UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS

POSTGRADO EN MODELOS ALEATORIOS

CARACTERIZACION DE DEUDORES DE TARJETAS DE CREDITO DE UNA INSTITUCION FINANCIERA A TRAVES DE TECNICAS DE APRENDIZAJE ESTADÍSTICO

Trabajo de Grado de Maestria presentado ante la ilustre Universidad Central de Venezuela por el Licenciado en Estadística Carlos David Salazar Paredes para optar al título de Magister Scientiarium en Modelos Aleatórios.

Tutor: Dra. Carenne Ludeña

Caracas – Venezuela
Marzo 2012

Contenido

Introducción

Las Tarjetas de Crédito (TDC) constituyen uno de los mecanismos de financiamiento que utilizan las Entidades Financieras a fin de ofrecer a sus clientes la posibilidad de adquirir bienes y servicios con la comodidad de pagarlos en un plazo de tiempo determinado con una tasa de interés establecida. Una vez que se produce un consumo o financiamiento con la TDC se genera un monto mínimo de pago, el cual se distribuye en 3 partes la primera es una fracción de este que se abona directamente al capital (monto del préstamo), otra parte son los intereses los cuales son la ganancia de la organización por proveer al cliente la facilidad de la adquisición financiada del bien o servicio, y una tercera parte conocida como la "mora", esta parte del pago mínimo se empieza a generar una vez que establecido el tiempo para dicho pago, el cliente no cumple con su pago y obliga a la institución financiera a invertir en recursos humanos y tecnológicos a fin de recordarle que tiene un pago pendiente y que debe ponerse al día para así continuar disfrutando los servicios que le ofrece poseer una TDC.

Este trabajo se centra en el estudio de un conjunto de deudores de TDC de una institución financiera los cuales una vez que incumplían con su pago mínimo en la fecha establecida eran gestionados a través de un Call Center de Cobranzas, constan de un periodo de observación de 15 meses y se obtuvieron cerca de 490.000 clientes que presentaron problemas de los pagos mínimos de sus tarjetas de crédito.

Los deudores fueron estudiados a través de técnicas de Aprendizaje Estadístico Supervisado y No Supervisado, por la parte no supervisada se aplico un Análisis de Clúster y una Análisis de Correspondencia Simple y Múltiple, en tanto que para el aprendizaje supervisado, se llevaron a cabo una Regresión Logística Multinomial y Arboles de Clasificación.

La aplicación del estas técnicas tienen como objetivo caracterizar de alguna forma a estos deudores y encontrar relaciones entre ellos, con lo cual las técnicas de aprendizaje no supervisado mostraron la relación entre grupos de deudores que se muestran en el análisis clúster y fueron validadas por medio de los análisis de correspondencia, una vez que se determina la relación entre grupos de deudores se procedió, por técnicas de aprendizaje supervisado a predecir cuantas veces estaría en

gestión un cliente lo que a la larga permitirá ofrecer alguna estrategia de cobranza para estos clientes.

El trabajo está dividido en 3 capítulos que contienen un planteamiento del problema así como la definición de los objetivos que forman parte de la investigación, el primer capítulo muestra el fundamento teórico de las técnicas de aprendizaje estadístico, luego en el segundo capítulo se describe los datos que conforman el estudio así como las estadísticas descriptivas de los mismos como fueron obtenidos y las manipulaciones hechas a algunas variables y finalmente en el tercer capítulo se presentan los resultados obtenidos en el proceso de aprendizaje. Finalmente las conclusiones obtenidas y las recomendaciones para futuros trabajos.

Identificación y Justificación del objeto del Problema.

Las instituciones financieras son las encargadas de manejar de una forma segura el capital de aquellas personas bien sea de naturaleza jurídica ó personal que en ellas confían sus haberes. La parte medular del negocio de estas instituciones consiste en generar a través de estos capitales rendimientos por medio de instrumentos que permitan una operatividad y un margen de utilidad rentable.

Una parte importante de dichos instrumentos la constituyen los créditos, siendo estos, préstamos en dinero donde las personas que los solicitan se comprometen a devolverlos en un período de tiempo definido según las condiciones establecidas para dichos préstamos, más los intereses devengados, seguros y costos asociados en caso de que los hubiera.

Los créditos permiten financiar de una manera segura la adquisición de bienes y servicios por parte de las personas con la garantía de pagarlos a lo largo de cierto tiempo.

Una de las modalidades de créditos más empleadas por las instituciones financieras son las tarjetas de crédito (TDC), las cuales no son más que una tarjeta de plástico con unas bandas previamente codificadas que permiten la compra de bienes y servicios según la capacidades de pago de las personas que la solicitan, y a la cual se les asigna un monto conocido como límite o cupo. Este límite o cupo es el monto total que la institución le da en calidad de préstamo y para la cual se le da un tiempo establecido de 36 meses para cancelar. Una vez que el cliente utiliza su tarjeta de crédito, se generan intereses por el concepto de financiar la compra y el cual debe cancelarse de manera mensual. Este monto se le conoce como Pago Mínimo que el cliente debe hacer para evitar el atraso en cuotas de su tarjeta de crédito y así poder disfrutar de manera continua los servicios que le ofrece la tarjeta.

El hecho de no pagar a tiempo el pago mínimo hace incurrir al cliente en un conjunto de faltas, las cuales se van agudizando según pase el tiempo y no se haya efectuado la cancelación. Cuando esto sucede se habla de altura de mora, lo cual es el número de días que el cliente tiene sin pagar su tarjeta de crédito y comúnmente corre en períodos de 30 días, es decir, un cliente se atrasa en su pago mínimo cae en una

altura de mora de 30 días, al atrasarse dos cuotas su altura de mora será de 60 días y así sucesivamente hasta llegar a las 360 días en la cual la tarjeta se declara como pérdida y sale de la contabilidad generando una perdida para la institución.

Uno de los problemas que viene atravesando las instituciones financieras, es que sus clientes en un status de "mora" (por lo menos 1 cuota atrasada en el pago mínimo de sus TDC), han presentando un crecimiento sostenido a lo largo del tiempo (durante los últimos 3 años), esto motivado a diversos factores, entre los cuales se reseña el hecho que las instituciones al captar clientes a fin de incrementar sus carteras de crédito y en consecuencia aumentar sus márgenes de utilidad a través del cobro de intereses, corren el riesgo de asignar una TDC a clientes cuya capacidad de pago es limitada por diversas razones. El ingreso al status mora se pueden suscitar 3 situaciones, a saber: 1) El Cliente sale de mora debido a que fue por causa de un descuido que no cancelo su pago mínimo, 2) El Cliente permanece moroso hasta el punto de llevar su plástico a perdida y 3) Estar variando entre las diversas alturas de mora y no llegar a ni a salir de ella pero tampoco perder su plástico por llegar al punto de castigo. Es entonces esta situación de crecimiento sostenido del número de deudores, lo que motiva a llevar a cabo un estudio de corte estadístico, a fin, de conocer a través de técnicas de Aprendizaje Estadístico, cuales son las características más relevantes que presentan los deudores de una institución financiera, con el objetivo de generar una caracterización, que permita, que una vez que ingrese un nuevo deudor estimar como podría ser el comportamiento de él durante el tiempo que esté en mora, a fin de tomar acciones que faciliten su recuperación y pueda así seguir disfrutando de los beneficios que le otorga ser un poseedor de una TDC, o simplemente determinar el hecho que no es necesario la inversión de recursos en él, pues es un deudor con muchas dificultades para pagar, y finalmente generar algún acuerdo de pago dado que será un cliente que se mantendrá en atraso por un periodo largo de tiempo y dar entonces respuestas a preguntas como:

- ¿Tiene algún efecto el número de TDC que posea un cliente sobre la cantidad de veces que cae en mora?
- ¿Podría hacer el monto del pago mínimo que un cliente permanezca en un sistema de gestión de cobro por algún tiempo?
- ¿será necesario hacer un agrupamiento en las variables para determinar clases homogéneas de deudores?

Objetivo General

- Caracterizar a través de técnicas de Aprendizaje Estadístico a los clientes con cuotas atrasadas en el pago de sus Tarjetas de Crédito.

Objetivos Específico

- Generar una clasificación a través de las características del deudor.
- Identificar las variables más relevantes en el proceso de clasificación.
- Generar a partir de los datos existentes variables que puedan tener Relevancia en el proceso de clasificación.

Capítulo I

Aprendizaje Estadístico

El aprendizaje estadístico comprende un conjunto de técnicas estadísticas a través de las cuales se desea conocer o predecir algún fenómeno por medio de algún modelo, este se divide en 2 tipos, el Aprendizaje Supervisado y el Aprendizaje No Supervisado, el primero es aquel en el cual se obtiene una variable respuesta que nos permite guiarnos a través del proceso de aprendizaje y el objetivo principal consiste en conseguir un modelo que nos permita hacer inferencia acerca de este proceso y predecir con la mayor precisión posible la posible respuesta. Luego en el segundo solo se tienen observaciones sin que exista alguna respuesta en cuyo caso el objetivo es encontrar la estructura que nos permita explicar de alguna manera la composición de este conjunto de datos y determinar asociaciones bien sea entre individuos o variables. Ambos tipos, pueden resumir el proceso de aprendizaje a través de tres etapas básicas a saber:

1) Observación de un fenómeno.
2) Construcción de un modelo para dicho fenómeno.
3) Hacer predicciones usando dicho modelo.

Entendemos en este sentido fenómeno como cualquier evento que sea de interés estudiar a través de la construcción de un modelo estadístico que permita llevar a cabo las etapas 2 y 3 dentro de los procesos de aprendizaje.

Una de las condiciones que hacen del aprendizaje estadístico una herramienta útil en el entorno de la estadística actual, es el hecho de considerar en la mayoría de sus casos y siempre y cuando la cantidad de observaciones lo permita, la división de los datos observados o disponibles en 2 grupos, un primer grupo denominado de ***Entrenamiento*** y un segundo llamado datos de ***Prueba o Validación*** con lo cual, el primer conjunto se destina a la construcción del modelo y el siguiente permite la evaluación del ajuste que el modelo construido presenta sobre los datos reales, y que permite tomar la decisión de acerca de la manipulación que se debe realizar sobre las variables a fin de lograr un buen ajuste.

Motivado a que la diferencia entre los tipos de aprendizaje es obvia, en el supervisado se tiene algún tipo de respuesta mientras que en el otro existe una ausencia

de ella, se tiene que dentro de cada tipo de aprendizaje están contenidas técnicas cuyos objetivos son acordes al tipo de aprendizaje que las contiene, Hastié, Tibshirani y Friedman (2008) las agrupan de la siguiente manera:

3.2. Principales Técnicas de Aprendizaje Estadístico Supervisado.

- Métodos lineales para Regresión
  - Modelos de Regresión
- Métodos lineales para Clasificación
  - Análisis Discriminante
  - Regresión Logística
  - Híper Planos Separadores
- Métodos No lineales.
  - Splines
  - Métodos de Suavizamiento del núcleo
  - Arboles de clasificación o regresión
  - Redes neuronales
  - Maquinas de Soporte Vectorial
  - Vecinos más cercanos

1.2 Principales Técnicas de Aprendizaje Estadístico No Supervisado.

- Reglas de Asociación
- Análisis Clúster
- Componentes Principales, curvas y superficies
- Análisis de Correspondencia simple y múltiple
- Escalonamiento Multidimensional
- Bosques Aleatorios

A lo largo de este trabajo nos concentraremos básicamente en 4 técnicas, 2 de Aprendizaje Estadístico supervisado y 2 de Aprendizaje no Supervisado como lo son: Regresión logística y Árboles de Clasificación por el lado supervisado y Análisis Clúster y Análisis de Correspondencia simple y múltiple en el caso no supervisado, y

finalmente la utilización de resultados obtenidos en la parte del aprendizaje No Supervisado para mejorar los modelos de Aprendizaje Supervisado.

A fin de ir mostrando a través de ejemplos el componente teórico de las diferentes técnicas de aprendizaje estadístico mostradas en este trabajo, vamos a definir un conjunto de variables que luego serán definidas con mayor precisión en el capítulo III.

- Variables Observadas Directamente de los Datos:
    - Monto Vencido Promedio **(PROM_VCDO)**
    - Monto Cobrado Promedio **(PROM_PAG)**
    - Cantidad de Tarjetas Créditos **(CANT_TDC)**
    - Número de veces en Gestión **(NGEST)**
    - Número de Pagos mientras estuvo en Gestión **(NPAG)**
    - Máxima Altura de Mora **(MAX_MORA)**
- Variables Ajenas al proceso de obtención:
    - Limite o Capacidad de Endeudamiento del Cliente **(LIMITE)**
    - Sexo del Cliente **(SEXO)**
- Variables construidas a partir del proceso de obtención de datos:
    - Frecuencia o Capacidad de Pago **(COMO_PAGA)**
    - Número de Ascensos en Mora **(NMOV)**
    - Cluster Definitivo **(cluster4def)**

1.1 Técnicas de Aprendizaje estadístico Supervisado

1.1.1 Regresión Logística:

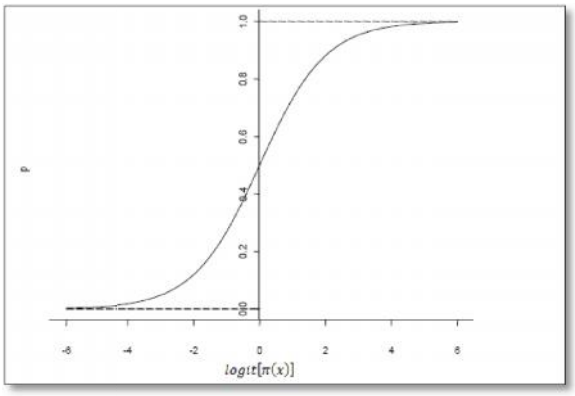
Regresión Logística Simple

Cuando dentro del aprendizaje estadístico supervisado nos encontramos con una variable repuesta *Y* que solo tiene 2 posibles resultados, por ejemplo y un conjunto de variables explicativas $\boldsymbol{X}=(X_1,X_2,\ldots X_n)$, la regresión logística simple o binaria es una de las técnicas que mayor difusión tiene para acatar este tipo de problemas, debido a su fácil interpretación y la capacidad predictora que esta posee.

Ahora bien si denominanos $\pi(x)$ como la probabilidad de éxito dado un valor de ***X,*** el modelo de regresión logística es de la forma:

$$\pi(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \quad .$$

Esta fórmula implica que $\pi(x)$ incrementa o decrece como una función con forma de S como se muestra en la grafica 1 que depende de $\boldsymbol{X}$, α es una constante y es el vector de parámetros estimados que corresponden según el número de variables predictivas se tengan.



Grafica 1.

Regresión Logística Caso Multinomial.

Si J denota el numero de de categorías de una variable respuesta Y, y sea $\{\pi_1, \pi_2, \ldots, \pi_j\}$, las respectivas probabilidades de respuesta las cuales satisfacen que $\sum_j \pi_j = 1$, con n observaciones independientes, la distribución de probabilidad para el numero de resultados del tipo J es Multinomial, la cual especifica que la probabilidad para cada posible resultado de las n observaciones pueden caer dentro de las J categorías de la variable respuesta Y. En el caso de los modelos de regresión logística Multinomial consideran una categoría de referencia bien puede ser la primera o la ultima, con lo cual se construyen los modelos de probabilidad de pertenencia de las j-1 categorías restantes de la forma siguiente.

$$P(G = j, X = x) = \frac{\exp(\beta_{j0} + \beta_j^T x)}{1 + \sum_{l=1}^{J-1} \exp(\beta_{l0} + \beta_l^T x)}, \quad J=1,\ldots,J-1$$

Donde j es una categoría del vector de respuesta de Y, y $\boldsymbol{X}$ es el vector de valores que toman las variables y igual que en el caso anterior β es el vector de parámetros estimados

Ajuste de los modelos de Regresión logística

En los modelos de regresión logística los parámetros son comúnmente estimados a través de métodos de Máxima Verosimilitud usando la verosimilitud condicional de G dado un valor de $\boldsymbol{X}$. con lo cual el logaritmo de la verosimilitud es de la forma:

$$l(\beta) = \sum_{i=1}^{N} \{y_i \beta^T x_i - \log(1 + \exp(\beta^T x_i))\}.$$

Donde $\beta = \{\beta_1, \beta_2, \ldots, \beta_{1j}\}$ y $\boldsymbol{x}$ es el vector de las observaciones y_i es el valor de variable categórica para la i-esima observación. Para maximizar el logaritmo de la verosimilitud se deben tomar las derivadas parciales de $l(\beta)$ con respecto a cada β_i y se igualan a cero.

$$\frac{\partial(\beta)}{\beta} = \sum_{i=1}^{N} x_i \left(y_i - \exp\left(\beta^T x_i\right)\right) = 0.$$

Para resolver estas ecuaciones se hace uso del algoritmo de Newton-Raphson, lo cual requiere de la segunda derivada o matriz Hessiana que está dada por.

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = \sum_{i=1}^{N} x_i x_i^T \exp\left(\beta^T x_i\right)\left(1 - \exp\left(\beta^T x_i\right)\right)$$

Ahora comenzando con β^o, una actualización del algoritmo de newton es

$$\beta^n = \beta^o - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right)^{-1} \frac{\partial l(\beta)}{\partial \beta}.$$

Donde las derivadas son evaluadas en β^o, y expresando las ecuaciones en forma matricial se tiene entonces que:

$$\frac{\partial l(\beta)}{\partial \beta} = \boldsymbol{X}^T(\boldsymbol{y} - \mu)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$$

Donde: $\boldsymbol{y}$ es el vector de los valores de la variable respuesta $\boldsymbol{X}$ una matriz N x $(p + 1)$ de los valores de x_i, $\boldsymbol{p}$ es el vector de probabilidades ajustadas con el *i-esimo* elemento $p(x_i;\beta^0)$, y $\boldsymbol{W}$ una matriz diagonal $\boldsymbol{N}$ **x** $\boldsymbol{N}$ de pesos cuya *i-esima* diagonal es $p(x_i;\beta^0)(1 - p(x_i;\beta^0))$.

Luego la solución se alcanza cuando se encuentra:

$$\beta^n \leftarrow \underset{\beta}{\operatorname{argmin}} (z - X\beta)^T W (z - X\beta)$$

Con: $z = X\beta^o + W^{-1}(y - p)$.

Una vez obtenidos los estimadores se determina la probabilidad de pertenencia de los individuos a cada una las *j* clases de la variable respuesta Y, con lo cual se estima la efectividad del modelo para la clasificación, y luego a partir de estas clasificaciones obtenidas se construye una tabla entre el valor (la categoría o clase) predicho por el modelo y el valor real, como se muestra en la tabla siguiente, en la cual los resultados se corresponden al problema que será estudiado más adelante.

Muestra	Valor Observado	Prediccion					
		1	2	3	4	5	% Correcta Clasificacion
Entrenamiento	1	97.062	341	0	0	25	99,6%
	2	5.553	70.180	117	3	201	92,3%
	3	91	5.588	49.825	388	1.090	87,4%
	4	34	708	7.821	38.501	5.242	73,6%
	5	23	152	1.892	10.527	190.072	93,8%
	% Total	21,2%	15,9%	12,3%	10,2%	40,5%	91,8%

Tabla 1.

Vemos en la Tabla 1 como en este caso los individuos que estaban en la categoría 1, 2 y 5 fueron ubicados por el modelo de de regresión logística de una manera casi exacta es decir más del 90 % de los datos que se encontraban en dichas categorías fueron estimados correctamente por el modelo, en tanto que se presenta una confusión entre los individuos pertenecientes a las categorías 3 y 4.

1.1.2 Árboles de Clasificación

Entre los métodos de aprendizaje supervisado nos encontramos muchas veces en que los modelos de regresión que se tienen disponibles tienden a fallar en situaciones de la vida real, y una de las principales causas de esto es que los efectos que ejercen las variables explicativas sobre la respuesta no son necesariamente lineales o sus errores no siguen distribuciones normales con lo cual, se necesitan otras técnicas para identificar y caracterizar los efectos no lineales de las variables explicativas. Los métodos basados en arboles son una de ellas, estas dividen el espacio en un conjunto de rectángulos donde luego ajustan dentro de cada uno de ellos un modelo simple como una constante, con la ventaja que son conceptualmente sencillos y de gran alcance.

Construcción de Arboles

El algoritmo parte de un conjunto de datos que son los denominados datos de entrenamiento los cuales son el resultado de dividir el conjunto inicial de datos en 2 partes estos primeros de entrenamiento y unos segundos llamados de prueba (siempre y cuando esto sea posible según la disponibilidad del numero de datos que se tengan), en el cual los datos están etiquetados con un conjunto de pertenencia, se comienza entonces con un nodo inicial y surge la pregunta de cómo dividir este conjunto de datos en 2 o más partes homogéneas haciendo uso de alguna de las variables disponibles, la cual es escogida de manera tal que la partición resulte lo más homogénea posible. Por ejemplo

seleccionamos la variable X_i la división seria de forma tal que en al lado izquierdo del nodo inicial estén los datos donde X_i <= c y del lado derecho estarán entonces aquellos datos donde la variable X_i sea > c, y luego se repite este proceso el cual termina cuando todas las observaciones hayan sido clasificadas correctamente en su grupo, finalmente una vez establecida las reglas de clasificación se aplican sobre los datos que se dejaron para validar el modelo y se compara la correcta clasificación a ver el potencial del modelo. De acuerdo con esto se pueden distinguir dos tipos de árboles de clasificación dependiendo de la estructura del árbol, esto es, del número de ramas que se puedan generar a partir de un nodo:

1. **Arboles construidos bajo el algoritmo CART** (Classification And Regression Trees), que permite generar únicamente dos ramas a partir de un nodo.
2. **Árboles construidos bajo el algoritmo CHAID (**Chi Square Automatic Interaction Detection), que genera un número distinto de ramas mayor o igual a 2 a partir de un nodo.

Una de las necesidades de los arboles de clasificación se hace en el hecho de controlar su crecimiento puesto que contar con arboles con muchos nodos (ramas) hace difícil su comprensión, con lo cual se aplican técnicas de poda las cuales se llevan a cabo mejorando una medida de impureza $Q_m(T)$ la cual está basada en las proporción de observaciones de la clase j en un nodo m que a su vez se define como:

$$\hat{p}_{mj} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = j).$$

Donde: R_m = es una región con N_m observaciones en un nodo m, e $I(y_i = j)$ una función indicatriz cuando $y_i = j$.

Existen diferentes medidas para la impureza:

- Error de pérdida de clasificación:

$$me = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq j(m)) = 1 - \hat{p}_{mj(m)}$$

- Índice de Gini:

$$gi = \sum_{j=1}^{J} \hat{p}_{mj}(1 - \hat{p}_{mj})$$

- Entropía cruzada:

$$ce = -\sum_{j=1}^{J} \hat{p}_{mj} log \hat{p}_{mj}$$

Los índices de entropía cruzada y de Gini son más sensibles en cambios a las probabilidades de los nodos que el Error de pérdida de clasificación (me), por lo cual los 2 primeros pudieran ser utilizados durante el crecimiento del árbol.

En los procesos de poda se puede usar cualquiera de los tres criterios, aunque comúnmente es el criterio de error de pérdida de clasificación el más utilizado.

Dentro de las ventajas de los árboles de clasificación, se tienen:

- Los resultados son invariantes por una transformación monótona de las variables explicativas.
- Es una técnica robusta.
- Se permiten mezclar variables continuas con discretas.
- Es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.
- El algoritmo tiende a una rápida convergencia lo cual lo hace computacionalmente eficiente.

Desventajas:

- Las variables predictoras (independientes) continuas se tratan de manera ineficiente como variables categóricas discretas.

- El árbol final puede que no sea óptimo. La metodología que se aplica sólo asegura que cada subdivisión es óptima.
- Los árboles grandes tienen poco sentido intuitivo y las predicciones tienen, a veces, cierto aire de cajas negras.
- Los arboles pudieran presentar varianzas altas.

A continuación se muestran 2 tablas a manera ilustrativa de una aplicación de un árbol de clasificación para el problema tratado en este trabajo, donde veremos cómo se confirma lo mostrado en la regresión logística.

Resumen del Arbol		
Especificaciones de Entrada	Método de Crecimiento	CHAID
	Variable Dependiente	NGEST2
	Variables Independientes	cluster_4, CANT_TDC, PROM_VCDO, PROM_PAG, LIMITE, NMOV, MAX_MORA, COMO_PAGA, NPAG_A
	Máxima Profundidad del arbol	3
	Minimo de casos en un Nodo Padre	100
	Minimo de casos en un Nodo Hijo	50
Resultados	Variables Independientes Incluidas	NPAG_A, NMOV, COMO_PAGA, MAX_MORA, cluster_4
	Numero de Nodos	30
	Numero de Nodos Terminales	21
	Profundidad	3

Muestra	Valor Observado	Prediccion					
		1	2	3	4	5	% Correcta Clasificacion
Entrenamiento	1	77.951	0	0	1	0	100,0%
	2	4.501	56.226	1	26	0	92,5%
	3	70	4.577	39.810	506	706	87,2%
	4	11	614	5.813	30.901	4.588	73,7%
	5	1	109	970	6.793	154.211	95,1%
	% Total	21,3%	15,8%	12,0%	9,8%	41,1%	92,5%
Prueba	1	19.476	0	0	0	0	100,0%
	2	1.128	14.156	0	16	0	92,5%
	3	16	1.128	9.890	109	170	87,4%
	4	5	139	1.393	7.652	1.190	73,7%
	5	0	23	247	1.798	38.514	94,9%
	% Total	21,3%	15,9%	11,9%	9,9%	41,1%	92,4%

Tabla 2.

La primera tabla (parte superior *Tabla 2*) muestra el resumen de la técnica de partición utilizada, en este caso tenemos particiones del tipo CHAID que son las que permiten más de 2 particiones en cada nodo, seguido a esto las variables tanto respuesta como las de entradas, luego vemos que la profundidad máxima del árbol es de 3 niveles, las variables que resultaron seleccionadas para la construcción del árbol y finalmente el numero de nodos terminales y la profundidad alcanzada, la segunda tabla nos muestra el

proceso de clasificación (valor observado vs valor pronosticado por el árbol) tanto para la muestra de entrenamiento (Training) como para la de validación (Test) así como, los porcentajes de correcta clasificación de cada uno del conjunto de datos, al comparar estos resultados utilizados con los de la regresión logística vemos comportamiento de clasificación similares para los grupos 1, 2 y 5 en los cuales se obtiene una correcta clasificación en ambos casos (entrenamiento y validación) y luego la misma confusión presente en los grupos 2 y 3, pero en general se logra una correcta clasificación por encima del 90 % de los valores observados.

1.2 Técnicas de Aprendizaje estadístico No Supervisado

Recordemos entonces que cuando estamos en presencia de un conjunto de observaciones (datos) en la cual no tenemos una variable respuesta la idea es tratar de hacer algún tipo de inferencia a partir de este conjunto de datos. Esto con el fin de determinar la existencia de algún tipo de asociación que pudieran tener las variables de este conjunto.

El aprendizaje Estadístico No Supervisado dispone de algunas técnicas como: componentes principales, escalonamiento multidimensional, análisis de correspondencia entre otras, las cuales nos dan ideas acerca de la posible asociación entre variables con el objetivo de la construir funciones que nos permiten reducir el espacio original, dicha reducción se puede hacer a través de la construcción de variables latentes o funciones que hacen más sencilla la comprensión de los datos.

Ahora bien otro posible interés se suscita cuando se tienen datos donde no existe una respuesta visible, entonces el objetivo estaría en determinar la existencia de individuos con características comunes y agruparlos en grupos con características similares a las de él, este tipo de procesos están contenidos en las técnicas de agrupamiento o clústerización.

En el desarrollo de este trabajo se utilizaron 2 técnicas de aprendizaje estadístico no supervisado, en una primera etapa se hace uso de una versión de las técnicas de agrupamiento (clúster) que permite la construcción de grupos usando variables continuas y categóricas por medio de un algoritmo denominado *"two steps clúster"* , y

luego se trabaja con un análisis de correspondencia sobre un conjunto de variables categóricas a fin de evaluar las asociación entre los niveles de estas variables que comprenden el conjunto de datos.

1.2.1 Análisis Clúster

El análisis clúster es también llamado segmentación de datos, cuenta una variedad de objetivos y todos se centran en agrupar en pequeños subconjuntos de observaciones llamados *"clústers"* a un conjunto de datos, con el fin de formar grupos homogéneos dentro de sí, pero heterogéneos entre ellos, otro de los objetivos de este tipo de análisis es usarlo como estadística descriptiva que permite ver la existencia de un conjunto de datos que pudieran tener características sustancialmente diferentes a los demás, básicamente la diferencia entre los tipos de técnicas de agrupamiento se basa es en como se mide la distancia entre las observaciones que comprenden los datos, pues este tipo de medida es quien determinará finalmente la ubicación de alguna observación entre un clúster u otro.

Varios son los algoritmos usados para el proceso de clasificación los cuales se pueden resumir en 3 grupos, a saber:

1) *Los algoritmos combinatorios:* Estos trabajan directamente en los datos observados sin referencia directa a un modelo subyacente de probabilidad.

2) *Los modelos de mezcla:* en los cuales se supone que los datos son observaciones independientes e idénticamente distribuidas descritos por una función de densidad de probabilidad, esta densidad puede ser el conjunto de una mezcla de densidades donde cada una puede representar a un clúster. El modelo es ajustado a través de métodos de máxima verosimilitud o Bayesianos.

3) *Los Algoritmos no paramétricos*: Los cuales intentan estimar directamente los distintos modos de la función de densidad de probabilidad. Las observaciones "más cercanas" a cada nodo respectivo, son entonces las que definen los clúster individuales.

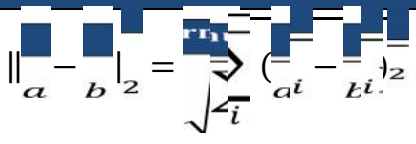
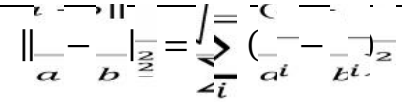
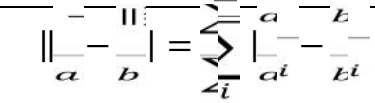
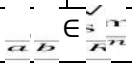
Los algoritmos de agrupamiento más populares asignan directamente cada observación a un grupo o a un clúster sin consideración de la probabilidad que describa

los datos, uno de ellos es el de ***K-medias***, un algoritmo iterativo en el cual se reúnen en un conjunto *K* de grupos o clústers a las *N* observaciones donde *K*<*N*. El procedimiento del análisis clúster de K-medias empieza con la construcción de unos centros de conglomerados iniciales , los cuales se pueden asignar de manera aleatoria o tener un procedimiento de selección de k observaciones bien situadas para formar los centros de conglomerados, una vez que se forman los centroides iníciales se asignan casos a los conglomerados basándose en la distancia de los centros de estos. Luego se vuelven a determinar los centroides con las observaciones incorporadas para ver la asignación de una nueva observación y se repite de manera iterativa hasta que se obtienen entonces grupos homogéneos entre si y heterogéneos entre ellos.

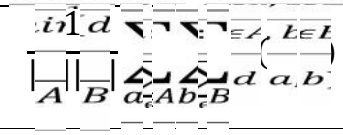
Otro de los algoritmos utilizados es el de ***Agrupación Jerárquica*** que se divide a su vez en 2 métodos, los aglomerativos y los divisivos, en los primeros se parte de una matriz de similaridades o distancias de tamaño *N x N*, con lo cual se tienen N clúster de partida, luego en el segundo paso se agrupan aquellos individuos que estén más cercanos entre sí en un nuevo clúster, con lo cual, se va reduciendo el espacio, luego se calcula la distancia entre este nuevo clúster y los demás y luego se repite de manera iterativa el procedimiento hasta que se forman un número menor de grupos a los N de partida, ahora bien, en los métodos divisivos se parte de un solo clúster de tamaño N y a partir de divisiones recursivas, se van obteniendo los clúster finalmente deseados y sus resultados son presentados en gráficos conocidos como Dendogramas.

Para la decisión acerca de hacia que clúster debe ser agregado un conjunto de observaciones en el caso de una agrupación aglomerativa o en que parte debe ser dividido un clúster cuando se trata de una agrupación divisiva, es requerida una medida de semejanza. En la mayoría de los métodos jerárquicos esto se hace a través de una distancia y un criterio de enlace el cual especifica la semejanza entre las observaciones como una función de los pares de distancia entre las observaciones.

A continuación se muestra una tabla de medidas de distancia que son utilizadas en los diferentes métodos de agrupación jerárquica.

Medida	Formula
Distancia Euclidea	
Distancia Euclidea al Cuadrado	
Distancia de Manhattan	
Distancia de Mahalanobis	 , con S matriz de Covarianza
	

Los criterios de enlace son mostrados en la siguiente tabla:

Criterio de Enlace	Formula
Máximo o Completo	
Mínimo o Simple	
Promedio	

Donde: d es la medida de distancia seleccionada, A y B subconjuntos de R^n.

Otras medidas de enlace son: la suma de toda la varianza interclúster, el incremento de la varianza para un nuevo clúster (principio de Ward), la probabilidad que un nuevo individuo a ser agregado en el clúster sea de la misma función de distribución (enlace en V).

Uno de los problemas se sucede cuando se requieren formar grupos partiendo de un conjunto de datos que contienen variables tanto continuas como categóricas, ya que en este caso establecer una medida de distancia y a su vez un criterio de enlace se hace difícil, más sin embargo existen algunas técnicas que permiten llevar a cabo este tipo de procedimientos una de ella es la que se empleo en el desarrollo de este trabajo, la cual es un algoritmo conocido como *"two steps clúster"* quien combina un proceso de arboles de clasificación y a partir de este se construyen clúster basados en una medida de distancia como se explica a continuación:

1.2.1.1 Two steps clúster

El algoritmo de Two steps Clúster es un algoritmo de análisis clúster de manera escalonada, el cual está diseñado para manejar de manera conjunta variables categóricas y variables continuas así como grandes volúmenes de información, y se lleva a cabo en 2 etapas a saber:

Etapa 1 (Pre Clúster): se produce a través de una agrupación previa de los datos en pequeños clúster a través de una aproximación secuencial en el cual se evalúa registro a registro y se determina si este puede ser agregado a alguno de los pequeños clúster formados previamente o bien puede formar algún nuevo sub clúster basado en algún criterio de distancia.

El procedimiento se lleva a cabo mediante la construcción de un clúster con características de árbol el cual tiene niveles de nodos donde cada uno contiene un número de entradas, con lo cual las hojas no nodo y sus entradas son utilizados para guiar a un registro de manera correcta a la hoja nodo correspondiente. Cada entrada está definida por las características de su clúster las cuales están resumidas en los estadísticos descriptivos de las variables continuas y la frecuencia observada para cada variable categoría, luego para cada registro sucesivo, comenzando desde la raíz este es guiado por las entradas de los nodos a lo largo del árbol hasta conseguir el nodo más cercano. Para este caso, se requiere que los registros estén distribuidos de manera aleatoria, para minimizar el efecto de ordenamiento.

Etapa 2 (Clúster): En esta etapa se toman los subclúster resultantes de la primera etapa y luego se agrupan dentro del número de clúster que se deseen tener como resultados finales, usando el método de agrupamiento jerárquico aglomerativo.

Una situación de interés en el algoritmo son las medidas de distancia, ya que en el caso en que existen variables continuas y categóricas al mismo tiempo se usa una distancia basada en el logaritmo de la verosimilitud luego de finalizada la primera etapa de clústerización preliminar a través de los clúster con características de árbol de la siguiente manera

Sean:

K^A: Número total de variables usadas en el procedimiento.
K^B: Número de Variables Categóricas usadas en el procedimiento.
L^k: Numero de Categorías de la K- esima variable Categórica.
R^k: Rango de la K- esima Variable Continua.
N Número total de Registros.
N_v: Número total de registros en el K- esimo clúster.
$\hat{\sigma}^2_{jk}$ Varianza estimada de la k-esima variable continúa en el j-esimo clúster.
$\hat{\sigma}^2_k$: Varianza estimada de la k-esima variable continua.
N_{jkl}: Numero de registros en el clúster j cuya k-esima variable categórica toma la l-esima categoría.
D (j,s): Distancia entre los Clúster J y S.
<j,s>: índice que representa el clúster formado por la combinación de los clústers j y s.

Entonces la medida de la distancia está definida por la siguiente expresión:

$$d(j,s) = \zeta_j + \zeta_s - \zeta_{js}. \quad (1)$$

Donde:

$$\zeta_j = N_v\left(\sum_{k=1}^{K^A} \frac{1}{2} log(\hat{\sigma}_k^2 + \hat{\sigma}_{kv}^2) + \sum_{k=1}^{k^B} \hat{E}_{vk}\right), \quad (2)$$

$$\hat{E}_{vk} = -\sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} \log\left(\frac{N_{vkl}}{N_v}\right) \quad (3)$$

Si $\hat{\sigma}_k^2$ es ignorada en la ecuación (2) entonces la distancia entre los clúster j y s podría ser exactamente el decrecimiento del logaritmo de la verosimilitud cuando se combinan los 2 clúster, por su parte el término $\hat{\sigma}_k^2$ es agregado a fin de eliminar el problema cuando $\hat{\sigma}_{vk}^2$ =0 pues en caso que esto suceda resultaría el logaritmo natural indefinido. Este tipo de casos ocurre cuando exista un clúster que tenga solo una observación.

Luego, si se da el caso en el cual todas las variables en el proceso son continuas, se utiliza como medida la distancia Euclidea para la formación de los clúster a partir de los centroides de cada sub-clúster, definido como el vector de medias de cada variable, formado en la etapa 1.

Otra de las opciones que presenta este algoritmo es el hecho de la manipulación de valores atípicos u outliers, los cuales pueden ser asignados a uno u otro clúster, para el caso en que no se decida que el algoritmo realice la manipulación de este tipo de datos se selecciona alguna medida de distancia para la asignación de las variables en los clústers.

Ahora si se decide manejar la asignación de los posibles valores atípicos en los clúster se lleva a cabo de dos maneras. La primara basada en una distancia calculada por el logaritmo de la verosimilitud, la cual suponiendo que estos valores tienen una distribución uniforme, se calcula entonces el logaritmo de la verosimilitud que resulta de asignar el valor atípico a uno de los clúster ya previamente formados o asignarlo a un clúster de observaciones outliers propiamente. Luego la otra manera de asignarlo es a través de la distancia Euclidea donde una observación es asignada a un clúster convencional si la distancia Euclidea es más pequeña que un valor *c* el cual se define como:

$$C = 2\sqrt{\sum_{l=1}^{K^1} \frac{\hat{\sigma}_{cl}^2}{K^A}}$$

Caso contrario la observación pasa a formar parte de un clúster de valores atípicos.

Las salidas mostradas a continuación son las que se obtienen al aplicar esta técnica al problema que trataremos más adelante, los resultados se muestran tanto para las variables continuas en cuyo caso se trata de los centroides de los clústers, así como

de gráficos de barras para la distribución de las variables categóricas dentro de cada clúster. Durante el desarrollo de este trabajo en una primera aproximación a la formación de los grupos con el algoritmo de Two Steps Clúster se obtuvieron 4 clúster en los cuales se aprecian diferencias significativas dentro de ellos, aun y cuando se tiene una varianza relativamente alta dentro los clúster en las variables continuas, esta es comprensible puesto que al momento de la determinación de los clúster las variables categóricas fueron las que predominaron durante el proceso de agrupamiento.

Salidas primera aproximación con la técnica de Two Steps Clúster

Centroides de las Variables continúas

Cluster	PROM_VCDO		PROM_PAG		LIMITE	
	Promedio	Desv. Estandar	Promedio	Desv. Estandar	Promedio	Desv. Estandar
1	995,93	1.034,67	876,19	1.156,27	9.197,70	9.145,44
2	2.254,45	3.229,40	401,87	693,18	8.282,26	8.795,12
3	955,79	1.298,63	1146.62	1.728,89	12.795,35	17.319,46
4	728,48	623,17	742,19	659,73	8.658,12	7.159,06
Total	**1.173,33**	**1.853,75**	**782,43**	**1.113,70**	**9.629,06**	**11.124,81**

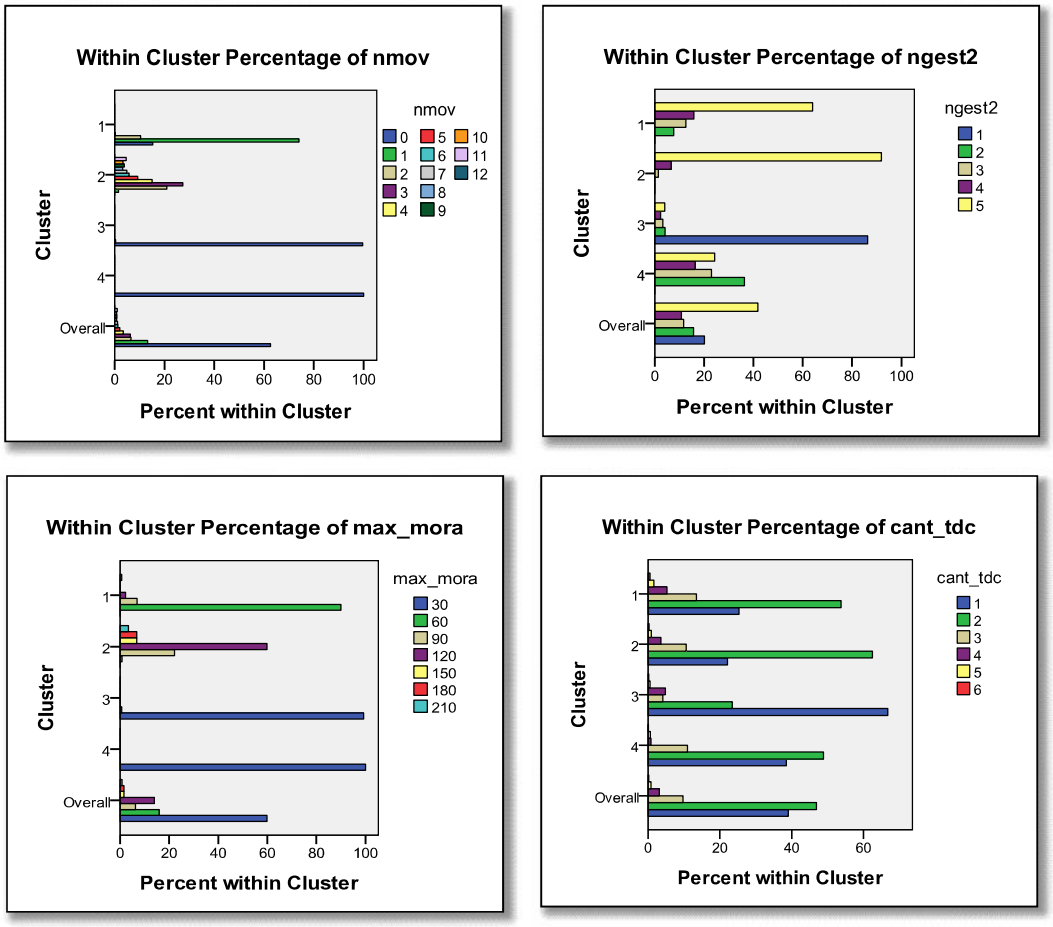
Tabla 3.

Como vemos en la *Tabla 3*, para este ejemplo se forman 4 clústers con características desde el punto de vista de las variables continuas, en el primero observamos a los individuos que pese a que pagan las cuotas de sus obligaciones, no lo hacen en la totalidad es decir no alcanzan a cubrir el promedio de sus pagos mínimos, el segundo contiene a aquellos individuos que se pueden considerar deudores potencialmente peligrosos puesto el promedio de los pagos de los individuos que los contiene no cubre ni la mitad de sus pagos mínimos, mientras que los individuos que pertenecen a los clústers 3 y 4 son los que logran cubrir el pago de sus obligaciones.

Distribución de las variables categóricas

Como en los clúster existen observaciones con variables continuas y categóricas, entonces las distribuciones de estas últimas en los clústers son presentadas a través de diagramas de frecuencia en la *Gráfica 2* donde se observa cómo está compuesta la variable dentro de cada clúster.

En la gráfica 2 se muestra como los clústers 3 y 4 están formados por clientes que solo alcanzaron una mora máxima de 30 días, mientras que el clúster 1 esta formado mayormente por clientes que alcanzaron una mora de 60 días, y el clúster 2 por clientes que alcanzaron moras superiores a los 90 días, diferenciándose estos del número de veces que están los clientes en gestión en el caso del clúster 3 está formado mayormente por clientes que estuvieron 1 vez en gestión en tanto que el clúster 4 pertenece a personas con más de 1 vez en gestión.



Grafica 2.

1.2.2 Análisis Factorial, Análisis de Correspondencia Simple y Múltiple

1.2.2 .1 Análisis Factorial

El análisis factorial es una técnica de interdependencia, donde el objetivo es estudiar la interrelación entre un conjunto de variables, con la finalidad tratar de explicar a través de espacios de menor dimensión al original y que contengan la máxima información la variabilidad propia de los datos, este tipo de análisis tiene por objetivo explicar según un modelo lineal un conjunto de datos que contengan un número grande de variables a través de un nuevo conjunto de menor dimensión formado por variables no observables, denominados factores los cuales no pueden ser observados a simple vista.

Supongamos que tenemos un conjunto de datos con p variables observables, entonces la idea es explicar la variabilidad mediante un numero k de factores menor que p, según un modelo lineal

$$X_1 = c_{11}F_1 + \cdots + c_{1k}F_k + d_1V_1$$

$$X_2 = c_{21}F_1 + \cdots + c_{2k}F_k + d_2V_2$$

$$X_p = c_{1p}F_1 + \cdots + c_{pk}F_k + c_pV_p$$

Donde $F_1, F_2, \ldots, F_k$ son los ***factores comunes*** y $V_1, V_2, \ldots, V_k$ se denominan los ***factores únicos*** pues tienen efecto específico solo sobre una variable, cada coeficiente c_{pk} se denomina ***carga factorial*** y representa la correlación lineal entre la variable X_i y el factor F_j, un valor alto en valor absoluto de un coeficiente indica que existe entre la variable y ese factor una estrecha relación.

Los factores comunes se obtendrán de modo que:

- El primero sea el que explique la mayor parte de variabilidad de los datos.
- El segundo esté no correlacionado con el primero (sea ortogonal) y sea el que explique la mayor parte de variabilidad restante que no fue explicada por el primero.

De este modo los factores comunes son linealmente no correlacionados y están ordenados de modo decreciente según el porcentaje de variabilidad total de los datos que vayan explicando cada uno de ellos.

1.2.2.2 Análisis de Correspondencia Simple (ACS)

El Análisis de Correspondencia Simple es una técnica la cual parte de una tabla de contingencia que agrupa 2 variables categóricas en sus diversas modalidades que guardan alguna relación entre ellas, donde cada casilla recoge la frecuencia en que se presentan las variables, luego a partir de este tabla se llevan a cabo técnicas similares a las de análisis factorial donde se consideran a los individuos como filas y a las columnas como variables, con la diferencia de que al hacer la trasposición de la tabla la información permanece sin alteración alguna, con el objetivo de tratar de explicar la dispersión que existe en la matriz de varianzas y covarianzas conocida como ***Matriz de Inercia***, en un número menor de factores, pero en este caso el análisis debe llevarse a cabo tanto para las filas como para las columnas con lo cual se estarían entonces realizando 2 análisis factoriales, uno para el espacio de las filas y otro para el espacio de las columnas.

Este análisis se basa en las siguientes consideraciones

- Tiene como punto de partida una tabla de contingencia de 2 entradas, donde se recogen las frecuencias de las observaciones y además se evidencia que existe una relación entre los niveles de las variables que la integran, puesto que no tendría ningún sentido buscar correspondencias donde no existe.
- Se trata de explicar la dispersión existente en la matriz de varianzas y covarianzas, a través de un número menor de factores, analizando tanto filas como columnas.
- La comparación entre filas y columnas se hace en términos relativos denominados perfiles fila y columna.
- Los n-puntos fila se representan es un espacio p-dimensional y los p-puntos columnas se representan en un espacio n-dimensional.

- Para cada espacio son calculados los autovalores y autovectores de la matriz de inercia los cuales sirven para definir el espacio de menor dimensión, que explique la información reflejada en la tabla de partida.
- Finalmente se puede llevar a cabo una representación grafica de las filas y columnas que permite analizar las correspondencias entre las categorías de las filas y las columnas.

- El número máximo de factores es del orden de:

$$(nrodefilas - 1)(nrodecolumnas - 1).$$

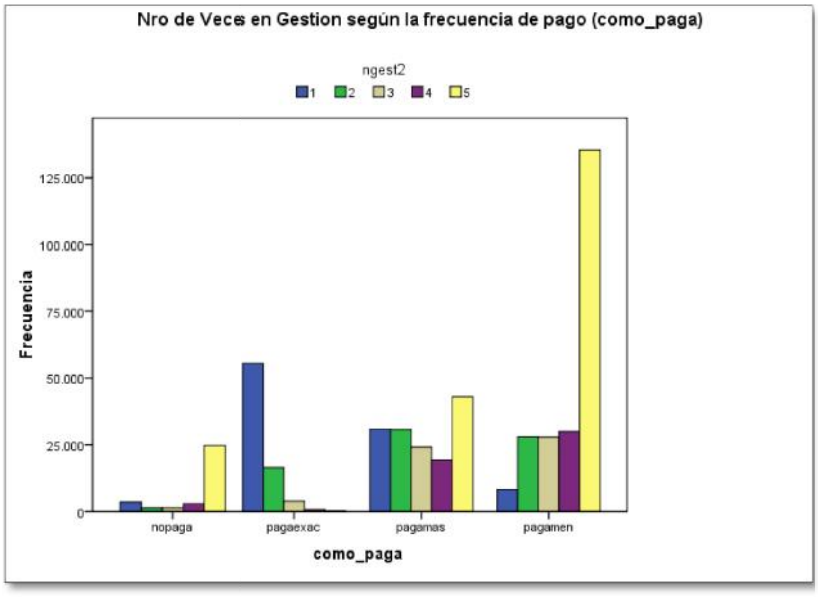
Veamos un ejemplo del conjunto de datos que hemos venido trabajando

En este caso veremos la asociación que existe entre la frecuencia de pago de los deudores, es decir, si estos pagan más, menos, el pago exacto o sencillamente no pagan y su relación con el número de veces que están en gestión.

Recordemos que el punto de partida es una tabla de contingencia de doble entrada:

	Número de Veces en Gestión				
Como Paga	**1**	**2**	**3**	**4**	**5**
No Paga	3.510	1.418	1.427	2.917	24.570
Paga Exacto	8.264	27.932	27.862	29.956	135.083
Paga Más	30.305	30.297	23.804	18.617	42.849
Paga Menos	55.349	16.407	3.889	816	164
Total	**97.428**	**76.054**	**56.982**	**52.306**	**202.666**

Tabla 4.



Grafica 3.

Desde un punto de vista grafico podemos ir obteniendo una idea acerca del problema, en este caso vemos en la *grafica 3* que aquellos individuos que pagan menos de lo que le corresponde están asociados a 5 o más veces en gestión, sin embargo hay una cantidad importante de clientes que pagan más del mínimo requerido y sin embargo están de igual forma 5 o más veces en gestión, mientras que una gran parte de aquellos clientes que están solo una vez en gestión hacen pagos exactos.

Una vez obtenida la tabla hacemos la prueba de independencia a fin de corroborar que no existe dependencia entre las variables y que estas son dependientes entre sí, lo que nos permite llevar a cabo el análisis de correspondencia, para ello vamos a usar una sencilla prueba chicuadrado donde la hipótesis nula es que las variables son independientes.

```
        Pearson's Chi-squared test

data:  table(datadef$como_paga, datadef$ngest2)
X-squared = 22858.2, df = 12, p-value < 2.2e-16
```

Tabla 5.

Este test nos permite rechazar la hipótesis nula y afirmar que existe dependencia entre las variables, y como se ve a continuación solo con 2 factores (dimensiones) explica la variabilidad total de la información.

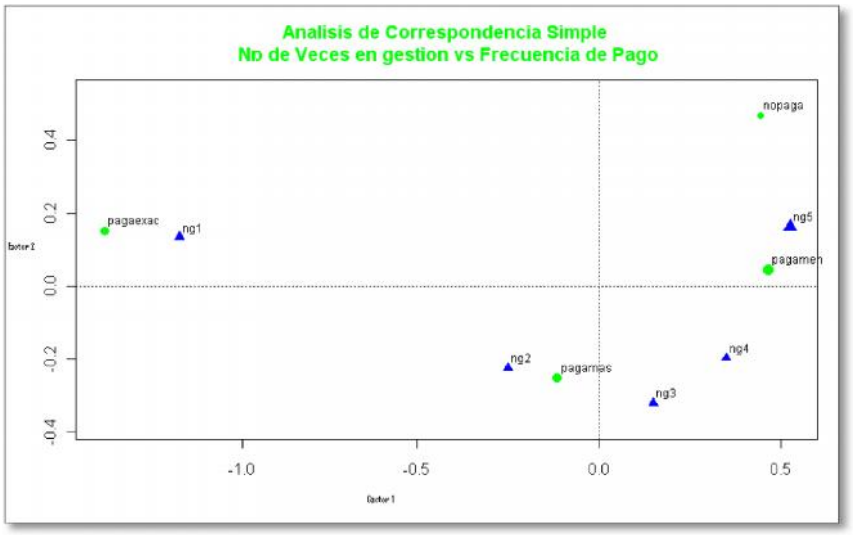
```
> summary(a)

Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.417134  91.4  91.4   *************************
 2      0.039024   8.6 100.0   **
 3      3e-05000   0.0 100.0
        -------- -----
 Total: 0.456188 100.0
```

Tabla 6.

Finalmente vemos como la *gráfica 4* se muestra la representación biespacial del análisis



Grafica 4.

Como vemos en la *grafica 4*, el primer factor separa a los individuos por su frecuencia de pago y opone a los que pagan al menos el pago mínimo, de los que no pagan o no alcanzan a pagarlo, en tanto que el factor 2 está asociado al número de veces en gestión y separa hacia el lado positivo a los que permanecen solo 1 y más de 5 veces en gestión mientras que del lado negativo están aquellos que están entre 2 y 4 veces en gestión.

1.2.2.3 Análisis de Correspondencia Múltiple (ACM)

En aquellos casos donde se requiere evaluar la relación entre 2 o más variables categóricas, se hace una expansión del ACS al ACM en este caso la aplicación se lleva a cabo sobre tablas de contingencia múltiples, en las que se tienen n filas y j variables, categóricas con $x_i\ i = 1,2,\cdots,j$ categorías, mutuamente excluyente y exhaustivas, de manera tal que cada está formada por 1 si el individuo se encuentra en la categoría en estudio y 0 en caso contrario.

Por tanto la tabla de datos es de la siguiente manera:

$$Z = [Z^1, Z^2, Z^3, \dots, Z^j].$$

Donde:

$$Z^i = \begin{bmatrix} Z^i_{11} & \cdots & Z^i_{1p_i} \\ \vdots & \ddots & \vdots \\ Z^i_{n1} & \cdots & Z^i_{np_i} \end{bmatrix}.$$

Con $Z^i_{wk} = \begin{cases} 1, & si\ el\ individuo\ w-esimo\ esta\ en\ la\ modalidad\ k \\ 0, & en\ otro\ Caso \end{cases}$

En el caso del análisis de correspondencia múltiple la matriz de análisis se basa en la matriz de Burt que se define como:

$$B = Z^t Z.$$

Con Z la matriz de datos originales. Entonces si D es la matriz de autovalores de B, la matriz de análisis se define por

$$V = \frac{1}{Q} D^{-1} B.$$

Donde: Q es el número total de categorías.

La matriz V es una matriz de varianzas y covarianzas o de inercia no centrada, una vez calculados los autovalores y autovectores de la matriz V se calculan las coordenadas de las modalidades de cada una de las variables y se obtiene el espacio factorial, así, como las puntuaciones empleadas y la interpretación de los resultados se hace de manera similar al ACS.

Veamos a través de un sencillo ejemplo con los datos que hemos venido trabajando, realizaremos el análisis con las variables: número de veces en gestión (NGEST), la cantidad de tarjetas que poseen los clientes CANT_TDC), la frecuencia de pago (COMO_PAGA).

En este caso vemos como solo con 2 dimensiones logramos explicar 99,6%, casi en la totalidad la variación en la matriz de inercias

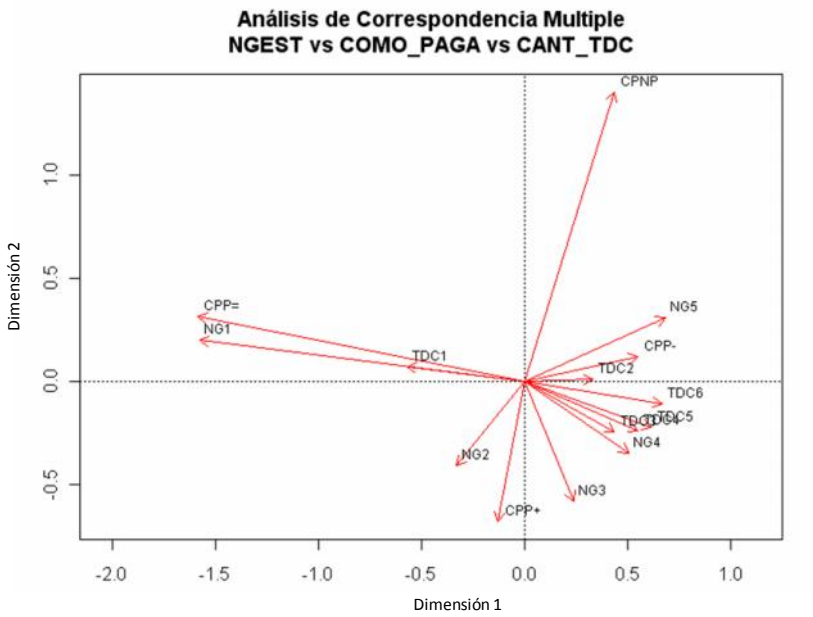
```
Principal inertias (eigenvalues):

     dim    value
[1,] 1      0.25294
[2,] 2      0.02201
[3,] 3      0.00298
[4,] 4      2.5e-050
[5,] 5      1e-05000
[6,]        --------
[7,] Total: 0.27179

 Diagonal inertia discounted from eigenvalues: 0.1133581
 Percentage explained by JCA in 2 dimensions: 99.6%
 (Eigenvalues are not nested)
 [Iterations in JCA: 50 , epsilon = 92.4274002]
```

Tabla 7.

Finalmente el diagrama de dispersión Biespacial



Grafica 5.

En este caso la gráfica 5 muestra como la dimensión 1 en su lado izquierdo opone perfectamente a aquellos individuos que estuvieron 1 vez en gestión y pagan el monto exacto que les corresponde pagar de aquellos es están más de 3 veces en gestión (lado superior derecho) y pagan menos o simplemente no cancelan sus obligaciones, mientras que el la dimensión 2 hace oposición sobre el número de tarjetas en la parte inferior de esta vemos los individuos con más de 3 TDC mientras que la parte superior retiene a los que poseen 1 y 2 TDC solamente.

Capítulo II

Los Datos

En el capitulo anterior se mostraron algunos ejemplos de aplicaciones sobre diferentes técnicas de Aprendizaje Estadístico, tales ejemplos fueron realizados a un conjunto de datos, los cuales serán descritos con más detenimiento en este capitulo, donde se hará una explicación de la manera en que fueron obtenidos, así como, de las manipulaciones que fueron necesarias para alcanzar los resultados que se mostraran más adelante.

2.1 Origen de Los datos

Los datos utilizados en este trabajo, provienen de Clientes deudores de Tarjetas de Crédito (TDC) de una Organización financiera de magnitud considerable en el país, los cuales presentaban atraso en la fecha tope al momento de cancelar sus pagos mínimos, y eran gestionados a través de un Call Center de Cobranzas, esto es, un lugar donde se realiza al cliente una llamada a fin de recordarle que se encuentra atrasado en la fecha tope para cancelar el pago mínimo y tomarle una promesa de pago sobre la deuda que tiene pendiente en su Tarjeta de Crédito, y así, seguir disfrutando los beneficios de financiamiento que ofrecen las TDC.

Los datos trabajados corresponden a un período de 15 meses desde Octubre de 2008 hasta diciembre de 2009 y en total se tuvieron cerca de 3 millones de registros, los cuales finalmente al ser agrupados y depurados se redujeron a cerca de 500 mil clientes que estuvieron siendo gestionados en el Call Center durante dicho periodo.

2.2 Obtención de los Datos.

Los datos presentados en este trabajo, fueron obtenidos a través del sistema de indicadores de gestión de la organización financiera, donde se lleva a cabo la medición de la efectividad en la recuperación de la TDC de clientes con atrasos en la fecha límite de su pago mínimo, este sistema toma registro de las gestiones telefónicas realizadas a los clientes de manera diaria, luego a través de un proceso de agrupamiento y

depuración va almacenando la evolución de la TDC mientras el cliente está en gestión, esto es, por ejemplo si varía su altura de mora o incrementa el monto por concepto de pago mínimo así como otros datos de interés. Una vez que se realiza la gestión telefónica, se otorga al cliente un plazo a fin de que cumpla la promesa de poner su obligación al día con la entidad financiera, finalmente al cerrar el mes se verifican los clientes que fueron gestionados y aquellos que realmente cumplieron con sus obligaciones y se determina la ***efectividad de la gestión de cobranza*** para dicho mes, el cual es un indicador que se obtiene del cociente entre el monto cancelado y el monto que se debía cancelar, y este se determina de la manera siguiente:

$$\%Efectividad = \frac{Monto\ Recuperado\ (Bs)}{Monto Asignado a Recuperar (Bs)} * 100.$$

Este indicador es el que permite evaluar la gestión en el proceso de recuperación de los deudores de TDC, además es el indicador que sirve de insumo en el proceso de recolección de los datos pues la información usada en este trabajo tiene su fuente en la determinación de la efectividad en la gestión de cobranza.

2.3 Determinación y construcción de las variables

De los datos generan los indicadores de la efectividad en la gestión de cobranza, se obtienen algunas de las variables que fueron usadas en el estudio que son: El Monto Asignado promedio, El Monto Cobrado Promedio, Cantidad máxima de TDC en gestión, Máximo nivel de mora registrado por el cliente, entre otras que serán explicadas de manera detallada más adelante.

Para la realización de este trabajo se tomaron en consideración 3 conjuntos de variables. El primer conjunto corresponde a aquellas variables que provenían directamente de la información para el cálculo de la efectividad, luego otro conjunto que era ajeno al proceso de gestión de cobranza pero que era de interés en el estudio y estaba vinculado directamente con el cliente, y un tercero constituido por las variables construidas a partir de los mismos datos obtenidos del proceso de gestión y que finalmente resultaron variables significativas durante el proceso de aprendizaje como se mostrara más adelante:

2.3.1 Variables Obtenidas Directamente de los Datos:

Monto Vencido Promedio (PROM_VCDO): Es el monto promedio de pago mínimo que tuvo el cliente mientras estuvo en gestión, este se calcula en base al promedio del monto total que debía cancelar el cliente para salir del sistema. En caso que tuviese 2 o más TDC en gestión en un mes se considera el monto total de la deuda a cancelar, luego se repetía este procedimiento para todos los meses que el cliente estuvo en gestión y finalmente se consideró el monto total a pagar entre el numero de meses que estuvo en gestión. Esta Variable esta medida en Bolívares.

$$PROM\_VCDO = \frac{\sum_{i=1}^{n} X_i}{n} \qquad i = 1,2,\ldots$$

Donde:

X_i: *Monto Asignado Total a recuperar en el i-esimo mes de gestión.*

N: Cantidad de veces que estuvo el cliente en Gestión.

Monto Cobrado Promedio (PROM_PAG): Es el monto promedio de pagos que tuvo el cliente mientras estuvo en gestión, se calcula en base al promedio de las cancelaciones que realizo el cliente durante el tiempo que estuvo en gestión en todas sus TDC, sin la necesidad que éstas lograran cubrir sus pagos mínimos, su forma de cálculo es similar al del monto promedio vencido. Esta Variable esta medida en Bolívares.

$$PROM\_PAG = \frac{\sum_{i=1}^{n} Y_i}{n} \qquad i = 1,2,\ldots$$

Donde:

Y_i: *Monto Total recuperado en el i-esimo mes de gestión.*

N: Cantidad de veces que estuvo el cliente en Gestión.

Cantidad de Tarjetas Créditos (CANT_TDC): Se refiere al número máximo de tarjetas que tuvo el cliente en gestión durante el tiempo que estuvo siendo gestionado por el Call Center de Cobranza.

$$CANT\_TCD = máx\{X_i, X_{i+1}, \dots, X_n\} \qquad i = 1,2,\dots$$

Donde:

X_i: Denota Cantidad de TDC Gestionadas en el i-esimo mes de gestión.

Número de veces en Gestión (NGEST): Esta Variable contabiliza el número de veces que estuvo el cliente durante los 15 meses del estudio.

Número de Pagos mientras estuvo en Gestión (NPAG): Esta Variable contabiliza el número de veces que el cliente realizó un pago aun y cuando no fuese suficiente para cumplir con el pago mínimo requerido, durante el tiempo que este estuvo en gestión.

Máxima Altura de Mora (MAX_MORA): Hace referencia al nivel máximo de mora que alcanzo el cliente en el periodo mientras estuvo en gestión. En caso que un cliente le estuvieran gestionando varias TDC en un mes se toma en cuenta aquella que tenga la máxima altura de mora para dicho meses y luego se considera la máxima mora que alcanzo durante el tiempo que estuvo siendo gestionado.

$$MAX\_MORA = máx\{X_i, X_{i+1}, \dots, X_n\} \qquad i = 1,2,\dots,15.$$

Donde:

X_i: Denota la Altura de Mora máxima alcanzada por el cliente en el i-esimo mes de gestión.

2.3.2 Variables Ajenas al proceso de obtención de los Datos:

Limite o Capacidad de Endeudamiento del Cliente (LIMITE): Para la determinación de esta variable se consideró la suma del los limites que tenía el cliente en las TDC que estuviesen en gestión, luego se consideró el máximo durante el tiempo que el cliente estuvo en gestión, es decir, que esta variable es el indicativo del monto total que disponía el cliente para hacer uso de un financiamiento en su TDC y su cálculo es ajeno al proceso de gestión de cobranza pues esta variable no tiene ninguna relevancia para dicho proceso. Esta Variable está medida en Bolívares.

$$LIMITE = máx\left\{\sum_{i=1}^{h} y_{ij}\right\} \qquad j = 1,2,\dots,K$$

Donde:

Y_{ij}: *Denota el Límite de crédito de la i- esima Tarjeta en el j- esimo mes de gestión.*

Sexo del Cliente (SEXO): Hace referencia al Sexo del cliente y esta se obtiene a través de los datos del cliente como tal y no del proceso de gestión de cobranza.

2.3.3 Variables construidas a partir del de obtención de los Datos:

Frecuencia de Pago (COMO_PAGA): Esta variable es el resultado de clasificar en 4 grupos o categorías la relación entre la variación que existe del Monto Cobrado Promedio (PROM_PAG) el cual no considera los reversos hechos a las TDC los cuales son negativos, y el Monto Vencido Promedio (PROM_VCDO). Según el valor de tal variación se realiza la clasificación a saber:

$$DIF = \frac{(PROM\_PAG) - (PROM\_VCDO)}{(PROM\_PAG)}, \qquad \square PROM\_PAG > 0$$

Entonces definida la relación entre la diferencia, para la variable COMO_PAGA la categorización es la siguiente:

- Si DIF > 0 → COMO_PAGA = PAGA MAS (P+).
- Si DIF = 0 → COMO_PAGA = PAGA IGUAL (P=).
- Si DIF < 0 → COMO_PAGA = PAGA MENOS (P-).
- Si PROM_PAG = 0 → COMO_PAGA = NO PAGA (NP).

Número de Ascensos en Mora (NMOV): Esta es una variable que contabiliza el número de veces que el cliente tuvo un nivel de mora mayor con respecto al mes anterior del estudio, y parte de una variable dicotómica construida a través de la siguiente expresión.

$$MOV_i = \begin{cases} 1, & ALT\_MORA_i > ALT\_MORA_{i-1} \\ 0, & OtroCaso \end{cases}$$

Y entonces finamente NMOV es de la forma:

$$NMOV = \sum_{i=1}^{n} MOV_i.$$

2.4 Análisis Descriptivo:

En esta sección se hará el análisis descriptivo de las variables que se utilizaron durante el proceso de aprendizaje.

Antes de Aclarar esto, es necesario inducir un par de conceptos conocidos como medidas de distribución que nos ayudaran un poco más en el entendimiento de los análisis descriptivos de las variables continuas que fueron usadas en el proceso de aprendizaje.

Asimetría

Esta medida nos permite identificar si los datos se distribuyen de forma uniforme alrededor del punto central o promedio, la asimetría puede ser de tres maneras, las cuales define de forma concisa como están distribuidos los datos respecto al eje de asimetría, es decir con respecto al centro de la distribución donde se encuentra la mayor concentración de observaciones.

Positiva: Cuando la mayoría de los datos se encuentran por encima del valor de la media aritmética.

Simétrica: Cuando se distribuyen aproximadamente la misma cantidad de valores en ambos lados de la media.

Negativa: Cuando la mayor cantidad de datos se aglomeran en los valores menores que la media.

El *Coeficiente de asimetría* g_1, se representa mediante la ecuación:

$$\frac{\bar{}}{\bar{}},$$

donde: x_i cada uno de los valores observados, $\overline{X}$ la media muestral y n_i la frecuencia de cada valor.

Relación entre en valor de g_1 y la simetría de los datos.

- $g_1 = 0$: La distribución es **Simétrica**, es decir, existe aproximadamente la misma cantidad de valores a ambos lados de la media, en el campo real es bastante difícil encontrar una asimetría exactamente igual a cero con lo cual existe una tolerancia y aquellos valores de simetría con una distancia ± 0.5 respecto del cero son consideradas distribuciones simétricas

- $g_1 > 0$**:** Se dice que la distribución es **Positiva** por lo que los valores se tienden a agruparse más en la parte izquierda de la media.

- $g_1 < 0$**:** La curva es asimétricamente **Negativa** por lo que los valores se tienden agruparse más en la parte derecha de la media.

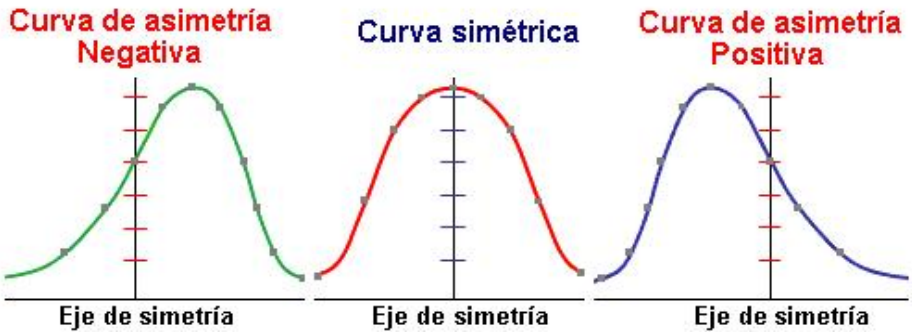


Figura 1.

Curtosis

Esta medida determina el grado de concentración que presentan los valores en la región central de la distribución. Por medio del Coeficiente de Curtosis, podemos identificar si existe una gran concentración de valores (Leptocúrtica), una concentración normal (Mesocúrtica) ó una baja concentración (Platicúrtica).

El coeficiente de Curtosis g_2 se calcula de la siguiente manera:

$$\frac{-}{-},$$

donde: x_i cada uno de los valores observados, $\bar{X}$ la media muestral y n_i la frecuencia de cada valor.

Relación entre en valor de g_2 y la concentración de los datos.

- ***g_2 = 0:*** la distribución es Mesocúrtica, al igual que en la asimetría, es bastante difícil encontrar un coeficiente de Curtosis de cero (0), por lo que se suelen aceptar los valores cercanos (± 0.5 aprox.).
- ***g_2 = 0:*** la distribución es Leptocúrtica.
- ***g_2 = 0:*** la distribución es Platicúrtica.

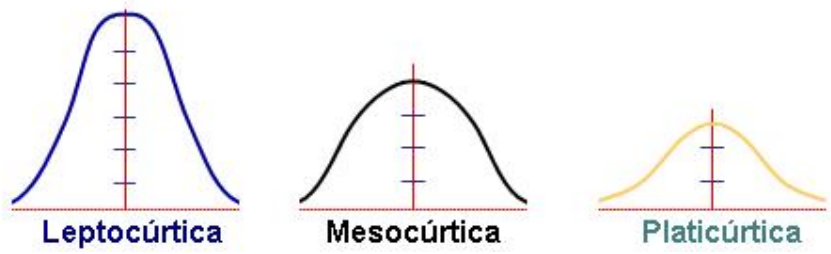


Figura 2.

2.4.1 Estadísticos Descriptivos para las variables Continuas

Del proceso de obtención de los datos se obtuvieron variables tanto cualitativas como cuantitativas, a continuación se presente el resumen de las estadísticas descriptivas para las variables cuantitativas, en las cuales se muestran los indicadores de: Promedio, Máximo, Mínimo, Desviación Estándar y las comentadas anteriormente Asimetría y Curtosis.

Variable	N	Mínimo	Máximo	Promedio	Desv Estándar	Asimetría	Curtosis
PROM_VCDO	485.436	10,00	23.060,52	1.148,36	1.592,39	4,65	33,78
PROM_PAG	485.436	0,00	50.716,97	781,37	1.113,70	5,55	75,72
LIMITE	485.436	250,00	136.800,00	9.517,98	10.372,77	3,58	21,21

Tabla 8

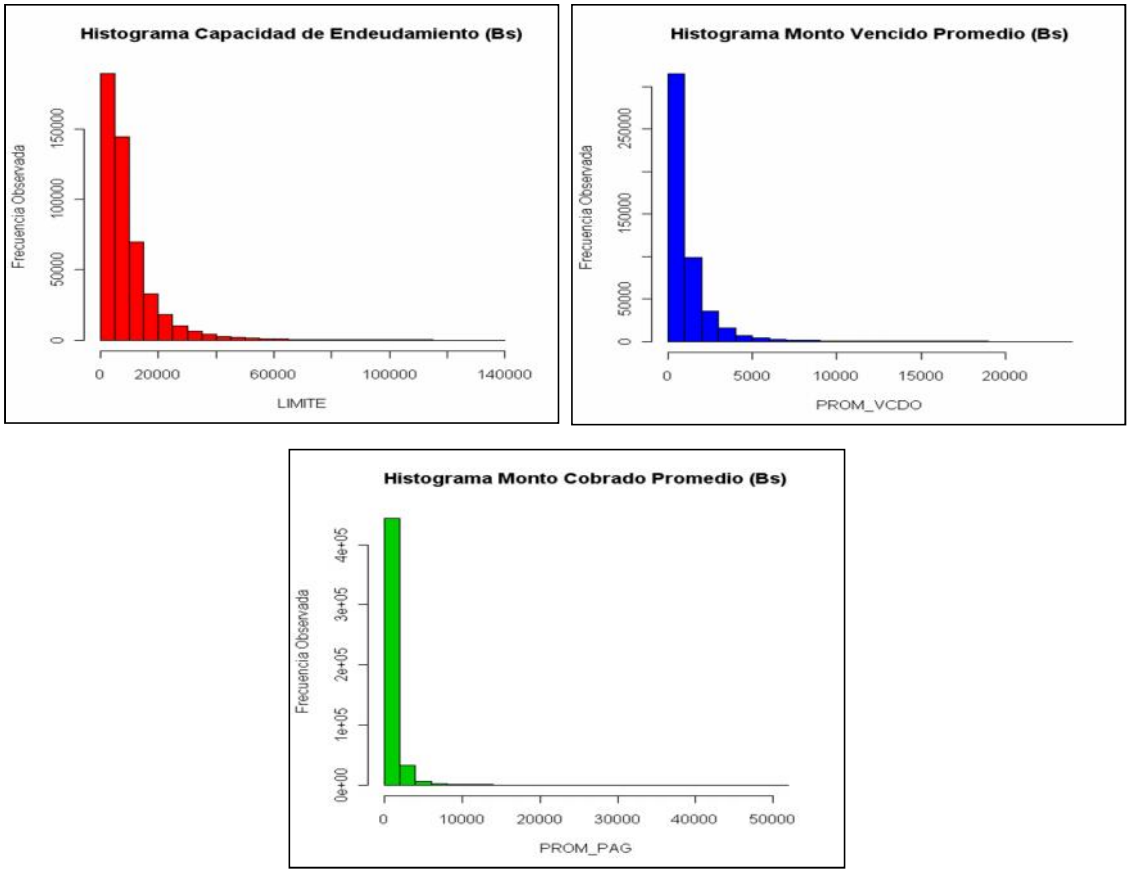
Luego del proceso de depuración se obtuvieron finalmente clientes 485.436 reflejados en la *Tabla 8*, que fueron gestionados durante el período de estudio por el Call Center de Cobranza, a fin que se pusieran al día en sus obligaciones, en el caso de las variables continuas vemos en la tabla 8 que todas son variables con desviaciones estándar altas lo que implica directamente una varianza alta y asimétricas positivas, es decir, que tienen la mayor concentración de los datos hacia el lado izquierdo de la distribución (asimetría mayor a 1) y con una forma Leptocúrtica bastante pronunciada como veremos en los gráficos más adelante debido al valor elevado de la Curtosis en las tres variables.

Por otro lado vemos en la *Tabla 8* que el monto promedio asignado es de 1.148, 36 Bs, el cual es muy superior al Monto Cobrado Promedio que es de 781.37 Bs. en tanto que los clientes que fueron gestionados por el Call Center de cobranza tienen una capacidad de endeudamiento promedio de 9.517,98 Bs. De igual forma se observa en la tabla hubo clientes que no cancelaron sus obligaciones por lo cual el monto mínimo cobrado es 0 Bs.

Vemos también que la desviación estándar de las variables es alta esto motivado al hecho que el rango (diferencia entre el valor mínimo y el máximo) de las variables en términos del problema que se está estudiando es si se quiere elevado, así como una gran diferencia entre el valor promedio de las variables y el máximo de las mismas, esta situación se sucede debido al hecho que al sistema de gestión de cobranza ingresaban

aquellos clientes que tuvieran un monto asignado a recuperar superior a cero aun y cuando el proceso de gestión involucrara un gasto mayor que la deuda en sí, actualmente se están tomando acciones a fin de evitar realizar gestiones a clientes cuyo costo de gestión sea superior a su deuda y se les hace la gestión a través de otros mecanismos que acarren costos menores en el proceso de gestión.

Histogramas para las variables Continuas



Grafica 6.

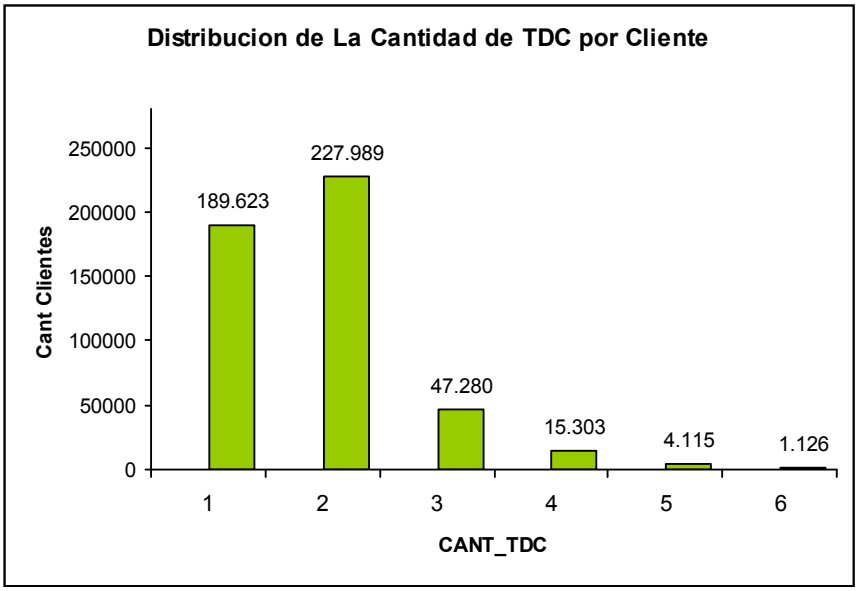
Como se observa en los histogramas de las variables continuas (*Gráfica 6*), se confirman las distribuciones positivas de las mismas así como una forma Leptocúrtica, ya que vemos que las variables están concentradas hacia el lado izquierdo de la distribución, con frecuencias observadas muy elevadas hacia dicho lado, sobre todo el monto cobrado, esto obedece a la gran varianza que presentan los datos y el hecho que un cliente no necesariamente puede cancelar el monto exacto que tiene asignado a recuperar sino que puede cancelar más si así lo desea como veremos más adelante en la variable frecuencia de pago, sin embargo aquellas

observaciones que parecieran ser atípicas desde un punto de vista individual en cada variable y hacen que las concentraciones en los histogramas se den hacia el lado izquierdo de la media, son observaciones que de manera conjunta es decir sus 3 variables continuas se comportan de manera razonable, lo cual les permite estar dentro del proceso de aprendizaje, ya que se trata de clientes con capacidades de endeudamiento elevada que se relacionan directamente con montos asignados a recuperar y montos pagados altos.

2.4.2 Variables Categóricas

Para el caso de las variables categóricas se mostraran las tablas de distribución de frecuencia de las variables conjunto con un grafico de barras que nos permitirá ver la distribución de cada una de las categorías de las variables usadas durante el proceso de aprendizaje.

2.4.2.1 Cantidad de TDC

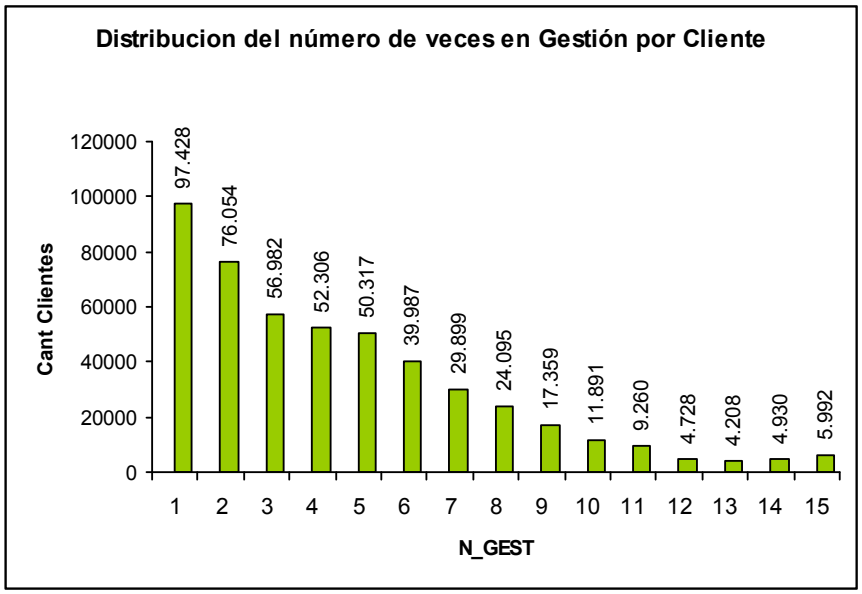


CANT_TDC	Cant Clientes	Frec Relativa	Frec Acumulada
1	189.623	39,06%	39,06%
2	227.989	46,97%	86,03%
3	47.280	9,74%	95,77%
4	15.303	3,15%	98,92%
5	4.115	0,85%	99,77%
6	1.126	0,23%	100,00%
Total	**485.436**	**100,00%**	

Figura 3.

En el caso de la distribución del numero de de tarjetas vemos en la *figura 3* que casi el 50% de los clientes que fueron gestionados en el Call Center de Cobranzas llegaron a tener 2 tarjetas en gestión, y cerca del 90% de los clientes tuvieron un máximo de 2 tarjetas en gestión luego vemos como además existen clientes que tuvieron 6 TDC en gestión durante un mes pero la mayor concentración de clientes está entre 1 y 2 TDC en gestión.

2.4.2.2 Número de veces en Gestión (NGEST):

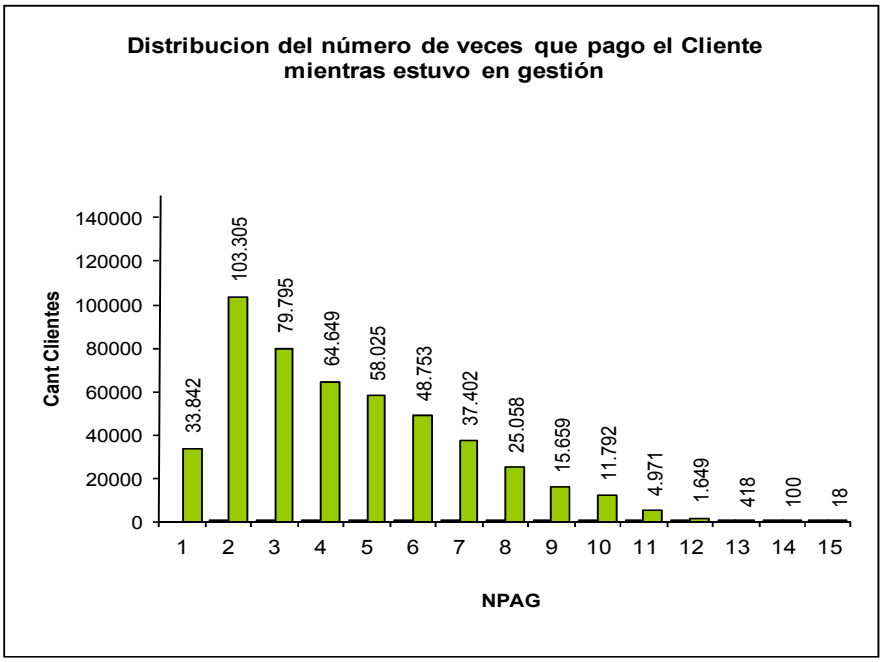


N_GEST	Cant Clientes	Frec Relativa	Frec Acumulada
1	97.428	20,07%	20,07%
2	76.054	15,67%	35,74%
3	56.982	11,74%	47,48%
4	52.306	10,78%	58,25%
5	50.317	10,37%	68,62%
6	39.987	8,24%	76,85%
7	29.899	6,16%	83,01%
8	24.095	4,96%	87,98%
9	17.359	3,58%	91,55%
10	11.891	2,45%	94,00%
11	9.260	1,91%	95,91%
12	4.728	0,97%	96,88%
13	4.208	0,87%	97,75%
14	4.930	1,02%	98,77%
15	5.992	1,23%	100,00%
Total	**485.436**	**100,00%**	

Figura 4.

Para el número de veces en gestión (NGEST) vemos en la *figura 4* que ocurre una mayor concentración entre 1 y 5 veces en gestión casi el 70% se encuentra entre este rango de entradas al sistema de gestión luego tenemos un poco más del 25 % de los clientes que estuvieron en gestión más de 6 veces hasta una pequeña proporción 1,02% de clientes que estuvieron durante todo el tiempo del estudio es decir los 15 meses.

2.4.2.3 Número de Pagos mientras estuvo en Gestión (NPAG):

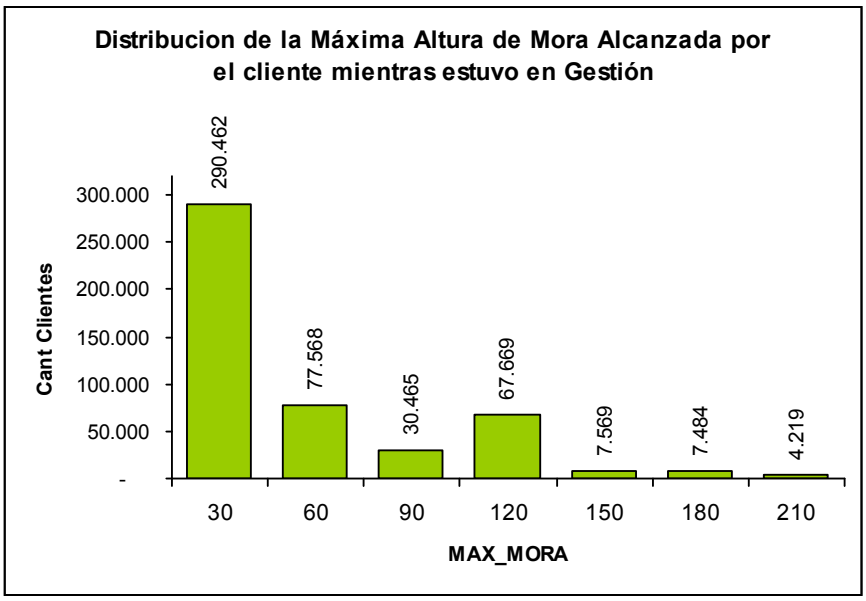


NPAG	Cant Clientes	Frec Relativa	Frec Acumulada
0	33.842	6,97%	6,97%
1	103.305	21,28%	28,25%
2	79.795	16,44%	44,69%
3	64.649	13,32%	58,01%
4	58.025	11,95%	69,96%
5	48.753	10,04%	80,00%
6	37.402	7,70%	87,71%
7	25.058	5,16%	92,87%
8	15.659	3,23%	96,10%
9	11.792	2,43%	98,53%
10	4.971	1,02%	99,55%
11	1.649	0,34%	99,89%
12	418	0,09%	99,98%
13	100	0,02%	100,00%
14	18	0,00%	100,00%
Total	**485.436**	**100,00%**	

Figura 5.

Para el numero de pagos que tuvieron los clientes mientras se encontraban en gestión, vemos reflejado en la *figura 5* como mas de la mitad de estos 52,77% pagaron al menos 3 veces mientras estuvieron en gestión, luego vemos como después que los clientes están más de 7 veces en gestión la proporción de los que pagan va disminuyendo conforme para el numero de veces en gestión al punto que no hubo clientes que si bien es cierto permanecieron 15 meses en gestión llegaron a pagar en todas las oportunidades puesto que el máximo número de pagos es 14 y es una proporción casi despreciable en relación a los que estuvieron igual cantidad de veces en gestión.

2.4.2.4 Máxima Altura de Mora (MAX_MORA):

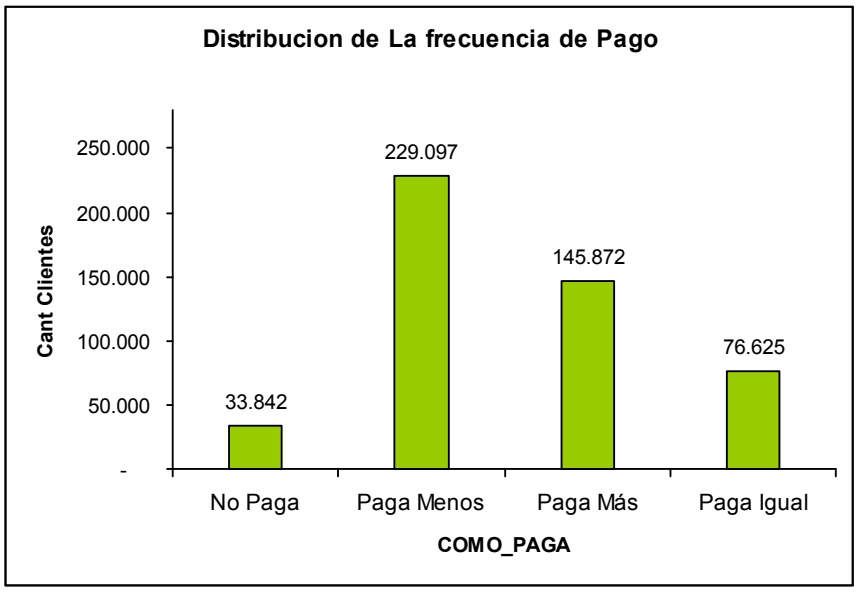


MAX_MORA	Cant Clientes	Frec Relativa	Frec Acumulada
30	290.462	59,84%	59,84%
60	77.568	15,98%	75,81%
90	30.465	6,28%	82,09%
120	67.669	13,94%	96,03%
150	7.569	1,56%	97,59%
180	7.484	1,54%	99,13%
210	4.219	0,87%	100,00%
Total	**485.436**	**100,00%**	

Figura 6.

De los clientes que estuvieron siendo gestionados durante el período de estudio casi el 60% de estos alcanzaron una altura de mora máxima de 30 días, y un poco más del 75% solo llego a estar atrasado en 2 cuotas de sus TDC (60 días), finalmente vemos (*Figura 6*) como existe un cambio brusco entre la altura de mora entre 120 y 150 días y como disminuye de manera considerable el número de clientes gestionados, esto obedece a una estrategia en las gestión de cobranza que se planteo para aquel entonces, en la cual aquellos clientes que llegaban a la altura de mora de 120 días les era ofrecido un plan de refinanciamiento a fin que pudiera cumplir con sus obligaciones y en algunos casos sus tarjetas volvían a ser llevadas a una altura de mora de 0 días una vez alcanzada la firma de una esquema de pago.

2.4.2.5 Frecuencia de Pago (COMO_PAGA):



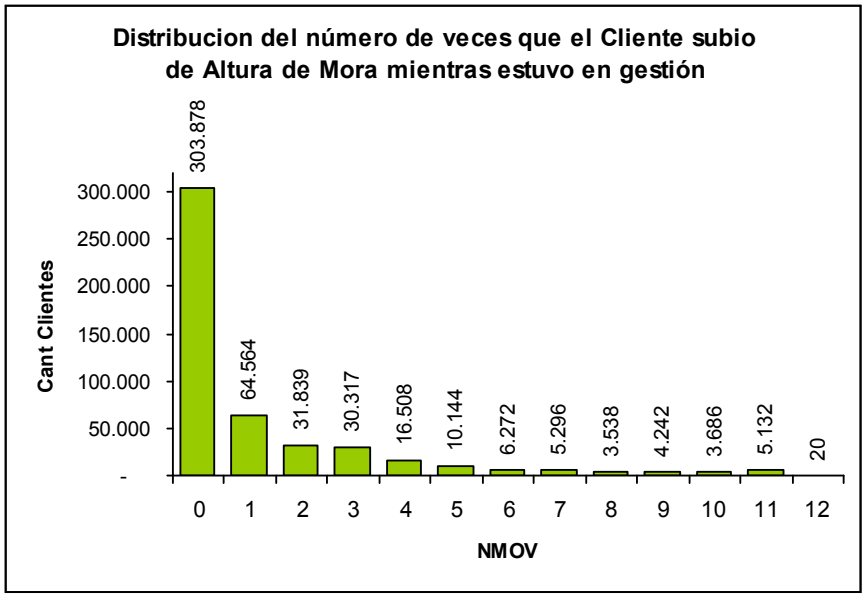
COMO_PAGA	Cant Clientes	Frec Relativa	Frec Acumulada
No Paga	33.842	6,97%	6,97%
Paga Menos	229.097	47,19%	54,17%
Paga Más	145.872	30,05%	84,22%
Paga Igual	76.625	15,78%	100,00%
Total	**485.436**	**100,00%**	

Figura 7

Para el caso de la frecuencia de pago vemos como en la *figura7* los clientes que pagan menos o no pagan ocupan más de la mitad de los que estuvieron en el estudio un total de 54,17% de estos se encuentran en esta condición, y solo un 15,78% de los clientes pagan exactamente del monto asignado a recuperar que tienen, en tanto que una

tercera parte de los clientes que estuvieron en gestión pagan más del monto asignado que tienen asignados.

2.4.2.6 Número de Ascensos en Mora (NMOV):

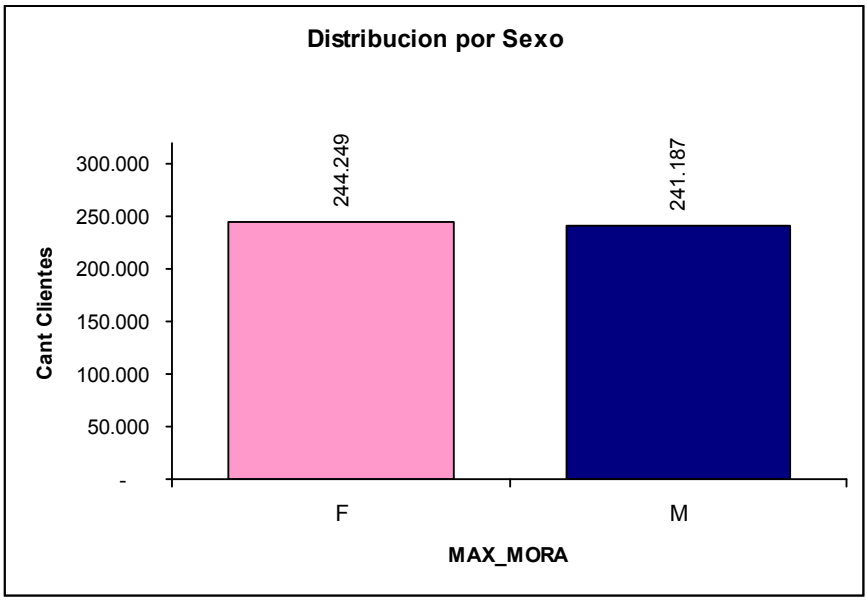


NMOV	Cant Clientes	Frec Relativa	Frec Acumulada
0	303.878	62,60%	62,60%
1	64.564	13,30%	75,90%
2	31.839	6,56%	82,46%
3	30.317	6,25%	88,70%
4	16.508	3,40%	92,10%
5	10.144	2,09%	94,19%
6	6.272	1,29%	95,49%
7	5.296	1,09%	96,58%
8	3.538	0,73%	97,31%
9	4.242	0,87%	98,18%
10	3.686	0,76%	98,94%
11	5.132	1,06%	100,00%
12	20	0,00%	100,00%
Total	**485.436**	**100,00%**	

Figura 8

Para el caso del número de veces que los clientes entran en gestión y luego suben de mora la *figura 8* muestra como solo un 25 % alcanzan en subir de mora más de una vez, mientras que los clientes que no se permiten ascensos en mora alcanzan una buena proporción de la población de estudio, es decir, son cliente que si bien pueden ingresar al sistema más de una vez se mantienen en la altura de mora que ingresa o salen del sistema lo cual no les permite registra ascensos en mora mientras están en gestión.

2.4.2.7 Sexo



SEXO	Cant Clientes	Frec Relativa
F	244.249	50,32%
M	241.187	49,68%
Total	**485.436**	**100,00%**

Figura 9.

En el caso de la variable sexo, se muestra en la *Figura 9* como no existe una diferencia entre el sexo del tarjetahabiente y se podría corroborar que esta proporción es casi igual para ambos sexo, en este caso y en vista que no existen diferencias significativas entre las proporciones de hombres y mujeres que son gestionadas en el Call Center de Cobranzas, más adelante no se hará uso de esta variable ya que no presenta ningún aporte en el estudio, salvo el entendimiento que para esta organización financiera no hay una diferencia marcada entre el sexo del tarjetahabiente y las características de estos en el proceso de aprendizaje estadístico.

Durante este proceso de recolección hubo un conjunto de variables como la edad o los ingresos de los clientes que se piensan pueden contribuir a los procesos de aprendizaje mas sin embargo no se encontraban actualizadas (ingreso) o era inconsistente (edad).

Capítulo III

Resultados

Luego de haber descritos las técnicas de Aprendizaje Estadístico que se aplicaron en este trabajo así como las variables a ser utilizadas, en este capítulo se llevará a cabo el desarrollo del proceso de aprendizaje estadístico el cual consta de la aplicación de las diversas técnicas introducidas en el capítulo 1 al grupo de datos descritos en el capitulo anterior, el proceso de aprendizaje se realizara en 2 etapas una primera de aprendizaje No Supervisado y una segunda de Aprendizaje Supervisado usando alguno de los resultados obtenidos en la etapa no supervisada.

3.1 Aprendizaje No Supervisado

3.1.1 TWO STEPS CLÚSTER

En la primera parte de las técnicas de aprendizaje no supervisado se realizo un proceso de clasificación usando la técnica de TWO STEPS CLÚSTERS, la cual permite a partir de un conjunto de variables de origen bien sean continuas y/o categóricas generar grupos con características similares.

En este caso, para la formación de los grupos se utilizaron las variables continuas: LIMITE, PROM_VCDO y PROM_PAG, y en el caso de las variables categóricas fueron usadas: NMOV, MAX_MORA, CANT_TDC, del agrupamiento de este conjunto de variables resultaron 4 grupos o clústers con características bien diferenciadas en sus variables categóricas, esto trae como consecuencia el hecho que dentro de los clústers exista una gran dispersión en las variables continuas puesto que las variables que tuvieron mayor peso en la formación de los clúster fueron las categóricas como se muestra en los resultados a continuación:

3.1.1.1 Distribución de Clientes por Clúster.

Cluster	Cant Clientes	Frec Relativa
1	140.150	28,9%
2	83.833	17,3%
3	150.217	30,9%
4	111.236	22,9%
Total	**485.436**	**100,0%**

Tabla 9.

Como observamos en la *Tabla 9* los clúster 1 y 3 son los que tienen mayor cantidad de clientes un porcentaje cercano al 60% se encuentra entre estos 2 clúster.

3.1.1.2 Centroides de las Variables Continuas:

Cluster	PROM_VCDO		PROM_PAG		LIMITE	
	Promedio	Desv. Estandar	Promedio	Desv. Estandar	Promedio	Desv. Estandar
1	447,29	514,32	501,65	599,99	6.400,98	6.745,93
2	985,43	1.016,06	889,13	1.161,61	9.234,49	9.291,60
3	1.158,30	1.086,86	1.268,75	1.467,93	13.663,31	13.600,73
4	2.141,03	2.595,10	394,40	674,20	8.060,84	7.727,27
Total	**1.148,36**	**1.592,39**	**781,37**	**1.113,70**	**9.517,98**	**10.372,77**

Tabla 10.

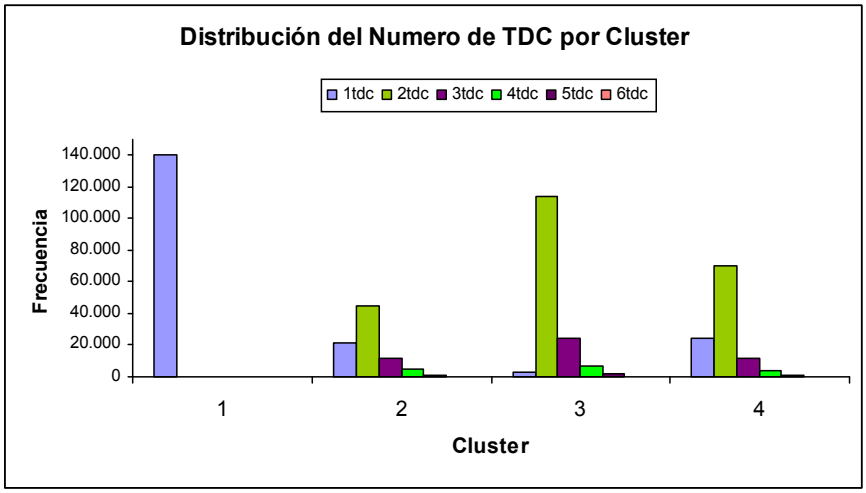
En el caso de las variables continuas vemos en la *tabla 10* que si bien es cierto que los clústers presentan una desviación estándar elevada los valores de Monto Promedio Vencido y Monto Cobrado promedio dentro de cada uno nos muestran comportamientos característicos dentro de ellos, a saber:

- El clúster 1 está formado por individuos que tienden a pagar un poco más del monto que les corresponde, y son individuos con montos asignados a recuperar de un promedio de 450 Bs. y cuyas capacidades de endeudamiento no superan los 6500 Bs.
- El clúster 2 está formado por individuos que en promedio cancelan menos del monto que les corresponde más sin embargo cancelan sus obligaciones aun y cuando en algunos casos no es suficiente para cubrir dicha deuda y sus capacidades de endeudamiento es un poco más elevada que los pertenecientes al clúster 1.

- El clúster 3 tiene comportamiento similar al 1 solo que los diferencia el hecho de presentar montos más elevados en las variables continuas sin embargo tienen la características fundamental que son clientes cuyo promedio de pagos es superior a los montos asignados a recuperar y su capacidad de endeudamiento es superior al promedio global del total de los clientes.
- El clúster 4 es aquel que contiene a los individuos cuyos pagos no alcanzan a cubrir sus obligaciones y su promedio de pago no alcanza a ser 20% del total que les corresponde calcular con cual se puede inferir que este clúster contiene a aquellos clientes con dificultades de recuperación de sus deudas y que presentaron mayores problemas al momento de ser gestionados.

Distribución de las Variables categóricas dentro de los clústers:

3.1.1.3 Cantidad de TDC (CANT_TDC) por Clúster

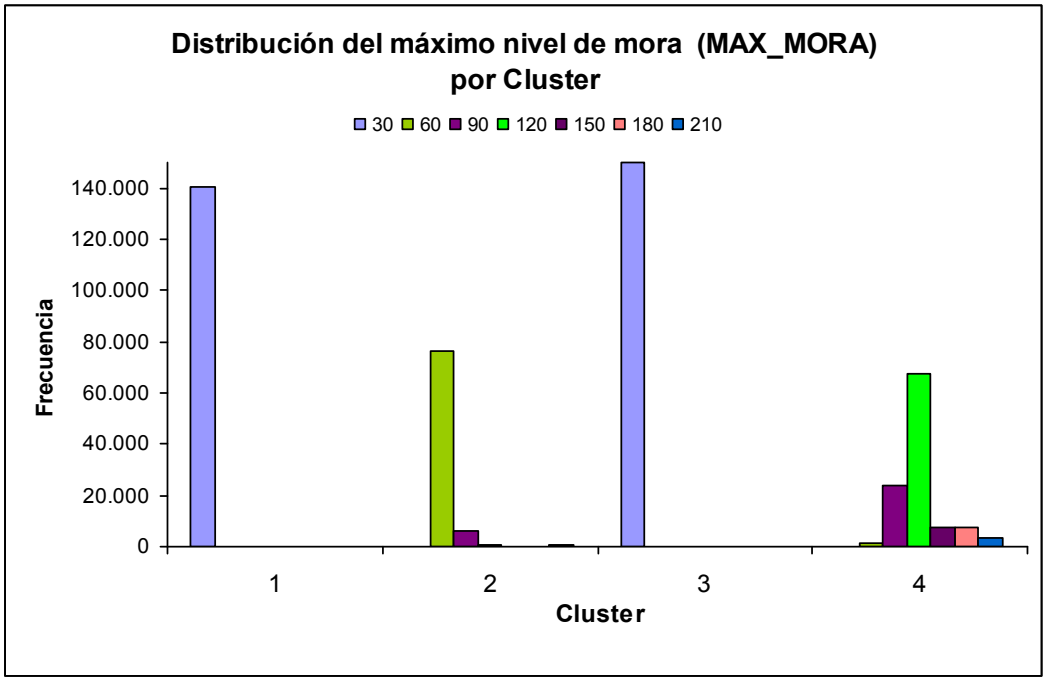


	CANT_TDC					
Cluster	**1tdc**	**2tdc**	**3tdc**	**4tdc**	**5tdc**	**6tdc**
1	140.150	0	0	0	0	0
2	21.804	44.521	11.256	4.406	1.368	478
3	3.091	113.724	24.244	6.964	1.811	383
4	24.578	69.744	11.780	3.933	936	265
Total	**189.623**	**227.989**	**47.280**	**15.303**	**4.115**	**1.126**

Figura 10.

Recordando los centroides de los clústers luego de ver en la *figura 10* la distribución del numero de TDC en los mismos, entendemos el comportamiento de estos centroides, ya que el clúster 1 está formado solo por clientes que tienen solo una TDC mientras que el clúster 3 que también son clientes que cancelan en promedio un pago mayor al que les corresponde contiene a aquellos clientes que tienen más de 2 TDC, en tanto que los clústers 3 y 4 tiene clientes que posen desde 1 hasta 6 TDC pero que no llegan a pagar el monto que tienen asignado.

3.1.1.4 Máximo nivel de mora (MAX_MORA) por Clúster



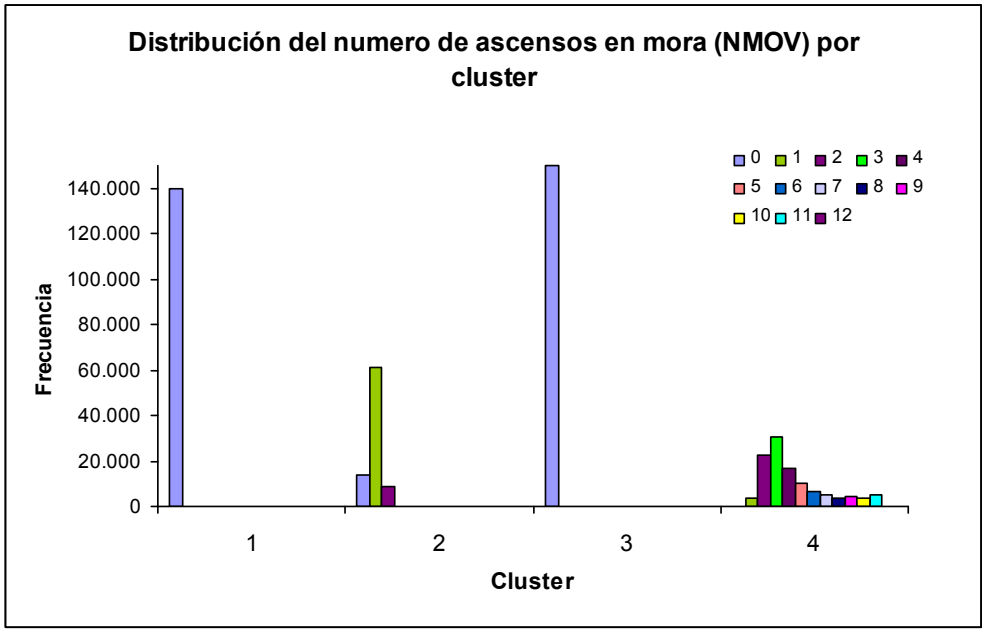
	MAX_MORA						
Cluster	**30**	**60**	**90**	**120**	**150**	**180**	**210**
1	140.150	0	0	0	0	0	0
2	331	76.323	6.239	383	11	21	525
3	149.978	143	37	10	48	1	0
4	3	1.102	24.189	67.276	7.510	7.462	3.694
Total	**290.462**	**77.568**	**30.465**	**67.669**	**7.569**	**7.484**	**4.219**

Figura 11.

Al ver la distribución de la máxima mora dentro de los clúster se observa en la *figura 11* como los clústers 1 y 3 que están asociados a clientes que se podrían denomina

buenos deudores en tanto que el clúster 2, conformado por aquellos clientes que pagan menos del monto que los corresponde, está formado por clientes entre 60 y 90 días de mora y finalmente el clúster 4 formado por clientes que posen alturas de moras superiores a los 90 días

3.1.1.5 Numero de Ascensos en Mora (NMOV)



	Cluster				
NMOV	1	2	3	4	Total
0	140.150	13.502	150.172	54	**303.878**
1	0	61.165	35	3.364	**64.564**
2	0	9.082	1	22.756	**31.839**
3	0	62	1	30.254	**30.317**
4	0	11	1	16.496	**16.508**
5	0	5	0	10.139	**10.144**
6	0	3	4	6.265	**6.272**
7	0	0	2	5.294	**5.296**
8	0	0	0	3.538	**3.538**
9	0	3	1	4.238	**4.242**
10	0	0	0	3.686	**3.686**
11	0	0	0	5.132	**5.132**
12	0	0	0	20	**20**

Figura 12.

Para el caso del número de ascensos en mora, vemos en la *figura 12* que los clúster 1 y 3 están formados por clientes que no tienen número de ascensos en mora, el clúster 2 correspondiéndose con la distribución de la variable MAX_MORA está formado por

clientes con hasta 2 movimientos es decir clientes lograron de manera consecutiva subir 2 niveles de mora, y finalmente el clúster 4 contiene a aquellos clientes que se mantienen en constante ascenso en mora y que finalmente salieron del estudio pues sus TDC terminaron llevándose a castigo.

3.1.2 Análisis de Correspondencia Simple y Múltiple

En esta parte del trabajo se mostraran los resultados obtenidos de los análisis de correspondencia simple y múltiple que se hicieron entre algunas variables, por ejemplo para el proceso de clasificación mostrado anteriormente no fue utilizada la variable número de veces en gestión, puesto que más adelante esta será utilizada en el proceso de aprendizaje estadístico supervisado, sin embargo necesitamos determinar cuál será la relación de esta variable con el clúster en el que finalmente quedo ubicado el cliente, por ejemplo, ya que esto nos servirá de interés para ver la importancia de esta variable durante el proceso de aprendizaje supervisado.

3.1.2.1 Análisis de Correspondencia Simple entre el número de veces en gestión y el Clúster de Pertenencia

Recordemos que este análisis parte de una tabla de contingencia en este caso de 2 entradas y una prueba Chi-cuadrado que nos permite determinar si existe asociación entre las variables, lo que haría que tenga sentido este tipo de análisis ya que de rechazar la hipótesis de no correlación el estudio no tendría mayor sentido.

NGEST	Cluster de Pertenencia				Total
	1	2	3	4	
1	70.246	1.048	26.096	38	97.428
2	31.888	6.252	37.703	211	76.054
3	14.724	9.913	30.040	2.305	56.982
4	9.002	12.646	22.738	7.920	52.306
5	5.489	13.943	14.901	15.984	50.317
6	3.462	12.986	9.120	14.419	39.987
7	2.281	10.467	5.090	12.061	29.899
8	1.380	7.381	2.487	12.847	24.095
9	764	4.570	1.197	10.828	17.359
10	467	2.549	524	8.351	11.891
11	238	1.226	197	7.599	9.260
12	114	550	81	3.983	4.728
13	56	214	33	3.905	4.208
14	23	69	9	4.829	4.930
15	16	19	1	5.956	5.992
Total	**140.150**	**83.833**	**150.217**	**111.236**	**485.436**

Estadístico	Valor	DF	p-value
Pearson Chi-Square	334308,278[a]	42	,000
Likelihood Ratio	356415,517	42	,000
Linear-by-Linear Association	140518,719	1	,000
N of Valid Cases	485436		

tabla 11.

En este caso las hipótesis vendrían dadas de la siguiente Manera:

H_0: Las Variables Número de Veces en Gestión y Clúster de Pertenencia son independientes.

H_1: Las Variables Número de Veces en Gestión y Clúster de Pertenencia son dependientes

Como vemos el valor p del estadístico Chi-cuadrado de Pearson, de la *tabla 11* es 0 con lo cual existe relación entre las variables ya que rechazamos la hipótesis nula, lo que da pie entonces a llevar a cabo el análisis de correspondencia simple entre estas variables

A continuación se muestran las salidas luego de llevar a cabo el ACS:

```
Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.533329  77.4  77.4   *************************
 2      0.129168  18.8  96.2   *****
 3      0.026179   3.8 100.0
        -------- -----
 Total: 0.688676 100.0
```

Perfiles	Categoria	Masa	ChiDist	Inercia	Dim. 1	Dim. 2
Filas	NG1	0,201	1,016	0,207	-1,216	1,339
	NG2	0,157	0,665	0,069	-0,865	-0,358
	NG3	0,117	0,558	0,037	-0,495	-1,117
	NG4	0,108	0,390	0,016	-0,018	-1,080
	NG5	0,104	0,458	0,022	0,544	-0,615
	NG6	0,082	0,610	0,031	0,753	-0,556
	NG7	0,062	0,730	0,033	0,918	-0,414
	NG8	0,050	0,911	0,041	1,224	0,164
	NG9	0,036	1,059	0,040	1,419	0,577
	NG10	0,024	1,196	0,035	1,564	0,979
	NG11	0,019	1,430	0,039	1,776	1,556
	NG12	0,010	1,476	0,021	1,813	1,668
	NG13	0,009	1,666	0,024	1,957	2,085
	NG14	0,010	1,786	0,032	2,050	2,322
	NG15	0,012	1,820	0,041	2,075	2,391
Columnas	C1	0,289	0,863	0,215	-1,060	1,035
	C2	0,173	0,687	0,081	0,628	-1,190
	C3	0,309	0,522	0,084	-0,491	-0,942
	C4	0,229	1,160	0,308	1,525	0,866

Figura 13.

Como se muestra en la primera parte de los resultados (*Figura 13*) vemos que con solo 2 dimensiones se logra explicar el 96,2% de la variabilidad entre las 2 categorías de las variables, por lo cual con este número de dimensiones es suficiente ver la relación entre las variables NGEST y CLÚSTER recordemos que esta última es obtenida a partir de una proceso de aprendizaje no supervisado. Se observa como la categoría NG1 y C1 son aquellas que tienen mayor masa en cada una de los perfiles esto motivado al hecho del peso total que tienen clientes que se encuentran en el cruce de estas categorías de sus respectivas variables, luego vemos como van descendiendo paulatinamente dentro del componente fila o columna.

Luego vemos la distribución en las dos dimensiones de cada una de las categorías.

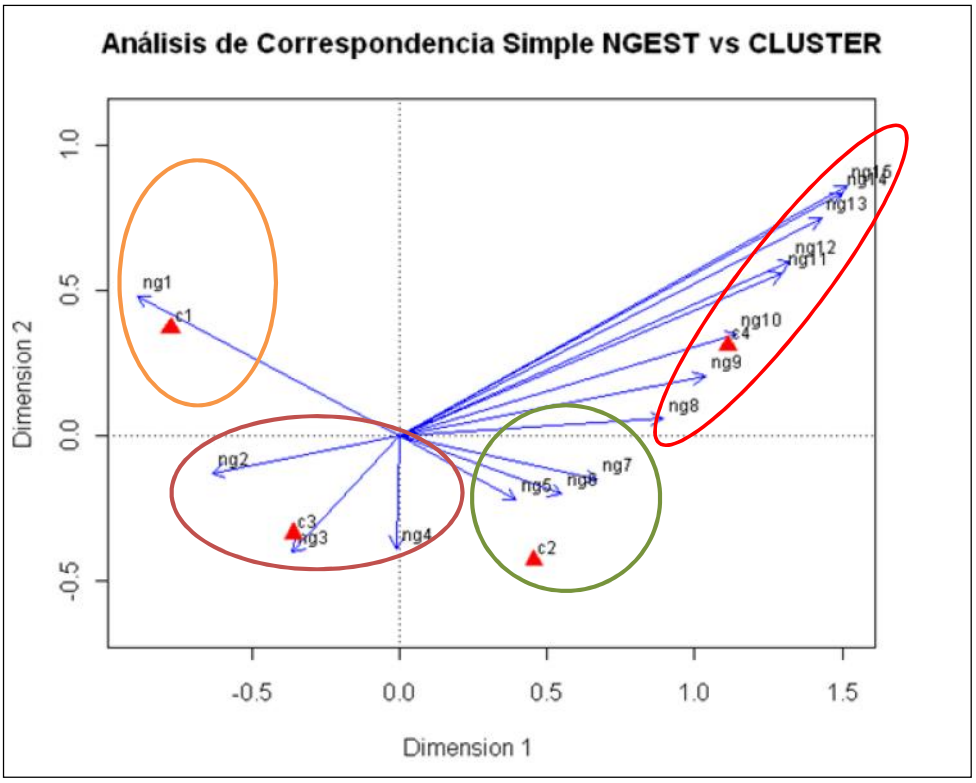


Grafico 8.

En este caso se muestra en el *Grafico 7* como la en la dimensión 1 se oponen los clientes pertenecientes a los clúster 1 y 3 del lado negativo, y los del clúster 2 y 4 del lado positivo recordando que estos últimos son los clústers que contienen a los clientes que no cancelan sus obligaciones en tanto que los primeros son los clientes que tienden a pagar un monto mayor al que les corresponde, luego en la segunda dimensión se oponen los clientes de acuerdo al número de veces en gestión, aun y cuando los clientes que se encuentran en el clúster 3 son clientes al día en sus obligaciones, estos están asociados con más de 2 entradas al sistema de gestión de cobro, mientras que los del clúster 1 solo están asociados con 1 sola entrada al sistema de gestión de cobro, luego del otro lado tenemos como los clientes que estuvieron entre 5 y 7 veces en gestión están asociados al clúster 2 que recordemos son clientes que cancelan sus obligaciones sin que sea suficiente, y finalmente los clientes con entradas en el sistema de gestión de

cobro mayores a 8 veces se asocian mayormente con el clúster 4, que recordamos son los clientes que escasamente llegan a cancelar alguna parte de sus obligaciones.

3.1.2.1 Análisis de Correspondencia Múltiple ACM

En el caso del análisis de correspondencia múltiple se consideraron las variables NGEST, NPAG y el clúster de pertenencia (clúster4def), a fin de evaluar el comportamiento de la relación entre el número de veces en gestión, pagos y el clúster de pertenencia.

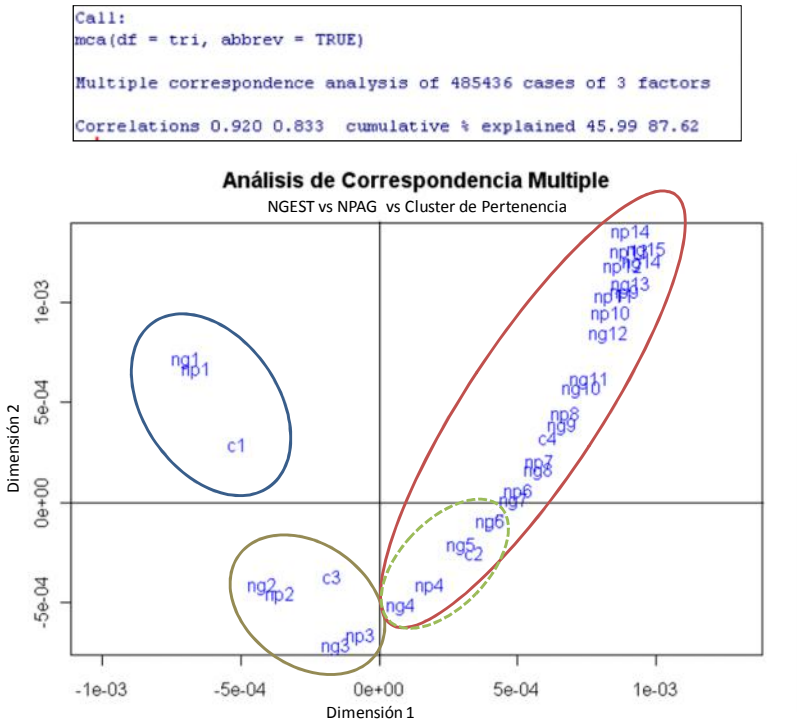


Figura 14.

Vemos en la *figura 14* que con 3 factores se logra explicar un 87,62% de la variabilidad del total y además se observan 3 grandes grupos los cuales se oponen de forma clara e interpretable y que darán pie más adelante a una reagrupación para la variable NGEST en función de esta correspondencia múltiple.

- El primer grupo es el formado por los clientes del clúster 1 que recordamos son los que pagan y estuvieron en gestión 1 vez.
- El segundo grupo de clientes que estuvieron 2 y 3 veces en gestión y que además pagaron la misma cantidad de veces y están asociados al clúster 3, en este caso

la dimensión 1 opone a los clientes que estuvieron entre 1 y 3 veces en gestión y que además pagaron de igual forma entre 1 y 3 veces, de aquellos clientes asociados a los clústers 2 y 4 que son los clientes que se consideran que presentan dificultades en la recuperación de sus obligaciones, ya que son aquellos que están con mas numero de movimientos además alcanzan mayores altura de mora y tienen como característica principal él hecho de no cumplir completamente el pago de sus obligaciones.

- Finalmente el tercer grupo está formado por clientes que han estado más de 4 en gestión y los cuales no logran ser bien separados por la segunda dimensión, y si bien da la impresión que se pueda hablar de un cuarto grupo (enmarcados en el ovalo punteado), este grupo no esta tan bien definido sobre alguno de los factores puesto que ambos extremos del mismo están muy cercanos hacia alguno de los ejes, lo que se traduce en el poco efecto que estos hacen sobre las dimensiones.

3.2 Aprendizaje Supervisado

Luego de haber realizado la clasificación y evaluar la relación entre algunas variables categóricas con cierta dependencia, pasamos al Aprendizaje Supervisado en este caso se utilizará como variable respuesta el número de veces que permanece el cliente en gestión NGEST, puesto que es una variable de interés en el estudio ya que a futuro permitirá la toma de decisiones acerca de las estrategias de cobranza que se le pueden aplicar a un cliente una vez que este ingrese al sistema de gestión.

En este caso se llevara a cabo una predicción del número de veces en gestión con 2 técnicas de Aprendizaje Supervisado: Arboles de Clasificación y Regresión Logística Multinomial, donde se tomara como respuesta el número de veces que permanece el cliente en gestión y como variables predictoras las demás que se disponen en el conjunto de datos exceptuando la variable sexo, pues esta no guarda correlación alguna ya que vimos anteriormente que la proporción de sexo es bastante similar en ambos casos.

3.2.1 Arboles de Clasificación

Para evaluar la capacidad de predicción de los arboles de clasificación se utiliza como método de división el Método CHAID y porcentaje Entrenamiento/Prueba de 80/20 del total de los datos respectivamente, además se consideran las siguientes variables:

- Respuesta:
    - Número de Veces en Gestión (NGEST).

- Predictoras:
    - Monto Vencido Promedio (PROM_VCDO).
    - Monto Cobrado Promedio (PROM_PAG).
    - Cantidad de Tarjetas Créditos (CANT_TDC).
    - Número de Pagos mientras estuvo en Gestión (NPAG).
    - Máxima Altura de Mora (MAX_MORA).
    - Limite o Capacidad de Endeudamiento del Cliente (LIMITE).
    - Frecuencia de Pago (COMO_PAGA).
    - Número de Ascensos en Mora (NMOV).
    - Clúster de Pertenencia (clúster4def).

Resultados del Árbol de Clasificación:

Resumen del Arbol		
Especificaciones de Entrada	**Método de Crecimiento**	CHAID
	Variable Dependiente	N_GEST
	Variables Independientes	PROM_VCDO, PROM_PAG, LIMITE, CANT_TDC, NMOV, cluster4def, COMO_PAGA,
	Máxima Profundidad del Arbol	**3**
	Minimo de Casos en un Nodo Padre	**100**
	Minimo de Casos en un Nodo Hijo	**50**
Resultados	**Variables Independientes Incluidas**	NMOV, PROM_VCDO, PROM_PAG, MAX_MORA, COMO_PAGA, cluster4def,
	Numero de Nodos	228
	Numero de Nodos Terminales	170
	Profundidad	3

Tabla 11.

En el Resumen del Árbol (*Tabla 11*) se muestran las especificaciones de entrada, vemos que el modo de crecimiento es CHAID que permite más de 2 particiones por nodo, las variables de entrada y las de salida, dado el método de crecimiento solo se permite un profundidad de 3 niveles o ramas, finalmente este árbol posee 228 nodos padres con 170 nodos hijos, y en este caso no todas las variables de entrada fueron utilizadas en la construcción del árbol ya que las variables CANT_TDC y LIMITE no se usaron en el proceso de construcción, sin embargo las variables predictoras se van mencionando conforme van siendo usadas en el lado de los resultados, vemos que en este caso las principales son NMOV y PROM_VCDO y además se observa como el clúster obtenido en el proceso de aprendizaje no supervisado (clúster4def) ayuda a la construcción del árbol.

Ahora bien al momento de observar la predicción del modelo, se mostrara que tanto en la muestra de entrenamiento como en la de validación la correcta clasificación solo es aceptable para las veces en gestión 1 y 15 en las cuales supera el 85% de correcta clasificación. Luego por otra parte para las veces en gestión 2 y 14 la clasificación es muy deficiente y no supera el 70%, llegando en algunos casos a ser inferior al 10% como se muestran en las tablas de clasificación a continuación,

Tabla de Clasificacion																	
Muestra	Valor Observado	Prediccion															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	% Correcta Clasificacion
Entrenamiento	1	70472	13131	3908	172		1										80,4%
	2	25353	26019	13269	1382	2252	95										38,1%
	3	9535	18793	14050	3072	5505	438	2									27,3%
	4	4576	13280	11905	4469	10640	1993	216	25								9,5%
	5	2457	8251	8475	3043	17031	4334	1269	320	62							37,6%
	6	1494	4897	5698	2280	9987	7751	3089	693	155							21,5%
	7	918	2744	3483	1594	7655	3811	4986	1354	367	5	32					18,5%
	8	503	1395	1855	1101	5125	2430	3868	4521	696	114	17					20,9%
	9	244	771	896	690	3094	1360	2828	2988	1929	761	84	15				12,3%
	10	134	404	431	375	1553	723	1606	1736	1209	2192	236	44				20,6%
	11	56	193	188	197	717	332	779	913	754	400	3560	194	50			42,7%
	12	31	88	64	94	332	123	278	337	369	177	510	1497	327	5		35,4%
	13	11	50	25	50	105	53	106	86	138	120	228	255	2572	27	15	67,0%
	14	3	19	6	17	32	24	35	32	32	31	61	61	1152	2878	69	64,6%
	15	3	12	1	6	4	5	15	6	8	6	19	5	223	284	4787	88,9%
	Porcentaje Total	26,5%	20,6%	14,7%	4,2%	14,7%	5,4%	4,4%	3,0%	1,3%	,9%	1,1%	,5%	1,0%	,7%	1,1%	38,6%

Tabla 12.

En el caso de la muestra de entrenamiento representada en la *tabla 12*, vemos como existe una gran confusión en el árbol. Por ejemplo, los clientes que están 4 veces en gestión solo clasifica correctamente el 9,2%, y de manera global la correcta clasificación solo alcanza el 38,6%, lo cual es muy deficiente.

Tabla de Clasificacion																	
Muestra	Valor Observado	Prediccion															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	% Correcta Clasificacion
Prueba	1	7801	1497	430	16												80,1%
	2	2828	2952	1474	150	270	10										38,4%
	3	1032	2055	1539	325	590	46										27,5%
	4	495	1411	1337	481	1218	233	23	4								9,2%
	5	270	912	968	319	1946	479	148	29	4							38,3%
	6	149	523	632	257	1106	818	350	92	16							20,7%
	7	87	290	371	197	803	449	542	167	43		1					18,4%
	8	67	167	218	119	557	250	497	506	79	8	2					20,5%
	9	29	72	102	68	333	127	319	315	225	100	6	3				13,2%
	10	16	41	51	51	174	88	203	212	149	234	26	3				18,8%
	11	5	26	15	18	79	34	87	99	83	53	394	23	11			42,5%
	12	1	11	16	12	37	14	41	30	42	24	45	185	37		1	37,3%
	13	1	1	4	6	6	3	4	13	16	5	24	20	260	3	1	70,8%
	14	2	4	0	2	1	3	3	3	5	6	4	8	104	328	5	68,6%
	15	0	1	0	1	1	0	0	0	0	2	2	2	24	35	540	88,8%
	Porcentaje Total	26,4%	20,6%	14,8%	4,2%	14,7%	5,3%	4,6%	3,0%	1,4%	,9%	1,0%	,5%	,9%	,8%	1,1%	38,7%

Tabla 13.

Para el caso de la muestra de prueba mostrada en la *tabla 13* el porcentaje global de correcta predicción es similar al de entrenamiento y el comportamiento de la correcta clasificación es superior o inferior a las de entrenamiento, pero sin mayores variaciones, salvo en los clientes con 13 y 14 veces en gestión que es superior al de los datos usados entrenamiento.

En vista de la clasificación deficiente que se obtiene al momento de predecir el número de veces que estará un cliente en gestión, y obedeciendo a la relación que se mostró en el aprendizaje no supervisado entre el clúster de pertenencia y el número de veces en gestión y el número de veces que paga el clientes como se mostro en el Análisis de Correspondencia Múltiple, se llevó a cabo un reagrupamiento de la variable NGEST y NPAG con la finalidad de refinar y mejorar el proceso de clasificación y predicción.

Este agrupamiento desde el punto de vista del problema en sí, de predecir el número de veces que estará el cliente en gestión, no afecta los resultados del problema. ya que la idea es que una vez ingrese el cliente al sistema de gestión de cobro, predecir el número de veces que este estará en gestión.

Con lo cual si un cliente es predicho como si solo estuviese una vez en gestión no será necesario una estrategia de cobranza muy agresiva puesto se sabe que saldrá del sistema sin mayor dificultad, en tanto que si un cliente estará más de 5 veces ya se torna como un cliente con ciertos problemas a quien se le ofrecerá algún plan de refinanciamiento o algo parecido para lograr la recuperación del monto, con lo cual el

agrupamiento lo que permitirá es la reducción del número de estrategias de cobranza sobre el deudor según su comportamiento.

En la siguiente tabla se muestra el agrupamiento de las variables NPAG y NGEST a fin de mejorar la predicción del número de veces en gestión.

NPAG	NPAG_A	NGEST	NGEST_A
1	1	1	1
Entre 2 y 3	2-3	Entre 2 y 3	2-3
Más de 4	4mas	Más de 4	4mas

Tabla 14.

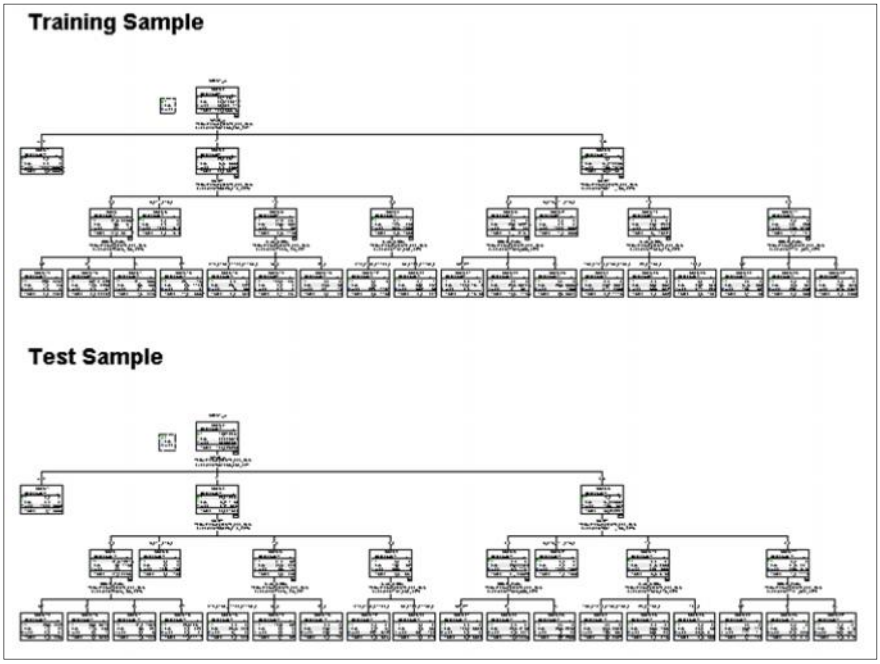
Una vez llevada a cabo la reclasificación mostrada en la tabla 14, vemos en las siguientes salidas (*Tabla 15*) la gran mejoría de la correcta clasificación del número de veces en gestión así como la selección de la variable numero de pagos agrupados NPAG_A es la principal variable en la formación del árbol.

Resumen del Arbol		
Especificaciones de Entrada	Método de Crecimiento	CHAID
	Variable Dependiente	NGEST_A
	Variables Independientes	cluster4def, CANT_TDC, PROM_VCDO, PROM_PAG, LIMITE, NMOV, MAX_MORA, COMO_PAGA, NPAG_A
	Máxima Profundidad del arbol	3
	Minimo de casos en un Nodo Padre	100
	Minimo de casos en un Nodo Hijo	50
Resultados	Variables Independientes Incluidas	NPAG_A, NMOV, COMO_PAGA, MAX_MORA, cluster4def
	Numero de Nodos	30
	Numero de Nodos Terminales	21
	Profundidad	3

Tabla 15.

En este caso se observa que en la construcción del árbol se muestra un número menor de variables que en el caso anterior y de hecho ninguna de las variables continuas es incluida, también se muestra que el numero de TDC no fue usado en la construcción, y que el clúster de pertenencia está dentro del proceso de construcción del árbol.

En la gráfica 7 se muestra ambos arboles tanto el de entrenamiento como el de prueba, en este caso ambos con un comportamiento muy similar.



Gráfica 7.

Finalmente se muestra la tabla de clasificación (*Tablas 16 y 17*) donde observamos la mejoría notable en relación con el primer caso cuando se tenía la variable NGEST con todas sus categorías.

Muestra	Valor Observado	Prediccion			
		1	2-3	mas4	% Correcta Clasificacion
Entrenamiento	1	78143	0	1	100,0%
	2-3	4547	100480	1192	94,6%
	mas4	13	7510	196595	96,3%
	Porcentaje Total	21,3%	27,8%	50,9%	96,6%

Tabla 16.

Muestra	Valor Observado	Prediccion			
		1	2-3	mas4	% Correcta Clasificacion
Prueba	1	19.284			100,0%
	2-3	1.168	25.309	340	94,4%
	mas4	4	1.839	49.011	96,4%
	Porcentaje Total	21,1%	28,0%	50,9%	96,5%

Tabla 17.

Luego del agrupamiento del número de veces en gestión vemos en las *Tablas 15 y 16* como en tanto en la muestra de entrenamiento como en la muestra de validación o prueba se alcanza una correcta clasificación en ambos casos superior al 95%, y se tiene una clasificación si se quiere perfecta para los clientes que solo están una vez en gestión.

3.2.2 Regresión Logística Multinomial (RLM)

En la parte anterior se mostro como los arboles de clasificación arrojaron resultados importantes en cuanto a la predicción del número de veces que permanece el cliente en gestión dado un agrupamiento posterior a la variable NGEST_A, en este caso verificaremos con otra técnica de aprendizaje estadístico supervisado como es la RLM, esto ya que la formación de los clústers que se utilizaron en los arboles tienen como formación principal (en la etapa 1 del procedimiento TWO STEPS CLÚSTER) un agrupamiento a través de arboles, lo cual podría llevar a pensar que se estarían usando la predicción de una variable con la misma técnica con la cual se construyen las predictoras. Entonces nuevamente tenemos:

Respuesta:

- Número de Veces en Gestión Agrupado (NGEST_A)

Predictoras:

- Monto Vencido Promedio (PROM_VCDO)
- Monto Cobrado Promedio (PROM_PAG)
- Cantidad de Tarjetas Créditos (CANT_TDC)
- Número de Pagos mientras estuvo en Gestión (NPAG)
- Máxima Altura de Mora (MAX_MORA)
- Limite o Capacidad de Endeudamiento del Cliente (LIMITE)
- Frecuencia de Pago (COMO_PAGA)
- Número de Ascensos en Mora (NMOV)
- Clúster de Pertenencia (clúster4def)

Conservando la partición porcentual 80/20 entrenamiento/prueba de los arboles se llevo a cabo el siguiente modelo de Regresión Logística Multinomial (RLM), considerando como grupo control el último grupo de la variable respuesta mas4.

$$NGEST\_A = CANT\_TDC + cluster4def + COMO\_PAGA + LIMITE + MAX\_MORA + NMOV + NPAG\_A + PROM\_PAG + PROM\_VCDO$$

Una vez planteado el modelo se obtienen las siguientes salidas en este caso usaremos un valor de significación del 5% para determinar que niveles de las variables predictoras son estadísticamente significativas, es decir si algún nivel de alguna variable tiene una significación superior a 0,05 se considerara que esta no tiene efecto sobre la respuesta ya que su coeficiente pudiera en algún momento ser cero.

Resultados NGEST_A = 1

Grupo NGEST_A	Var Ind.	β	Error Std.	Wald	gl	Sig.	Exp(B)	Intervalos al 95%	
								Lim Inf	Lim Sup
1	PROM_VCDO	,001	,000	417,01	1	,000	1,00	1,001	1,001
	PROM_PAG	,000	,000	23,12	1	,000	1,00	1,000	1,000
	LIMITE	,000	,000	18,86	1	,000	1,00	1,000	1,000
	[cluster4def=1]	,029	,594	,00	1	,961	1,03	,321	3,301
	[cluster4def=2]	3,529	,520	45,97	1	,000	34,07	12,287	94,488
	[cluster4def=3]	-,162	,591	,08	1	,784	,85	,267	2,707
	[cluster4def=4]	0[b]	.	.	0			.	.
	[COMO_PAGA=NP]	-1,034	,534	3,75	1	,053	,36	,125	1,012
	[COMO_PAGA=P-]	-5,649	,501	127,08	1	,000	,00	,001	,009
	[COMO_PAGA=P+]	-7,242	,499	210,71	1	,000	,00	,000	,002
	[COMO_PAGA=P=]	0[b]	.	.	0	.	.	.	.
	[NPAG_A=1]	23,703	4,591	26,65	1	,000	1,969E+10	2,434E+06	1,593E+14
	[NPAG_A=2-3]	3,245	9,747	,11	1	,739	25,66	1,297E-07	5,080E+09
	[NPAG_A=mas4]	0[b]	.	.	0				
	[NMOV=0]	-7,521	673,260	,00	1	,991	,00	,000	.[c]
	[NMOV=1]	-12,988	673,261	,00	1	,985	,00	,000	.[c]
	[NMOV=2]	-18,309	673,263	,00	1	,978	,00	,000	.[c]
	[NMOV=3]	-26,170	673,361	,00	1	,969	,00	,000	.[c]
	[NMOV=4]	-25,866	673,526	,00	1	,969	,00	,000	.[c]
	[NMOV=5]	-23,282	673,529	,00	1	,972	,00	,000	.[c]
	[NMOV=6]	-10,032	673,633	,00	1	,988	,00	,000	.[c]
	[NMOV=7]	-7,695	673,866	,00	1	,991	,00	,000	.[c]
	[NMOV=8]	-9,738	673,979	,00	1	,988	,00	,000	.[c]
	[NMOV=9]	-9,147	673,831	,00	1	,989	,00	,000	.[c]
	[NMOV=10]	-12,014	674,208	,00	1	,986	,00	,000	.[c]
	[NMOV=11]	-6,498	674,081	,00	1	,992	,00	,000	.[c]
	[NMOV=12]	0[b]	.	.	0				
	[CANT_TDC=1]	2,283	,692	10,88	1	,001	9,81	2,526	38,085
	[CANT_TDC=2]	1,976	,686	8,31	1	,004	7,21	1,882	27,644
	[CANT_TDC=3]	1,533	,688	4,97	1	,026	4,63	1,204	17,820
	[CANT_TDC=4]	1,208	,702	2,96	1	,085	3,35	,846	13,235
	[CANT_TDC=5]	1,088	,809	1,81	1	,179	2,97	,608	14,491
	[CANT_TDC=6]	0[b]	.	.	0				
	[MAX_MORA=30]	11,891	1,343	78,44	1	,000	146006,04	10508,073	2,029E+06
	[MAX_MORA=60]	5,458	1,273	18,38	1	,000	234,63	19,353	2844,543
	[MAX_MORA=90]	4,683	1,290	13,18	1	,000	108,10	8,626	1354,660
	[MAX_MORA=120]	2,647	1,302	4,14	1	,042	14,11	1,101	180,987
	[MAX_MORA=150]	3,772	1,354	7,76	1	,005	43,45	3,058	617,384
	[MAX_MORA=180]	5,583	33,261	,03	1	,867	265,90	1,297E-26	5,452E+30
	[MAX_MORA=210]	0[b]	.	.	0				

Tabla 18.

En el Caso del grupo 1 vemos en la *Tabla 18* que la variable NMOV no es significativa en ninguno de sus niveles, de igual forma la categoría 1 de la variable clúster4def que es la que se asocia con este número de veces en gestión, tampoco es estadísticamente significativa, el nivel 1 de la variable de número de pagos agrupados

NPAG_A que se asocia linealmente con la categoría 1 de NGEST_A, es estadísticamente significativo así como las categorías 1,2 y 3 del numero de TDC del cliente, y los niveles de mora 30,60 y 90 y de igual manera todas las variables continuas son estadísticamente significativas.

Resultados NGEST_A = 2-3

Grupo NGEST_A	Var Ind.	β	Error Std.	Wald	gl	Sig.	Exp(B)	Intervalos al 95%	
								Lim Inf	Lim Sup
2-3	Intercept	-13,954	581,205	,00	1	0,98			
	PROM_VCDO	,000	,000	335,47	1	0,00	1,00	1,000	1,000
	PROM_PAG	,000	,000	157,75	1	0,00	1,00	1,000	1,000
	LIMITE	,000	,000	1,63	1	0,20	1,00	1,000	1,000
	[cluster4def=1]	-,374	,290	1,66	1	0,20	,69	,389	1,216
	[cluster4def=2]	-1,275	,114	125,17	1	0,00	,28	,224	,349
	[cluster4def=3]	0	0,2889104	2,75	1	0,10	,62	0,351427	1,09063
	[cluster4def=4]	0b	.	.	0		.	.	.
	[COMO_PAGA=NP]	-2,138	,508	17,72	1	0,00	,12	,044	,319
	[COMO_PAGA=P-]	-5,106	,498	105,08	1	0,00	,01	,002	,016
	[COMO_PAGA=P+]	-6	0,497833	140,71	1	0,00	,00	0,001027	0,00723
	[COMO_PAGA=P=]	0b	.	.	0		.	.	.
	[NPAG_A=1]	20,101	,052	147317,15	1	0,00	5,367E+08	4,843E+08	5,947E+08
	[NPAG_A=2-3]	19	0	.	1		140083662	1,4E+08	1,4E+08
	[NPAG_A=mas4]	0b	.	.	0		.	.	.
	[NMOV=0]	1,543	581,205	,00	1	1,00	4,68	,000	.c
	[NMOV=1]	,868	581,205	,00	1	1,00	2,38	,000	.c
	[NMOV=2]	-2,709	581,205	,00	1	1,00	,07	,000	.c
	[NMOV=3]	-14,613	581,309	,00	1	0,98	,00	,000	.c
	[NMOV=4]	-15,315	581,372	,00	1	0,98	,00	,000	.c
	[NMOV=5]	-15,669	581,478	,00	1	0,98	,00	,000	.c
	[NMOV=6]	-13,654	581,719	,00	1	0,98	,00	,000	.c
	[NMOV=7]	,835	582,118	,00	1	1,00	2,30	,000	.c
	[NMOV=8]	,851	582,499	,00	1	1,00	2,34	,000	.c
	[NMOV=9]	,841	582,282	,00	1	1,00	2,32	,000	.c
	[NMOV=10]	-,249	582,441	,00	1	1,00	,78	,000	.c
	[NMOV=11]	0	582,2113	,00	1	1,00	1,64	0	.c
	[NMOV=12]	0b	.	.	0		.	.	.
	[CANT_TDC=1]	,512	,188	7,41	1	0,01	1,67	1,154	2,414
	[CANT_TDC=2]	,569	,186	9,39	1	0,00	1,77	1,228	2,542
	[CANT_TDC=3]	,512	,187	7,49	1	0,01	1,67	1,156	2,409
	[CANT_TDC=4]	,402	,191	4,42	1	0,04	1,50	1,028	2,176
	[CANT_TDC=5]	0	0,2121601	1,13	1	0,29	1,25	0,826384	1,898295
	[CANT_TDC=6]	0b	.	.	0		.	.	.
	[MAX_MORA=30]	1,910	,366	27,30	1	0,00	6,75	3,299	13,830
	[MAX_MORA=60]	1,493	,250	35,60	1	0,00	4,45	2,725	7,266
	[MAX_MORA=90]	1,067	,252	17,94	1	0,00	2,91	1,774	4,762
	[MAX_MORA=120]	-2,285	,264	74,85	1	0,00	,10	,061	,171
	[MAX_MORA=150]	-2,304	,329	48,92	1	0,00	,10	5,237E-02	1,905E-01
	[MAX_MORA=180]	-,544	,586	,86	1	0,35	,58	1,841E-01	1,831E+00
	[MAX_MORA=210]	0b	.	.	0		.	.	.

Tabla 19.

Ya para el grupo 2-3 en la *Tabla 19* se muestra como la variable clúster4def es significativa en su categoría 2, la variable LIMITE no es significativa en tanto que las otras 2 variables continuas si lo son, nuevamente la variable NMOV no es significativa en ninguna de sus categorías, más sin embargo en este caso los niveles de mora son significativos en su mayoría, la variable que mide la frecuencia de pago COMO_PAGA es significativa en todos sus niveles de igual forma el numero de pagos agrupados NPAG_A.

Una vez evaluados el modelo en los datos de entrenamiento y prueba se obtiene la siguiente tabla de clasificación.

La Tabla de Clasificación

Muestra	Valor Observado	Predicción				
		1	2-3	mas4	Total	% Correcta Clasificacion
Entrenamiento	1	78.115	28	1	78.144	100,0%
	2-3	4.531	100.613	1.075	106.219	94,7%
	mas4	23	7.697	196.398	204.118	96,2%
	%Total	21,3%	27,9%	50,8%	388.481	96,6%
Prueba	1	19.281	3	-	19.284	100,0%
	2-3	1.164	25.382	271	26.817	94,6%
	mas4	4	1.896	48.954	50.854	96,3%
	%Total	5,3%	7,0%	12,7%	96.955	96,6%

Tabla 20.

Vemos en la *Tabla 20* como los resultados no tienden a diferir mayormente con los obtenidos en los arboles, sin embargo la variable NMOV que fue utilizada entre las principales para la formación del árbol, en el análisis RLM no es significativa en ninguno de sus niveles.

Conclusiones y Recomendaciones

Una vez llevado a cabo el proceso de caracterización de deudores de TDC de una institución financiera por medio de las técnicas de aprendizaje estadístico se puede llegar a las siguientes conclusiones:

- Las técnicas de aprendizaje estadístico constituyen una herramienta importante en el análisis estadístico de datos ya que tiene la posibilidad de estudiar un problema según el origen de este, bien sea cuando se tiene conocimiento de lo que queremos obtener del estudio (una respuesta) o bien la obtención de información cuando los datos no tienen una estructura definida y se desea sacar algún tipo de información de interés de estos.
- El uso de las técnicas de aprendizaje no supervisado fue de suma importancia al momento de llevar a cabo la caracterización en el proceso de aprendizaje supervisado ya que fue e través de este que se le dio sentido a la variable respuesta que fue estudiada en la parte supervisada.
- Finalmente se logró obtener a través de las técnicas de aprendizaje estadístico no supervisado un conjunto de 4 grupos de deudores con características claramente definidas.
- Por el área del aprendizaje supervisado se consiguió un clasificador que permitió predecir con un alto porcentaje de acierto el tiempo que estaría un cliente en gestión dada alguna de sus características.
- Las técnicas aplicadas en este trabajo lograron establecer una especie de mezcla entre técnicas de aprendizaje no supervisado y supervisado, ya que la primera indujo la reducción de un espacio de respuesta que permitió mejorar los procesos de clasificación usados en la segunda.

Como recomendación se podría establecer el hecho de realizar el proceso de clasificación por medio de otras técnicas de aprendizaje supervisado como las redes Neuronales a fin de evaluar que tan eficiente fueron las técnicas aquí aplicadas, luego otra posible recomendación sería la realización de un algoritmo de clasificación en la cual se permita la mezcla de variables cualitativas y cuantitativas en la formación de clúster y que opere distinto al algoritmo de two steps clúster mostrado en este trabajo.

Bibliografía

- *Venables W. N.; Ripley B. D.*. ***Modern Applied Statistics with S***. *4° Edición. New York. Springer, 2002.*
- *Peter Dalgaard.* ***Introductory Statistics with R,.*** New York. Springer, 2002.
- *Agresti Alan. An Introduction to Categorical Data Analysis. 2° Edicion. New Jersey, Wiley. 2002.*
- *Friedman J; Hastie Trevor; Tibshiriani R.* ***The Elements of Statistical Learning.*** *2° Edicion. California. Springer 2008.*
- *Bousquet O; Boucheron S; Lugosi G.* ***Introduction to Statistical Learning Theory.*** *Lecture Notes in Artificial Intelligence, 3176:169–207, 2004.*
- *Huang, Z. 1998.* ***Extensions to the k-means algorithm for clústering large data sets with categorical values****. Data Mining and Knowledge Discovery, 2, 3, 283-304.*
- *Salvador, M (2003).* ***"Análisis de Correspondencias"****, 5campus.com Estadística <http://www.5campus.com/leccion/correspondencias> 20/10/2010.*
- *Fox J. (2005). "****Maximum-Likelihood Estimation of the Logistic Regression Model",*** *<http://socserv.mcmaster.ca/jfox/Courses/UCLA/logistic-regression-notes.pdf> 16/12/2010.*
- *SPSS Inc. (2004).* ***TwoStep Clúster Analysis.*** *Technical report, Chicago. <http://support.spss.com/tech/stat/Algorithms/12.0/twostep_clúster.pdf.> 22/05/2010.*
- *Bacher, J; Knut, W; Melanie, V:* ***SPSS TwoStep Clúster – A First Evaluation****. Arbeits- und Diskussionspapiere 2004-2.*
- *Czepiel, S. A.,* ***Maximum likelihood estimation of logistic regression models: theory and implementation.*** *< http://czep.net/stat/mlelr.pdf>. 16/01/2011.*
- *Ferrari D.* ***R Programming II: Data Manipulation and Functions.*** *UCLA Department of Statistics, Statistical Consulting Center. 2010.*
- *Everitt B.; Hothorn T.* ***A Handbook of Statistical Analyses Using R.*** *2° Edicion. Florida. Chapman & Hall/CRC.2010.*
- *Tapia Jesús Manuel.* ***Introducción al Análisis de Datos Multivariantes****. Barinas. Ediciones de la Universidad Ezequiel Zamora. 2007.*