UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN
Departamento Académico de Estadística e Informática

Trabajo Monográfico

"IDENTIFICACIÓN DE UN MODELO EXPLICATIVO DE RETENCIÓN DE CLIENTES CON RIESGO DE FUGA PARA UNA ENTIDAD BANCARIA APLICANDO REGRESIÓN LOGÍSTICA Y ÁRBOLES DE CLASIFICACIÓN CART"

Presentado para optar al título de Ingeniero Estadístico e Informático

MARÍA ALEJANDRA HUAMANÍ MIRANDA

Modalidad de Examen Profesional

LIMA-PERÚ

2014

DEDICATORIA

El presente trabajo dedico a mi madre
Marina, por haberme inculcado la ética de trabajo y superación.
A mi esposo Gumer Enrique, a mis hijos Matías y Gumer Ignacio
por todo el apoyo para lograr el objetivo propuesto.

ÍNDICE GENERAL

1. Introducción 1
2. El Problema de Investigación 2
   2.1. Fundamentación del Problema de Investigación 2
   2.2. Formulación del Problema de Investigación 3
   2.3. Objetivos de la Investigación 3
      2.3.1. Objetivo General 3
      2.3.2. Objetivos Específicos 4
   2.4. Justificación de la investigación 4-5
3. Marco Teórico 6
   3.1. Modelos Estadísticos 6
      3.1.1. Regresión Logística 6
         3.1.1.1. El Modelo de Regresión Logística 6
         3.1.1.2. El Modelo multivariante de Regresión Logística binaria 6-7
         3.1.1.3. Coeficiente de regresión 8
         3.1.1.4. Medidas de bondad de ajuste 9
            3.1.1.4.1. Coeficiente Pseudo R2 9-10
            3.1.1.4.2. Prueba de bondad de ajuste Hosmer y Lemeshow 11
            3.1.1.4.3. Curva ROC 12
            3.1.1.4.4. Tablas de clasificación 12
      3.1.2. CART 13
         3.1.2.1. Árboles de Clasificación 13-14
         3.1.2.2. El Modelo CART 15
         3.1.2.3. Medida de impureza 16-17
         3.1.2.4. Estimación de la tasa de error del árbol 18-19

4. Metodología de la investigación 20
4.1. Tipo de investigación 20
4.2. Formulación de la hipótesis 20
4.3. Identificación de Variables 21
4.3.1. Variable dependiente 21
4.3.2. Variables independientes 21
4.4. Diseño de investigación 22
4.4.1. Fuente de información 22
4.5. Población y Muestra 23
4.6. Diseño de la muestra 24
5. Procedimiento de análisis de datos 25
5.1. Análisis Exploratorio 25-26
5.2. Análisis de Regresión Logística 27-31
5.3. Análisis del algoritmo de Árbol de Clasificación CART 32-38
5.4. Comparación del Análisis de Regresión logística vs el algoritmo de Árbol de Clasificación CART 39-42
6. Conclusiones 43
7. Recomendaciones 44
8. Bibliografía 45-46
9. Anexo 47-48

RESUMEN

La realidad competitiva que en estos días enfrentan las entidades bancarias ha provocado que éstas no sólo concentren sus esfuerzos de marketing exclusivamente en estrategias de captación de clientes, sino también en estrategias de retención y fidelización; la fuga de clientes es una situación que afecta la rentabilidad de la gran mayoría de las instituciones bancarias dado que se invierte mucho más en la captación de clientes que en campañas para la retención, por ello, es un tema de intensivo estudio científico en los últimos años.

Las entidades bancarias requieren contar con herramientas que les permitan estimar probabilidades de fuga para su cartera de clientes y así decidir sobre que clientes concentrar sus esfuerzos de retención.

En el presente trabajo se utilizó la regresión logística de respuesta binaria y el algoritmo de árbol de clasificación CART para predecir y clasificar a los clientes con riesgo de fuga y así identificar el mejor modelo explicativo de retención de clientes con riesgo de fuga para una entidad bancaria.

El modelo que mejor explica el riesgo de fuga de un cliente fue la Regresión Logística binaria que obtuvo como variables predictoras número de transacciones, ingreso bruto, número de tarjetas usadas y línea de crédito.

Las variables identificadas permitirán a la entidad bancaria reorientar las estrategias en las campañas de retención de clientes.

Palabras claves: retención de clientes, captación de clientes, cartera de clientes, regresión logística binaria, arboles de clasificación CART.

1. INTRODUCCIÓN

El presente trabajo monográfico constituye el desarrollo de un modelo de Retención de Clientes, para identificar a los clientes con mayor probabilidad de fuga y en la determinación de las estrategias o procedimientos que aumenten el grado de fidelización y bajen los índices de fuga en la cartera.

El estudio nace como una necesidad de implementar acciones enfocadas a la retención de clientes lo cual implica racionalizar y reorientar las inversiones, al poner el foco en las experiencias que verdaderamente importan a las entidades financieras, para que sus clientes continúen comprando con la tarjeta de crédito.

La implementación de acciones de retención de clientes supone un efecto económico positivo a largo plazo, pero se debe destacar que también se logran otros a muy corto plazo, esto justifica el desarrollo de un modelo de retención.

Por tanto en este trabajo se busca determinar un modelo a partir de variables socio demográficas y variables de comportamiento bancario, que permitan identificar y predecir a los clientes con riesgo de fuga para retenerlos, a través de un modelo estadístico de regresión logística binaria o utilizando árboles de clasificación CART.

Finalmente, comparar los resultados obtenidos para determinar el mejor modelo explicativo para la retención de clientes con riesgo de fuga.

2. EL PROBLEMA DE INVESTIGACIÓN

2.1. Fundamentación del Problema de Investigación

La cartera de clientes es uno de los activos más importantes para las entidades financieras, ya que están estrechamente relacionadas con las utilidades del negocio.

En la entidad bancaria viene realizando innumerables campañas para mantener a los clientes activos realizando transacciones pero la actividad competitiva es muy fuerte y la tasa de fuga aumenta esto genera pérdidas a la entidad bancaria.

Actualmente la entidad bancaria realiza la clasificación de los clientes a retener en base al conocimiento del negocio más no utiliza una herramienta cuantitativa que le permita determinar con mayor precisión la predicción de los clientes con riesgo de fuga, como son la Regresión logística y Árboles de Clasificación.

La predicción de fuga es un elemento importante para la retención de clientes, tanto en la identificación de los clientes con probabilidad razonable de fuga como en la determinación de su rentabilidad futura por lo que se requiere focalizar los esfuerzos en la retención de los clientes más apropiados.

La Regresión Logística es un método estadístico paramétrico que se aplica para predecir la probabilidad que ocurra un evento de interés, representado por una variable dicotómica, en función de un conjunto de variables predictoras.

Los Árboles de Clasificación es un método de clasificación no paramétrico, representado por gráficos estadísticos que ilustran reglas de decisión, que parten de un nodo raíz que contiene todas las observaciones de una muestra. A

medida que se desarrolla el árbol los datos se dividen en ramas de subconjuntos de datos exclusivos. Los Árboles de Clasificación son los más utilizados en la actualidad para la segmentación, estratificación y predicción de clientes.

2.2. Formulación del Problema de Investigación

El problema de investigación se puede resolver respondiendo a las siguientes interrogantes:

- ¿Se puede identificar que variables influyen en la fuga de clientes en una entidad bancaria usando Regresión Logística Binaria teniendo en cuenta para cada cliente las variables predictoras, edad, sexo, estado civil, ingreso bruto, número de tarjetas usadas, Monto, Saldo, número de transacciones y línea de crédito?
- ¿Se puede predecir y clasificar qué clientes tienen riesgo de fuga para retenerlos en una entidad bancaria identificando un modelo explicativo usando un modelo de Regresión Logística Binaria y Árboles de clasificación CART?
- ¿Cuál es el método de los dos utilizados en el estudio que brinde mejores resultados en la identificación y predicción de clientes con riesgo de fuga?

2.3. Objetivos de la Investigación

2.3.1. Objetivo General

- Determinar un modelo predictivo a partir del resultado de la comparación de la Regresión Logística con los Árboles de clasificación CART, para la retención de los clientes con riesgo de fuga de una entidad bancaria.

2.3.2. Objetivos Específicos

- Identificar las variables predictoras más relevantes para la predicción de fuga de clientes aplicando una Regresión Logística Binaria.
- Obtener un modelo que permita predecir y clasificar qué clientes tienen riesgo de fuga en una entidad bancaria aplicando una Regresión Logística Binaria.
- Elaborar un modelo para predecir y clasificar a los clientes con riesgo de fuga en una entidad bancaria aplicando Árboles de Clasificación con el Algoritmo CART.
- Comparar los modelos obtenidos con ambas metodologías a través de las tablas de clasificación.

2.4. Justificación de la Investigación

Se consideró varios motivos para desarrollar la presente investigación:

Justificación Práctica. Dentro del marco económico se sabe que en la actualidad la retención de clientes es una actividad netamente competitiva, debido a la oferta de diversos productos y servicios bancarios enfocados en clientes que generan alta rentabilidad, existe una gama de promociones y si fuera poco la competencia de tasas e impuestos que generan movimientos de clientes entre bancos es variable día a día, los bancos compran deudas de otros bancos y los clientes pasan de banco en banco según convenga.

Para las entidades bancarias es importante invertir en la retención de sus clientes con el precedente de que actualmente realizan distintas promociones para fidelizar a sus clientes enviando diversas campañas; sin embargo, la tasa

de fuga va en aumento es por ello la relevancia del estudio, a su vez teniendo en consideración que es más rentable a largo plazo retener un cliente que arriesgarse a captar uno nuevo siendo el costo por captarlo más alto.

Justificación Metodológica. La entidad bancaria hoy en día no cuenta con una metodología que determine con precisión la predicción y clasificación de los clientes con riesgo de fuga, debido a que solo se basan en el conocimiento del negocio para implementar planes de acción en sus campañas. Es por ello que utilizando herramientas estadísticas se propone utilizar Regresión Logística o Árboles de Clasificación para resolver el problema de investigación.

Es en este contexto, se plantea el presente trabajo de investigación, cuyo objetivo principal se concentrara en clasificar y predecir a los clientes con riesgo de fuga, mediante la determinación de patrones que determinen la situación de cada cliente.

3. MARCO TEÓRICO

3.1. Modelos Estadísticos

El objetivo del trabajo se centra en predecir el comportamiento de una variable categórica con dos grupos: ser un cliente con riesgo de fuga o no.
En el presente estudio se trabaja con dos técnicas estadísticas: Regresión Logística y Árboles de clasificación.

3.1.1. Regresión Logística

3.1.1.1. El Modelo de Regresión Logística

El problema de discriminación se convierte en prever el valor de la variable ficticia Y en un nuevo elemento del que conocemos variables X . Si el valor previsto está más próximo a cero que a uno, se clasificará al elemento en la primera población. En otro caso, se hará en la segunda. Para modelar este tipo de relaciones se utilizan los modelos de respuesta cualitativa. El modelo de esta clase más utilizado es el modelo logístico.

Este modelo requiere realizar menos supuestos, lo que permite obtener resultados más robustos y es flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de escala y categóricas. Permite estudiar el impacto que tiene cada una de las variables independientes en la probabilidad de que ocurra el suceso de estudio.

3.1.1.2. El Modelo multivariante de Regresión Logística binaria

Para este modelo se considera que la variable respuesta, es una variable dicotómica es decir que toma dos valores. Para nuestra investigación, la

variable respuesta sería si el cliente fuga o no fuga, en donde a los clientes fuga se les asigna el valor de 1 y a los clientes no fuga el valor de 0.
Para estos modelos dicotómicos, las dos categorías deben de ser mutuamente excluyentes, es decir que el cliente debe estar adscrito a una, y solamente una, de esas dos alternativas. En nuestro caso, el cliente no puede ser fuga y no fuga a la vez, sólo puede optar por una de las dos alternativas.

La variable respuesta se puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, \mathrm{P}(Y_i = 1) = P_i \\ 0, \mathrm{P}(Y_i = 0) = 1 - P_i \end{cases}$$

Tenemos que: $E(Y_i) = 1xP_i + 0x(1 - P_i) = P_i$

Se puede apreciar que la media teórica es igual a la probabilidad que Y_i tome el valor de 1. Vamos a considerar que Y_i es explicado por las variables independientes X_{2i}, X_{3i}, …, X_{ki}, designamos Z_i como:

$$Z_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} = \begin{bmatrix} 1 & X_{2i} & \ldots & X_{ki} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Entonces la esperanza de Y_i dado las variables independientes, es:

$$E(Y_i / X_{2i}, \cdots, X_{ki}) = F(\beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}) = F(Z_i)$$

El modelo quedaría como:

$$Y_i = E(Y_i / X_{2i}, \cdots, X_{ki}) + U_i = F(Z_i) + U_i$$

U_i es una perturbación aleatoria.

Según lo cual, la función F puede generar distintos modelos dicotómicos, como el modelo de probabilidad lineal, el modelo probit y el modelo logit. Para nuestro caso, usaremos el modelo logit.

Para nuestro caso tomamos como función F a la función logística, obteniéndose el modelo logit:

$$P_i = E(Y_i / X_{2i}, \cdots, X_{ki}) + U_i = F(Z_i) = \frac{1}{1 + e^{-Z_i}} = \Lambda(Z_i)$$

3.1.1.3. Coeficientes de Regresión

Para contrastar la significatividad global en los modelos logit, se utiliza el estadístico de razón de verosimilitud (RV). Normalmente se usa la prueba ómnibus.

En este caso la hipótesis nula que se desea contrastar es que todos los coeficientes de las variables independientes son iguales a cero (significancia de modelo), es decir, determinar si al menos una de las variables independientes es significativa o no, para ello se plantea las siguientes hipótesis:

$$H_o : \beta_1 = \beta_2 = \beta_3 = \ldots \beta_k = 0$$

$$H_a : \textit{Al menos uno de los coeficientes es distinto de cero}$$

En un modelo con término independiente β_0, es decir no rechazar H_o, sería tanto como admitir que el modelo que solo incluye la constante predice mejor los valores observados de Y que el modelo ajustado en cuestión con n variables predictoras. Por el contrario, si H_o fuese rechazado, esto indicaría que al menos uno de los coeficientes es distinto de cero.

En la regresión logística este se realiza por medio del test G o prueba de la razón de verosimilitud. El estadístico G de razón de verosimilitud. Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico G sigue una χ^2 (no se asume normalidad). La ausencia de significación implica que el modelo sin la predictora no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha predictora debe ser eliminada del modelo ya que no aporta nada al mismo. Esta prueba no asume ninguna distribución concreta, por lo que es la más recomendada para estudiar la significación de los coeficientes.

$$G = \frac{-2\ln\ verosimilitud\_del\_modelo\_solo\_con\_la\_constante(Lo)}{verosimilitud\_del\_modelo\_solo\_seleccionado(Lp)}$$

Se distribuye como una Chi-cuadrado con $p-1$ grados de libertad, donde p representa el número de parámetros en el modelo sometido al estudio. Este estadístico se basa en la función de verosimilitud de cada modelo y, en definitiva, compara la probabilidad de que los datos estimados por cada uno de los modelos representan a los valores realmente observados de la variable respuesta (Hosmer y Lemescow 2000), (Ruiz y Maya L.1995).

3.1.1.4. Medidas de bondad de ajuste

3.1.1.4.1. Coeficiente Pseudo R^2

En vez de utilizar el R^2 de la regresión lineal, en los modelos logit se utilizan los Pseudo R^2, el cual es un estadístico que mide la bondad de ajuste del modelo a los datos.

R cuadrado de Mac Fadden

Es una aproximación basada en una comparación de la verosimilitud del modelo solo con la constante, con la verosimilitud del modelo con todos los parámetros.

$$PR^2_{MacFadden} = 1 - \frac{LnL_1}{LnL_0}$$

Siendo "L_1" es el estimador de máxima verosimilitud del modelo con todas las variables explicativas y "L_0" es el estimador de máxima verosimilitud del modelo sin variables explicativa. El Pseudo R^2 es una medida útil del ajuste del modelo a los datos, y puede servir para comparar la capacidad explicativa de modelos distintos.

R cuadrado de Cox y Snell

El coeficiente R^2 de Cox y Snell es un estadístico basado en el logaritmo de la verosimilitud, pero toma en cuenta el tamaño N de la muestra. Esta medida para la regresión logística alcanza el valor máximo de 1.

$$PR^2_{Cox-Snell} = 1 - e^{-\frac{2}{N}(LnL_0 - LnL)} = 1 - e^{-\frac{RV_0}{N}}$$

Finalmente, el **R cuadrado de Nagelberke**

Nagelberke (1991) propuso una modificación de la R^2 de Cox y Snell de tal manera que se pudiera alcanzar el valor 1.

$$PR^2_{Nagelberke} = \frac{PR^2_{Cox-Snell}}{PR^2_{MaxCox-Snell}} = \frac{PR^2_{Cox-Snell}}{1 - e^{-\frac{2}{N}LnL_0}}$$

3.1.1.4.2. Prueba de bondad de ajuste de Hosmer y Lemeshow

El Test de Hosmer y Lemeshow es muy utilizado en Regresión logística. Se trata de un test de bondad de ajuste al modelo propuesto. Un Test de bondad de ajuste lo que hace es comprobar si el modelo propuesto puede explicar lo que se observa. Es un Test donde se evalúa la distancia entre un observado O_i y un esperado E_i.

$$HL = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \text{, donde } E_i = n_i \bar{p}_i$$

Donde n_i es el número de patrones de predictoras del grupo i-esimo, es decir el numero de respuestas afirmativas registradas para la variable respuesta ($Y = 1$) para los n_i patrones de predictoras.

Donde:

O_i : número de unos en el decil i-ésimo

$\bar{p}_i$: media de probabilidades en el decil i-ésimo

n_i : número de observaciones en el decil i-ésimo

Las hipótesis nula y alternante son:

H_0 : El modelo de regresión logística se ajusta a los datos.

H_a : El modelo de regresión logística no se ajusta a los datos.

"HL" se distribuye como una Chi-cuadrado, el criterio de decisión será sí $\chi^2 \geq \chi^2_{\alpha,i}$, no se rechaza la hipótesis nula y se concluye que le modelo de ajusta a los datos. Si rechazamos la "H_0", implica que el modelo ajustado no es el adecuado.

3.1.1.4.3. Curva ROC

La curva ROC (del inglés *Receiver Operating Characteristic curves*) indica que cuanto más alejada este de la diagonal principal mejor es el método de diagnóstico, ya que la curva ROC ideal sería la que con una especificidad de 1 tuviera una sensibilidad de 1 (la especificidad y sensibilidad se explicarán en el punto 3.1.1.4.4), y cuanto más cercana esté a dicha diagonal peor será el método de diagnóstico. Cabe recordar que la diagonal principal es la que corresponde al peor test de diagnóstico y que tiene un área bajo de ella de 0.5. Esto diagnóstico se puede afirmar con el área bajo la curva ROC, que se utiliza como medida de discriminación y representa para todos los pares posibles de individuos formados por un individuo en el que ocurrió el evento y otro en el que no, la proporción de los que el modelo predice una mayor probabilidad para el que tuvo el evento.

Las hipótesis nula y alternante son:

H_0: El área bajo la curva ROC es igual a 0.5

H_a: El área bajo la curva ROC no es igual a 0.5

Si rechazamos la "H_0" asociado a un "p-value", implica que el modelo ajustado es el adecuado. A partir de un área de 0,7 la discriminación del modelo se considera aceptable.

3.1.1.4.4. Tablas de clasificación

La tabla de clasificación muestra la distribución de valores observados y estimados. Los valores estimados se obtienen a partir del modelo.

La regla de clasificación predeterminada para un caso es que si la probabilidad estimada de pertenencia en el grupo de respuesta con el valor más alto es mayor o que igual a 0.5, entonces predice la pertenencia a ese grupo. De lo contrario, predice la pertenencia al grupo con el valor de respuesta más bajo.

Por otro lado, la capacidad de que nuestro modelo estime el suceso de interés cuyo valor es 1, se denomina sensibilidad. Por el contrario, la capacidad de que nuestro modelo no estime el suceso de interés cuyo valor es 0, se denomina especificidad.

3.1.2. CART

3.1.2.1. Árboles de Clasificación

El análisis del árbol de clasificación llamado también de decisión o de identificación, es una técnica de segmentación diseñada para dividir a una población en dos o más grupos basándose en sus atributos, por ejemplo: se desea conocer que segmento de personas se encuentran predispuestas a tener hipertensión (género, edad, antecedentes familiares, etc.).

El modelo de árbol de clasificación es una técnica no paramétrica que presentan un estructura en forma de árbol, en donde las ramas representan conjuntos de decisiones; estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos de datos disjuntos y exhaustivos. Las ramificaciones se generan de forma recursiva hasta que se cumplan ciertos criterios de parada.

Los árboles de decisión son empleados para clasificar y pronosticar, es decir identificar el resultado categórico atendiendo a una serie de criterios dados y

pronosticar el resultado según una futura serie de criterios o variables independientes.

El objetivo de este método es obtener individuos u objetos más homogéneos con respecto a la variable discriminadora dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas a partir de las cuales se va a realizar la discriminación de la población en subgrupos.

Dentro de los métodos basados en árboles se pueden distinguir dos tipos dependiendo de tipo de variable a discriminar:

Árboles de clasificación. Este tipo de árboles se emplea para variables categóricas, tanto nominales como ordinales.

Árboles de regresión. Este tipo de discriminación se aplica a variables continuas.

Diferentes algoritmos pueden ser usados para construir arboles de decisión tales como la Detección Automática de Interacciones (CHAID) y Árboles de Regresión (CART), QUEST y C5.0.

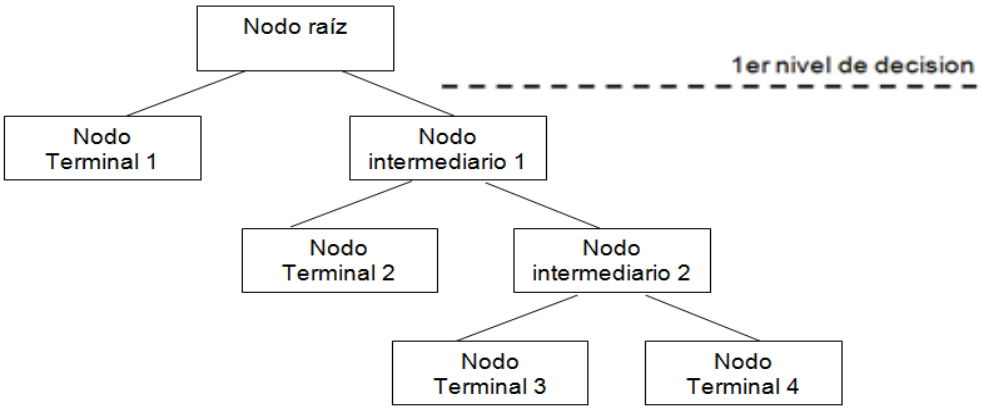
El Árbol de Clasificación comienza con un nodo al que pertenecen todos los casos de la muestra a clasificar (nodo raíz), el resto de nodos se dividen en nodos intermedios o no terminales y nodos hojas o nodos terminales, ver Gráfico N°1.

Un árbol de decisión consta de los siguientes elementos:

Nodo intermediario: se generan dos o más segmentos descendientes inmediatos (dependiendo del método empleado). También llamados segmentos intermedios.

Nodo terminal: Es un nodo que no se puede dividir más. También denominado segmento terminal. Rama de un nodo t: Consta de todos los segmentos descendientes de t, excluyendo t.

Gráfico N°1 Árbol completo generado por el algoritmo CART



3.1.2.2. El Modelo de CART

El procedimiento para este modelo no utiliza un modelo estadístico formal fue desarrollado por matemáticos de la universidad de Berkeley y Stanford (Breiman, Friedman, Olshen y Stone) a mediados de los 80, para clasificar, se utilizan particiones binarias sucesivas de los valores de una variable, trabaja con variables de todo tipo. El corte en cada nodo viene dado por reglas de tipo binario. Se pueden formular como preguntas: ¿ $X_K < a$? ¿Pertenece X_K a un subconjunto E de estados?

3.1.2.3. Medidas de Impureza

Impureza de un nodo

Una medida cuantitativa de la homogeneidad es la noción de impureza. La idea es la siguiente:

$$\text{Impureza de un nodo} = \frac{\text{Número de sujetos que cumplen la característica en el nodo}}{\text{Número total de sujetos en el nodo}}$$

Para decidir qué variable va a utilizarse para hacer la partición en un nodo se calcula primero la proporción de observaciones que pasan por el nodo para cada uno de los grupos. Si se denomina a los nodos como $t = 1, 2, \ldots, T$ y $p(g/t)$ a las probabilidades de que las observaciones que lleguen al nodo t pertenezcan a cada una de las clases, se define la impureza del nodo t como:

$$i(t) = \phi \quad p(1/t), p(2/t), \ldots . p(G/t)$$

Donde: ϕ es la función de impureza y, $p(g/t)$ puede calcularse empíricamente como la proporción de casos de clase g en el nodo t:

Es decir:
$$p(g/t) = \frac{n_g(t)}{n(t)}$$

La variable que se introduce en un nodo es la que minimiza la heterogeneidad o impureza que resulta de la división en el nodo. La clasificación de las observaciones en los nodos terminales se hace asignando todas las observaciones del nodo al grupo más probable en ese nodo, es decir, el grupo con máxima $p(g/t)$. Si la impureza del nodo es cero, todas las observaciones pertenecerían al mismo nodo, en caso contrario puede haber cierto error de clasificación. Cuando el número de variables es grande, el árbol puede contener un número excesivo de nodos por lo que se hace necesario definir procedimientos de poda o simplificación del mismo.

Bondad de una partición – Índice de Gini

Se tiene el índice de Gini en el nodo t, $i(t)$, definida como:

$$i(t) = g(t) = 1 - \sum_{g=1}^{G} p(g/t)^2$$

Este índice es una medida de impureza en la clasificación de los datos, a medida que ve van clasificando correctamente los datos, el índice de Gini va tomando valores cercanos a 0.

La función del criterio Gini $\phi(s,t)$ para la división s en el nodo t se define como:

$$\phi(s/t) = g(t) - p_L g(t_L) - p_R g(t_R)$$

Así, como conocemos cómo calcular $g(t)$, podemos calcular $\phi(s,t)$ para cada partición s y seleccionar la mejor partición como la que proporciona la mayor bondad $\phi(s,t)$. Para establecer el efecto que produce la selección de la mejor partición en cada nodo sobre el árbol final necesitamos una medida de la impureza global del árbol. Este valor, ponderado por la proporción de todos los casos del nodo t, es el valor del que se informa en el árbol como "mejora ".

Construcción del árbol de clasificación con el algoritmo CART tiene los siguientes pasos:

1. Para llevar a cabo un análisis CART, comenzado por el nodo raíz $t = 1$, buscar la división, de entre todos los candidatos posibles S, que dé lugar a la mayor reducción de la impureza:

   $$\phi(s^*,1) = \max_{s \in S} \phi(s,1)$$

   Luego dividir el nodo $g(t = 1)$ en dos nodos, $t = 2$ y $t = 3$, utilizando la división S^*.

2. Repetir el proceso de búsqueda de divisiones para uno de los nodos $t = 2$ y $t = 3$, y así sucesivamente.
3. Continuar con el proceso de desarrollo del árbol hasta alcanzar al menos una de las reglas de parada.

Se detiene el proceso de desarrollo del árbol cuando se cumple una de las diversas reglas de parada disponibles. Un nodo no se dividirá si se cumplen alguna de las siguientes condiciones:

a) Se ha alcanzado la máxima profundidad del árbol permitida.
b) No se pueden realizan más particiones, porque se ha verificado alguna de las siguientes condiciones:
   - No hay variables explicativas significativas para realizar la partición del nodo.
   - El número de elementos en el nodo terminal es inferior al número mínimo de casos permitidos para poder realizar la partición.
   - El nodo no se podrá dividir en el caso en el cual el número de casos en uno o más nodos hijos sea menor que el mínimo número de casos permitidos por nodo.

3.1.2.4. Estimación de la tasa de error del árbol

La elección de un árbol respecto de otro dependerá en general de una estimación de su tasa de error R(T). El problema es cómo realizar la estimación de dicha tasa, por ello existen diversas formas de calcular la estimación con una serie de ventajas e inconvenientes que se detallan a continuación:

Estimador por resustitución (estimación intramuestral): Es el estimador más simple. Consiste en dejar caer por el árbol la misma muestra que ha servido para construirlo, pero como los árboles tienen gran flexibilidad para

adaptarse a la muestra se puede obtener una estimación sesgada inferiormente de la tasa de error, y por tanto desconocer realmente el error real del árbol.

Estimador por muestra de validación (muestra de contraste): Consiste en dejar caer por el árbol una muestra distinta a la empleada para la realización del árbol. Por ello éste no se ha podido adaptar a dichos registros como ocurría en el estimador anterior. Tenemos de esta forma un estimador de R(T) insesgado, sin embargo este tiene el inconveniente de forzar a reservar, para su uso en la validación, una parte de la muestra la cual podía haberse empleado en la construcción del árbol. Por lo que hay cierta pérdida de información. Este estimador es empleado cuando se tiene tamaño de muestra muy grande.

Estimación por validación cruzada: consiste en estimar R(T) procediendo de forma reiterada similar al estimador por muestra de validación. Se deja fuera de la muestra a una fracción m^1 del tamaño muestral total para la construcción del árbol. Obteniéndose de esta forma m estimaciones $R^{(1)}(T) + ... + R^{(m)}(T)$ y promediándolas de la siguiente forma:

$$R^{VC}(T) = \frac{R^{(1)}(T) + ... + R^{(m)}(T)}{m}$$

Observar que el árbol realizado para cada una de las submuestras podría ser distinto a los demás, en este caso la expresión anterior no sería válido.

Estimador bootstrap: Recientemente se ha propuesto esta técnica de remuestreo para la estimación de la tasa de error. Ripley (1996).

4. METODOLOGÍA DE LA INVESTIGACIÓN

4.1. Tipo de Investigación

La presente investigación es de tipo correlacional, ya que se comprueba si existe alguna relación entre las variables independientes y la variable dependiente de manera individual, después se halla que variables influyen en la variable independiente (riesgo de fuga de un cliente en la entidad bancaria).

Por otro lado el estudio es del tipo predictivo ya que se obtuvo una regla que permita identificar si un cliente con ciertas características tiene riesgo de fuga o no en la entidad bancaria.

4.2. Formulación de Hipótesis

En el presente trabajo se busca determinar la predicción y clasificación de fuga de clientes mediante una regresión logística binaria y Árboles de Clasificación. Podemos formular las siguientes hipótesis:

- El riesgo de fuga de un cliente puede ser explicado con las variables predictoras: edad, sexo, estado civil, ingreso bruto, número de tarjetas usadas, monto, saldo, número de transacciones y línea de crédito.

- La regresión logística binaria es el método estadístico que mejor predice y clasifica los clientes fuga en una entidad bancaria en comparación con el método árbol de clasificación CART.

4.3. Identificación de Variables

4.3.1. Variable dependiente

Para el presente trabajo se define como variable dependiente a la variable fuga de clientes, que identifica si un cliente está en situación de fuga o no de la entidad bancaria.

Tipo	Variable	Descripción	Valores	Etiquetas
Dicotómica	Situación	Riesgo de fuga de la entidad bancaria	0 1	No Fuga Fuga

Esta variable se obtuvo al identificar a los clientes para No Fuga (clientes activos sin restricción y los clientes activos que tienen una mora de 5 – 89 días) para Fuga (clientes inactivos que tienen mora mayor a 90 días)

4.3.2. Variables independientes

Variables Sociodemográficas

Tipo	Variable	Descripción	Valores	Etiquetas
Continua	Edad	Edad del cliente de 20 hasta 90 años	Número de años	
Dicotómica	Sexo	Sexo del cliente	0 1	F M
Categórica	Estado Civil	Estado Civil del cliente	0 1 2 3	C D S V
Continua	Ingreso_Bruto	Sueldo que el cliente declara percibir	Cantidad de sueldo	

Variables de Comportamiento Bancario

Tipo	Variable	Descripción
Discreta	Nro_Tarjetas_usadas	Número de tarjetas usadas en el periodo en estudio sin incluir la tarjeta de la entidad bancaria.
Continua	Monto	Monto que consumió el cliente
Continua	Saldo	Saldo del cliente en la tarjeta de crédito
Discreta	Trxs_Total	Cantidad de transacciones realizadas en el periodo
Discreta	LineaCredito	La línea de crédito en la tarjeta de crédito

4.4. Diseño de Investigación

La datos se obtuvieron a través de Data Warehouse de la Entidad Bancaria, de toda la cartera de clientes se seleccionó para los clientes “No fuga” aquellos que cumplan con ser clientes activos sin restricción y los clientes activos que tienen una mora de (5 a 89 días), para Fuga se considero a clientes inactivos que tienen mora mayor a 90 días.

Debido a la naturaleza dicotómica de la variable respuesta se utilizó las metodologías: Regresión Logística Binaria y los Árboles de Clasificación CART, es decir, se obtuvo un modelo con cada método y se escogió el que mejor se ajusta a los datos, por tanto puede predecir y clasificar mejor si un cliente tiene riesgo de fuga o no.

4.4.1. Fuente de información

La recolección de datos se obtuvo a través de Data Warehouse de la Entidad Bancaria la cual está orientada a un determinado ámbito (empresa, organización, etc.), es integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se trata, sobre todo, de un expediente completo de una organización, más allá de la información

transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (especialmente OLAP, procesamiento analítico en línea). El almacenamiento de los datos no debe usarse con datos de uso actual. Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas dependiendo del subsistema de la entidad del que procedan o para el que sea necesario.

4.5. Población y Muestra

Población

Toda la cartera de clientes que poseen una tarjeta de crédito en la entidad financiera en Mayo del 2012, clientes activos sin restricción y los clientes activos que tienen una mora de (5 a 89 días); así mismo, a clientes inactivos que tienen mora mayor a 90 días .Finalmente se obtuvo N=67,654

Muestra

Se realizó un muestreo aleatorio simple, para la selección de la muestra ver Cuadro N°1 . Para el presente trabajo se seleccionó a 8,410 clientes.

Cuadro N°1: Selección del tamaño de muestra

Población	$N = 67,654$
1% de error de muestreo	$Error = 0.01$
95% de confianza	$z = 1.96$
Variación máxima	$p = 0.5$

$$n_0 = \frac{z^2 p(1-p)}{(error)^2} = \frac{(1.96)^2(0.5)(1-0.5)}{(0.01)^2} = 9,604$$

Ajustando el tamaño de la muestra

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{9,604}{1+\frac{9,604}{67,654}} = 8,410.11 \cong 8,410$$

4.6. Diseño de la muestra

Gráfico N°2: Diseño de la muestra

De la muestra de 8,410 clientes de la cartera, haremos una partición una aleatoria de 6,653 es decir 80% para la base de entrenamiento (base para identificar las variables que influyen o tienen asociación con la variable respuesta y elaborar el modelo) y 1,757 es decir 20% para la base de validación (base para probar y validar el modelo obtenido).

5. PROCEDIMIENTO DE ANÁLISIS DE DATOS

5.1. Análisis Exploratorio

Se realizo el análisis de regresión logística binaria para explorar las posibles asociaciones entre la variable respuesta y las distintas variables predictoras para analizar si alguna de las variables predictoras no se encuentra asociada a la variable dependiente. Se construyeron tablas de contingencia entre cada una de las variables predictoras con la variable respuesta. Así como histogramas de frecuencias y ploteo de los gráficos de dispersión, para explorar cuáles son las posibles variables que tienen un efecto significativo en la variable respuesta y si la relación es directa o inversa.

Cuadro N°2:Estadísticos descriptivos de las variables cuantitativas

	N	Mínimo	Máximo	Media	Desv. típ.
Edad	6653	20	90	43.66	12.748
Ingreso_Bruto	6653	750	4500	1909.92	943.258
Nro_Tarjetas_usadas	6653	.00	4.00	1.0120	.97172
Monto	6653	0	7756	172.16	447.306
Saldo	6653	0	3500	394.44	684.916
Trxs_Total	6653	0	39	1.86	3.442
LineaCredito	6653	0	8000	1585.81	1937.681
N válido (según lista)	6653				

Fuente: Elaboración propia

Cuadro N°3 :Pruebas de Chi-cuadrado de Pearson para la prueba de asociación de la variable respuesta SITUACION vs las variables predictoras

	Valor	gl	Sig. asintótica (bilateral)
Sexo	18.780	1	.000
Estado_Civil	15.739	3	.001

Fuente: Elaboración propia

Todas las pruebas son significativas por lo que concluimos que existe una asociación entre la variable dependiente y las predictoras.

Posteriormente se realizó un análisis de correlación entre las variables predictoras, para garantizar la inclusión de variables predictoras no correlacionadas en el análisis de regresión.

Cuadro N°4: Variables que tienen asociación con la variable Situación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1a	LineaCredito	-.005	.000	1091.369	1	.000	.995
	Constante	4.134	.120	1181.229	1	.000	62.429
Paso 2b	Trxs_Total	-4.103	.317	167.701	1	.000	.017
	LineaCredito	-.005	.000	520.122	1	.000	.995
	Constante	5.210	.191	743.080	1	.000	183.150
Paso 3c	Nro_Tarjetas_usadas	-.756	.086	77.064	1	.000	.470
	Trxs_Total	-3.609	.310	135.737	1	.000	.027
	LineaCredito	-.005	.000	469.984	1	.000	.995
	Constante	5.729	.215	711.269	1	.000	307.806
Paso 4d	Ingreso_Bruto	.000	.000	4.567	1	.033	1.000
	Nro_Tarjetas_usadas	-.753	.086	76.080	1	.000	.471
	Trxs_Total	-3.640	.315	133.412	1	.000	.026
	LineaCredito	-.005	.000	471.559	1	.000	.995
	Constante	5.379	.264	414.631	1	.000	216.728

a.Variable(s) introducida(s) en el paso 1: LineaCredito.
b.Variable(s) introducida(s) en el paso 2: Trxs_Total.
c.Variable(s) introducida(s) en el paso 3: Nro_Tarjetas_usadas.
d.Variable(s) introducida(s) en el paso 2:Ingreso_Bruto.

Fuente: Elaboración propia

El Cuadro N°4 indica que variables de todas las inicialmente seleccionadas tienen alguna asociación a la variable Situación.

Podemos ver que las variables predictoras asociadas son Ingreso_Bruto, Nro_Tarjetas_usadas, Trxs_Total y LineaCredito.

5.2. Análisis de Regresión Logística

Se realizó un modelo de regresión logística binaria mediante el método "adelante wald" al grupo de variables que tienen alguna asociación con la variable dependiente "Situación", identificados con el análisis exploratorio realizado en el punto (5.1), para buscar el mejor grupo de variables independientes. Con lo cual obtuvimos los siguientes resultados:

Cuadro N°5:Codificación de la variable dependiente

Valor original	Valor interno
No Fuga	0
Fuga	1

El Cuadro N°5 especifica la codificación de la variable dependiente que es dicotómica. Se asigna el valor "0" a los clientes que no tienen riesgo de fuga y " 1 " a los clientes que si tienen riesgo de fuga de la entidad financiera, lo cual es correcto.

Cuadro N°6: Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 4	Paso	4.672	1	.031
	Bloque	8136.028	4	.000
	Modelo	8136.028	4	.000

En el Cuadro N°6 tenemos la Prueba ómnibus la cual se somete a la hipótesis que los coeficientes de las variables introducidas en el último paso son iguales a 0.

$$H_0 : \beta_i = 0 \qquad i = 1, 2, 3, \ldots, k$$
$$H_a : \beta_i \neq 0$$

Si observa el p valor, se puede determinar que al menos algún $\beta_i \neq 0$, dado que es significativo nos indica que el modelo con las nuevas variables mejora el ajuste con respecto a lo que se tenía con solo el parámetro constante.

Cuadro N°7: Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
4	1086.804[a]	.706	.941

a. La estimación ha finalizado en el número de iteración 11 porque las estimaciones de los parámetros han cambiado en menos de .001.

En el Cuadro N°7 vemos el -2 log de la verosimilitud al ser un valor alto eso indica que el modelo se ajusta a los datos. La siguiente medida es el Pseudo R cuadrado de Cox y Snell es 0.706, lo cual indica que el 70.6% de la variabilidad de la variable respuesta explicada por las variables predictoras, por otro lado el R cuadrado de Nagelkerke lo explica en 94.1% es decir que en este modelo las variables independientes explican el 94.1% del comportamiento de la variable dependiente.

Bondad de Ajuste

Prueba de bondad de ajuste de Hosmer y Lemeshow

Cuadro N°8: Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.
4	9.678	8	.288

El Cuadro N° 8 es la Prueba de Hosmer y Lemeshow

H_0 : El modelo de regresión logística se ajusta a los datos.

H_a : El modelo de regresión logística no se ajusta a los datos.

Prueba: $HL = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

Como resultado se tiene que resulta no significativa con 0.288, con lo que se llega a la conclusión de que los datos se ajustan al modelo de Regresión Logística.

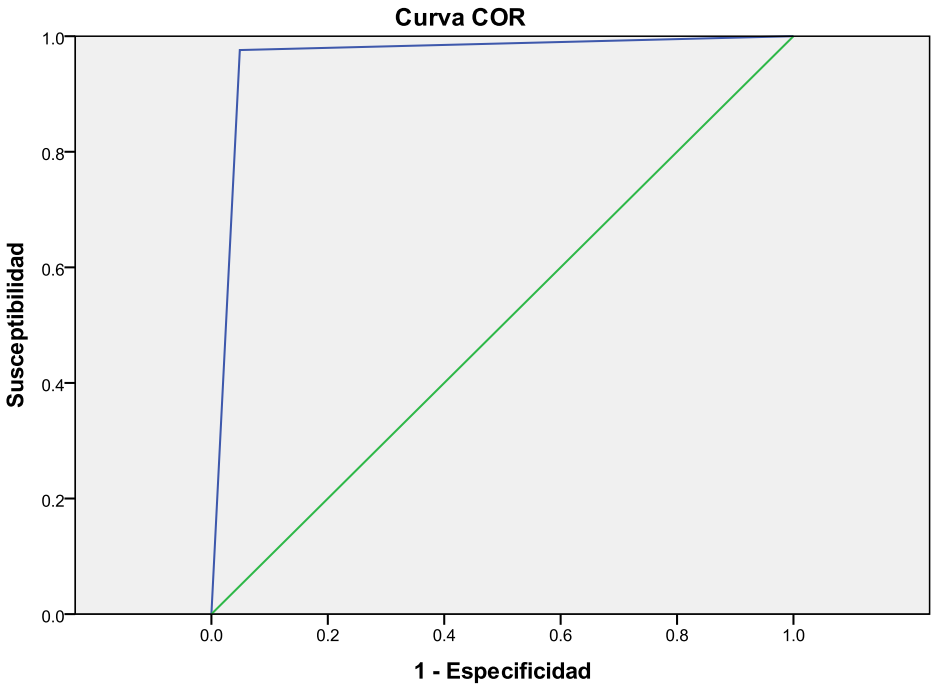
Curva ROC

Tenemos la hipótesis

H_0 : El área bajo la curva ROC es igual a 0.5

H_a : El área bajo la curva ROC no es igual a 0.5

Grafico N°3



Los segmentos diagonales son producidos por los empates.

Cuadro N°9: Área bajo la curva

Variables resultado de contraste: Grupo pronosticado

Área
.964

La variable (o variables) de resultado de contraste: Grupo pronosticado tiene al menos un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Los estadísticos pueden estar sesgados.

El área es de 0.964 por tanto dado que es mayor que 0,7 se concluye que la discriminación del modelo se considera aceptable.

Tablas de clasificación

Cuadro N°10 Tabla de clasificación[a]

Observado			Pronosticado		
			Situación		Porcentaje correcto
			No Fuga	Fuga	
Paso 4	Situación	No Fuga	3200	144	95.7
		Fuga	69	3240	97.9
	Porcentaje global				96.8

a. El valor de corte es .500

En el Cuadro N°10 se determina la eficacia predictiva del modelo, es aceptable dado que clasifica correctamente a los cliente Fuga al 97.9% con un porcentaje global de clasificación correcta de 96.8%.

Cuadro N°11:Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 1a	LineaCredito	-.005	.000	1091.369	1	.000	.995	.994	.995
	Constante	4.134	.120	1181.229	1	.000	62.429		
Paso 2b	Trxs_Total	-4.103	.317	167.701	1	.000	.017	.009	.031
	LineaCredito	-.005	.000	520.122	1	.000	.995	.994	.995
	Constante	5.210	.191	743.080	1	.000	183.150		
Paso 3c	Nro_Tarjetas_usadas	-.756	.086	77.064	1	.000	.470	.397	.556
	Trxs_Total	-3.609	.310	135.737	1	.000	.027	.015	.050
	LineaCredito	-.005	.000	469.984	1	.000	.995	.994	.995
	Constante	5.729	.215	711.269	1	.000	307.806		
Paso 4d	Ingreso_Bruto	.000	.000	4.567	1	.033	1.000	1.000	1.000
	Nro_Tarjetas_usadas	-.753	.086	76.080	1	.000	.471	.398	.558
	Trxs_Total	-3.640	.315	133.412	1	.000	.026	.014	.049
	LineaCredito	-.005	.000	471.559	1	.000	.995	.994	.995
	Constante	5.379	.264	414.631	1	.000	216.728		

a.Variable(s) introducida(s) en el paso 1: LineaCredito.
b.Variable(s) introducida(s) en el paso 2: Trxs_Total.
c.Variable(s) introducida(s) en el paso 3: Nro_Tarjetas_usadas.
d.Variable(s) introducida(s) en el paso 2:Ingreso_Bruto.
Fuente: Elaboración propia

Por último el Cuadro N°11 muestra a las variables independientes: LINEACREDITO, TRXS_TOTAL, NRO_TARJETAS_USADAS y INGRESO_BRUTO.

De Cuadro N°11 se obtiene la ecuación para el modelo de la regresión logística con los pesos y significación de cada variable independiente en la predicción del cliente Situación Fuga.

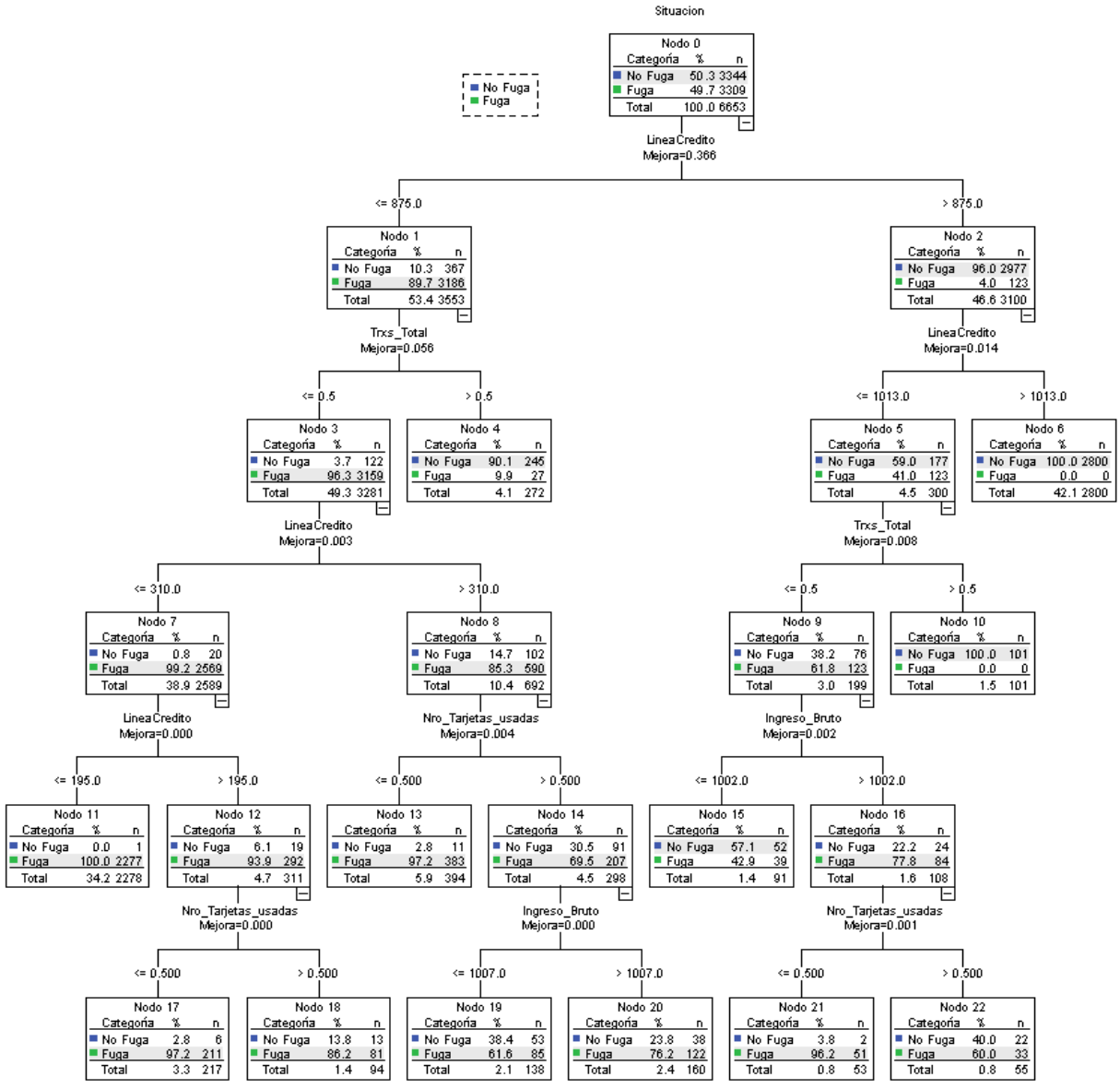
$$\text{Logit}\left[P_{(Situación=1)}\right] = 5.379 + 0.00025*\text{Ingreso\_Bruto} - 0.753*\text{Nro\_Tarjetas\_usadas}$$

$$-3.640*\text{Trxs\_Total} - 0.0052*\text{LineaCredito}$$

5.3. Análisis del Algoritmo de Árbol de Clasificación CART

Se generó el Árbol de Clasificación utilizando el algoritmo CART, para poder predecir las variables predictoras con análisis de sensibilidad basado en Gini y sistema de validación por muestra de contraste.

Gráfico N°4: Árbol de Clasificación CART con las variables predictoras



Fuente: Elaboración propia

El Grafico N°4 muestra el Árbol de Clasificación del cual se puede interpretar:

Primer nivel

- La variable LineaCredito es el mejor predictor para la variable dependiente Situación de los clientes de la entidad financiera.
- El 89.7% de los clientes fuga tienen LineaCredito menor igual 875 soles.
- Mientras que el 96.0% de los clientes no fuga tienen LineaCredito mayor a 875 soles.

Segundo nivel

- Se puede observar que como segundo predictor en significancia se tiene a la variable Trxs_Total, en este nodo los clientes con mayor igual a 1 transacción tienen 90.1% de probabilidad de fuga.
- Mientras que el 100% de los clientes que tienen LineaCredito mayor a 1013 soles no fuga.

Tercer nivel

- De los clientes que tienen LineaCredito mayor a 875 soles y menor igual a 1013 soles y presentan Trx_Total mayor a 0.5 el 100% de clientes no fuga.

Cuarto nivel

- En este nivel ingresan dos variables como predictoras estas son Nro_Tarjetas_usadas e Ingreso_Bruto.
- El 100% de los clientes que tienen una LineaCredito menor igual a 195 soles se Fuga.
- Para los que tenían la LineaCredito mayor a 310 soles y no tienen Nro_Tarjetas_usadas el 97.2% se Fuga.

- También vemos que los clientes que tienen LineaCredito mayor a 875 soles y menor igual a 1013 soles y presentan Trx_Total menor igual a 0.5 y su Ingreso_Bruto es menor igual a 1002 soles el 57.1% de clientes No fuga.

Quinto nivel

- Las variables significativas en este nivel son Nro_Tarjetas_usadas e Ingreso_Bruto.
- Donde el 97.2% de clientes que no tiene Nro_Tarjetas_usadas son clientes fugados.
- El 76.2% de los clientes que tiene Ingreso_Bruto mayor a 1007 presenta situación de fuga para la entidad financiera.
- De los clientes que tienen Ingreso_Bruto es mayor a 1002 soles y el Nro_Tarjetas_usadas es es menor igual a 0.5 es decir no tienen tarjetas usadas el 96.2% de estos clientes son Fuga.

Del Grafico N°4 también podemos concluir el análisis del algoritmo de árbol de clasificación CART para determinar los patrones :

No Fuga:

Se registra con más claridad en los nodos 4, 6 y 10 el patrón de clientes que deciden quedarse:

Nodo 4: Son los clientes que tienen una Línea de Crédito menor igual 875 soles y realizan 1 a mas transacciones quienes representan un total de 7.3% de clientes que deciden quedarse sobre el total de clientes No Fuga de la entidad bancaria.

Nodo 6: Son los clientes que tienen una Línea de Crédito mayor a 1013 soles quienes representan un total de 83.7% de clientes que deciden quedarse sobre el total de clientes No Fuga de la entidad bancaria.

Nodo 10: Son los clientes que tienen una Línea de Crédito mayor a 1013 soles y realizan 1 a mas transacciones quienes representan un total de 3% respecto al total de clientes No Fuga de la entidad bancaria.

Fuga:

Se registra con más claridad en los nodos 11,13,17,20 y 21 el patrón de clientes que deciden irse de la entidad bancaria.

Nodo 11: Son los clientes que tienen una Línea de Crédito menor igual 875 soles ,que no presentan transacciones y de estos quienes tienen Línea de Crédito mayor a 195 soles representan un total de 69% de clientes que deciden irse sobre el total de clientes Fuga de la entidad bancaria.

Nodo 13: Son los clientes que tienen una Línea de Crédito menor igual 875 soles ,que no presentan transacciones y de estos quienes tienen Línea de Crédito mayor a 310 soles y que no han utilizado ninguna tarjeta quienes representan un total de 11.6% de clientes que deciden irse sobre el total de clientes Fuga de la entidad bancaria.

Nodo 17: Son los clientes que tienen una Línea de Crédito menor igual 875 soles ,que no presentan transacciones y de estos quienes tienen Línea de Crédito menor igual a 195 soles y que no han utilizado ninguna tarjeta quienes representan un total de 7% de clientes que deciden irse sobre el total de clientes Fuga de la entidad bancaria.

Nodo 20: Son los clientes que tienen una Línea de Crédito menor igual 875 soles ,que no presentan transacciones y de estos quienes tienen Línea de Crédito mayor a 310 soles y que utiliza más de una tarjeta y su ingreso bruto es mayor a 1007 soles representan un total de 4% de clientes que deciden irse sobre el total de clientes Fuga de la entidad bancaria.

Nodo 21: Son los clientes que tienen una Línea de Crédito mayor a 875 y menor igual a 1013 soles, no tienen transacciones ,su ingreso bruto es mayor a 1002 soles y no usan tarjetas quienes representan un total de 2% de clientes que deciden irse sobre el total de clientes Fuga de la entidad bancaria.

Cuadro N°12:Ganancias para los nodos

	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
11	2278	34.2%	2277	68.8%	100.0%	201.0%
17	217	3.3%	211	6.4%	97.2%	195.5%
13	394	5.9%	383	11.6%	97.2%	195.4%
21	53	.8%	51	1.5%	96.2%	193.5%
18	94	1.4%	81	2.4%	86.2%	173.3%
20	160	2.4%	122	3.7%	76.3%	153.3%
19	138	2.1%	85	2.6%	61.6%	123.8%
22	55	.8%	33	1.0%	60.0%	120.6%
15	91	1.4%	39	1.2%	42.9%	86.2%
4	272	4.1%	27	.8%	9.9%	20.0%
6	2800	42.1%	0	.0%	.0%	.0%
10	101	1.5%	0	.0%	.0%	.0%

Métodos de crecimiento: CRT

Variable dependiente: Situación

Fuente: Elaboración propia

El Cuadro N°12 muestra los nodos terminales, aquellos en los que se tiene el crecimiento del árbol.

Cuadro N°13:Riesgo

Muestra	Estimación	Típ. Error
Entrenamiento	.032	.002
Contraste	.035	.004

Métodos de crecimiento: CRT Variable dependiente: Situación

Fuente: Elaboración propia

Este Cuadro N°13 muestra una estimación de riesgo de 0.032 indica que la categoría pronosticada por el modelo (Fuga o No Fuga) es errónea para el 3.2% de los casos. Es decir el riesgo de clasificar erróneamente a un cliente de la entidad financiera es de 3.2% aproximadamente. Esto nos permite hacer una evaluación rápida de la bondad del funcionamiento del modelo.

Cuadro N°:14 Importancia de las variables independientes

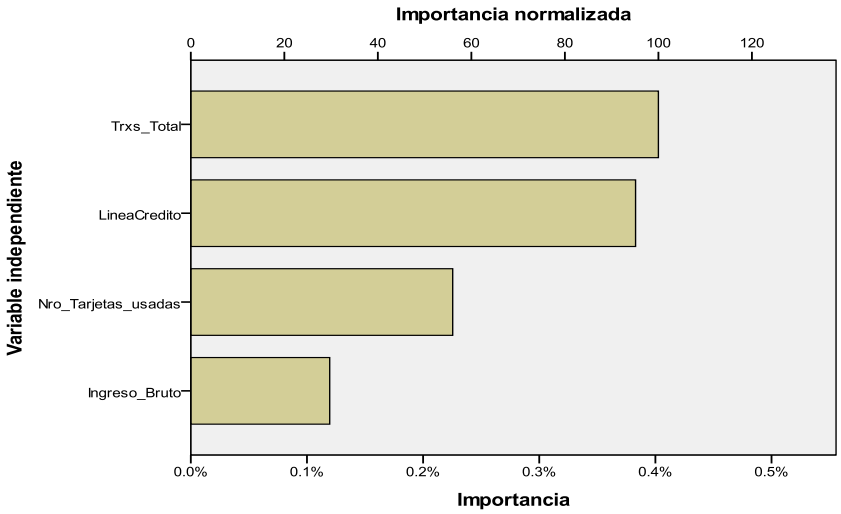
Variable independiente	Importancia	Importancia normalizada
Trxs_Total	.403	100.0%
LineaCredito	.383	95.1%
Nro_Tarjetas_usadas	.225	56.0%
Ingreso_Bruto	.120	29.7%

Métodos de crecimiento: CRT Variable dependiente: Situación

Fuente: Elaboración propia

El Cuadro N°14 muestra el grado de importancia de cada variable predictora, siendo la variable Trx_Total la variables con mayor importancia para el modelo. Así mismo podemos verlo gráficamente en el siguiente Grafico N°4 que muestra la importancia normalizada.

Gráfico N°5: Importancia Normalizada



Método de crecimiento:CRT

Variable dependiente:Situacion

Bondad de Ajuste

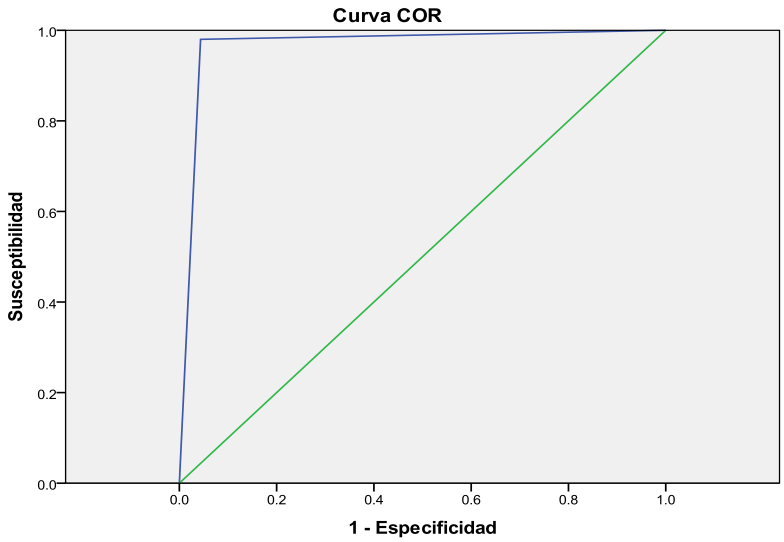
Curva ROC

Tenemos la hipótesis

H_0 : El área bajo la curva ROC es igual a 0.5

H_a : El área bajo la curva ROC no es igual a 0.5

Grafico N°6



Los segmentos diagonales son producidos por los empates.

Cuadro N°15: Área bajo la curva

Variables resultado de contraste: Grupo pronosticado

Área
.968

La variable (o variables) de resultado de contraste: Grupo pronosticado tiene al menos un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Los estadísticos pueden estar sesgados.

Fuente: Elaboración propia

El área es de 0.968 por tanto dado que es mayor que 0,7 se concluye que la discriminación del modelo se considera aceptable.

Tablas de clasificación

Cuadro N°16: Clasificación

	Pronosticado		
Observado	No Fuga	Fuga	Porcentaje correcto
No Fuga	3198	146	95.6%
Fuga	66	3243	98.0%
Porcentaje global	49.1%	50.9%	96.8%

Métodos de crecimiento: CRT Variable dependiente: Situación

Fuente: Elaboración propia

En el Cuadro N°16 se determina la eficacia predictiva del modelo, es aceptable dado que clasifica correctamente a los cliente Fuga al 98% con un porcentaje global de clasificación correcta de 96.8%.

5.4. Comparación del Análisis de Regresión Logística vs el Algoritmo de Árbol de Clasificación CART

Para identificar el mejor modelo se comparó las predicciones de cada modelo versus la base de validación (contraste) y dicha operación se comprobó hasta en 5 iteraciones para sacar una tasa de clasificación promedio.

Tabla de clasificación_1

Observado		Pronosticado		
		Situacion		Porcentaje correcto
		No Fuga	Fuga	
Situacion	No Fuga	839	42	95.2
	Fuga	18	858	97.9
Porcentaje global				96.6

Tabla de clasificación_2

Observado		Pronosticado		
		Situacion		Porcentaje correcto
		No Fuga	Fuga	
Situacion	No Fuga	795	35	95.8
	Fuga	13	821	98.4
Porcentaje global				97.1

Tabla de clasificación_3

Observado		Pronosticado		
		Situacion		Porcentaje correcto
		No Fuga	Fuga	
Situacion	No Fuga	774	51	93.8
	Fuga	19	802	97.7
Porcentaje global				95.7

Tabla de clasificación_4

Observado		Pronosticado		
		Situacion		Porcentaje correcto
		No Fuga	Fuga	
Situacion	No Fuga	842	40	95.5
	Fuga	19	823	97.7
Porcentaje global				96.6

Tabla de clasificación_5

Observado		Pronosticado		
		Situacion		Porcentaje correcto
		No Fuga	Fuga	
Situacion	No Fuga	794	38	95.4
	Fuga	19	827	97.8
Porcentaje global				96.6

Fuente: Elaboración propia

Cuadro N°17: Tasas de Clasificación para Regresión Logística

Comprobación	Tasa de Clasificación
1	96.6%
2	97.1%
3	95.7%
4	96.6%
5	96.6%
Promedio	**96.5%**

Clasificación_1

	Pronosticado		
Observado	No Fuga	Fuga	Porcentaje correcto
No Fuga	840	41	95.3%
Fuga	20	856	97.7%
Porcentaje global	48.9%	51.1%	96.5%

Clasificación_2

	Pronosticado		
Observado	No Fuga	Fuga	Porcentaje correcto
No Fuga	786	44	94.7%
Fuga	10	824	98.8%
Porcentaje global	47.8%	52.2%	96.8%

Clasificación_3

	Pronosticado		
Observado	No Fuga	Fuga	Porcentaje correcto
No Fuga	771	54	93.5%
Fuga	23	798	97.2%
Porcentaje global	48.2%	51.8%	95.3%

Clasificación_4

	Pronosticado		
Observado	No Fuga	Fuga	Porcentaje correcto
No Fuga	840	42	95.2%
Fuga	19	823	97.7%
Porcentaje global	49.8%	50.2%	96.5%

Clasificación_5

	Pronosticado		
Observado	No Fuga	Fuga	Porcentaje correcto
No Fuga	776	56	93.3%
Fuga	7	839	99.2%
Porcentaje global	46.7%	53.3%	96.2%

Métodos de crecimiento: CRT

Variable dependiente: Situación

Fuente: Elaboración propia

Cuadro N°18: Tasas de Clasificación Árbol CART

Comprobación	Tasa de Clasificación
1	96.5%
2	96.8%
3	95.3%
4	96.5%
5	96.2%
Promedio	**96.3%**

El Cuadro N° 17 muestra las 5 tablas de comprobación y la tasa de clasificación global promedio de 96.5% para el modelo de regresión logística pronostica a los clientes con situación Fuga en un 97.9% y los No Fuga en un 95.2%.

El Cuadro N° 18 muestra las 5 tablas de comprobación y la tasa de clasificación promedio de 96.3% para que el Algoritmo de Árbol de Clasificación CART el cual pronostica a los clientes con situación Fuga en un 98.1% y los No Fuga en un 94.4%.

Al comparar ambas tablas de clasificación vemos que la regresión logística proporciona un mejor pronóstico individual para los clientes Fugados tenemos un 97.9% de clientes correctamente clasificados.

Por lo tanto, para este caso, la regresión logística binaria, predice mejor a la variable respuesta Situación.

6. CONCLUSIONES

- Se ha identificado a las variables predictoras Ingreso_bruto, Nro_de_tarjetas usadas, Trxs_Total y LineaCredito. como aquellas que influyen significativamente en la situación de fuga de un cliente de una entidad bancaria aplicando una Regresión Logística Binaria.

- La variable predictiva que tiene mayor importancia significativa en los modelos fue la variable cantidad de transacciones realizadas. Dado que tiene un mayor coeficiente de regresión en el modelo de Regresión logística y en el caso del árbol es la variable que mejor divide a la muestra en el primer nivel.

- Con los resultados obtenidos del diagrama de Árboles de Clasificación CART se puede concluir que las variables que explican la situación de fuga de los clientes en la entidad bancaria con tarjetas de crédito presentan diferentes características es así que podemos clasificarlos según el patrón que cumplan, por ejemplo será un cliente con riesgo de fuga si:
    - No Presenta tarjetas usadas.
    - La línea de crédito es mayor a 310 soles y menor igual a 875 soles.
    - Su ingreso bruto es mayor a 1007 soles.
    - No presenta transacciones.

- Al comparar ambas metodologías con las tablas de clasificación se concluye que la Regresión Logística Binaria da un mejor pronóstico individual para los clientes Fugados tenemos un 97.9% de clientes correctamente clasificados. Por lo tanto la regresión logística binaria, predice mejor a la variable respuesta Situación.

7. RECOMENDACIONES

- Priorizar la limpieza de datos así como valores atípicos y falta de datos antes de comenzar con el análisis.
- Se debe establecer una adecuada estrategia de aumento de línea de crédito dado que los clientes con menor línea tienden a buscar otros banco que les brinden mayor línea de crédito y esto los motiva a fugarse de la entidad bancaria.
- Focalizar las estrategias de retención, teniendo campañas que motiven al cliente a realizar más de una transacción mensual.
- Realizar Indicadores de Control para Fuga e Inactividad de Clientes.
- Definir campañas focalizadas para retener clientes en función de su número de tarjetas usadas, de su línea de crédito, de su ingreso bruto y de las transacciones de compra que haya realizado, es vital buscar estrategias comerciales adecuadas para el grupo de clientes con perfil de riesgo fuga.

8. BIBLIOGRAFÍA

a. Breiman,L.,Friedman,R.Olshen,A. y Stone C.(1984).Classication and regression trees.

b. Cortijo, F. (2001) Técnicas supervisadas II: Aproximación no paramétrica. Publicado por Computer Based Learning Unit, University of Leeds.

c. C. Roberto Hernández Sampieri.1991.Metodología para la investigación – Primera Edición

d. David W. Hosmer y Stanley Lemeshow. Applied Logistic Regression Second Edition.

e. Dobson, A. 2002. An Introduction to Generalizad Linear Models. Chapman & Hall/CRC, 2da Edición. USA.

f. Marín. J. (2009). Análisis de Cluster y Árboles de Clasificación. Publicado por Universidad Carlos III de Madrid. Obtenido el 27/10/08 desde http:// halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf.

g. Puerta, A. (2002) IMPUTACIÓN BASADA EN ÁRBOLES DE CLASIFICACIÓN. Publicado por Eustat. Obtenido el 22/10/08 desde http:// www.eustat .es/document/datos/ct_04_c.pdf.

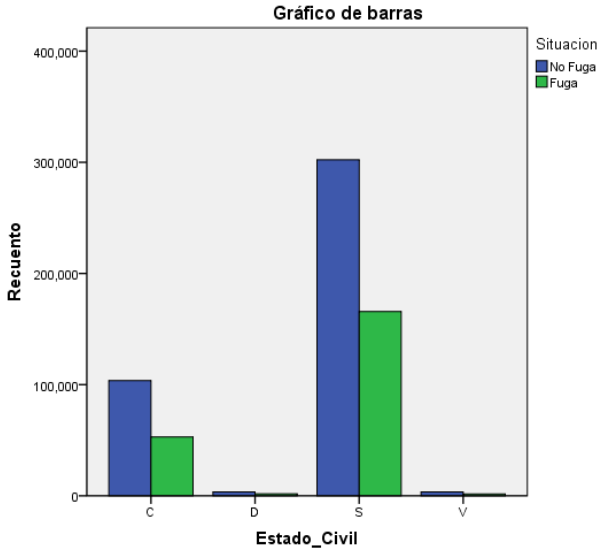
h. Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge: Cambridge.

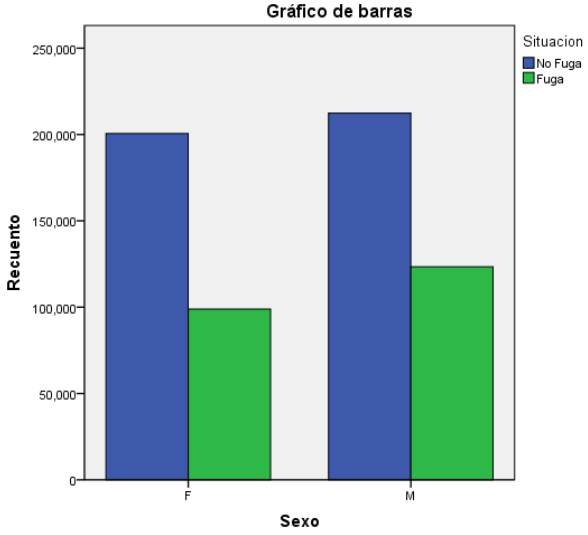
i. Sinnexus. Datawarehouse. Dirección URL: <http://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos>.

j. Scott Menard. Applied Logistic Regression Analysis – Second Edition

9. ANEXOS

Gráfico N°6: Histogramas de la variable respuesta vs las predictoras





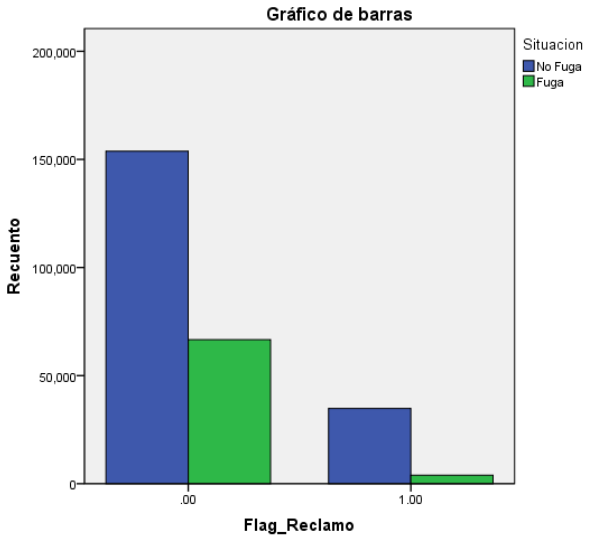


Gráfico de barras
Situacion
No Fuga
Fuga
Recuento
200,000
150,000
100,000
50,000
0
.00
1.00
Flag_Reclamo