxima por etapas sucesivas un cierto número (prefijado) de clusters haciendo uso de los centroides de los puntos que deben representar.

El algoritmo se compone de los siguientes pasos:

- Sitúa K puntos en el espacio en el que "viven" los objetos que se quieren clasificar.
- Estos puntos representan los centroides iniciales de los grupos.
- Asigna cada objeto al grupo que tiene el centroide más cercano.
- Tras haber asignado todos los objetos, recalcula las posiciones de los K centroides.
- Repite los pasos 2 y 3 hasta que los centroides se mantengan estables. Esto produce una clasificación de los objetos en grupos que permite dar una métrica entre ellos.

### **2.1.2 Supuestos**

1. K-means utiliza la distancia Euclidiana para asignar los objetos a los clusters. Se esperaría que los datos tengan una escala similar para el uso de distancias.
2. Para buscar la partición óptima se busca reducir la varianza dentro de los cluster, esto se puede lograr minimizando la suma de errores al cuadrado (SEE). La debilidad de este supuesto es que cada grupo tendría aproximadamente el mismo SSE, lo que significa que la matriz de variancia/covariancia entre objetos del mismo grupo será igual.

### 2.1.3 Tratamiento de variables

#### Análisis de correlación

Si dos variables están muy correlacionadas es recomendable que se tome solo una de ellas, para no caer en redundancia y alterar las distancias del algoritmo.

#### Estandarización de variables categóricas en Clementine

Si las variables son medidas en escalas distintas, las variables con grandes valores contribuyen en mayor medida a los cálculos de distancias que las variables con valores pequeños. (Norusis 2011) [8]

Yaghini (2010) [9] menciona que para compensar el efecto de la escala, los campos de rango son transformados, de manera que todos tienen la misma escala.

En Clementine, los campos de rango se reajustarán a tener valores entre 0 y 1. La transformación utilizada es:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Clementine recodifica un campo simbólico como un grupo de campos numéricos con un campo numérico para cada categoría o valor del campo original. Para cada registro, el valor del campo correspondiente a la categoría del registro se establece en 1, y todos los otros valores de campo derivados se establecen a 0. Tales campos derivados son a veces llamados campos de indicadores.

Por ejemplo, considere los siguientes datos, en donde  $x$  es un campo simbólico con posibles valores A, B, y C:

#	$X$	$x'_1$	$x'_2$	$x'_3$
1	B	0	1	0
2	A	1	0	0
3	C	0	0	1

Para algoritmos que utilizan la distancia euclídea para medir las diferencias entre los registros, la diferencia entre dos registros con diferentes valores de  $i$  y  $j$  para el conjunto es:

$$\sqrt{\sum_{k=1}^J (x_{k1} - x_{k2})^2}$$

Donde  $J$  es el número de categorías, y  $x_{kn}$  es el valor del indicador derivado de la categoría  $k$  para el registro  $n$ .

Los valores serán diferentes en dos de los derivados indicadores,  $x_i$  y  $x_j$ . La suma será:

$$\sqrt{(1-0)^2 + (0-1)^2} = \sqrt{2} \approx 1.414$$

El cual es mayor que 1.

Eso quiere decir que en base a esta codificación, si los nuevos campos derivados se codifican con un valor de 1, tenderán a dominar la solución del cluster comparando con los campos numéricos. Para corregir este sesgo, Clementine aplica una escala factor a los campos de conjunto de derivados, tales que una diferencia de valores en un campo de juego produce una distancia euclidiana de 1.

El factor de escala por defecto es:

$$\sqrt{\frac{1}{2}} \approx 0.707$$

Este valor da el resultado deseado:

$$\sqrt{\left(\sqrt{\frac{1}{2}} - 0\right)^2 + \left(\sqrt{\frac{1}{2}} - 0\right)^2} = \sqrt{\frac{1}{2} + \frac{1}{2}} = 1$$

En consecuencia, el valor 0.707 es usado en lugar de la raíz cuadrada de 1.

Principales pasos del algoritmo

1. Seleccione centros iniciales de los conglomerados
2. Asignar cada registro a la agrupación más cercana
3. Actualización de los centros de los conglomerados en base a los registros asignado a cada clúster
4. Repita los pasos 2 y 3 hasta que:

En el paso 3, no hay ningún cambio en los centros de los conglomerados de la iteración anterior, o el número de iteraciones excede las iteraciones máximas parámetro.

## **2.2 Segmentación Bietápico**

Es una técnica de segmentación no jerárquica exploratoria la cual permite revelar grupos naturales de un conjunto de datos. Fue propuesta para resolver algunos problemas que enfrentaban las otras técnicas de Segmentación, en particular el que pueda trabajar con variables mixtas (continuas y categóricas) y el que pueda determinar el número de clusters automáticamente. (Bacher et. al. 2004) [1].

### **2.2.1 Algoritmo de Segmentación Bietápico**

Norusis (2011) [8] describe los pasos de la siguiente forma:

Paso 1: Consiste en la formación de los preclusters. Los preclusters son tan solo los clusters de los casos originales que son usados en lugar de la data cruda en el clusters jerárquico. El objetivo de este paso es reducir el tamaño de la matriz de distancias y el algoritmo que decide, basado en la medida de distancia, si el caso actual debe unirse con algún precluster para formar un nuevo precluster. Luego de completado este paso, el tamaño de la matriz de distancias ya no es dependiente del número de casos sino en el número de precluster.

Paso 2: Consiste en el agrupamiento de los precluster mediante la segmentación jerárquica. Formando clusters mediante la metodología jerárquica permite explorar una variedad de soluciones con un diferente número de clusters.

### **2.2.2 Supuestos**

El algoritmo de segmentación Bietápica está basado en la medida de distancia que da los mejores resultados si todas las variables son independientes, variables continuas que tienen una distribución normal y variables categóricas con distribución multinomial.

Sin embargo, el algoritmo está pensado para trabajar razonablemente bien cuando las suposiciones no se cumplan. Esto debido a que las técnicas de segmentación no involucran una prueba de hipótesis y cálculo de niveles de significancia, solo están presentes para hacer seguimiento descriptivo. Es perfectamente aceptable segmentar datos cuando no se cumplan estos supuestos. Solamente el investigador determinará si los resultados son satisfactorios para sus necesidades (Norusis 2011) [8].

### **2.2.3 Medida de cohesión y silueta**

Esta medida se utiliza como referencia para indicar si los resultados de los conglomerados son pobres, correctos o buenos (Kauffman y Rousseeuw 1990). Esta medida permite comprobar la calidad del conglomerado.

Un resultado “bueno” indica que los datos reflejan una evidencia razonable o sólida de que existe una estructura de conglomerados, de acuerdo con la valoración Kaufman y Rousseeuw. Un resultado “correcto” indica que esta evidencia es débil, y un resultado “pobre” significa que, según esa valoración no hay evidencias.

Las medias de medida de silueta, en todos los registros está dada por  $(B-A)/\max(A,B)$ , donde A es la distancia del registro al centro de su conglomerado y B es la distancia del registro al centro del conglomerado más cercano al que no pertenece.

Un coeficiente de silueta de 1 podría implicar que todos los casos están ubicados directamente en los centros de sus conglomerados. Un valor de -1 significaría que todos los casos se encuentran en los centros de los conglomerados de otro conglomerado. Un valor de 0 implica, de media, que los casos están equidistantes entre el centro de su propio conglomerado y el siguiente conglomerado más cercano. (SPSS 2011).

### **2.3 Segmentación Kohonen**

En 1982 T. Kohonen presentó un modelo de red denominado mapas auto-organizados o SOM (Self-Organizing Maps), basado en ciertas evidencias descubiertas a nivel cerebral. Este tipo de red posee un aprendizaje no supervisado competitivo.

No existe ningún maestro externo que indique si la red neuronal está operando correcta o incorrectamente porque no se dispone de ninguna salida objetivo hacia la cual la red neuronal deba tender.

La red auto-organizada debe descubrir rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones. Se dice, por tanto, que las neuronas deben auto-organizarse en función de los estímulos (datos) procedentes del exterior.

En el aprendizaje competitivo las neuronas compiten unas con otras con el fin de llevar a cabo una tarea dada. Se pretende que cuando se presente a la red un patrón de entrada, sólo una de las neuronas de salida (o un grupo de vecinas) se active. Por tanto, las neuronas compiten por activarse, quedando finalmente una como neurona vencedora y otra como anulada del resto, que son forzadas a sus valores de respuesta mínimos.

El objetivo de este aprendizaje es categorizar los datos que se introducen en la red. Se clasifican valores similares en la misma categoría y, por tanto, deben activar la misma neurona de salida.

Las clases o categorías deben ser creadas por la propia red, puesto que se trata de un aprendizaje no supervisado, a través de las correlaciones entre los datos de entrada.

Kohonen afirma: *I just wanted an algorithm that would effectively map similar patterns (pattern vectors close to each other in the input signal space) onto contiguous locations in the output space.* (Kohonen, 1995, p. VI.)

### 2.3.1 Algoritmo de Kohonen

El proceso de aprendizaje del SOM es el siguiente (Kaski, Kangas y Kohonen 1998) [7]:

Paso 1. Un vector  $x$  es seleccionado al azar del conjunto de datos y se calcula su distancia (Similitud) a los vectores del codebook, usando, por ejemplo, la distancia euclídea:

$$\|x - m_c\| = \min_j \{\|x - m_j\|\}$$

Paso 2. Una vez que se ha encontrado el vector más próximo o BMU (best matching unit) el resto de vectores del codebook es actualizado. El BMU y sus vecinos (en sentido topológico) se mueven cerca del vector  $x$  en el espacio de datos. La magnitud de dicha atracción está regida por la tasa de aprendizaje.

Mientras se va produciendo el proceso de actualización y nuevos vectores se asignan al mapa, la tasa de aprendizaje decrece gradualmente hacia cero. Junto con ella también decrece el radio de vecindad también.

La regla de actualización para el vector de referencia dado  $i$  es la siguiente:

$$m_j(t+1) = \begin{cases} m_j(t) + \alpha(t)(x(t) - m_j(t)) & j \in N_c(t) \\ m_j(t) & j \notin N_c(t) \end{cases}$$

Los pasos 1 y 2 se van repitiendo hasta que el entrenamiento termina. El número de pasos de entrenamiento se debe fijar antes a priori, para calcular la tasa de convergencia de la función de vecindad y de la tasa de aprendizaje.



Una vez terminado el entrenamiento, el mapa ha de ordenarse en sentido topológico:  $n$  vectores topológicamente próximos se aplican en  $n$  neuronas adyacentes o incluso en la misma neurona.

## 2.4 Regresión Logística Multinomial

La regresión logística multinomial (Hosmer y Lemeshow, 1989) [5], es utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías (politómica) y es una extensión mul-tivariante de la regresión logística binaria clásica. Las variables independientes pueden ser tanto continuas (regresores) como categóricas (factores).

Tradicionalmente las variables dependientes politómicas han sido modeladas mediante análisis discriminante pero, gracias al creciente desarrollo de las técnicas de cálculo, cada vez es más habitual el uso de modelos de regresión logística multinomial, ya implementados en paquetes estadísticos.

### 2.4.1 Calidad del Ajuste

Al igual que en la regresión logística binaria, la calidad del ajuste en la regresión logística multinomial se mide mediante coeficientes de determinación conocidos como Pseudo- $R^2$ . De entre todos ellos tenemos a los más clásicos, que son los que proporciona el paquete estadístico S.P.S.S.

#### Pseudo – $R^2$ de Mc Fadden

Se basa en la función auxiliar  $\Lambda$  utilizada en el ajuste y viene dado por:

$R^2_{MF} = 1 - \frac{\Lambda_f}{\Lambda_0}$ ; su rango teorico de valores es  $0 \leq R^2_{MF} \leq 1$ , pero muy rara vez su valor se

aproxima a 1. Suele considerarse una buena calidad de ajuste cuando  $0.2 \leq R^2_{MF} \leq 0.4$ , y excelente para valores superiores.

## Pseudo – R<sup>2</sup> de Cox y Snell

Se basa directamente en la verosimilitud  $L$ , definido como:

$$R^2_{cs} = 1 - \frac{(\sqrt{L_0})^2}{(\sqrt{L_f})^2} = 1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{n}\right), \text{ siendo } L_0 = \exp(-\Lambda_0/2) \text{ y } L_f = \exp(-\Lambda_f/2). \text{ El}$$

rango teórico para este coeficiente es  $0 \leq R^2_{cs} \leq 1 - (\sqrt{L_0})^2$ , lo que le hace poco interpretableal depender de  $L_0$ . Por este motivo es preferible el seudo  $R^2$  de Nagelkerke

## Pseudo – R<sup>2</sup> de Nagelkerke

Se define como:  $R^2_{N=} = \frac{R^2_{cs}}{1 - (\sqrt{L_0})^2} = \frac{1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{n}\right)}{1 - \exp\left(\frac{-\Lambda_0}{n}\right)}$  y su rango de valores es  $0 \leq R^2_N \leq 1$ ,

por lo que puede interpretarse del mismo modo que el coeficiente de determinación de la regresión lineal clásica, aunque es más difícil que alcance valores próximos a 1.

## Calidad en la predicción

Si, a partir del modelo ajustado, clasificamos cada observación en la categoría más probable, podemos construir una matriz de clasificación observados - pronosticados y utilizar el porcentaje de clasificaciones correctas como una medida de la calidad de predicción, del mismo modo que se hace en el análisis discriminante.

### **III. MATERIALES Y MÉTODOS**

#### **3.1 Procedimiento de análisis de datos**

El procesamiento de los datos se realizó mediante la aplicación de tres técnicas de segmentación: K-means, Bietápico y Kohonen. Dado que no tenemos un antecedente del número de conglomerados, dicho número estará en el rango de 2 conglomerados como mínimo y 15 conglomerados como máximo.

Se comparó los resultados de las tres técnicas, una de las comparaciones se basó en los valores de medida de silueta y cohesión. Y por otro lado, también se comparó la proporción de los conglomerados obtenidos. Finalmente, para obtener las reglas de clasificación se aplicó la técnica de regresión logística multinomial. Previamente, se realizó una partición de la muestra, es decir, una muestra de entrenamiento que representa el 70% y otra de comprobación que representa el 30%.

Para el procesamiento de los datos se utilizó la herramienta de modelamiento SPSS Clementine 12.

#### **3.2 Formulación de la hipótesis**

La hipótesis que corresponde a la investigación es la siguiente:

1. Existen perfiles diferenciados de los clientes que realizaron transacciones en el primer trimestre del año, según las variables de historial crediticio y las variables originadas de acuerdo a sus transacciones.

#### **3.3 Definición operacional de variables**

Las variables  $X_8$ ,  $X_9$ ,  $X_{10}$  no se considerarán para el análisis de segmentación por la alta correlación que presentan con otras variables que se consideran indispensables en el modelo. Se omiten estas variables con el fin de no sobredimensionar los grupos. (Anexo 1).

Finalmente, entran al modelo las variables cuantitativas  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ ,  $X_6$  y  $X_7$ , y las variables  $X_{11}$  y  $X_{12}$  de naturaleza categórica.

Las variables de partida que se tienen para este estudio son:

**Cuadro Nro.1 Descripción de variables cuantitativas**

$X_i$	Variables	Descripción
$X_1$	CONSUMO_SOLES	Consumo acumulado en el primer trimestre
$X_2$	SCORE	Score
$X_3$	CAPITAL_ADEUDADO	Capital adeudado en el primer trimestre
$X_4$	DEUDA_SF	Deuda total en el Sistema Financiero
$X_5$	ANTIGÜEDAD	Antigüedad de la cuenta en meses
$X_6$	LINEA_DISPONIBLE	Línea de crédito disponible
$X_7$	SF_I	Deuda en el Sistema Financiero/ingreso
$X_8$	CC_I	Cuota comprometida/ingreso
$X_9$	LD_LC	Línea disponible / Línea de crédito
$X_{10}$	CA_LC	Capital adeudado/ Línea de crédito

**Cuadro Nro.2 Descripción de variables cualitativas**

$X_i$	Variables	Descripción	Valores
$X_{11}$	CLASIF_MAX6	Clasificación crediticia por la SBS en los últimos 6 meses	0 "Normal"
			1 "Cliente con problemas potenciales"
			2 "Deficiente"
			3 "Dudoso"
			4 "Pérdida"
$X_{12}$	Tipo de Tarjeta	Tipo de Tarjeta	1 "Golden"
			2 "Platinum"
			3 "Clásica"

### **Definición de Clasificación crediticia por la SBS en los últimos 6 meses:**

Categoría normal (0): Son aquellos deudores que vienen cumpliendo con el pago de sus créditos de acuerdo a lo convenido o con un atraso de hasta ocho 8 días calendario.

Categoría con problemas potenciales (1): Son aquellos deudores que registran atraso en el pago de sus créditos de nueve a treinta días calendario.

Categoría deficiente (2): Son aquellos deudores que registran atraso en el pago de sus créditos de treinta y uno a sesenta días calendario.

Categoría dudoso (3): Son aquellos deudores que registran atraso en el pago de sus créditos de sesenta y uno a ciento veinte días calendario.

Categoría pérdida (4): Son aquellos deudores que muestran atraso en el pago de sus créditos de más de ciento veinte días calendario.

### **3.4 Diseño de la Investigación**

El diseño de la investigación es no experimental, del tipo transversal y descriptivo, debido a que tomamos los datos del primer trimestre del presente año como un solo momento, y a la vez la información de los clientes crediticios, con el objeto de describir los grupos que se obtienen en dicha muestra de acuerdo a las variables más influyentes.

### **3.5 Población y muestra**

La población está definida por los clientes activos, que realizan transacciones con alguna de las tres tarjetas principales de la empresa en estudio. Dado que este trabajo está orientado a la minería de datos, es decir el descubrimiento de conocimiento útil en base a grandes volúmenes de datos, se tomará una muestra de 8, 504 clientes que efectuaron transacciones en el primer trimestre del presente año.

## IV. RESULTADOS Y DISCUSIÓN

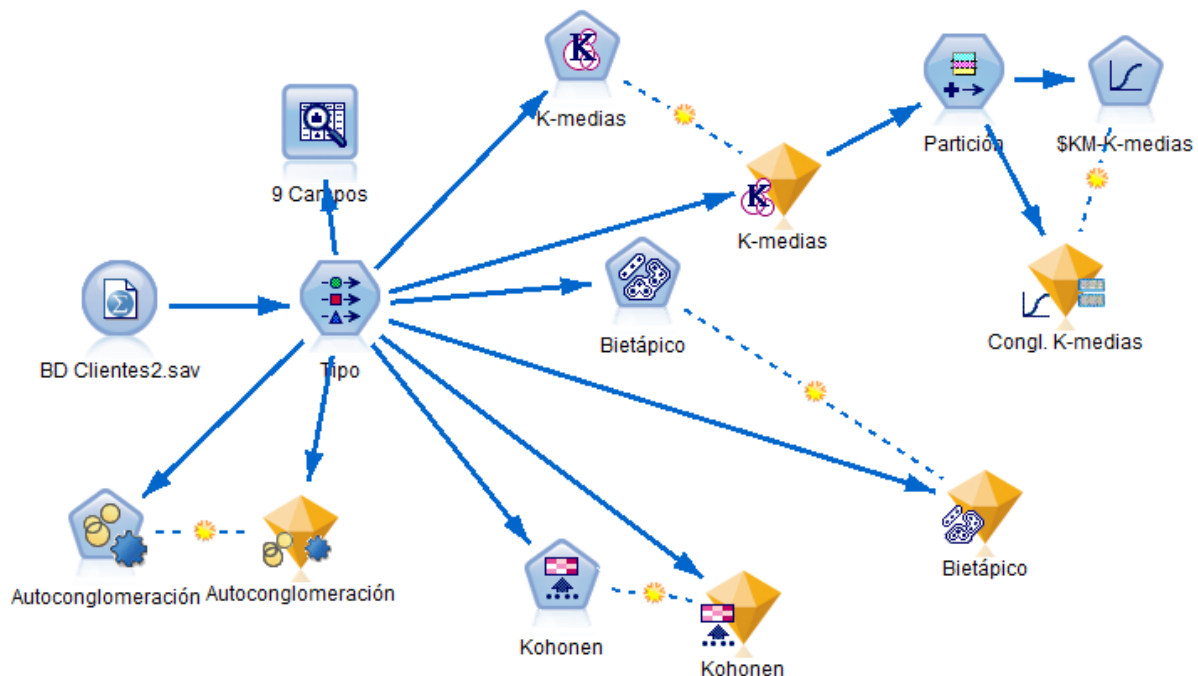
### 4.1 Aplicación de Análisis Segmentación

Para seleccionar el mejor conjunto de conglomerados, que permita diferenciar los perfiles de clientes que realizaron transacciones en el primer trimestre del año, se aplicaron tres tipos de conglomeración Bietápico, K- medias y Kohonen.

Las siguientes variables entraron al proceso de segmentación: CONSUMO\_SOLES, SCORE, CAPITAL\_ADEUDADO, DEUDA\_SF, ANTIGÜEDAD, LINEA\_DISPONIBLE, SF\_I, CLASIF\_RCC y TIPO\_TARJETA.

En el proceso de iteración y generación de la mejor alternativa de agrupación de los clientes, se procesaron los siguientes modelos en el software SPSS Clementine 12.

**Gráfico Nro. 1: Modelamiento de las tres técnicas de Conglomerados en Clementine**



A continuación detallaremos el modelamiento de cada conglomerado.

#### 4.1.1 Aplicación de Segmentación K - means

Según la metodología descrita por Yaghini (2010) [9], para evitar el problema del uso de campos con rangos y desviaciones típicas muy diferentes, Clementine recodifica las variables categóricas antes de que comience la agrupación, obteniéndose nuevos valores de entrada para el cálculo de las distancias. Los datos categóricos se transforman en un equivalente numérico. Por ejemplo, para la variable “Tipo de Tarjeta”, que tiene tres valores, tendrá tres nuevas entradas transformadas:

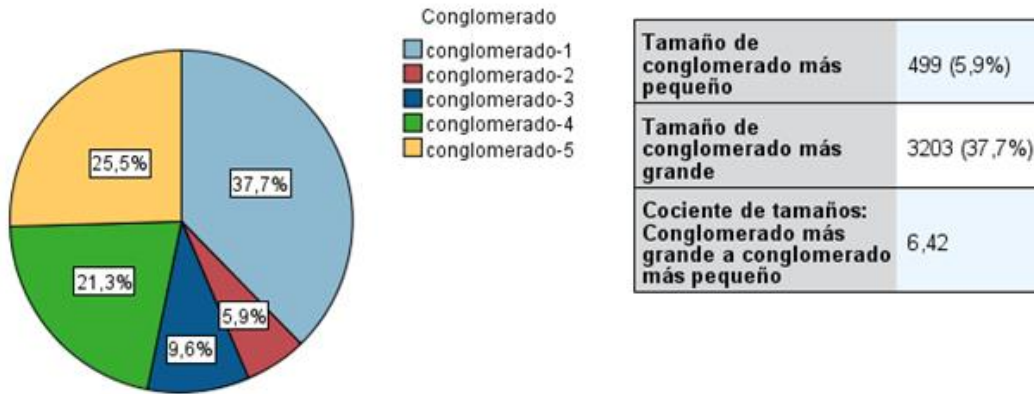
- Tarjeta Golden (1 0 0)
- Tarjeta Platinum (0 1 0)
- Tarjeta Clásica (0 0 1)

La misma metodología se aplica para la variable categórica “Clasificación crediticia en la SBS, en los últimos seis meses”. Entonces, con la transformación de las variables categóricas y uso de las variables numéricas se aplicó la técnica K-means.

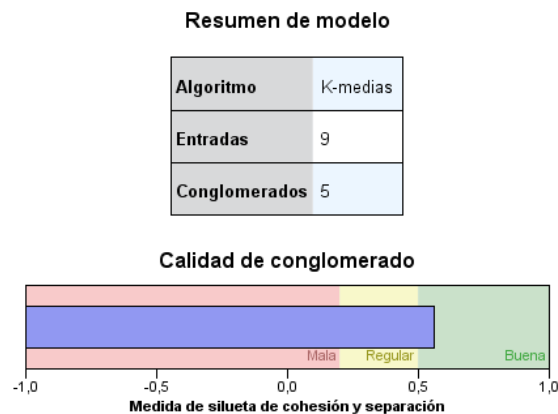
Con la aplicación de ésta técnica se obtuvieron 5 conglomerados, del gráfico Nro. 3, se puede observar que la medida de cohesión y silueta tiene un valor de 0.559, lo cual evidencia un resultado “bueno”, es decir indica que los conglomerados al interior tienen características similares y que se diferencian entre ellos. También se muestra el cociente del tamaño del conglomerado más grande al más pequeño de 6.42.

**Gráfico Nro. 2: Modelo K –Means**

**Tamaños de conglomerados**



**Gráfico Nro. 3: Medida de silueta de K- means**



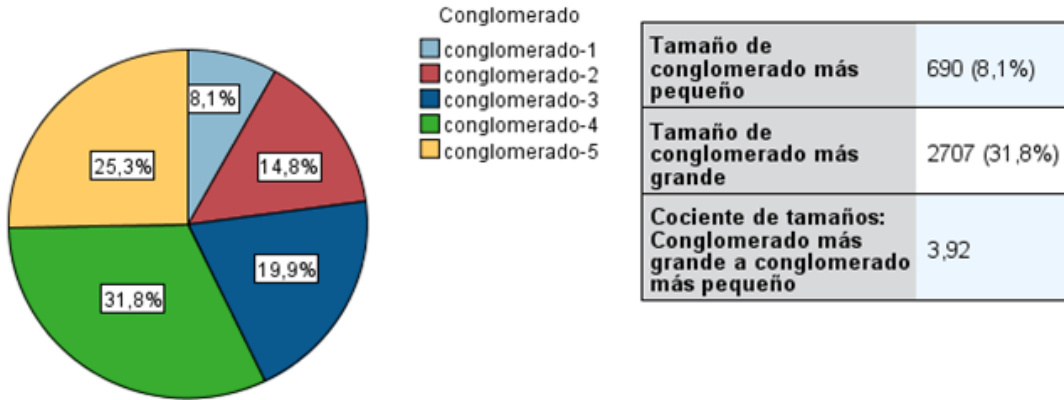
#### 4.1.2 Aplicación de Segmentación Bietápica

Con la aplicación de ésta técnica de igual manera que con la técnica de k-means se obtuvieron 5 conglomerados, sin embargo la distribución de los conglomerados refleja una distribución más equitativa. Por otro lado, según el gráfico Nro. 5, podemos observar un menor valor de silueta y cohesión, pero este valor de 0.518 también evidencia un resultado “bueno”, indicando que los conglomerados al interior tienen características similares y que se diferencian entre ellos.



### Gráfico Nro. 4: Modelo Bietápico

Tamaños de conglomerados

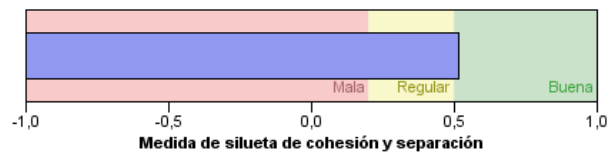


### Gráfico Nro. 5: Medida de silueta de Bietápico

Resumen de modelo

Algoritmo	Bietápico
Entradas	9
Conglomerados	5

Calidad de conglomerado

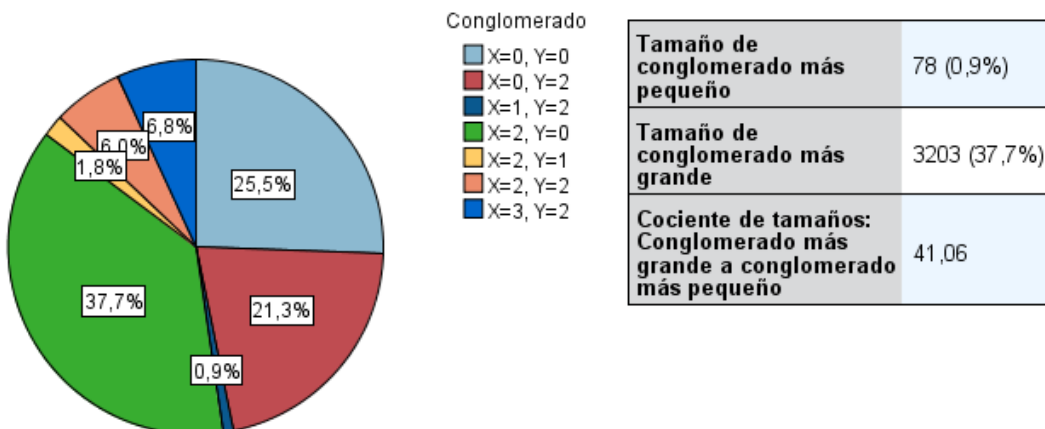


### 4.1.3 Aplicación de Segmentación Kohonen

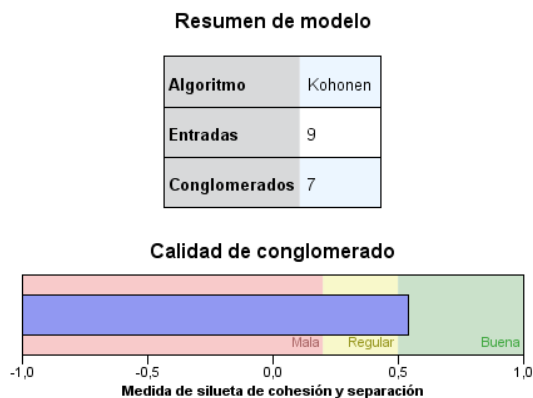
La aplicación de esta técnica identifica 7 conglomerados, sin embargo el tamaño de cada conglomerado se encuentra más diversificado, siendo el cociente del conglomerado más grande al conglomerado más pequeño de 41,06. Por otro lado, según el gráfico Nro. 7, podemos observar un valor adecuado de medida de silueta y cohesión, con este valor de 0.557 también se evidencia un resultado “bueno”, indicando que los conglomerados al interior tienen características similares y que se diferencian entre ellos.

**Gráfico Nro. 6: Modelo Kohonen**

**Tamaños de conglomerados**



**Gráfico Nro. 7: Medida de silueta de Kohonen**




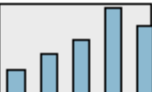



## Comparación de modelos

Al realizar la comparación de los tres modelos obtenidos, se observa que; los conglomerados obtenidos con las técnicas de K-medias y Kohonen tienen mayores valores de silueta y cohesión, 0.559 y 0.557, respectivamente. Y con la técnica Bietápica se observa el menor valor. Por tanto, de acuerdo a la silueta podemos decidir entre K-medias y Kohonen. Sin embargo al observar las proporciones de los conglomerados, se observa que con la técnica de Kohonen, estos se encuentran distribuidos entre muy bajas (1.8%, 6.0%, 6.8%, 0.9%) y altas proporciones (37.7%, 25.5%, 21.3%), a diferencia de lo obtenido con K-means, donde tenemos proporciones distribuidas en 25.5%, 21.3%, 37.7%, 9.6% y 5.9%.

Por tanto, concluimos que la mejor alternativa de conglomerados se da por los resultados obtenidos con la técnica K-means.

**Gráfico Nro. 8: Comparativo Medida de Silueta y Cohesión**

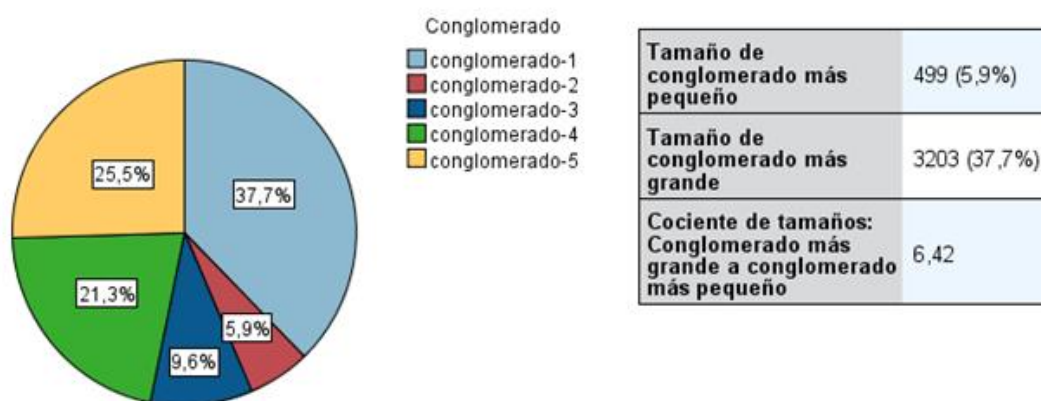
¿Uso?	Gráfico	Modelo	Tiempo de generación	Silueta	Número de conglomerados
<input checked="" type="checkbox"/>		 K-medias 1	< 1	0,559	5
<input type="checkbox"/>		 Kohonen 1	< 1	0,557	7
<input type="checkbox"/>		 Bietápico 1	< 1	0,518	5

## Conglomerados Finales - Técnica K-means

Los conglomerados más adecuados se obtuvieron con la técnica de K-means. Del gráfico Nro. 9 se puede apreciar la distribución de cada conglomerado. Por otro lado, del gráfico Nro. 10 se puede observar la importancia de cada predictor incluido en el modelo del conglomerado.

**Gráfico Nro. 9: Conglomerados Finales –K-means**

**Tamaños de conglomerados**



## Gráfico Nro. 10: Importancia de cada predictor

### Conglomerados

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

Conglomerado	conglomerado-1	conglomerado-2	conglomerado-3	conglomerado-4	conglomerado-5
<b>Etiqueta</b>					
<b>Descripción</b>					
<b>Tamaño</b>	37,7% (3203)	5,9% (499)	9,6% (817)	21,3% (1815)	25,5% (2170)
<b>Entradas</b>	CLASIFICACION_RC C	CLASIFICACION_RC C	CLASIFICACION_RC C	CLASIFICACION_RC C	CLASIFICACION_RC C
	COD_GRUPO_TEMP Platinum (100,0%)	COD_GRUPO_TEMP Platinum (68,1%)	COD_GRUPO_TEMP Clásica (60,2%)	COD_GRUPO_TEMP Golden (100,0%)	COD_GRUPO_TEMP Clásica (100,0%)
	LINEA_DISPONIBLE 6.232,37	LINEA_DISPONIBLE 6.535,92	LINEA_DISPONIBLE 2.006,81	LINEA_DISPONIBLE 13.125,15	LINEA_DISPONIBLE 2.380,78
	SCORE 659,20	SCORE 495,59	SCORE 310,74	SCORE 687,07	SCORE 647,81
	CAPITAL_ADEUDAD O	CAPITAL_ADEUDAD O	CAPITAL_ADEUDAD O	CAPITAL_ADEUDAD O	CAPITAL_ADEUDAD O
	DEUDA_SF 68.704,28	DEUDA_SF 90.968,08	DEUDA_SF 50.191,28	DEUDA_SF 105.962,13	DEUDA_SF 33.160,99
	ANTIGUEDAD 44,52	ANTIGUEDAD 40,78	ANTIGUEDAD 42,86	ANTIGUEDAD 25,08	ANTIGUEDAD 39,02
	CONSUMO_SOLES 777,25	CONSUMO_SOLES 741,23	CONSUMO_SOLES 374,21	CONSUMO_SOLES 820,90	CONSUMO_SOLES 424,67
	SF_I 24,24	SF_I 23,77	SF_I 23,72	SF_I 22,71	SF_I 17,55

A continuación se detallan los perfiles encontrados mediante la segmentación K-means:

**Conglomerado 1: Clientes antiguos con medianos niveles de consumo y con buena calificación crediticia.**

Los clientes pertenecientes a este conglomerado, presentan en promedio un consumo de S/. 777.25 en el 1er trimestre del año, un score de 659.20, que significa “bajo riesgo” en el sistema financiero, un capital adeudado S/.5, 168.74, una deuda en el sistema financiero de S/.68, 704.28, además poseen una línea disponible de S/. 6, 232.37, tienen una antigüedad de 4 años, un ratio de deuda en el sistema financiero respecto al ingreso de 24.24, poseen la tarjeta “Platinum” y tienen una calificación crediticia máxima de “0” , es decir un comportamiento “normal” en el sistema financiero dentro de los 6 últimos meses.

**Conglomerado 2: Clientes antiguos con medianos niveles de consumo y con posibles problemas crediticios.**

Los clientes pertenecientes a este conglomerado, presentan en promedio un consumo de S/. 741.23 en el 1er trimestre del año, un score de 495.59, que significa “alto riesgo” en el sistema financiero, un capital adeudado S/.6, 705.86, una deuda en el sistema financiero de S/.90, 968.68, además poseen una línea disponible de S/. 6, 535.92, tienen una antigüedad de 3.5 años, un ratio de deuda en el sistema financiero respecto al ingreso de 23.77, poseen la tarjeta “Platinum” y tienen una calificación crediticia máxima de “1”, es decir un cliente con problemas potenciales de crédito, según el sistema financiero dentro de los 6 últimos meses.

**Conglomerado 3: Clientes antiguos con bajos niveles de consumo y con posibles problemas crediticios.**

Los clientes pertenecientes a este conglomerado, presentan en promedio un consumo de S/. 374.23 en el 1er trimestre del año, un score de 310.74, que significa “muy alto riesgo” en el

sistema financiero, un capital adeudado S/.3, 504.43, una deuda en el sistema financiero de S/.50, 191.28, además poseen una línea disponible de S/. 2, 006.81, tienen una antigüedad de 4 años, un ratio de deuda en el sistema financiero respecto al ingreso de 23.72, poseen la tarjeta “Clásica” y tienen una calificación crediticia máxima de “1”, es decir un cliente con problemas potenciales de crédito, según el sistema financiero dentro de los 6 últimos meses.

**Conglomerado 4: Clientes nuevos con medianos niveles de consumo y con buena calificación crediticia.**

Los clientes pertenecientes a este conglomerado, presentan en promedio un consumo de S/. 820.90 en el 1er trimestre del año, un score de 687.07, que significa “bajo riesgo” en el sistema financiero, un capital adeudado S/.7, 367,49, una deuda en el sistema financiero de S/.105, 962.13, además poseen una línea disponible de S/.13, 125.15, tienen una antigüedad de 2 años, un ratio de deuda en el sistema financiero respecto al ingreso de 22.71, poseen la tarjeta “Golden” y tienen una calificación crediticia máxima de “0” , es decir un comportamiento “normal” en el sistema financiero dentro de los 6 últimos meses.

**Conglomerado 5: Clientes nuevos con bajos niveles de consumo y con buena calificación crediticia**

Los clientes pertenecientes a este conglomerado, presentan en promedio un consumo de S/. 424.67 en el 1er trimestre del año, un score de 647.81, que significa “bajo riesgo” en el sistema financiero, un capital adeudado S/.1, 634.30, una deuda en el sistema financiero de S/.33, 160.99, además poseen una línea disponible de S/. 2, 380.78, tienen una antigüedad de 3 años, un ratio de deuda en el sistema financiero respecto al ingreso de 17.55, poseen la tarjeta “Clásica” y tienen una calificación crediticia máxima de “0”, es decir un comportamiento “normal” en el sistema financiero dentro de los 6 últimos meses.

## 4.2 Análisis de Regresión Multinomial

Se realizó un análisis de regresión logística multinomial para obtener una regla de clasificación que permita asignar nuevos clientes a cada uno de los conglomerados.

Para obtener y validar el modelo obtenido, se realizó una partición de la muestra, es decir un 70% de entrenamiento y un 30% de validación.

Los modelos obtenidos, tanto con la muestra de entrenamiento y comprobación presentan una significancia alta ( $<0.05$ ), lo cual nos indica que en el modelo final de cada partición, los coeficientes del parámetro son diferentes de cero (considerando la  $H_0: \beta_i=0$ ,  $H_a= \beta_i \neq 0$ ), es decir las variables se ajustan al modelo.

**Cuadro Nro.3 Resultados del ajuste del modelo**

<b>Información del ajuste del modelo – Muestra de entrenamiento</b>						
<b>Modelo</b>	Criterio de ajuste del modelo			Contrastes de la razón de verosimilitud		
	AIC	BIC	-2 log verosimilitud	Chi-cuadrado	gl.	Sig.
<b>Sólo la intersección</b>	17032.3	17059.1	17024.3			
<b>Final</b>	10749.7	11124.1	10637.7	6386.7	52	0.000
<b>Información del ajuste del modelo – Muestra de comprobación</b>						
<b>Modelo</b>	Criterio de ajuste del modelo			Contrastes de la razón de verosimilitud		
	AIC	BIC	-2 log verosimilitud	Chi-cuadrado	gl.	Sig.
<b>Sólo la intersección</b>	7429.7	7453.2	7421.7			
<b>Final</b>	3273.9	3601.9	3161.9	4259.8	52	0.000



Por otro lado, el pseudo R-cuadrado nos indica que las variables independientes están explicando aproximadamente en un 60% a la variable dependiente, es decir a los cinco Conglomerados, sin embargo podría considerarse la posibilidad de revisar el modelo para tratar de hacer mejores predicciones.

**Cuadro Nro.4 Pseudo R- Cuadrados del Modelo**

<b>Pseudo R-cuadrado – Muestra de Entrenamiento</b>	
<b>Cox y Snell</b>	0.606
<b>Nagelkerke</b>	0.669
<b>McFadden</b>	0.375
<b>Pseudo R-cuadrado – Muestra de Comprobación</b>	
<b>Cox y Snell</b>	0.808
<b>Nagelkerke</b>	0.856
<b>McFadden</b>	0.574

**Eficiencia del Modelo:**

Según el Cuadro Nro.5, observamos que a nivel general se obtuvo un 94.1% de eficiencia para la muestra de entrenamiento, y un 94.0% para la muestra de comprobación, lo que nos indica una alta eficiencia en el modelo obtenido, lo que nos puede indicar que podemos realizar predicciones a los futuros clientes.

**Cuadro Nro.5 Eficiencias de los Modelos**

<b>Pronosticado – Muestra de Entrenamiento</b>						
<b>Observado</b>	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	<b>Porcentaje correcto</b>
Cluster 1	2217	0	0	0	0	100.0%
Cluster 2	0	0	347	0	0	0.0%
Cluster 3	0	0	566	0	0	100.0%
Cluster 4	0	0	5	1258	0	99.6%
Cluster 5	0	0	0	0	1528	100.0%
<b>Porcentaje global</b>	37.40%	0%	15.50%	21.20%	25.80%	94.1%
<b>Pronosticado – Muestra de Comprobación</b>						
<b>Observado</b>	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	<b>Porcentaje correcto</b>
Cluster 1	986	0	0	0	0	100.0%
Cluster 2	0	0	152	0	0	0.0%
Cluster 3	0	0	251	0	0	100.0%
Cluster 4	0	0	5	550	0	99.6%
Cluster 5	0	0	0	0	642	100.0%
<b>Porcentaje global</b>	38.20%	0%	15.70%	21.30%	24.90%	94.0%

Las reglas obtenidas se encuentran en el Anexo2.

## V. CONCLUSIONES Y RECOMENDACIONES

### Conclusiones

De los resultados obtenidos se concluye:

- 1) Con el algoritmo K-means se determinó que los clientes que realizaron transacciones en el primer trimestre del presente año, se pueden distribuir en cinco conglomerados, ya que presenta el mayor valor de medida de silueta y cohesión, siendo 0.559. Aunque la técnica Kohonen, también nos mostró un valor alto de silueta (0.557), al observar el cociente de tamaños de conglomerados, el valor de 41.06 evidenció una distancia muy elevada entre el conglomerados más grande al más pequeño, a diferencia del cociente con K-means de 6.42. Por tanto la decisión se enfocó en los cinco conglomerados obtenidos con K-means, al observar el valor de silueta y a su vez la proporción de los conglomerados obtenidos.
  
- 2) Según la caracterización de los Conglomerados, se logró identificar dos conglomerados que necesitan una gestión anticipada, tenemos a los “Clientes antiguos con medianos niveles de consumo y con posibles problemas crediticios” (Conglomerado 2) y a los “Clientes antiguos con bajos niveles de consumo y con posibles problemas crediticios (Conglomerado 3), además engloban a las tarjetas “Platinum” y “Clásica”, respectivamente. Los clientes que se encuentran en dichos conglomerados deben ser gestionados de manera temprana ya que evidencian futuros problemas crediticios.
  
- 3) Mediante la aplicación de la técnica de Análisis de Regresión Logística Multinomial, tomando como variable dependiente a los conglomerados obtenidos, se obtuvo como resultado una tasa aproximada del 94% de buena clasificación tanto en la muestra de entrenamiento como validación. Por lo cual, podemos concluir la eficiencia de nuestro modelo para predicciones futuras e implementar las reglas de clasificación obtenidas.

## **Recomendaciones**

- 1) Se recomienda agregar al análisis una técnica comparable a la Regresión Logística Multinomial, como el algoritmo de Árboles de Clasificación, para evaluar la eficiencia de nuestro modelo.
- 2) Realizar el proceso de segmentación en otros periodos de tiempo, e ir evaluando periódicamente los conglomerados, para saber si se incorpora algún nuevo grupo de interés a los conglomerados. Ya que a la empresa le interesa conocer aquellos conglomerados, cuyas características definan la necesidad de gestiones para anticiparse a los eventos de no pago de sus clientes
- 3) Tratar de tener siempre datos de entrenamiento y comprobación, para obtener una correcta validación de los resultados obtenidos.

## VI. BIBLIOGRAFÍA

1. Bacher, J. Wenzing, K. Vogler, M. SPSS Two Step Cluster: A first Evaluation Universitat Erlangen-Nurnberg. Alemania. 2004.
2. C. Beltrán. Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos. Biometría, Filosofía y Lingüística informática. 2011.
3. Cheng Y, Church GM. “Biclustering of expression data”. 2000. Proceeding of the 8<sup>th</sup> International Conference on Intelligent Systems for Molecular Biology; 93-103.
4. Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. Statist Med. 2008; 27:4238–53.
5. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley, 1989.
6. Karthikeyani Visalakshi, N; Thangavel, K. Impact of Normalization in Distributed K –means Clustering. s.I International Journal of Soft Computing, v.4, no. 4, p.168-172. 2009.
7. Kaski, S., Kangas, J. and Kohonen, T. Bibliography of self-organizing map (SOM) papers: 1981-1997. Neural Computing Surveys.1998, 1: 102-350.
8. Norusis. M, IBM: SPSS Statistics Guides 2005-2011 (en línea). Consultado 29 oct. 2013. Disponible en [http://www.norusis.com/pdf/SPC\\_v19.pdf](http://www.norusis.com/pdf/SPC_v19.pdf).
9. Yaghini, M. 2010. Data Mining: SPSS Clementine 12 (en línea). Disponible en <http://webpages.iust.ac.ir/yaghini/>.

## VII. ANEXOS

### Anexo 1: Correlaciones Variables Cuantitativas

		CONSUMO _SOLES	SCORE	CAPITAL_AD EUDADO	DEUDA_SF	ANTIGUED AD	LINEA_DISPO NIBLE	SF_I	CC_I	LD_LC	CA_LC
CONSUMO_SOLES	Corr. de Pearson	1.00	-,05**	,21**	,08**	.02	,04**	,08**	,03**	-,08**	,10**
SCORE	Corr. de Pearson	-,05**	1.00	-,31**	.00	,15**	,44**	-.01	,07**	,68**	-,66**
CAPITAL_ADEUDA DO	Corr. de Pearson	,21**	-,31**	1.00	,35**	,17**	-.02	,15**	.02	-,42**	,52**
DEUDA_SF	Corr. de Pearson	,08**	.00	,35**	1.00	,26**	,24**	,46**	,08**	-.02	,05**
ANTIGUEDAD	Corr. de Pearson	.02	,15**	,17**	,26**	1.00	,17**	,16**	,07**	.01	-.01
LINEA_DISPONIBLE	Corr. de Pearson	,04**	,44**	-.02	,24**	,17**	1.00	.02	,03**	,54**	-,47**
SF_I	Corr. de Pearson	,08**	-.01	,15**	,46**	,16**	.02	1.00	,51**	-,05**	,06**
CC_I	Corr. de Pearson	,03**	,07**	.02	,08**	,07**	,03**	,51**	1.00	,08**	-,07**
LD_LC	Corr. de Pearson	-,08**	,68**	-,42**	-.02	.01	,54**	-,05**	,08**	1.00	-,88**
CA_LC	Corr. de Pearson	,10**	-,66**	,52**	,05**	-.01	-,47**	,06**	-,07**	-,88**	1.00

\*\* . La correlación es significativa en el nivel 0,01 (2 colas).

## Anexo 2:

### Reglas de Clasificación – Modelo Regresión Logística Multinomial

#### (Muestra de Entrenamiento)

##### Ecuación para conglomerado-1

Categoría base

$$+ 0,00000000000000000000$$

##### Ecuación para conglomerado-2

$$\begin{aligned} & -0,00001074 * CONSUMO_SOLES + \\ & 0,0004692 * SCORE + \\ & -0,000002562 * CAPITAL_ADEUDADO + \\ & 0,0000004615 * DEUDA_SF + \\ & 0,001105 * ANTIGUEDAD + \\ & -0,000007392 * LINEA_DISPONIBLE + \\ & -0,0005685 * SF_I + \\ & -1,613 * [CLASIFICACION_RCC=0] + \\ & 11,45 * [CLASIFICACION_RCC=1] + \\ & -0,1614 * [CLASIFICACION_RCC=2] + \\ & -0,02155 * [CLASIFICACION_RCC=3] + \\ & 2,112 * [COD_GRUPO_TEMP=2] + \\ & -0,3342 * [COD_GRUPO_TEMP=3] + \\ & + -2,244 \end{aligned}$$

##### Ecuación para conglomerado-3

$$\begin{aligned} & 0,000007009 * CONSUMO_SOLES + \\ & -0,0001686 * SCORE + \\ & -0,0000009336 * CAPITAL_ADEUDADO + \\ & -0,00000005148 * DEUDA_SF + \\ & -0,000359 * ANTIGUEDAD + \\ & 0,000003231 * LINEA_DISPONIBLE + \\ & 0,00006732 * SF_I + \\ & -11,49 * [CLASIFICACION_RCC=0] + \\ & 15,53 * [CLASIFICACION_RCC=1] + \\ & 0,1161 * [CLASIFICACION_RCC=2] + \\ & -0,4583 * [CLASIFICACION_RCC=3] + \\ & -1,057 * [COD_GRUPO_TEMP=2] + \\ & -3,314 * [COD_GRUPO_TEMP=3] + \\ & + 10,72 \end{aligned}$$

Ecuación para conglomerado-4

$$\begin{aligned}
 &0,000003142 * CONSUMO_SOLES + \\
 &0,0003365 * SCORE + \\
 &-0,0000007845 * CAPITAL_ADEUDADO + \\
 &0,000000256 * DEUDA_SF + \\
 &-0,0002439 * ANTIGUEDAD + \\
 &0,000004301 * LINEA_DISPONIBLE + \\
 &-0,0005157 * SF_I + \\
 &-0,6173 * [CLASIFICACION_RCC=0] + \\
 &-0,3649 * [CLASIFICACION_RCC=1] + \\
 &0,4502 * [CLASIFICACION_RCC=2] + \\
 &2,924 * [CLASIFICACION_RCC=3] + \\
 &4,149 * [COD_GRUPO_TEMP=2] + \\
 &-2,346 * [COD_GRUPO_TEMP=3] + \\
 &+ -0,2169
 \end{aligned}$$

Ecuación para conglomerado-5

$$\begin{aligned}
 &-0,00001351 * CONSUMO_SOLES + \\
 &0,0004294 * SCORE + \\
 &0,0000002994 * CAPITAL_ADEUDADO + \\
 &0,0000002119 * DEUDA_SF + \\
 &0,0003894 * ANTIGUEDAD + \\
 &-0,000005414 * LINEA_DISPONIBLE + \\
 &-0,0001977 * SF_I + \\
 &0,4215 * [CLASIFICACION_RCC=0] + \\
 &0,1365 * [CLASIFICACION_RCC=1] + \\
 &0,3389 * [CLASIFICACION_RCC=2] + \\
 &-0,08236 * [CLASIFICACION_RCC=3] + \\
 &-3,229 * [COD_GRUPO_TEMP=2] + \\
 &-5,58 * [COD_GRUPO_TEMP=3] + \\
 &+ 2,215
 \end{aligned}$$

## Reglas de Clasificación – Modelo Regresión Logística Multinomial

### (Muestra de Comprobación)

Ecuación para conglomerado-1

Categoría base

$$+ 0,00000000000000000000$$

Ecuación para conglomerado-2

$$\begin{aligned}
 &0,00003484 * CONSUMO_SOLES + \\
 &0,0006026 * SCORE + \\
 &0,000008063 * CAPITAL_ADEUDADO + \\
 &-0,0000001134 * DEUDA_SF + \\
 &0,00005151 * ANTIGUEDAD + \\
 &-0,000007027 * LINEA_DISPONIBLE + \\
 &-0,0006214 * SF_I + \\
 &-1,834 * [CLASIFICACION_RCC=0] + \\
 &11,19 * [CLASIFICACION_RCC=1] + \\
 &-0,1388 * [CLASIFICACION_RCC=2] + \\
 &-0,211 * [CLASIFICACION_RCC=3] + \\
 &2,067 * [COD_GRUPO_TEMP=2] + \\
 &-0,1735 * [COD_GRUPO_TEMP=3] + \\
 &+ -2,191
 \end{aligned}$$



■ Ecuación para conglomerado-3

-0,00003007 \* CONSUMO\_SOLES +  
-0,0005276 \* SCORE +  
-0,00000288 \* CAPITAL\_ADEUDADO +  
0,0000001166 \* DEUDA\_SF +  
0,00003766 \* ANTIGUEDAD +  
0,000004626 \* LINEA\_DISPONIBLE +  
0,0001864 \* SF\_I +  
-11,15 \* [CLASIFICACION\_RCC=0] +  
10,88 \* [CLASIFICACION\_RCC=1] +  
0,5179 \* [CLASIFICACION\_RCC=2] +  
-0,1927 \* [CLASIFICACION\_RCC=3] +  
-0,934 \* [COD\_GRUPO\_TEMP=2] +  
-3,281 \* [COD\_GRUPO\_TEMP=3] +  
+ 10,61

■ Ecuación para conglomerado-4

-0,00003536 \* CONSUMO\_SOLES +  
0,0003443 \* SCORE +  
0,000006167 \* CAPITAL\_ADEUDADO +  
0,00000005091 \* DEUDA\_SF +  
-0,000312 \* ANTIGUEDAD +  
0,000004844 \* LINEA\_DISPONIBLE +  
-0,0002028 \* SF\_I +  
-2,594 \* [CLASIFICACION\_RCC=0] +  
-2,013 \* [CLASIFICACION\_RCC=1] +  
-1,733 \* [CLASIFICACION\_RCC=2] +  
-0,351 \* [CLASIFICACION\_RCC=3] +  
4,267 \* [COD\_GRUPO\_TEMP=2] +  
-2,273 \* [COD\_GRUPO\_TEMP=3] +  
+ 1,692

■ Ecuación para conglomerado-5

-0,0000005643 \* CONSUMO\_SOLES +  
0,0002155 \* SCORE +  
0,000005045 \* CAPITAL\_ADEUDADO +  
-0,00000001727 \* DEUDA\_SF +  
-0,0001444 \* ANTIGUEDAD +  
-0,000003558 \* LINEA\_DISPONIBLE +  
-0,0002527 \* SF\_I +  
0,328 \* [CLASIFICACION\_RCC=0] +  
0,02795 \* [CLASIFICACION\_RCC=1] +  
0,5858 \* [CLASIFICACION\_RCC=2] +  
-0,05237 \* [CLASIFICACION\_RCC=3] +  
-3,37 \* [COD\_GRUPO\_TEMP=2] +  
-5,615 \* [COD\_GRUPO\_TEMP=3] +  
+ 2,495