UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"CLASIFICACIÓN DE FUGA DE CLIENTES EN UNA ENTIDAD FINANCIERA UTILIZANDO EL ALGORITMO SMOTE PARA DATOS DESBALANCEADOS EN UNA REGRESIÓN LOGÍSTICA"

Presentada por:

JEFFERSON CLAUSS PARIONA HUARHUACHI

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima - Perú

2017

DEDICATORIA

A mis padres Aurea y Claudio,
por todo su esfuerzo y amor depositado en mí.
A mis hermanos Romario, Misael y Edith
por todo el cariño, comprensión y apoyo
durante este tiempo.

A mis amigos de la vida: Jimmy , Martin y Álvaro
Por estar ahí apoyándome hasta el final.
Y a Leslie por esa cuota de ternura, tranquilidad y
apoyo incondicional en todo momento.

A Jehová Dios, por su bendición y protección
ya que sin él nada sería posible.

AGRADECIMIENTO

Un agradecimiento singular al profesor Jesús Salinas Flores que, como asesor de esta tesis, me ha orientado, apoyado y corregido en mi labor científica con un interés y una entrega que ha sobre pasado con mucho, todas mis expectativas.

ÍNDICE

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

La retención de clientes ha tomado mucha importancia en los últimos años en las entidades financieras debido a la competencia agresiva por parte del sector, así como la autonomía del cliente en buscar mejores beneficios dentro de todas las ofertas que existen en el mercado bancario lo que se ve reflejado en el aumento de la tasa de clientes fugados. Ante esto se ha visto necesaria la implementación de técnicas estadísticas y/o técnicas de minería de datos, con la finalidad de construir un clasificador predictivo que pueda ayudar a identificar a clientes potenciales a fugarse. En muchos casos cuando se aplican técnicas de clasificación, es común que la clase a predecir ocurra con menor frecuencia que la otra clase: la presencia de datos desbalanceados. Es decir, se tiene menor número de clientes fugados que no fugados, lo cual representa un inconveniente debido a que el clasificador necesita datos suficientes de ambas clases para poder aprender de ellas y así alcanzar una buena predicción. En esta investigación se propone el algoritmo Syntetic Minority Over-sampling Technique (SMOTE) como solución a este problema. SMOTE crea instancias nuevas a partir de un sobre-muestreo de las instancias existentes, llevando la clase minoritaria a un número suficiente para ser considerada balanceada y la clase mayoritaria si es necesaria reducirla mediante sub-muestreo aleatorio. En la presente investigación se validarán tales beneficios con la construcción de un modelo de regresión logística binaria con datos desbalanceados con y sin la aplicación del algoritmo de SMOTE; con el fin predecir la fuga de clientes en una entidad financiera. Se usarán para medir la precisión, la curva ROC y elementos de la comprobación de tabla cruzada como la especificidad y la sensibilidad.

Palabras clave: Fuga, CTS, SMOTE, sub-muestreo, sobre-muestreo, sensibilidad, disimilaridad.

ABSTRACT

Customer retention has taken much importance in recent years in financial institutions due to aggressive competition from the sector, as well as the autonomy of the client to seek better benefits within all offers that exist in the banking market, which is reflected in the increase in the rate of customers escaped. It has been necessary the implementation of statistical or technical techniques of data mining, in order to build a predictive classifier that can help identify potential customers to abscond. In many cases when classification techniques are applied, it is common to predict class to occur less frequently than other kind: the presence of unbalanced data. I.e. you have fewer customers escaped than not escapees, which represents a drawback since the classifier needs sufficient both kinds of data to be able to learn from them and thus achieve a good prediction. This research proposes the Syntetic Minority Over-sampling algorithm Technique (SMOTE) as a solution to this problem. SMOTE creates instances new starting from a sobre-muestreo of them instances existing, carrying the class minority to a number enough to be considered balanced and the class majority if is necessary reduce it by sub-sampling random. In the present study are validated such benefits with the construction of a model of binary logistic regression with unbalanced data with and without the application of the algorithm of SMOTE; in order to predict the flight of clients in a financial institution. They will be used to measure the precision, the ROC curve and elements of table cross as the specificity and sensitivity testing.

I. INTRODUCCIÓN

Los clientes son uno de los activos más importantes para cualquier negocio, ya que estos tienen una relación directa con las utilidades del negocio. Por esta razón, en los últimos años en el ámbito del marketing empresarial ha tenido una fuerte acogida una estrategia de negocio abocada directamente a la gestión de clientes; CRM (Customer RelationShip Costumer).

El CRM se centra en las relaciones con el cliente para conocer sus necesidades con el objetivo de fidelizarlo. Dos de las acciones comerciales más importantes del CRM tienen como objetivo mantener y mejorar la cartera de clientes: la captación de clientes nuevos y la retención de clientes existentes. La captación de nuevos clientes implica aumentar el número de clientes a través de la definición e incorporación de nuevos segmentos objetivos. Esta captación se realiza principalmente a través de elaboradas estrategias de publicidad, alta inversión en fuerza de ventas y la generación de ofertas focalizadas. La retención de clientes consiste en la identificación de los clientes con mayores tendencias a la 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.

Dentro del ámbito financiero; existen dos tipos de fuga de clientes: la fuga voluntaria y la fuga no voluntaria. La fuga voluntaria es la desafiliación del cliente por iniciativa propia sin injerencia directa por parte de la entidad financiera; es decir, cancelaciones de productos pasivos (ahorros), en productos activos (deudas con el banco) son los pagos por adelantado antes del plazo pactado. La fuga no voluntaria es la desafiliación del cliente cuando la entidad financiera es responsable directo del término de los acuerdos contractuales, donde el cliente no posee ninguna injerencia, generalmente se da por acciones delincuenciales o por mala utilización de los productos por parte del cliente. Un ejemplo de fuga de clientes no voluntaria puede ser efectuar algún tipo de fraude financiero con cheques del banco o la

clonación de tarjetas de crédito lo que genera la cancelación de su cuenta por parte de la entidad financiera.

En el mercado financiero peruano, durante los últimos años las entidades financieras han concentrado la mayor parte de sus esfuerzos en estrategias de captación y retención de clientes, esto se explica principalmente porque el mercado ha madurado y se ha convertido muy competitivo. Es por ello, el gran auge que en los últimos años vienen alcanzando los modelos estadísticos como herramientas de clasificación y predicción de la fuga de clientes mediante el reconocimiento de patrones comunes de comportamiento y el uso de probabilidades respectivamente.

Para identificar estos patrones, las entidades financieras utilizan principalmente técnicas de minería de datos, específicamente técnicas de clasificación, que les permite construir un clasificador que pueden aplicar sobre la cartera actual de clientes para predecir el perfil de cada uno de sus clientes y determinar su probabilidad de fuga de la entidad financiera. Dentro de las técnicas de clasificación binarias utilizadas para estos fines se encuentran: Regresión Logística, Support Vector Machines, Redes Neuronales, Arboles de Clasificación, Cadenas de Markov y Redes Bayesianas, entre otras (Haddena y otros, 2007; Neslin y otros, 2006; Van den Poel y Lariviere, 2004; Hsieh, 2004; Chiang y otros, 2003; Ngai y otros, 2009; Miranda y otros, 2005; Hung y otros, 2006; Zhao y otros, 2005; Coussement y den Poel, 2008).

La regresión logística es la técnica más usada por sus ventajas: similitud con el modelo de regresión múltiple, no cumplimiento estricto de los supuestos de normalidad multivariante ni la igualdad de matrices de varianzas covarianzas entre los grupos, sumado a ello la manera sencilla de explicar las estimaciones de sus coeficientes. Sin embargo, esta técnica ante datos con clases desbalanceadas presenta problemas en su tarea de clasificador y disminuye su calidad predictiva. Serna Pineda (2009), en su investigación menciona que la regresión logística presenta una tendencia de clasificación hacia la clase mayoritaria, minimizando de esta manera el error de clasificación y clasificando correctamente instancias de clase mayoritaria en detrimento de instancias de clase minoritaria. Por ejemplo, esto sucede en el caso de clasificar un cliente fugado o no, donde el número de clientes fugados es mucho menor al de clientes no fugados.

Para estos casos existen distintos tipos de metodologías para poder mejorar este desbalance de casos como; *Oversampling* que consiste en duplicar al azar instancias de la clase minoritaria. *Undersampling* que consiste en reducir al azar instancias de la clase mayoritaria. *Boosting***;** consiste en asociar pesos a cada instancia que se van modificando en cada iteración del clasificador. Así mismo, se han desarrollado diferentes algoritmos para solucionar el problema del desbalanceo de datos. Marco Altini (2015).

En la presente investigación se utiliza la metodología Oversampling con el algoritmo SMOTE (Syntetic Minority Over-sampling Technique). Genera instancias "sintéticas" o artificiales para equilibrar la muestra de datos minoritaria, basado en la regla del vecino más cercano. La generación se realiza extrapolando nuevas instancias en lugar de duplicarlas como hace el algoritmo de Oversampling. Para cada una de las instancias minoritarias se buscan las instancias minoritarias vecinas (más cercanas) y se crean N instancias entre la línea que une la instancia original y cada una de las vecinas. El valor de N depende del tamaño de oversampling deseado.

Según lo expuesto hasta ahora, es necesario analizar los datos desbalanceados y mejorar la clasificación de la clase minoritaria a predecir. Es por ello que en esta investigación se trató de identificar entre los modelos de Regresión Logística Binaria con datos desbalanceados sin y con aplicación del algoritmo SMOTE, cuál de ellos proporciona una mejor clasificación de la fuga de clientes en una entidad financiera.

Para la aplicación de lo señalado, se tomó en cuenta la fuga de clientes en una entidad financiera durante 12 meses en todas las oficinas distribuidas en el país. El producto financiero pasivo a analizar es la CTS (Compensación por Tiempo de Servicio) y se tomaron en cuenta a los clientes que poseen una cuenta CTS con un monto mayor a S/ 350 y que sus cuentas estén activas.

1.1 JUSTIFICACIÓN DE LA INVESTIGACIÓN

Una tarea fundamental dentro de una estrategia de retención, es la predicción de fuga de clientes, es decir poder clasificar un cliente como fugado antes que realmente lo sea. Kotler (2000) señala la importancia de las estrategias de retención de clientes. Una de las razones

para concentrar esfuerzos en retención, es el costo de adquirir nuevos clientes en comparación con el costo de retener clientes existentes. Por lo menos es cinco veces más costoso adquirir un nuevo cliente que retener uno existente.

Por otro lado, Reichheld y Sasser (2000) concluyen que aumentos de un 5% en la tasa de retención de clientes puede generar incrementos de utilidades del orden de un 25% a un 85% promedio, dependiendo del tipo de negocio o industria. Según Reichheld y Sasser (2000), a medida que aumenta la permanencia de los clientes en el tiempo, aumentan los beneficios que estos entregan a la empresa. Esto se explica por aumentos en los niveles de transacciones con los clientes, posibilidad de aplicar estrategias de cross-selling y upselling, es decir ventas cruzadas, reducciones en los costos operacionales, referencias que atraen nuevos clientes, entre otros factores.

El problema de clasificar un cliente como fugado o no, se ha abordado en diversos trabajos, principalmente utilizando técnicas de minería de datos. Haddena (2007) revisa las principales técnicas de minería de datos aplicadas para predecir fuga de clientes que han tenido buenos resultados. Sin embargo, los autores discuten la necesidad de concentrar esfuerzos de investigación en crear clasificadores de fuga de clientes más poderosos, exactos y robustos, que permitan entender mejor el comportamiento de los clientes y las causas que provocan las fugas.

El actor principal y ejecutante del plan de retención es la entidad financiera que intenta retener a sus clientes, sin embargo, el proceso de fuga es más complejo y considera la interacción de la entidad financiera, su cartera de clientes y su competencia. En este contexto, resulta interesante seguir desarrollando clasificadores; modelos de predicción de fuga de clientes para utilizar sus resultados como parte de una estrategia de retención de clientes más completo, pero que también se aproveche el potencial de los datos que posee la entidad financiera interesada en utilizar el clasificador.

Por otro lado, Haibo He y Yunqian Ma (2013) mencionan que los recientes desarrollos en la ciencia y la tecnología han permitido el crecimiento y la disponibilidad de datos que ocurren en un ritmo explosivo para la investigación de ciencias de la ingeniería de

conocimientos y datos, estos juegan un papel esencial en una amplia gama de aplicaciones. Sin embargo, todos estos datos no siguen una proporción adecuada para poder ser trabajada, es decir las clases que quieren ser analizadas normalmente se encuentras desbalanceadas, y esto obedece a su propia naturaleza, es decir es común que existan pocos individuos de una clase que otra según el ámbito y circunstancias de donde se las tome, por ejemplo; fraudes, fuga de clientes, toma de productos.

En los últimos años, el problema del aprendizaje de datos desbalanceados ha atraído una cantidad importante de interés en diferentes ámbitos, académicas así como empresas interesadas en análisis de datos. La cuestión clave con el problema de aprendizaje de datos desbalanceados es su capacidad de comprometer significativamente el rendimiento de aprendizaje desde el algoritmo y/o modelos estadísticos de regresión hasta los algoritmos de Machine Learning. Por lo tanto, cuando se presentan datos desbalanceados, los algoritmos no representan adecuadamente la distribución de características de los datos y, como resultado, proporcionan desfavorables precisiones a través de las clases de los datos. (Serna Pineda, 2009)

Es por ello que nace la necesidad de proponer una solución a este problema, en este trabajo de investigación se presentará una alternativa a ello; el algoritmo de SMOTE, que consiste en crear instancias artificiales de la clase minoritaria de tal forma que permite que las clases se encuentren en proporciones suficientemente balanceadas y así poder realizar un análisis adecuado de los datos.

1.2 OBJETIVOS DE LA INVESTIGACIÓN

1.2.1 Objetivo General

- Identificar entre las técnicas de balanceo de sub-muestreo simple y algoritmo SMOTE aplicados en una Regresión Logística Binaria, cual proporciona una mejor clasificación de la fuga de clientes en una entidad financiera.

1.2.2 Objetivos Específicos

- Hallar las curvas ROC, el área bajo la curva (AUC), la sensibilidad y la especificidad para elegir el mejor modelo.
- Encontrar el perfil del cliente fugado.

II. REVISIÓN DE LITERATURA

2.1 RETENCIÓN DE CLIENTES

El mayor costo de adquirir un nuevo cliente con respecto a retener uno existente se explica porque el consumidor debe tomar la decisión de abandonar una entidad financiera competidora a la que ya se encuentra afiliado y esta decisión tiene costos importantes para el cliente. Un cliente que decide cambiarse de entidad financiera debe estar convencido que el beneficio que obtendrá en la nueva entidad financiera será superior al costo asociado al cambio más los beneficios que obtendría manteniéndose en la misma entidad financiera. Se define el *switching cost* como el costo que debe enfrentar un cliente cuando decide abandonar una entidad financiera y escoger una empresa rival. Klemperer (1995) clasifica los *switching costs* en distintos tipos:

- **Costos por compatibilidad**

  Al cambiar de proveedor de algún producto y/o servicio el cliente debe evaluar si el nuevo proveedor entregará productos compatibles con otros que el consumidor posee en otros ámbitos. Por ejemplo, si un cliente decide cambiar algún software que utilice en su computador, debe evaluar si el nuevo software es compatible con su sistema operativo, si es capaz de comunicarse con los otros software que utiliza, si es compatible con la impresora que posee, etc.

- **Costos de transacción por cambio de proveedor**

  Existen costos asociados a la transacción, al trámite y al cumplimiento de cláusulas contractuales asociadas a la decisión de dejar de ser cliente de una entidad financiera y afiliarse a una nueva. Estos costos pueden ser, por ejemplo: el tiempo invertido por el cliente en la transacción, el pago de multas por dar término por adelantado a un contrato y los costos notariales de firmar un nuevo contrato. Algunos de este tipo de costos se incurren por ejemplo cuando un cliente decide cerrar una cuenta corriente en un banco y abrir una nueva en otro.

- **Costos de aprendizaje**

  Es el costo asociado a tener que aprender a utilizar un nuevo producto y/o servicio. Si un cliente se cambia de banco, debe aprender a utilizar el nuevo sitio web del banco, conocer los nuevos teléfonos de atención al cliente, conocer las nuevas políticas del banco, etc. Otro ejemplo es el usuario que decide cambiar su sistema operativo de Windows a Linux, quien debe aprender un nuevo lenguaje, funciones y estructura del sistema.

- **Incertidumbre acerca de la calidad de productos no probados**

  Cuando un consumidor escoge un producto y/o servicio que satisface sus necesidades es reacio a cambiarse a un competidor por el riesgo de que este no cumpla sus expectativas. Por ejemplo, las personas que utilizan un medicamento de un laboratorio determinado no quieren tomar el mismo medicamento comercializado por otro laboratorio, ya que tienen dudas con respecto a su calidad y a que no tenga la misma efectividad del que ya conoce.

- **Pérdidas de programas de beneficios**

  Las entidades financieras premian a sus clientes frecuentes con beneficios a través de sus programas de lealtad. Si un consumidor decide cambiar de empresa perdería sus beneficios acumulados. Por ejemplo, los programas de viajero frecuente de las aerolíneas incentivan a sus pasajeros a seguir prefiriéndola para seguir acumulando kilómetros o millas que luego podrán canjear por pasajes en futuros viajes.

- **Costos psicológicos**

  Referidos a la lealtad no económica que las empresas generan sobre los consumidores, principalmente a través del marketing. Las marcas a través del marketing generan una personalidad y se asocian con determinadas causas con la que los consumidores se identifican. Esto genera una preferencia a priori del consumidor hacia determinada entidad financiera. Un consumidor frente a dos productos equivalentes escogerá el que ofrece la empresa de su preferencia.

  A través de esfuerzos de marketing las empresas pueden crear *switching costs* sobre su cartera actual de clientes. Sobre los clientes que tienen intenciones de abandonar

la empresa, se pueden aplicar estrategias de retención para crear o aumentar estos costos asociados al cambio. El objetivo es hacer más difícil la decisión del cliente de abandonar la empresa.

Los programas de fidelidad que las empresas ofrecen buscan que los clientes obtengan beneficios por preferirla recurrentemente. Si una empresa identifica que un cliente tiene intenciones de abandonar la empresa, esta puede ofrecer por ejemplo descuentos en el precio o algún beneficio especial por mantenerse en la empresa. Los *switching costs* relacionados a costos económicos son fácilmente compensables por las otras empresas que buscan atraer a un cliente. El competidor que busca captar a un cliente de una empresa rival puede asumir y pagar los *switching costs* que enfrenta el consumidor. Un competidor fácilmente puede igualar el descuento ofrecido por otra entidad financiera como oferta de retención. Si el cliente enfrenta un costo de aprendizaje, la entidad competidora puede ofrecer las facilidades para que el nuevo cliente conozca cómo utilizar su producto y/o servicio. En caso que exista incertidumbre con respecto a la calidad de producto ofrecido por la competencia, estas pueden hacer demostraciones y dar un periodo de prueba para que el consumidor se entere de las bondades de sus productos. Por esta razón es que las entidades financieras no pueden escoger estrategias de retención que se basen únicamente en crear *switching costs* asociados a costos económicos.

Literatura especializada pone énfasis en que las estrategias de retención que utilicen las empresas tienen que buscar crear la lealtad de los clientes además de poner barreras económicas para evitar su fuga (Peppers y Rogers, 2011; Buttle, 2008; DeBonis y otros, 2003; Thompson, 2004).

Como se plantea en Peppers y Rogers (2011) un cliente retenido no es necesariamente un cliente leal a la empresa. Las empresas deben invertir en construir relaciones individuales más fuertes para potenciar los niveles de lealtad de sus clientes. Para esto es necesario que las empresas pongan énfasis en la calidad de sus productos y servicios y demostrar un interés real en construir una relación de largo plazo con cada uno de sus clientes.

La lealtad de los clientes está asociada a la calidad de la relación que mantiene la empresa con el cliente y al nivel de satisfacción que siente cada cliente a lo largo del tiempo. La lealtad de un cliente entonces es un intangible que resulta del nivel de identificación del cliente con la empresa y de los niveles de satisfacción que el cliente perciba como resultado de una relación de largo plazo. Por esta razón, la lealtad de los clientes representa un *switching costs* que las empresas rivales difícilmente pueden compensar al momento de intentar atraer un cliente. Las empresas con altos niveles de satisfacción en sus clientes, tienen clientes más leales, mayores tasas de retención y en consecuencia obtienen mayores utilidades a lo largo del tiempo.

2.2 FUGA DE CLIENTES EN UNA ENTIDAD FINANCIERA

- **Panorama actual**

  La mala reputación del sector financiero en los últimos años ha hecho que disminuya la confianza de los clientes y deteriorado su relación con las entidades financieras. Y esto se ha visto reflejado mucho en los productos pasivos, es decir los ahorros de los clientes. A diferencia de los productos activos, en los productos pasivos como la CTS, el cliente tiene casi el total manejo de este, es decir puede cancelarla, retirarla para usarla (si tuviera el disponible que estipula la ley vigente N° 30408 de la constitución peruana: el exceso de 4 sueldo brutos), o llevarla a otra entidad financiera que le pudiera ofrecer mejores condiciones financieras. A eso hay que sumarle las ofertas competitivas de las demás entidades financieras: bancos y sobre todo cajas rurales de ahorros, traducidas en; tasas anuales de interés muy atractivas, incluyendo beneficios como: descuentos en restaurantes, préstamos de efectivos con mejores tasas, puntos para ser canjeados por viajes, etc.

- **Fuga de clientes**

  Teniendo en cuenta esta situación, las entidades financieras se ven obligadas a plantear un plan de retención completo e integrado. La entidad financiera debe contar con una definición exacta y conveniente para la fuga de clientes. Este punto es muy importante dado que la definición de fuga de clientes debe estar alineada

con los objetivos e intereses de la entidad financiera. Ahora bien, existen dos tipos de fuga de clientes: no voluntaria y voluntaria.

- **Fuga no voluntaria**

  Es la situación cuando la entidad financiera cancela la cuenta del cliente por algún motivo delincuencial o fraudulenta que no pone en peligro los intereses de la entidad y de los demás clientes. Ejemplo: clonación de tarjetas, cobros de cheques sin fondos, creación de cuentas de ahorros con dinero proveniente de acciones fuera de la ley: narcotráfico, terrorismo, etc.

- **Fuga voluntaria**

  Es la situación cuando el cliente por iniciativa propia decide cancelar su producto financiero dentro del banco: pudiendo ser productos activos (prestamos, tarjetas de crédito, etc.) o productos pasivos; ahorros. En su gran mayoría los clientes recurren en esto por dos motivos: motivos propios y motivos ajenos a la entidad financiera. Motivos propios a la entidad financiera son todas las situaciones que vienen de parte de la entidad financiera hacia el cliente como: mala experiencia en el servicio al cliente, cobros indebidos, aumento de tasa de interés sin previo aviso, mala información de parte de la entidad financiera hacia el cliente, que normalmente se siente engañado, etc. Motivos ajenos a la entidad financiera son todas las situaciones que la entidad financiera no puede controlar como: mala situación financiera del cliente que le hace incapaz de seguir teniendo el producto adquirido, mejores ofertas de parte de otras entidades financieras.

2.3 REGRESIÓN LOGÍSTICA

- **Introducción**

  Según Agresti (2002), este es el modelo más importante para los datos de las respuestas categóricas. Se utiliza cada vez más en una amplia variedad de aplicaciones. Los primeros usos fueron en estudios biomédicos, pero los últimos 20 años también han visto mucho uso en la investigación de las ciencias sociales y el marketing.

  La Regresión Logística es un método de clasificación que se utiliza comúnmente en el Credit Scoring. Según Nieto (2010), el modelo de regresión logística no requiere

de los supuestos de la regresión lineal, como son el supuesto de normalidad de los errores, homocedasticidad y el supuesto que las variables involucradas sean continuas. En este sentido, la regresión logística, se aplica tanto a datos que se distribuyen como normal, como a datos que no lo son, y por lo tanto, el modelo de regresión logística es útil cuando la variable de respuesta "y" no está distribuida normalmente y tanto las variables predictoras como de respuesta tienen valores discretos, categóricos, ordinales o no ordinales.

Los modelos de regresión logística son una herramienta que permite explicar el comportamiento de una variable respuesta discreta (binaria o con más de dos categorías) a través de una o varias variables independientes explicativas de naturaleza cuantitativa y/o cualitativa. Según el tipo de variable respuesta estaremos hablando de regresión logística binaria (variable dependiente con 2 categorías), o de regresión logística multinomial (variable dependiente con más de 2 categorías), pudiendo ser esta última de respuesta nominal u ordinal. Los modelos de respuesta discreta son un caso particular de los modelos lineales generalizados formulados por Nelder y Wedderburn en 1972 (Nelder y Wedderburn ,1972), al igual que los modelos de regresión lineal o el análisis de la varianza.

- **El modelo logístico**

  Según Johnson y Wichern (2007) el modelo logit o logístico se aplica a una amplia gama de situaciones donde las variables explicativas no tienen una distribución conjunta normal multivariante. Por ejemplo, si algunas son categóricas, podemos introducirlas en el modelo logit mediante variables ficticias, como se hace en el modelo de regresión estándar. En particular, si todas las variables son binarias independientes y se llaman a los parámetros de la primera población P_1 y $P_2 = (P_{21},\ldots,P_{2p})$ a los de la segunda y observamos un elemento $X_i = (X_{ii},\ldots,X_{ip})$, se tendrá que, suponiendo que las probabilidades a priori son las mismas donde (1):

$$c = \frac{P(y=1)}{P(x_i)} \tag{1}$$

  La transformación logística (2) será:

$$g_i = \log \frac{P(y=1 \mid x_i)}{1-P(y=1 \mid x_i)} = \sum x_{ij} \log(p_{1j} / p_{2j}) + \sum (1-x_{ij}) \log\left[(1-p_{1j})/(1-p_{2j})\right] \quad (2)$$

que es una función lineal en las variables (3), que se puede escribir como:

$$g_i = \beta_0 + \beta_1^{'} x_i \quad (3)$$

Donde (4)

$$\beta_0 = \sum \log\left[(1-p_{1j})/(1-p_{2j})\right] \quad \text{y} \quad \beta_1^{'} = \sum \log\left[(p_{1j})(1-p_{2j})/(p_{2j}(1-p_{2j})\right] \quad (4)$$

Por tanto, se espera que este modelo se comporte bien cuando todas las variables de clasificación son binarias y aproximadamente independientes.

Una ventaja adicional de este modelo es que cuando las variables son normales también verifican el modelo logístico. En efecto, suponiendo dos poblaciones normales multivariantes con distinta media pero la misma matriz de varianzas covarianzas y suponiendo las probabilidades a priori de ambas poblaciones iguales (5):

$$p_i = P(y=1 \mid x_i) = \frac{f_1(x_i)}{f_1(x_i) + f_2(x_i)} \quad (5)$$

utilizando la transformación logit, (6):

$$g_i = \log \frac{f_1(x_i)}{f_2(x_i)} = -\frac{1}{2}(x_i - \mu_1)^{'} V^{-1}(x_i - \mu_1) + \frac{1}{2}(x_i - \mu_2)^{'} V^{-1}(x_i - \mu_2) \quad (6)$$

y simplificando en (7)

$$g_i = \frac{1}{2}(\mu_2 V^{-1} \mu_2 - \mu_1 V^{-1} \mu_1) + (\mu_1 - \mu_2)^{'} V^{-1} x_i \quad (7)$$

Por tanto, g_i es una función lineal de las variables x, que es la característica que define el modelo logit. Comparando con (8) la ordenada en el origen, β_0, es igual:

$$\beta_0 = \frac{1}{2}(\mu_2 V^{-1} \mu_2 - \mu_1 V^{-1} \mu_1) = -\frac{1}{2} w^{'}(\mu_1 + \mu_2) \quad (8)$$

Donde $w = V^{-1}(\mu_1 - \mu_2)$, el vector de pendientes (9)

$$\beta_1 = w \quad (9)$$

Se observa que la estimación de w mediante el modelo logístico no es eficiente en el caso normal. En efecto, en lugar de estimar los $p(p+1)/2$ términos de la matriz $\hat{V}$ y los $2p$ de las medias $\bar{x}_1$ y $\bar{x}_2$, con el modelo logístico se estima únicamente $p+1$ parámetros $\beta_0, \beta_1, \ldots, \beta_p$. En el caso de normalidad se obtiene un mejor procedimiento con la regla de Fisher, que estima $\hat{V}$, $\bar{x}_1$ y $\bar{x}_2$, la distribución completa de las x, mientras que el modelo logístico estima sólo los $p+1$ parámetros de la distribución de y condicionada a x;

Como (10):

$$f(x,y) = f(y \mid x) f(x) \tag{10}$$

Se pierde información al considerar sólo la condicionada $f(y \mid x)$, como hace el modelo logístico, en lugar de la conjunta $f(x,y)$. Efron (1975) demostró que cuando los datos son normales multivariantes y se estiman los parámetros en la muestra, la función de discriminación lineal de Fisher funciona mejor que la regresión logística.

En resumen, en el caso de normalidad la regla discriminante es mejor que el modelo logístico. Sin embargo, la función logística puede ser más eficaz cuando las poblaciones tengan distinta matriz de covarianzas o sean marcadamente no normales. En el campo de la concesión automática de créditos (Credit Scoring) existen numerosos estudios comparando ambos métodos. La conclusión general es que ninguno supera al otro de manera uniforme. Rosenberg y Gleit (1994) y Hand y Henley (1997) han presentado estudios sobre este problema.

- **Interpretación del modelo logístico**

Agresti (2002) menciona que los parámetros del modelo son β_0, la ordenada en el origen, y $\beta' = (\beta_1, \ldots, \beta_p)'$, las pendientes. A veces se utilizan también como parámetros e^{β_0} y e^{β_i}, que se denominan los odds ratios o ratios de probabilidades, e indican cuánto se modifican las probabilidades por unidad de cambio en las variables z. En efecto, de (8) se deduce que

$$O_i = \frac{p_i}{1-p_i} = e^{\beta_0} \prod_{j=1}^{p} e^{(\beta_j) x_j} \tag{11}$$

Suponiendo dos elementos, i,k con todos los valores de las variables iguales excepto la variable h y $X_{ih} = X_{jh} + i$. El cociente de los ratios de probabilidades (odds ratio) para estas dos observaciones es (12):

$$\frac{O_i}{O_k} = e^{\beta_h} \tag{12}$$

e indica cuánto se modifica el ratio de probabilidades cuando la variable X_h aumenta una unidad. Sustituyendo $p_i = 0.5$ en el modelo logit (13), entonces,

$$\log \frac{p_i}{1-p_i} = \beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = 0, \tag{13}$$

Es decir (14),

$$x_{i1} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{p} \frac{\beta_{ij} x_{ij}}{\beta_1}, \tag{14}$$

Y X_{ii} representa el valor de X que hace igualmente probable que un elemento, cuyas restantes variables son $X_{i2},\ldots,X_{ip}$, pertenezca a la primera a la segunda población.

2.4 EVALUACIÓN DE CLASIFICACIÓN

- **Matriz de confusión**

  La matriz de confusión o también llamada tabla cruzada, contiene información acerca de las predicciones realizadas por un método o sistema de clasificación, comparando para el conjunto de individuos en la tabla de aprendizaje la predicción dada versus la clase a la que estos realmente pertenecen. (Fawcett, 2004)

  Para decir que un modelo es mejor que otro, se tiene que demostrar que uno es realmente es mejor que el otro. Para eso, se necesita una métrica que contabilice el número de errores cometidos.

Un modelo de clasificación es una función que permite decidir cuáles de un conjunto de instancias están relacionadas o no por pertenecer a un mismo tipo o clase. El resultado del clasificador o del diagnóstico puede ser un número real (valor continuo), en cuyo caso el límite del clasificador entre cada clase debe determinarse por un valor umbral (por ejemplo para determinar si una persona tiene hipertensión basándonos en una medida de presión arterial), o puede ser un resultado discreto que indica directamente una de las clases. (Fawcett, 2004)

- **Componentes de la matriz de confusión**

  Considerando un problema de predicción de clases binario, en la que los resultados se etiquetan positivos (p) o negativos (n). Hay cuatro posibles resultados a partir de un clasificador binario como el propuesto (Cuadro 1). Si el resultado de una exploración es p y el valor dado es también p, entonces se conoce como un Verdadero Positivo (VP); sin embargo, si el valor real es n entonces se conoce como un Falso Positivo (FP). De igual modo, tenemos un Verdadero Negativo (VN) cuando tanto la exploración como el valor dado son n, y un Falso Negativo (FN) cuando el resultado de la predicción es n pero el valor real es p. (Fawcett, 2004)

  Un ejemplo aproximado de un problema real es el siguiente: consideremos una prueba diagnóstica que persiga determinar si una persona tiene una cierta enfermedad. Un falso positivo en este caso ocurre cuando la prueba predice que el resultado es positivo, cuando la persona no tiene realmente la enfermedad. Un falso negativo, por el contrario, ocurre cuando el resultado de la prueba es negativo, sugiriendo que no tiene la enfermedad cuando realmente sí la tiene. La siguiente tabla describe el comportamiento de un sistema de diagnóstico con dos únicos posibles resultados (positivo o negativo).

Cuadro 1: Matriz de confusión

Clase Predecida	
No Fuga	Fuga

Clase Real	No Fuga	Verdadero Negativo (VN)	Falso Positivo (FP)
	Fuga	Falso Negativo (FN)	Verdadero Positivo (VP)

FUENTE: Elaboración Propia

En resumen, la terminología es:

Verdadero positivo (VP): Un ejemplo que es fuga y ha sido clasificado correctamente como fuga.

Verdadero Negativo (VN): Un ejemplo que es no fuga y ha sido clasificado correctamente como fuga.

Falso Positivo (FP): Un ejemplo que es no fuga, pero ha sido clasificado erróneamente como fuga.

Falso negativo (FN): Un ejemplo que es fuga, pero que ha sido clasificado erróneamente como no fuga.

- **Métricas para la evaluación de una clasificación**

  Swets (1996) definió los siguientes conceptos:

  - **Sensibilidad:** Proporción de casos clasificados como positivos, a partir del criterio establecido, en los que se comprueba que efectivamente sucede el estado que se pretende detectar. (1)

$$Sensibilidad = \frac{VP}{VP+FN} \quad (1)$$

  - **Especificidad:** Proporción de casos clasificados como negativos, a partir del criterio establecido, en los que se comprueba que efectivamente no sucede el estado que se pretende detectar. (2)

$$Especificidad = \frac{VN}{VN+FP} \quad (2)$$

  - **Correcta Clasificación:** Proporción de casos correctamente clasificados (VP y VN) como positivos y negativos, a partir del criterio establecido respecto a toda la clasificación. (VP, VN, FP, FN) (3)

$$\frac{VP+VN}{FN+FP+VP+VN} \quad (3)$$

- **Mala Clasificación:** Proporción de casos mal clasificados (FP y FN), a partir del criterio establecido respecto a toda la clasificación. (VP, VN, FP, FN) (4)

$$\frac{FP+FN}{FN+FP+VP+VN} \quad (4)$$

- **Curva ROC: Receiver Operating Characteristic**
  - **Historia**

    García (2012) menciona que la curva ROC se comenzó a utilizar durante la Segunda Guerra Mundial para el análisis de señales de radar, a partir de lo cual se desarrolló la Teoría de Detección de Señales. Después del ataque a Pearl Harbor en 1941, el ejército de los Estados Unidos comenzó un programa de investigación para detectar correctamente los aparatos japoneses a partir de sus señales de radar.

    En los años 50, las curvas ROC se utilizaron en Psicofísica para evaluar la capacidad de detección de humanos (y también de no humanos) en señales débiles. En medicina el análisis ROC se ha utilizado de forma muy extensa en epidemiología e investigación médica, de tal modo que se encuentra muy relacionado con la medicina basada en la evidencia. En Radiología, el análisis ROC es la técnica de preferencia para evaluar nuevas técnicas de diagnóstico por imagen. (García, 2012)

    Más recientemente, las curvas ROC se han mostrado muy útiles para la evaluación de técnicas de aprendizaje automático. La primera aplicación de las ROC en esta área fue por Spackman en 1989, quien demostró el valor de las curvas ROC para la comparación de diferentes algoritmos de clasificación.

  - **Definición**

    La curva ROC es una representación gráfica de la tasa de correctamente clasificación frente a la tasa de mala clasificación para situaciones donde se detecten solo dos resultados posibles. (De Ullibarri, 1998)

La tabla de contingencia puede proporcionar varias medidas. Para dibujar una curva ROC sólo son necesarias las razones de Verdaderos Positivos (VP) y de falsos positivos (FP). La razón de VP mide hasta qué punto un clasificador o prueba diagnóstica es capaz de detectar o clasificar los casos positivos correctamente, de entre todos los casos positivos disponibles durante la prueba. La FP define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles durante la prueba. (De Ullibarri, 1998)

Un espacio ROC se define por FP y VP como ejes x e y respectivamente, y representa los intercambios entre verdaderos positivos (en principio, beneficios) y falsos positivos (en principio, costes). Dado que VP es equivalente a sensibilidad y FP es igual a 1-especificidad, el gráfico ROC también es conocido como la representación de sensibilidad frente a (1-especificidad). Cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio ROC. (Fawcett, 2004)

El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). A este punto (0,1) también se le llama una clasificación perfecta. Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación, desde el extremo inferior izquierdo hasta la esquina superior derecha (independientemente de los tipos de base positiva y negativa). Un ejemplo típico de adivinación aleatoria sería decidir a partir de los resultados de lanzar una moneda al aire, a medida que el tamaño de la muestra aumenta, el punto de un clasificador aleatorio de ROC se desplazará hacia la posición (0.5, 0.5). (Fawcett, 2004)

La diagonal divide el espacio ROC (Figura 1). Los puntos por encima de la diagonal representan los buenos resultados de clasificación (mejor que el azar), puntos por debajo de la línea de los resultados pobres (peor que al azar).

Nótese que la salida de un predictor consistentemente pobre simplemente podría ser invertida para obtener un buen predictor.

Considérense los siguientes cuatros resultados de 100 instancias positivas y otras 100 negativas (Cuadro 2):

Cuadro 2: Matriz de confusión

A		
VP=63	FP=28	91
FN=37	VN=72	109
100	100	200
VP = 0.63		
FP = 0.28		
Precision = 0.68		

B		
VP=77	FP=77	154
FN=23	VN=23	46
100	100	200
VPR = 0.77		
FPR = 0.77		
Precision = 0.50		

C		
VP=24	FP=88	112
FN=76	VN=12	88
100	100	200
VPR = 0.24		
FPR = 0.88		
Precision = 0.18		

C′		
VP=76	FP=12	88
FN=24	VN=88	112
100	100	200
VPR = 0.76		
FPR = 0.12		
Precision = 0.82		

FUENTE: ROC Graphs, Notes and Practical Considerations for Researchers (Fawcett, 2004).

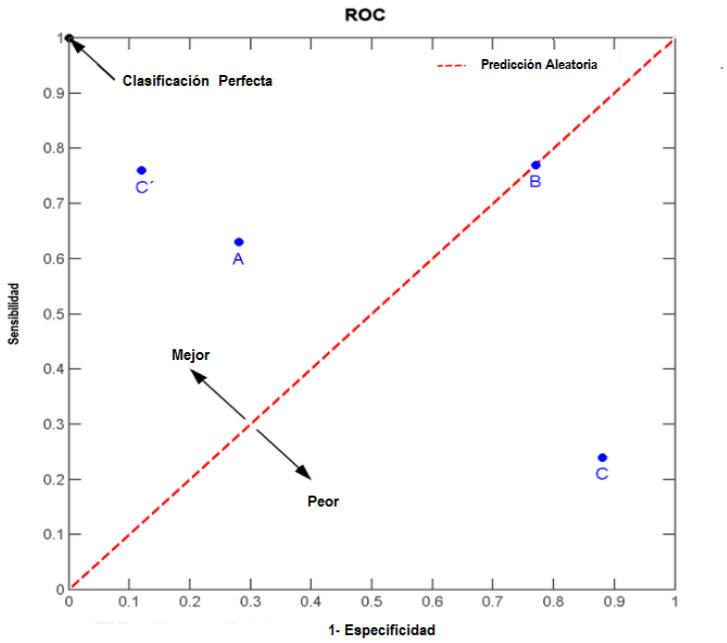


Figura 1: Curva ROC

FUENTE: ROC Graphs, Notes and Practical Considerations for Researchers (Fawcett, 2004).

En la figura 1 se muestran los puntos que los cuatro ejemplos anteriores en el espacio ROC. El resultado del método A muestra claramente ser el mejor de entre los métodos A, B Y C. El resultado de B se encuentra sobre la línea de estimación aleatoria (diagonal); en la tabla 1 se puede ver que la precisión de

este método es del 50%. El método C aparece como el peor de los tres, con un resultado muy pobre.

Sin embargo, se considera ahora la construcción de un cuarto método de predicción C' que simplemente invierte los resultados predichos por el método C. Este nuevo método mostrará una tabla de contingencia opuesta a la de C y su punto en el espacio ROC estará ahora por encima de la diagonal, y más próximo al punto de clasificación perfecta que el método A. Mientras C presentaba un pobre poder de predicción, a partir de él se ha construido un predictor mejor que todos los demás. Cuando el método C predice 'n' o 'p', el método C predice 'n' o 'p' respectivamente. Siempre que un método presente un punto en el espacio ROC por debajo de la diagonal habrá que invertir sus predicciones para aprovechar su capacidad de predicción. (Fawcett, 2004)

Cuanto más cerca esté un método de la esquina superior izquierda (clasificación perfecta) mejor será, pero lo que en realidad marca el poder predictivo de un método es la distancia de este a la línea de estimación aleatoria, da igual si por arriba o por abajo.

El rendimiento global de una prueba de diagnóstico se suele resumir el área bajo la curva ROC, el cual según (Cuadro 3):

Cuadro 3: Regla de Indicadores Curva ROC

%	Resultado
ROC = 50	Esto no sugiere discriminación
70 < ROC < 80	Esto se considera una discriminación aceptable
80 < ROC < 90	Esto se considera una discriminación excelente
ROC > 90	Esto se considera una discriminación sobresaliente

FUENTE: Hosmer y Lemeshow (2000).

2.5 MANEJO DE DATOS DESBALANCEADOS

En general, cualquier conjunto de datos puede considerarse como conjunto de datos desbalanceado si el número de observaciones entre las clases no es igual. Sin embargo, el

entendimiento común dentro del ámbito analítico de minería de datos es cuando un conjunto de datos posee un extremo y significativo desbalance. La relación de desequilibrio es de por lo menos 1:10. Aunque hay varios casos de conjuntos de datos multiclase, en esta investigación se consideran casos binarios (o dos clases).

De preferencia, dado cualquier conjunto de datos, por lo general se requiere un clasificador estándar que proporcione pesos equilibrados de precisión predictiva para la minoría y la mayoría de las clases. En la práctica, el clasificador estándar tiende a proporcionar una precisión extrema de predicción desequilibrada, generalmente la clase minoritaria tiene una exactitud menor del 10 por ciento, mientras que la clase mayoritaria tiene una precisión de cerca del 100% (Chawla et al., 2002). Sin embargo, en la aplicación de datos desequilibrados conjunto (Chawla et al., 2002), se espera que un clasificador sea sensible a la clase minoritaria pues es la clase que interesa predecir, aunque se pague cierta penalización por clasificar erróneamente la clase de mayoritaria. Por lo tanto, se prefiere un clasificador que identifique a la mayoría de observaciones de la clase minoritaria, aunque sigue siendo relativamente alta la precisión de la clase mayoritaria. Además, la medición común como las métricas de evaluación convencionales no proporciona suficiente información útil cuando se trata de conjunto de datos desequilibrada. Más métricas de evaluación de información son necesarias como la curva de características de funcionamiento del receptor (ROC).

Comúnmente se contempla o acepta como algo intrínseco el desbalance a los datos, es decir, el desbalance es el resultado de la naturaleza del espacio de los datos. Sin embargo, no solo se puede atribuir el desbalance a lo intrínseca de la variedad. El conjunto de datos puede considerarse como desbalanceado por factores extrínsecos, si se tienen variables como fechas, datos acumulativos o históricos. Por desbalance extrínseco, el desbalance no es directamente resultado de la naturaleza del espacio de los datos. Por ejemplo, un flujo continuo de datos balanceados puede mantener los datos de entrenamiento equilibrado por un tiempo hasta que, durante un intervalo de tiempo, el flujo es interrumpido temporalmente, a consecuencia de ello los datos sufren un desbalance. La figura 2 es un ejemplo de ello. Durante el tiempo interna de 4.85 y 7.99, la clase 2 se convierte en clase minoritaria como su flujo de datos ha sido interrumpida. (Tianxiang Gao, 2015).

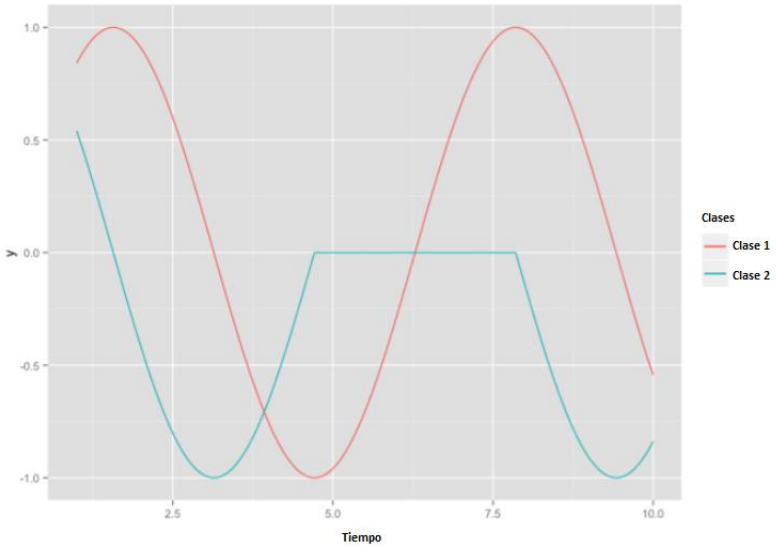


Figura 2: Flujo datos en un periodo de tiempo

FUENTE: Hybrid classification approach of SMOTE and instance selection for imbalanced datasets. Tianxiang Gao (2015).

El desbalance absoluto es debido a observaciones raras, mientras el desbalance relativo es el desequilibrio con relación a otra clase. Por ejemplo, un conjunto de datos dado con proporción de desequilibrio 1: 100 contiene 200,000 observaciones Es indudable que el número de clase mayoritaria domina el número de clase minoritaria. Sin embargo, 2000 observaciones de la clase de la minoría no son realmente raras. En cambio, el número de la clase de la minoría es raro con relación a la clase de la mayoría. Varios artículos han mostrado que la clase minoritaria no es tan difícil de identificar como conjunto de datos desbalanceado relativo. (Batista et al., 2004). Estos resultados son muy sugerentes. De ahí que la proporción de desequilibrio es sólo uno de factores que dificultan a los clasificadores. Como resultado, la complejidad del conjunto de datos juega un papel fundamental en el deterioro de la clasificación.

La complejidad del conjunto de datos obedece a la superposición y a pequeñas separaciones (pequeños grupos dentro de una clase). Es frecuente que un conjunto de datos contenga más de una clase superpuesta, es decir, se traslapan varios datos de las clases en la zona de decisión límite. En la figura 3, hay conjuntos de datos superpuestas entre la clase de la mayoría que es el círculo y la clase de la minoría que es graficada como el triángulo. Por otra parte, la clase minoritaria contiene un subconjunto de datos que son de color azul puro. Esto es otra situación de desbalance, desbalance dentro de la clase (Jo y

Japkowicz, 2004). La existencia del desbalance interno está estrechamente relacionada con pequeñas separaciones, que deprecian mucho el rendimiento de clasificación (Jo y Japkowicz, 2004). Para interpretar estos pequeños grupos, un clasificador necesita detectar la clase de minoría (o mayoría clase) creando múltiples reglas disjuntas es decir reglas por separado que describa individualmente los subconjuntos específicos de clase de la minoría. Por ejemplo, en la figura 3 se observa que al eliminar todos los datos del subconjunto de clase de la minoría, un clasificador por lo general fácilmente creará grandes grupos que cubren un gran pedazo de observaciones asociadas con el conjunto principal. Sin embargo, debido a la existencia de subconjunto de la clase minoritaria, el clasificador en cambio deberá probar necesariamente al menos dos pequeños grupos que individualmente cubran el espacio de datos específicos de la clase minoritaria. Por otra parte, el ruido puede influir más si existen subconjuntos pequeños dentro de la clase minoritaria. Por lo tanto, la validez de los datos se convierte en una cuestión esencial. En la figura 3, el ruido de los triángulos morados es considerado como clase minoritaria catalogada como pérdida.

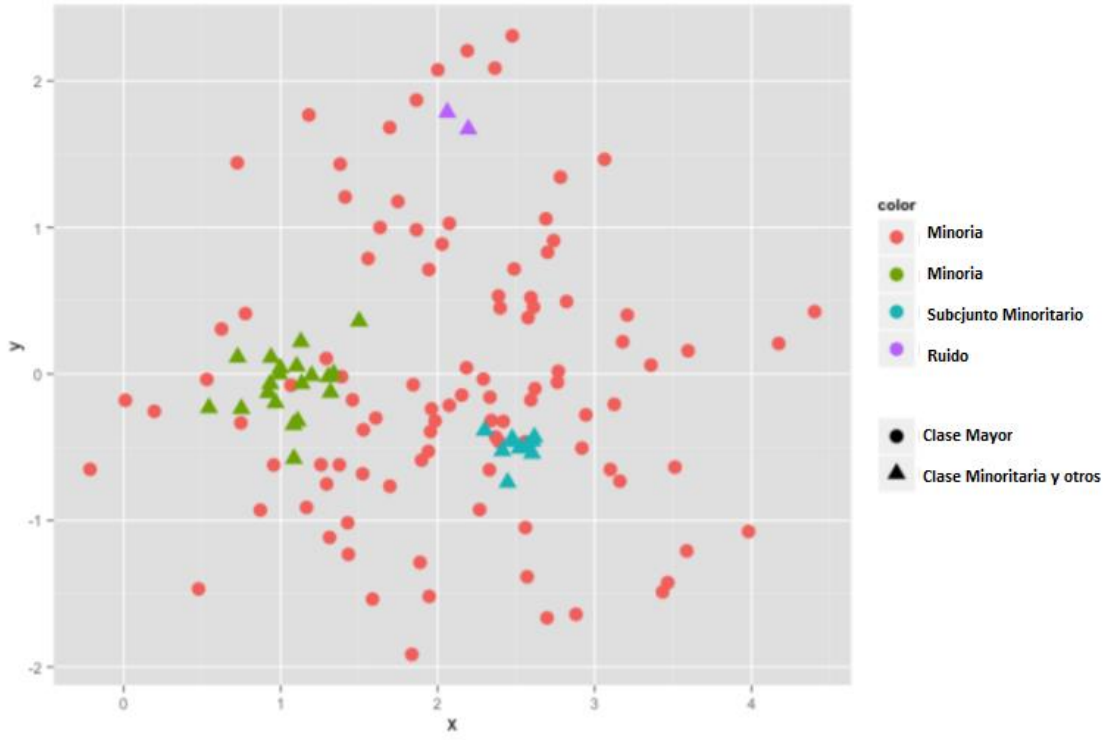


Figura 3: Conjunto de datos desbalanceados

FUENTE: Hybrid classification approach of SMOTE and instance selection for imbalanced datasets Tianxiang Gao (2015).

La combinación de datos desbalanceados y el pequeño tamaño de la muestra es un problema completamente asociado con el conjunto de datos (base de datos). En la realidad es inevitable que el número de variables predictoras sean mucho más que el número de casos, es decir, i.e p >> n. Así, el primer problema está relacionado a que el número de muestras es bastante limitado, todos los problemas están asociados con el desbalance absoluto y desbalance interno de una clase. En segundo lugar, es más difícil para el clasificador encontrar un papel inductivo cubriendo la clase de la minoría. Además, el clasificador va a sobre ajustar si genera la regla específica debido a muestras limitadas de la clase minoritaria en el espacio de datos específico.

- **El problema de los datos desbalanceados**

  El aprendizaje a partir de datos no balanceados es uno de los desafíos que actualmente se enfrenta el aprendizaje automático, debido al mal funcionamiento de los algoritmos frente a conjuntos de este tipo. La ocurrencia de sucesos pocos frecuentes ha dado lugar a que exista una desproporción considerable entre el número de ejemplos en cada clase lo que se conoce como clases no balanceadas o desbalanceadas. En numerosas situaciones aparece desbalance entre las clases, dentro de las que sobresalen: diagnóstico de enfermedades con condiciones médicas poco frecuentes como tiroides, detección de llamadas telefónicas fraudulentas, detección de derrames de petróleo a partir de imágenes de radar, fuga de clientes en entidades financieras (bancos), entre otras. (Martínez, E. R., et al, 2009)

  Los clasificadores logran muy buenas precisiones con la clase más representada (mayoritaria), mientras que en la menos representada (minoritaria) ocurre todo lo contrario.

  En los no balanceados el conocimiento más novedoso suele residir en los datos menos representados, sin embargo, muchos clasificadores pueden considerarlos como rarezas o ruido, pues los mismos no tienen en cuenta la distribución de los datos, únicamente se centran en los resultados de las medidas globales. Esto se ilustra mejor a continuación con un ejemplo.

  Dado un conjunto de datos de los datos de la transacción, interesa saber qué clientes son fraudulentos y que clientes no lo son. Ahora bien, el costo puede llegar a ser

muy alto para la empresa de comercio electrónico, si una transacción fraudulenta llega a efectuarse y tener éxito, pues afecta a nuestros clientes que confían en nosotros, y cuesta dinero. Por lo que se quieren identificar las transacciones fraudulentas como sea posible.

Si hay un conjunto de datos que consta de 10000 transacciones no fraudulentas, es decir genuinas y 10 transacciones fraudulentas, el clasificador tendrá la tendencia a clasificar las transacciones fraudulentas como transacciones genuinas. La razón puede explicarse fácilmente por los números. Suponga que el algoritmo o modelo tiene dos salidas, posiblemente, de la siguiente manera:

- Modelo 1, clasifica 7 de cada 10 transacciones fraudulentas como transacciones genuinas y 10 de cada 10.000 transacciones genuinas como transacciones fraudulentas.
- Modelo 2, clasifica 2 de cada 10 transacciones fraudulentas como transacciones genuinas y 100 de 10000 transacciones genuinas como transacciones fraudulentas.

Si el rendimiento del clasificador se determina por el número de errores, entonces claramente el Modelo 1 es mejor, ya que sólo hace un total de 17 errores mientras que el Modelo 2 hizo 102 errores. Sin embargo, como se quiere minimizar el número de transacciones fraudulentas se debe escoger el Modelo 2 que sólo hizo 2 errores clasificación de las transacciones fraudulentas. Por supuesto, esto podría venir a expensas de más transacciones genuinas que serán clasificadas como operaciones fraudulentas, pero será un coste que se puede soportar por ahora. Esto es un problema dado que al final un algoritmo automático escogerá el menor error, es decir, el modelo; 1. En la práctica, esto significa que se va a dejar que una gran cantidad de transacciones fraudulentas pasen a pesar de que se podría haber evitado mediante el uso del modelo 2. Esto se traduce en clientes insatisfechos y dinero perdido para la empresa.

- **Trabajos anteriores sobre poblaciones desbalanceadas**

  Kubat y Matwin (1997) aplicaron selectivamente el sub-muestreo a la clase mayoritaria, mientras que mantenían la población original de la clase minoritaria. Ellos utilizaron la media geométrica como una medida de mejora para el clasificador, el cual se puede relacionar a un simple punto en la curva ROC. Se dividieron los ejemplos de la clase minoritaria en cuatro categorías: ruido que traslapa la región de decisión de la clase positiva, ejemplos fronterizos, ejemplos redundantes y ejemplos seguros. Los ejemplos frontera fueron detectados usando el concepto Tomek links. (Tomek, 1976).

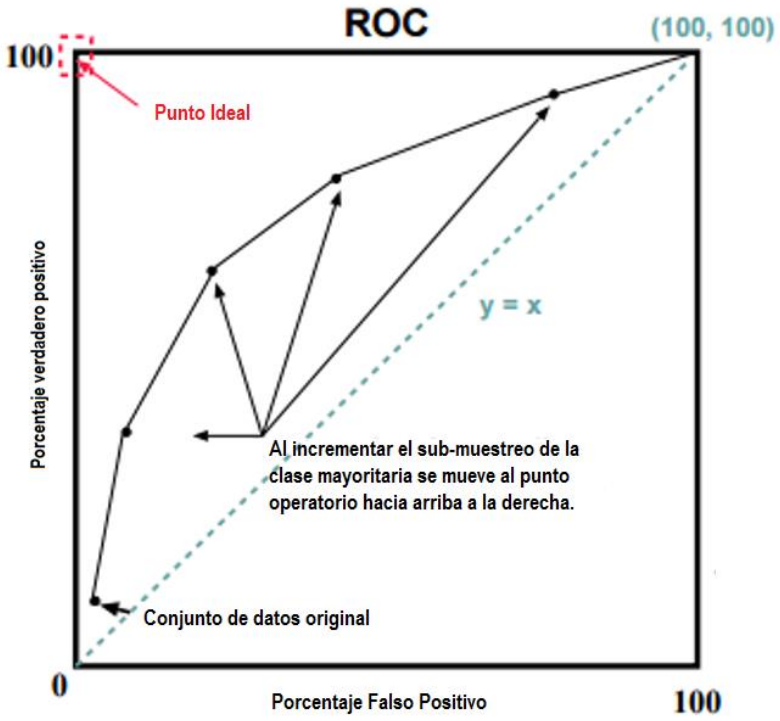


Figura 4: Modificación de la curva ROC mediante el sub-muestreo

FUENTE: SMOTE, Synthetic Minority Over-sampling Technique. Nitesh V. Chawla (2002).

Otro trabajo propuso el sistema SHRINK que clasifica una región de solapamiento de la clase minoritaria (positiva) y la clase mayoritaria (negativa) como positiva. Este sistema busca "la mejor región positiva" (Kubat et al., 1998). En la figura 4 se muestra que al incrementar el sub-muestreo de la clase mayoritaria (negativa), la mejora se moverá desde el punto abajo-izquierda al punto arriba-derecha.

Japkowicz (2000) discutió el efecto del desbalance en el conjunto de datos. La autora evalúo tres estrategias: sub-muestreo, re-muestreo y esquema de inducción por reconocimiento. La presente investigación se concentra en sus enfoques sobre el muestreo. La autora experimentó con datos 1D artificiales para calcular y construir con facilidad el concepto de complejidad. Se consideraron dos métodos de re-muestreo. El re-muestreo aleatorio consistía en re-muestrear la clase más pequeña de forma aleatoria hasta obtener tantas muestras como la clase mayoritaria y "el re-muestreo enfocado" (focused resampling) consistía en re-muestrear solo los ejemplos de la clase minoritaria que se encontraban en el límite de las clases mayoritarias y minoritarias. Se tomó en cuenta el sub-muestreo aleatorio, el cual significa sub-muestrear las muestras de la clase mayoritaria de forma aleatoria hasta que sus números igualen a los números de muestras de la clase minoritaria. El sub-muestreo enfocado consistía en sub-muestrear las muestras de la clase mayoritaria que estén más alejados. La autora notó que ambas estrategias de muestreo eran efectivas y también observó que utilizar las técnicas de muestreo sofisticadas no le daban ninguna ventaja clara en el dominio considerado (Japkowicz, 2000).

Un enfoque que es particularmente relevante es el que desarrollaron Ling & Li (1998). Ellos combinaron el sobre-muestreo de la clase minoritaria con el sub-muestreo de la clase mayoritaria. Además, utilizaron el análisis "lift" en lugar de precisión para medir la mejora de un clasificador. Propusieron que los ejemplos de prueba sean ordenados por una medida de confianza y que lift se use como un criterio de evaluación. Una curva lift es similar a una curva ROC, pero es más precisa para el problema de análisis de marketing (Ling & Li, 1998). En un experimento, realizaron el sub-muestreo a la clase mayoritaria y notaron que el mejor índice lift se obtiene cuando las clases son representadas igualitariamente (Ling & Li, 1998). En otro experimento, realizaron el sobre-muestreo a los ejemplos positivos (minoritarios) con reemplazo con el fin de igualar el número de ejemplos negativos (mayoritarios) con el número de ejemplos positivos. La combinación de sobre-muestreo y sub-muestreo no entregó una mejora significante en el índice lift. Sin embargo, el enfoque de sobre-muestreo difiere con lo que se maneja en esta investigación.

Solberg (1996) consideraron el problema de los conjuntos de datos desbalanceados en la clasificación de fuga de petróleo de las imágenes de SAR. Utilizaron las técnicas de sobre-muestreo y sub-muestreo para mejorar la clasificación de fugas de petróleo. Sus datos de aprendizaje tuvieron una distribución de 42 fugas de petróleo y 2471 similares, entregando una probabilidad previa de 0.98 para los similares. Este desbalance llevará al aprendiz (sin funciones de pérdida apropiadas o una metodología para modificar información previa) a clasificar correctamente casi todos los que parezcan similares a costa de clasificar de forma equivocada muchas de las muestras de fuga de petróleo (Solberg, 1996). Para solucionar este problema de desbalance, realizaron el sobre-muestreo (con reemplazo) en 100 muestras de la fuga de petróleo, y muestrearon de forma aleatoria 100 muestras de la clase que no pertenecen a la fuga de petróleo para crear un nuevo conjunto de datos con probabilidades iguales. Los autores adquirieron información de un árbol clasificador en este conjunto de datos balanceados y lograron un 14% de tasa de error en las fugas de petróleo en un método dejando uno fuera para la estimación de error. En los similares, ellos lograron una tasa de error de 4% (Solberg, 1996).

Otro enfoque que es similar al que se maneja en esta investigación es el que propone Domingos (1999) donde compara el enfoque "metacosto" para cada sub-muestreo mayoritario y sobre-muestreo minoritario. Encontró que el metacosto mejora en cualquiera de los dos, y que es preferible utilizar el sub-muestreo a usar el sobre-muestreo minoritario. Los clasificadores basados en el error son hechos para ser sensitivos al costo. Se estima la probabilidad de cada clase para cada ejemplo, y se reetiquetan los ejemplos de manera óptima tomando en cuenta la clasificación de forma equivocada de los costos. El reetiquetado de los ejemplos expande el espacio de decisión ya que crea nuevas muestras de las cuales el clasificador puede aprender (Domingos, 1999).

Una red neuronal pre alimentada entrenada en un conjunto desbalanceado puede no aprender a discriminar bien entre clases (DeRouin, Brown, Fausett, & Schneider, 1991). Los autores propusieron que la tasa de aprendizaje de la red neuronal sea adaptada a las estadísticas de representación de las clases en los datos y calcularon un factor de atención utilizando la proporción de muestras presentada a la red

neuronal para entrenamiento. La tasa de aprendizaje de los elementos de la red se ajustó basada en el factor de atención. Experimentaron en un conjunto de entrenamiento generado artificialmente y en un conjunto de entrenamiento del mundo real, ambos con múltiples clases (más de dos). Compararon esto con la estrategia de replicación de muestras de la clase minoritaria para balancear el conjunto de datos utilizado para el entrenamiento. Se mejoró la precisión de la clasificación en la clase minoritaria.

Lewis y Catlett (1994) examinaron el muestreo heterogéneo de incertidumbre para el aprendizaje supervisado (heterogeneous uncertainty sampling for supervised learning). Este método es útil para muestras de entrenamiento con clases inciertas. Las muestras de entrenamiento se etiquetan progresivamente en dos fases y las instancias inciertas se pasan a la siguiente fase. Modificaron el C4.5 con el fin de incluir un ratio de pérdida para determinar los valores de clase en las hojas. Se determinaron los valores de clase por comparación con un umbral de probabilidad de LR (LR+1) donde LR es el ratio de pérdida (Lewis & Catlett, 1994).

El dominio de recuperación de información (IR, por sus siglas en inglés) (Dumais et al., 1998; Mladenić & Grobelnik, 1999; Lewis & Ringuette, 1994; Cohen, 1995a) también enfrenta el problema de la clase desbalanceada en el conjunto de datos. Un documento o una página web se convierte en una representación de bolsa de palabras, esto significa que se construye un vector característico que refleja las ocurrencias de palabras en la página. Generalmente, hay muy pocas instancias de la categoría de interés en la categorización de texto. Esta sobre representación de la clase negativa en los problemas de recuperación de información puede causar problemas en la evaluación de la mejora de los clasificadores. Dado que la tasa de error no es una buena métrica para conjuntos de datos sesgados, el desempeño de la clasificación de los algoritmos en la recuperación de información se mide generalmente por sensibilidad (1) y exhaustividad (2):

$$\text{Sensibilidad} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Exhaustividad} = \frac{TP}{TP+FN} \quad (2)$$

Mladenić y Grobelnik (1999) propusieron un enfoque de selección del subconjunto característico para enfrentar la distribución de clase desbalanceada en el dominio IR. Experimentaron con varios métodos de selección característicos, y encontraron que el ratio de probabilidades (Van Rijsbergen, Harper, & Porter, 1981) se desempeña mejor en su dominio cuando se combina con un clasificador bayesiano ingenuo. El ratio de probabilidades es una medida probabilística utilizada para clasificar documentos de acuerdo a su relevancia para la clase positiva (clase minoritaria). La adquisición de información por palabra, por otro lado, no pone atención a una clase meta particular; se calcula por palabra para cada clase. En un conjunto de datos desbalanceados de texto (asumiendo que el 98% al 99% es la clase negativa), muchas de las características se relacionarán con la clase negativa. El ratio de probabilidades incorpora la información de clase meta en su métrica dando mejores resultados cuando se compara con la adquisición de información para la categorización de textos.

Provost & Fawcett (1997) introdujeron el método de la envolvente convexa de ROC con el fin de calcular el desempeño del clasificador para conjuntos de datos desbalanceados. Los autores notaron que los problemas de distribución desigual de clases y los costos desiguales de error se relacionan y que se ha hecho poco para solucionar cualquiera de éstos problemas (Provost & Fawcett, 2001). En el método de la envolvente convexa de ROC, el espacio de ROC se usa para separar el desempeño de clasificación de la información sobre distribución de costo y clase.

Para resumir la bibliografía, el sub-muestreo de la clase mayoritaria permite que se construyan mejores clasificadores, más que el sobre-muestreo de la clase minoritaria. Una combinación de los dos antes mencionados, como se realizó en una investigación previa, no lleva a encontrar a clasificadores que superen a aquellos construidos utilizando solamente el sub-muestreo. Sin embargo, se ha realizado el sobre-muestreo de la clase minoritaria utilizando el muestreo con reemplazo de los datos originales. El enfoque utilizado en esta investigación es un método diferente de sobre-muestreo.

- **Enfoques vigentes más usados para tratar datos desbalanceados**

  Entre los principales trabajos que han afrontado el desbalance de clases en problemas de clasificación supervisada se tienen: Matrices de costo (Lomax y Vadera, 2013; López et al., 2014a; Miñardi y Torelli, 2014; Fernandez et al, 2013), modificación de algoritmos, enfoques basados en muestreos y entre otros que se explican a continuación.

  - **Los enfoques basados en matrices de costo**

    Estas soluciones, utilizando matrices de costo asignan un alto costo en los errores de clasificación para los objetos que pertenecen a la clase minoritaria. Estas incluyen estrategias a nivel de datos, de algoritmos, o mixtas, con el principal objetivo de minimizar el costo total. Se han propuesto varios trabajos, los cuales se pueden resumir en los siguientes enfoques generales:

    - **Métodos directos**

      La idea fundamental es construir clasificadores que introducen y utilizan un costo asociado a una mala clasificación. Por ejemplo, en el contexto de los árboles de decisión, la estrategia de construcción es adaptada para minimizar el costo total. De esta manera, la información del costo es usada para seleccionar divisiones candidatas o cual es la mejor rama a ser podada. Por otra parte, los métodos basados en algoritmos genéticos incorporan el uso de costos asociados a la función de aptitud y las Redes Neuronales utilizan un método de puntuación de riesgo en la combinación de varios modelos. De una manera parecida, los algoritmos basados en reglas incorporan el costo al momento de construir las reglas o crean pesos para cada regla. De esta forma, cada algoritmo incluye el costo total dentro de sus objetivos a minimizar. (Lomax y Vadera, 2013)

    - **Meta-Clasificadores**

      Esta metodología integra mecanismos de pre-procesamiento para el conjunto de entrenamiento o un post-procesamiento en el resultado, en ambos mecanismos se utiliza un clasificador (denominado clasificador-base) sin ser

modificado. Los meta-clasificadores sensitivos al costo pueden ser agrupados en:

- ✓ **Umbralización:** tiene como base la teoría básica de decisión que le asigna, a un objeto, la clase que minimice el costo esperado. Algunos de los algoritmos más populares que utilizan este tipo de técnica son MetaCost y Cost-Sensitive Classifier (CSC), que asignan una nueva clase a los objetos en dependencia de la clase que minimice el costo esperado. (Domingos, 1999)

- ✓ **Re-muestreo:** está basado en modificar el conjunto de entrenamiento teniendo en cuenta la matriz de costo asociada a cada clase. La técnica más popular es balancear la distribución de clases del conjunto de entrenamiento mediante el uso de una de las técnicas de re-muestreo (Zadrozny et al., 2003) o asignándole pesos a los objetos (Ting, 2002). Estas modificaciones han demostrado ser eficaces y también pueden aplicarse a cualquier algoritmo de clasificación que no sea tolerante al desbalance (Zhou y Liu, 2006).

- **Modificación de algoritmos**

  Este tipo de soluciones adaptan o crean algoritmos de clasificación para reforzar la predicción de la clase minoritaria, sin utilizar re-muestreo o matrices de costo.

  Para los árboles de decisión, las estrategias más utilizadas son: ajustar la estimación probabilística en las hojas, crear divisiones candidatas que tienen en cuenta la proporción por clases e introducir nuevas técnicas de poda (Liu et al., 2010), para favorecer la predicción en la clase minoritaria. En el caso de las Maquinas de Vectores de Soporte (SVM), se adaptan diferentes constantes de penalización para diferenciar las clases o se ajustan las fronteras entre las clases usando un kernel de alineación de fronteras. En la extracción de reglas de asociación se especifican diferentes soportes mínimos para cada una de las clases. Otro de los algoritmos afectados por el desbalance de clases suele ser el k-NN, donde una de las formas de mitigar este problema es transformar las

probabilidades a priori por probabilidades a posteriori empleando modelos de redes bayesianas para estimar los pesos de confianza en cada una de las clases (Liu y Chawla, 2011).

Una de las estrategias que se han utilizado, a nivel de algoritmos, para mitigar los problemas de clasificación en bases de datos con clases desbalanceadas, son los sistemas de múltiples clasificadores. Estos tratan de mejorar el rendimiento de los clasificadores individuales mediante la inducción de varios clasificadores y la combinación de ellos para obtener un nuevo clasificador que supera a cada uno de los clasificadores individuales. Una taxonomía reciente de estos clasificadores, para el aprendizaje con clases desbalanceadas, se puede encontrar en (Galar et al., 2012), la cual se resume en la Figura 5. Principalmente, los autores distinguen cuatro familias diferentes de sistemas de múltiples clasificadores para base de datos con clases desbalanceadas.

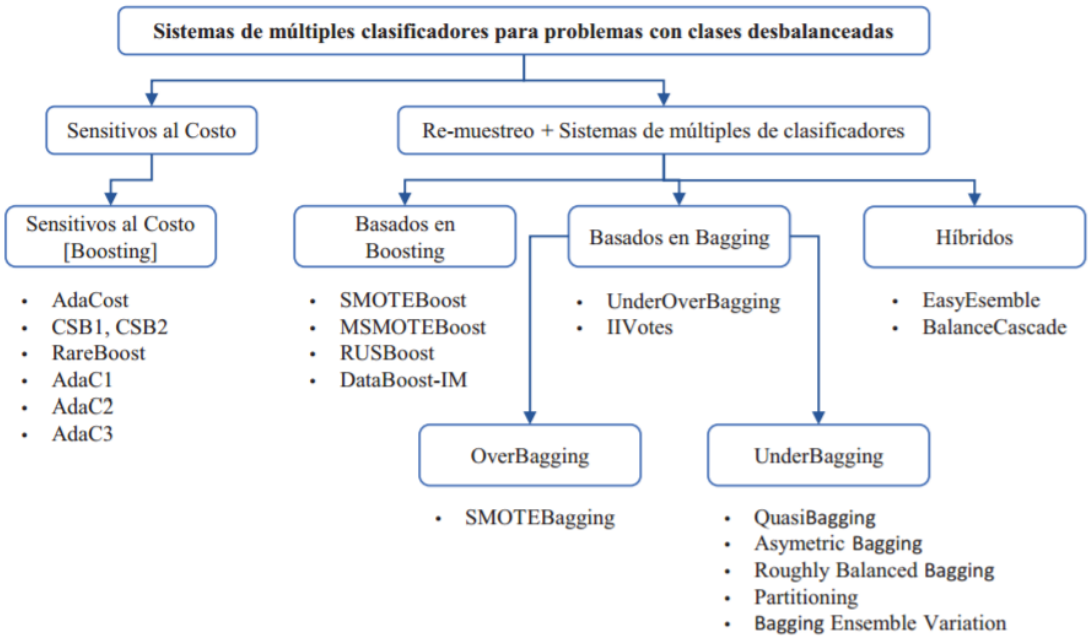


Figura 5: Taxonomía de los sistemas de múltiples clasificadores para problemas con clases desbalanceadas

FUENTE: Clasificadores Supervisados basados en Patrones Emergentes para Bases de Datos con Clases Desbalanceadas. Octavio, et al. (2014).

- **Enfoques basados en muestreos**

- **Sub-muestreo**

  Consiste en eliminar aleatoriamente elementos de la clase mayoritaria hasta obtener el mismo tamaño que la clase minoritaria. Por sub-muestreo, se podría correr el riesgo de eliminar algunas de las instancias de la clase mayoritaria que son más representativas, descartando así información útil. Esto se puede ilustrar como sigue (Figura 6):

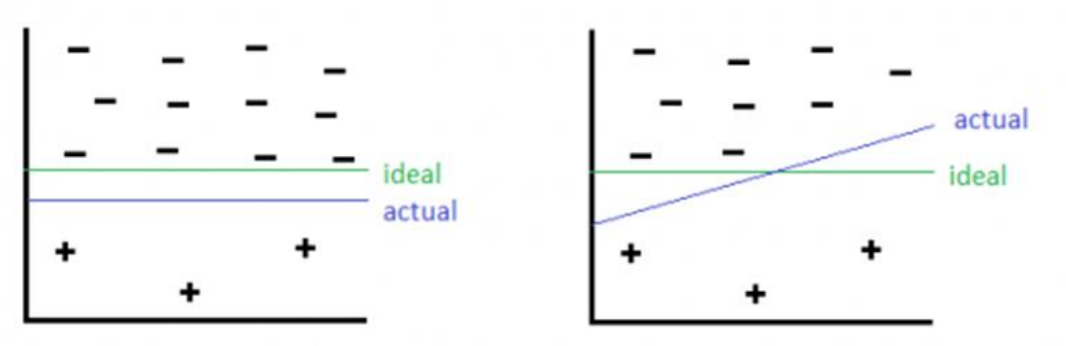


Figura 6: Sub-muestreo

FUENTE: Imbalance Problem, Lo (2013).

Aquí la línea verde es la frontera de decisión ideal que nos gustaría tener, y el azul es el resultado real. En el lado izquierdo está el resultado de aplicar un algoritmo de aprendizaje automático general sin el uso del sub-muestreo. A la derecha, se sub-muestrea la clase negativa, pero se remueve algo de información de clase negativa, lo que inclinó la frontera de decisión azul, causando que algunas clases negativas sean clasificadas como positivas erróneamente.

- **Sobre-muestreo**

  Consiste en generar ejemplos de la clase minoritaria aleatoriamente hasta tener tantos ejemplos como la otra clase con muestreo con reemplazo. Por sobre-muestreo, el solo duplicar las clases minoritarias podría llevar a problemas de sobreajuste del clasificador, las cuales son ilustradas a continuación (Figura 7):

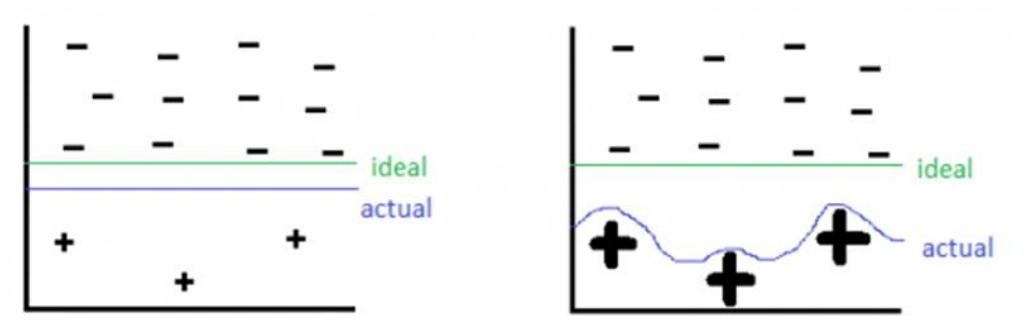


Figura 7: Sobre-muestreo

FUENTE: Imbalance Problem, Lo (2013).

El lado izquierdo se encuentra antes del sobre-muestreo, en el lado derecho el sobre-muestreo ya se ha aplicado. En el lado derecho, los signos positivos gruesos indican que hay múltiples copias repetidas de la instancia de datos. El algoritmo de aprendizaje automático ve estos casos muy a menudo y por lo tanto se diseña para ajustarse a estos ejemplos específicamente, lo que resulta en una línea azul limite como la anterior.

- **Enfoque híbrido**

  Estos métodos combinan los algoritmos de sobre-muestreo y sub-muestreo. Se pueden aplicar de forma secuencial o de conjunto. La idea es utilizar una forma inteligente de combinar las técnicas de re-muestreo expuestas anteriormente (Ramentol et al., 2011; Li et al., 2010). Estos métodos tienen las mismas limitaciones expuestas que los algoritmos de sobre-muestreo y sub-muestreo.

- **Enfoques más recientes**

  En 2002, un algoritmo de muestreo informativo llamado SMOTE (Synthetic minority over-sampling technique) se introdujo para abordar el problema de desequilibrio clase. Este es uno de los enfoques más adoptados, debido a su simplicidad y eficacia. Es una combinación entre sobre-muestreo y sub-muestreo, pero en este caso el sobre-muestreo no se hace mediante la replicación de clase minoritaria sino al crear nuevas instancias de datos de clase minoritaria a través de un algoritmo. En el sobre-muestreo tradicional, la clase minoritaria se replica. En SMOTE, nuevos datos minoritarios se construyen.

Otro enfoque que mencionan Moreno, Rodriguez, Sicilia, Riquelme y Ruiz (2009) es Boosting, que consiste en asociar pesos a cada instancia que se van modificando en cada iteración del clasificador. Inicialmente todas las instancias tienen el mismo peso y después de cada iteración, en función del error cometido en la clasificación se reajustan los pesos con objeto de reducir dicho error:

- ✓ AdaBoost: Implementa el algoritmo de Boosting descrito. En cada iteración AdaBoost genera nuevas instancias utilizando Resampling.
- ✓ SMOTEBoost: Es similar a AdaBoost pero usa SMOTE en lugar del Resampling para generar nuevas instancias.

- ✓ RUSBoost: Aplica AdaBoost pero en cada iteración utiliza RUS (Random Undersampling) que reducen el tamaño de la muestra de datos y simplifican y aumentan el rendimiento del clasificador.

2.6 CRITERIOS BASADOS EN DISTANCIAS COMO INDICADORES DE DISIMILARIDAD

Para medir lo similares (o disimilares) que son los individuos existe una enorme cantidad de índices de disimilaridad o divergencia.

La mayor parte de estos índices serán o bien, indicadores basados en la distancia (considerando a los individuos como vectores en el espacio de las variables) (en este sentido un elevado valor de la distancia entre dos individuos indicará un alto grado de disimilaridad entre ellos); o bien, indicadores basados en coeficientes de correlación; o bien basados en tablas de datos de posesión o no de una serie de atributos.

Se da, en general, el nombre de distancia o disimilaridad entre dos individuos i y j a una medida, indicada por $d_{(i,j)}$, que mide el grado de semejanza, o a mejor decir de desemejanza, entre ambos objetos o individuos, en relación a un cierto número de características cuantitativas y/o cualitativas. El valor de $d_{(i,j)}$ es siempre un valor no negativo, y cuanto mayor sea este valor mayor será la diferencia entre los individuos i y j.

Toda distancia debe verificar, al menos, las siguientes propiedades:

$$d_{(i,j)} = > 0 \text{ (no negatividad) (1)}$$

$$d_{(i,j)} = 0 \qquad (2)$$

$$d_{(i,j)} = d_{(j,i)} \text{ (simetría)} \qquad (3)$$

Una distancia es euclidiana cuando pueda encontrarse un espacio vectorial de dimensión igual o inferior a la dimensión del espacio de las variables en el que podamos representar a los individuos por puntos cuya distancia euclídea ordinaria coincida con la distancia utilizada.

Es decir si existe un espacio vectorial R^m, con $m < n$ (siendo n el número de variables consideradas para representar a los individuos) y dos puntos de ese espacio, P_i y P_j de coordenadas*:* $P_i = (P_{i1}, P_{i2}, \ldots P_{i\,m})$ *y* $P_j = (P_{j1}, P_{j2}, \ldots, P_{j\,m})$ verificándose que la distancia que estamos considerando entre los individuos i y j es igual a la distancia euclídea entre los puntos P_i y P_j en R^m ; esto es: Si $d_{(i,j)} = (P_i - P_j)^o$, diremos que la distancia $d_{(i,j)}$ es euclidiana.

Cuando la distancia es euclidiana se verifica además que:

$$d_{(i,j)} < d_{(i,t)} + d_{(j,t)} \text{ (Desigualdad triangular)} \qquad (4)$$

$$d_{(i,j)} > 0 \text{ " i } \neq \text{ j} \qquad (5)$$

Cualquier distancia que verifica la propiedad (4) es llamada distancia métrica.

Cumpliéndose, en consecuencia, que las distancias euclidianas son un subconjunto de las distancias métricas. Si además de verificar la propiedad (4) una distancia verifica la propiedad:

$$d_{(i,j)} < \max [d_{(i,t)}, d_{(j,t)}] \qquad (6)$$

(Desigualdad triangular ultra métrica) se dice que la distancia es ultra métrica.

Existe una gran cantidad de distancias e indicadores de disimilaridad y no se puede disponer de una regla general que nos permita definir una disimilaridad conveniente para todo tipo de análisis. De las propiedades de que goce, de la naturaleza de las variables utilizadas y de los individuos estudiados y de la finalidad del análisis dependerá la adecuada elección de una u otra.

Recordando que los datos de partida del análisis son las observaciones de n variables y N individuos.

Teniendo en cuenta esto, se pueden representar a los individuos en el espacio de las variables de manera que representaremos al individuo i-ésimo por el vector (7):

$$W_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix} \text{ y al individuo j-simo: } W_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad (7)$$

- **Distancias para variables numéricas**

  - **Distancia euclídea**

    La distancia euclídea es la disimilaridad más conocida y más sencilla de comprender, pues su definición coincide con el concepto más común de distancia. El cálculo de esta se muestra en la figura 8.

    Su expresión es la siguiente (1):

$$d_{(i,j)} = (W_i - W_j)' (W_i - W_j) \quad (1)$$

    La distancia euclídea, a pesar de su sencillez de cálculo y de que verifica algunas propiedades interesantes tiene dos graves inconvenientes:

- El primero de ellos es que la euclídea es una distancia sensible a las unidades de medida de las variables: las diferencias entre los valores de variables medidas con valores altos contribuirán en mucha mayor medida que las diferencias entre los valores de las variables con valores bajos. Como consecuencia de ello, los cambios de escala determinarán, también, cambios en la distancia entre los individuos. Una posible vía de solución de este problema es la tipificación previa de las variables, o la utilización de la distancia euclídea normalizada.

- El segundo inconveniente no se deriva directamente de la utilización de este tipo de distancia, sino de la naturaleza de las variables. Si las variables utilizadas están correlacionadas, estas variables nos darán una información, en gran medida redundante. Parte de las diferencias entre los valores individuales

de algunas variables podrían explicarse por las diferencias en otras variables. Como consecuencia de ello la distancia euclídea inflará la disimilaridad o divergencia entre los individuos.

La solución a este problema pasa por analizar las componentes principales (que están no correlacionadas) en vez de las variables originales. Otra posible solución es ponderar la contribución de cada par de variables con pesos inversamente proporcionales a las correlaciones, lo que nos lleva, como veremos a la utilización de la distancia de Mahalanobis. La distancia euclídea será, en consecuencia, recomendable cuando las variables sean homogéneas y estén medidas en unidades similares y/o cuando se desconozca la matriz de varianzas.

Ejemplo:

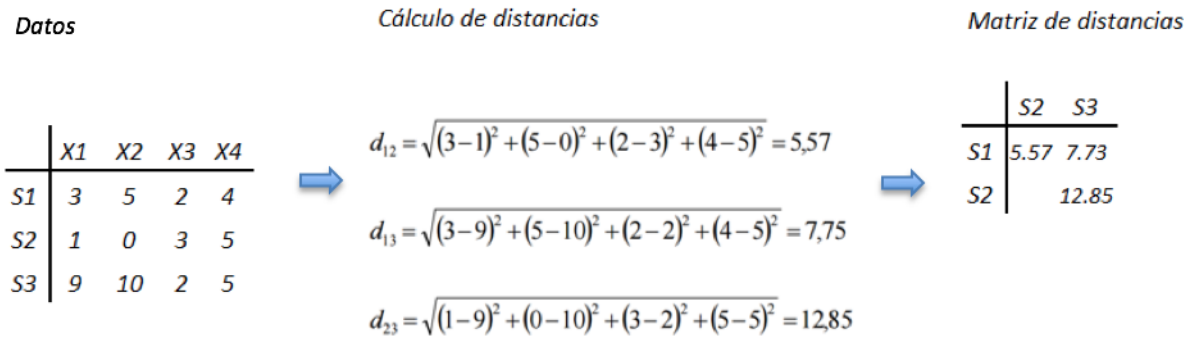


Figura 8: Cálculo de las distancias euclidianas

FUENTE: Elaboración propia.

- **Distancia euclídea normalizada**

  La distancia euclídea normalizada entre los individuos i y j la definiremos como:

$$d_{(i,j)} = (W_i - W_j)' S^{-1} (W_i - W_j) \qquad (1)$$

Donde S es una matriz diagonal con las varianzas en su diagonal principal y ceros en el resto de sus elementos. Obviamente S^{-1} será su inversa: la matriz diagonal que tendrá los valores recíprocos de las varianzas en su diagonal. Utilizar como matriz de la forma cuadrática distancia la matriz S^{-1}, en vez de la

matriz identidad, I, es, claramente, equivalente a utilizar como valores de partida los valores de las variables cambiados de escala a la desviación típica de las variables.

Empleando este tipo de distancia solventamos el inconveniente de los efectos de unidades de medida distintas de las variables y obtenemos una distancia que no dependerá de las unidades de medida.
Sin embargo, la alta correlación entre algunas variables puede seguir siendo un grave inconveniente.

- **Distancia de Mahalanobis**

  La distancia de Mahalanobis entre los individuos i y j la definimos por la expresión (1):

$$d_{(i,j)} = (W_i - W_j)' V^{-1} (W_i - W_j) \quad (1)$$

  Donde la matriz asociada a la forma cuadrática V^{-1} es la inversa de la matriz de varianzas *V*.
  Esta distancia presenta las ventajosas propiedades de solventar los dos inconvenientes de la aplicación de la distancia euclídea: Por un lado, es invariante ante los cambios de escala y no depende, por tanto, de las unidades de medida.
  En efecto: Si se consideran las variables originales ***x*** representadas por el vector de variables (2):

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad (2)$$

  y considerando su transformación lineal a otras nuevas variables, ***y,*** representadas por el vector de variables (3):

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (3)$$

que vendrá dada por la relación Y=CX; la matriz de varianzas de *Y* será: *V= C' V C.*

En el espacio de las nuevas variables los individuos vendrán representados por un nuevo vector: *W* = C' W*

La distancia de Mahalanobis sobre las nuevas variables será entonces (4):

$$d_{(i,j)} = (W_i^* - W_j^*)' V^{-1}(W_i^* - W_j^*) = (W_i - W_j)'C[(C' V^{-1} C)]C' (W_i - W_j)$$
$$= (W_i - W_j)'V^{-1}(W_i - W_j) \quad (4)$$

que es la distancia de Mahalanobis calculada sobre las variables originales.

Por otro lado, al utilizar la matriz *V*, se consideran las correlaciones entre las variables y se corrige el efecto de la redundancia.

Sin embargo, se debe tener en cuenta dos observaciones:

- ✓ Si las variables están no correlacionadas, la distancia de Mahalanobis coincide con la distancia euclídea normalizada. En efecto: Si las variables están no correlacionadas la matriz *V* coincide con la matriz *S*, y, por tanto, la inversa de *V* coincidirá con la inversa de *S*.
- ✓ La distancia de Mahalanobis coincide con la distancia euclídea calculada sobre el espacio de las componentes principales.

La distancia de Mahalanobis es invariante respecto de los cambios de escala. En particular, será invariante respecto de la tipificación. De forma que podemos partir de la distancia de Mahalanobis sobre el espacio de las variables tipificadas. En consecuencia, se representa cada individuo por el vector (5):

$$W_i^* = \begin{pmatrix} F_{i1} \\ F_{i2} \\ \vdots \\ F_{in} \end{pmatrix} \text{ teniendo que } W_i^* = A'W_i \quad (5)$$

La distancia de Mahalanobis entre los individuos i y j vendrá dada por (6):

$$d_{(i,j)} = (W_i - W_j)' R^{-1} (W_i - W_j) \quad (6)$$

Por otro lado, $R = A'A$ (donde A es la matriz factorial); y de la relación entre componentes principales y variables originales, tenemos que: $Z = A\,F$.

De manera que si el i-ésimo individuo puede describirse en función de las componentes principales como:

De forma que la distancia de Mahalanobis quedaría (7):

$$d_{(i,j)} = (W_i - W_j)'R^{-1}\,(W_i - W_j) = (W_i - W_j)'(AA')^{-1}\,(W_i - W_j) = (W_i - W_j)'\,AA'\,(W_i - W_j)$$
$$= (W_i^* - W_j^*)'\,(W_i^* - W_j^*) \qquad (7)$$

es decir, la distancia euclídea entre los individuos considerados en función de las componentes principales.

Para los casos en los que existan relaciones lineales entre las variables, y, por tanto, la matriz V sea singular, la distancia de Mahalanobis puede generalizarse como (8):

$$d_{(i,j)} = (W - W)'\,G\,(W - W) \qquad (8)$$

donde G es una **g-inversa** que verifica que $VGV = V$.

- **Otras distancias**

  Además de las tres distancias citadas, que son las más utilizadas, cabe mencionar, entre otras:

  La distancia Manhattan o Ciudad

  $d_{(i,j)} = S|\,x_{ki} - x_{kj}\,|$

  La distancia de Chebyshev

  $d_{(i,j)} = Max|\,x_{ki} - x_{kj}\,|$

  Las distancias de MinKowski

  $d_{(i,j)} = (S(x_{ki} - x_{kj})^r)^{1/r}$

  donde al ir variando el valor de r se van generando distintas distancias.

- **Distancias para variables categóricas**

  - **Distancia euclídea binaria**

Versión de la distancia euclídea para variables dicotómicas. Este índice, que representaremos por $d_{(i,j)}$, se calcula a partir de una tabla de frecuencias 2x2 elaborada para cada par de sujetos o variables a clasificar. Como las variables son dicotómicas podemos codificar con 1 (presencia) a una categoría de la variable y con 0 (ausencia) a la otra categoría. Para cada par de sujetos se obtiene la siguiente tabla (Cuadro 4).

Cuadro 4: Frecuencia de doble entrada

		Sj	
		1	0
Si	1	a	b
	0	c	d

FUENTE: Elaboración propia.

En la tabla:

a es la frecuencia de acuerdos en el valor 1 para el conjunto de variables.

d: es la frecuencia de acuerdos en el valor 0 para el conjunto de variables.

b: es la frecuencia de desacuerdos. En las variables en las que el sujeto i tiene un 1 el sujeto j tiene un 0.

c: es la frecuencia de desacuerdos. En las variables en las que el sujeto i tiene un 0 el sujeto j tiene un 1.

La distancia euclídea binaria (1) se obtiene de la tabla de frecuencias con la expresión

$$d_{(i,j)} = \sqrt{b + c} = \sqrt{desacuerdos} \qquad (1)$$

La distancia así definida es un índice de la disimilaridad entre sujetos. El valor mínimo de 0 ocurre cuando todo son acuerdos y el valor máximo cuando no hay acuerdos en los sujetos comparados. Un ejemplo sencillo aclarará estos conceptos (Figura 9).

a)

	X1	X2	X3	X4	X5
S1	1	1	0	1	0
S2	1	1	1	0	0
S3	0	0	0	1	1
S4	0	0	0	0	1
S5	1	0	0	1	0

b)

		S2		S3		S4		S5	
		1	0	1	0	1	0	1	0
S1	1	2	1	1	2	0	3	2	1
	0	1	1	1	1	1	1	0	2
S2	1			0	3	0	3	1	2
	0			2	0	1	1	1	1
S3	1					1	1	1	1
	0					0	3	1	2
S4	1							0	1
	0							2	2

c)

	S2	S3	S4	S5
S1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$	$\sqrt{1}$
S2		$\sqrt{5}$	$\sqrt{4}$	$\sqrt{3}$
S3			$\sqrt{1}$	$\sqrt{2}$
S4				$\sqrt{3}$

d)

	S2	S3	S4	S5
S1	1.41	1.73	2	1
S2		2.23	2	2.73
S3			1	1.41
S4				1.73

Figura 9: Cálculo de las distancias euclidianas binarias

FUENTE: Elaboración propia.

a) Ejemplo de Datos: 5 individuos *(S1, ..., S5)* con 5 variables *(X1, ..., X5)*, las variables son discretas de naturaleza dicotómica.

b) Se calcula la matriz de frecuencias de doble entrada para cada individuo.

c) Se calculan las distancias mediante la fórmula dada.

d) Matriz de distancias entre todos los individuos

- **Distancia para variables Mixtas**

  - **Coeficiente de similitud de Gower**

    El coeficiente de similitud de Gower propuesto por Gower en 1971 permite la manipulación simultánea de variables cuantitativas y cualitativas en una base de datos, mediante la aplicación de este coeficiente se logra hallar la similitud entre individuos a los cuales se les han medido una serie de características en común. Una similaridad alta, es decir cercana a 1, indicará gran homogeneidad entre los individuos; por el contrario, una similaridad cercana a cero indica que los individuos son diferentes.

    Se probó en que la definición del coeficiente de similitud de Gower al cuadrado tiene estructura métrica y que, por tanto, es más que un coeficiente de similitud, es una distancia (1).

$$d_{i,i'} = 1 - s(i,i') \quad (1)$$

Dado un conjunto I de individuos $I=\{i_1, i_2, \dots, i_n\}$, descritos por K variables $X_1, X_2, \dots, X_K$ de diferentes tipos, siendo x_{ik} el valor que la variable $X_k (k = 1{:}K)$ toma para el individuo $i \in I$, el coeficiente de similitud sugerido por Gower define la similitud entre el par de individuos (i, i') como (2):

$$s(i,i') = \frac{\sum_{k=1}^{K} w_k(i,i') s_k(i,i')}{\sum_{k=1}^{K} w_k(i,i')} \quad (2)$$

donde:

- K es el número de variables.

- $w_k(i,i')$ toma el valor de 1 o 0 indicando si la comparación de los individuos i, i' para la variable X_k, es o no es posible:
  - ✓ 0 si (x_{ik}= missing) o $x_{i'k}$=missing); es decir si posee un valor perdido.
  - ✓ 0 si (la variable x_{ik} es binaria) y (x_{ik} = false) y ($x_{i'k}$= false) y queremos excluir la ausencia negativa de la variable X_k
  - ✓ 1 de lo contrario

- $s_k(i,i')$ indica el indica el grado de similitud entre los individuos (i, i') para la variable k-ésima. Este grado de similitud, que abarca tanto variables numéricas como categóricas, está definido en el intervalo [0, 1] como:
  - ✓ $1 - \frac{|x_{ik} - x_{i'k}|}{R_k}$, si (X_k es numérica)
  - ✓ 1, si (X_k es numérica) y ($x_{ik} = x_{i'k}$ es numérica)
  - ✓ 0, si (X_k es numérica) y ($x_{ik} \neq x_{i'k}$ es numérica)

  Dado de que las unidades de medida de cada variable estén normalizadas por el rango de la variable X_k en lugar de la desviación estándar se debe a dos motivos, según Gower: el rango de la variable es más fácil de calcular y la desviación estándar tiene poco significado en el caso de trabajar con variables de tipos diversos.

Cabe señalar que existen dos tipos de variables cualitativas binarias: simétricas y asimétricas. Simétricas, en el cual el 0 equivale al mismo valor en las demás variables, esto siempre en cuando usemos el mismo tipo de variables, en la mayoría de casos usamos distintos tipos de variables y pudieras tener otro tipo de significado o valor. Por ello por defecto en la mayoría de casos se usa la distancia de Gower para variables asimétricas, es decir el 0 puede significar ausencia u otro tipo de valor según la variable en estudio.

Otra forma de expresar la distancia de Gower es como lo señala Aurea Grané hace referencia a $d^2{}_{i,i'} = 1 - s(i,i')$ para variables mixtas con variables binarias asimétricas, donde es el coeficiente de similaridad de Gower (3).

$$s(i,i') = \frac{\sum_{k=1}^{K}\left(1-\frac{|x_{ik}-x_{i'k}|}{Gh}\right)+a+\alpha}{p_1+(p_2-d)+p_3} \qquad (3)$$

- $\boldsymbol{p_1}$ es el número de variables cuantitativas continuas
- $\boldsymbol{p_2}$ es el número de variables binarias
- $\boldsymbol{p_3}$ es el número de variables cualitativas (no binarias)
- $\boldsymbol{a}$ es el número de coincidencias (1, 1) en las variables binarias
- $\boldsymbol{d}$ es el número de coincidencias (0, 0) en las variables binarias
- $\boldsymbol{\alpha}$ es el número de coincidencias en las variables cualitativas (no binarias) y $\boldsymbol{Gh}$ es el rango (o recorrido) de la h-ésima variable cuantitativa

Chávez (2010) menciona que, autores como (LONDOÑO et al, 2007) señalan algunas de las propiedades del coeficiente de similitud de Gower (GOWER, 1971) que resultan especialmente ventajosas para la taxonomía de las especies. Las mismas son las siguientes:

- ✓ Usando el coeficiente de similitud de Gower (GOWER, 1971), es posible trabajar con bases de datos en las que faltan observaciones de algunas variables, sin prescindir de todo el vector que representa a la unidad muestral ni usar ningún método de imputación. Esta

propiedad resulta muy útil en estudios taxonómicos pues a menudo aparecen observaciones faltantes. (LONDOÑO et al, 2007).

- ✓ Mediante el uso de este coeficiente es posible ponderar las variables de manera diferencial, dependiendo del papel que se quiera que cada una juegue en la ordenación. En este sentido, es posible asignar las ponderaciones dando mayor peso a las variables que en estudios precedentes han mostrado alta capacidad discriminante. (LONDOÑO et al, 2007).

2.7 SMOTE: TÉCNICA DE SOBRE-MUESTREO MINORITARIO SINTÉTICO

- **Sobre-muestreo minoritario con reemplazo**

  Investigaciones previas (Ling & Li, 1998; Japkowicz, 2000) han discutido el sobre-muestreo con reemplazo y han notado que no mejora considerablemente el reconocimiento de la clase minoritaria. En esta investigación se interpreta el efecto subyacente en términos de regiones de decisión en espacio característico. Esencialmente, a medida que se realiza el sobre-muestreo de la clase minoritaria cada vez más y más, el efecto es identificar como región de decisión para la clase minoritaria, las regiones similares, pero más específicas en el espacio característico.

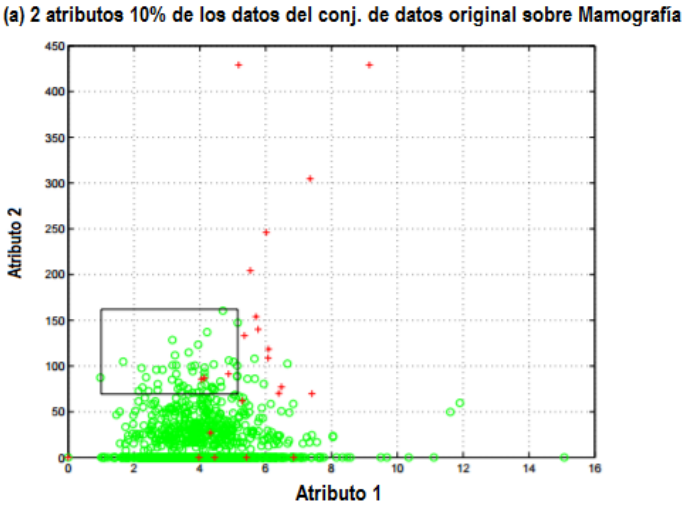


Figura 10: Muestras de clases minoritarias

FUENTE: SMOTE, Synthetic Minority Over-sampling Technique. Nitesh V. Chawla (2002)

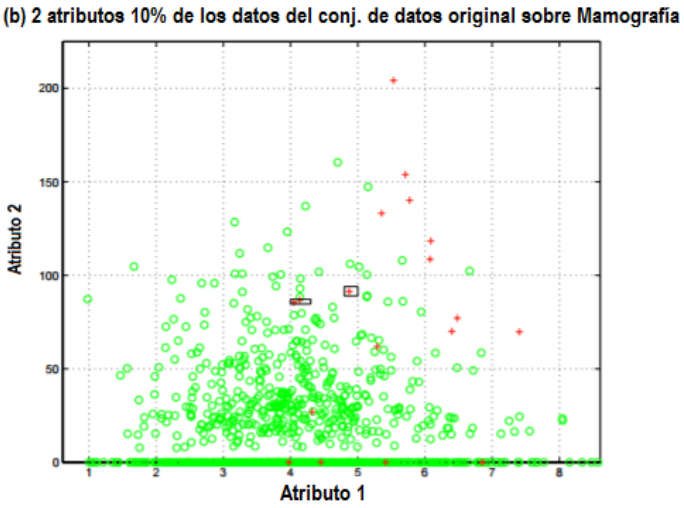


Figura 11: Vista aumentada del sobre-muestreo con replicación

FUENTE: SMOTE, Synthetic Minority Over-sampling Technique. Nitesh V. Chawla (2002)

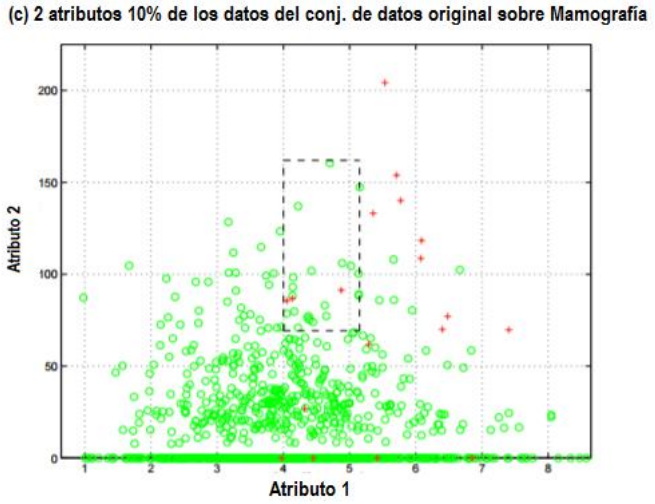


Figura 12: Vista aumentada de la región de decisión del sobre-muestreo con generación sintética

FUENTE: SMOTE, Synthetic Minority Over-sampling Technique. Nitesh V. Chawla (2002).

La región de decisión en la que tres muestras de clases minoritarias (expuestas por "+") residen después de construir un árbol de decisión. Esta región de decisión se indica con el rectángulo de línea gruesa, representa la región de decisión en la que residen tres muestras de clases minoritarias (Figura 10). Una vista aumentada de las

muestras elegidas de la clase minoritaria para el mismo conjunto de datos. Los rectángulos pequeños de línea gruesa muestran las regiones de decisión como un resultado de realizar el sobre-muestreo de la clase minoritaria con replicación (Figura 11). Una vista aumentada de las muestras elegidas de la clase minoritaria para el mismo conjunto de datos. Líneas punteadas muestran la región de decisión después de realizar el sobre-muestreo de la clase minoritaria con generación sintética (Figura 12).

Los datos de los gráficos de las figuras 10,11 y 12 se extrajeron de un conjunto de datos sobre la Mamografía (Woods et al., 1993). Las muestras de la clase minoritaria se muestran como + y las muestras de la clase mayoritaria aparecen como o en el gráfico. En la Figura 10 la región indicada por el rectángulo de línea gruesa es una región de decisión de la clase mayoritaria. Sin embargo, contiene tres muestras de la clase minoritaria que se representan por "+" como falsos negativos. Si replicamos la clase minoritaria, la región de decisión para la clase minoritaria se volverá muy específica y causará nuevas divisiones en el árbol de decisión. Esto conducirá a más nódulos terminales (hojas), ya que el algoritmo de aprendizaje trata de adquirir información de más y más regiones específicas de la clase minoritaria; en esencia, sobreajuste (overfitting). La replicación de la clase minoritaria no causa que su límite de decisión se expanda a la región de la clase mayoritaria. Entonces, en la Figura 11, las tres muestras que antes se encontraban en la región de decisión de la clase mayoritaria, ahora tienen regiones de decisión muy específicas.

- **Algoritmo SMOTE: descripción general.**

  Esta investigación propone utilizar un enfoque de sobre-muestreo en la cual se realiza el sobre-muestreo a la clase minoritaria creando ejemplos "sintéticos" en lugar de utilizar el sobre-muestreo con reemplazo. Este enfoque se inspira en una técnica que demostró ser exitosa en el reconocimiento de caracteres escritos a mano (Ha & Bunke, 1997). Ellos crearon datos de entrenamiento adicionales al realizar ciertas operaciones en datos reales. En su caso, operaciones como rotación e inclinación eran formas naturales que se utilizaban para perturbar los datos de entrenamiento. Esta investigación se enfoca en la generación de ejemplos sintéticos

en una manera específica de aplicación, ya que operamos en "el espacio característico" en lugar de en "el espacio de datos". Se realiza el sobre-muestreo en la clase minoritaria tomando cada muestra de clase minoritaria e introduciendo ejemplos sintéticos a lo largo de los segmentos de línea que se unen a todos los k vecinos más cercanos de la clase minoritaria. Dependiendo de la cantidad de sobre-muestreo requerido, se escogen los vecinos que pertenezcan a los k vecinos más cercanos de manera aleatoria. Actualmente, nuestra implementación utiliza cinco vecinos más cercanos. Por ejemplo, si la cantidad de sobre-muestreo necesario es 200%, solo se escogen dos vecinos de los cinco vecinos más cercanos y se genera una muestra en la dirección de cada uno. Las muestras sintéticas se generan de la siguiente forma: se toma la diferencia que existe entre el vector característico bajo consideración (muestra) y su vecino más cercano. Se multiplica esta diferencia por un número aleatorio entre 0 y 1, y se suma al vector característico bajo consideración. Esto genera la selección de un punto aleatorio a lo largo del segmento de línea entre dos características específicas. Este enfoque obliga de manera efectiva a la región de decisión de la clase minoritaria a volverse más general.

- **Aplicación para variables continuas**

  El algoritmo SMOTE, es el pseudocódigo para SMOTE. En la figura 13 se muestra un ejemplo del cálculo de muestras sintéticas aleatorias. La cantidad de sobre-muestreo es un parámetro del sistema, además, se pueden generar series de curvas ROC para poblaciones diferentes y se puede realizar el análisis ROC.

Los ejemplos sintéticos provocan que el clasificador cree regiones de decisión menos específicas y más grandes como se muestra en las líneas punteadas de la Figura 12, en vez de regiones más específicas y más pequeñas. Ahora la clase minoritaria adquiere información de las regiones más generales en lugar de tomarla de aquellas que la clase mayoritaria subsume y que se encuentran a su alrededor. Esto da como resultado que los árboles de decisión generalicen mejor. Se comparan el muestreo minoritario con reemplazo (clásico Sub-muestreo) y SMOTE. Los experimentos se llevaron a cabo en el conjunto de datos sobre mamografía. Originalmente, había 10,923 individuos en la clase mayoritaria y 260 ejemplos en

la clase minoritaria. Aproximadamente, tenemos 9,831 individuos en la clase mayoritaria y 233 ejemplos en la clase minoritaria para el conjunto de entrenamiento utilizado en la validación cruzada de 10 iteraciones. Se realizó el sobre-muestreo de la clase minoritaria al 100%, 200%, 300%, 400% y 500% de su tamaño original. Los gráficos muestran que los tamaños de los árboles para el sobre-muestreo minoritario con reemplazo, en grados más altos de replicación, son más grandes que aquellos para SMOTE, y que la técnica del reconocimiento de la clase minoritaria de sobre-muestreo minoritario con reemplazo, en grados más altos de replicación, no es tan buena como SMOTE.

```
Algoritmo SMOTE (T, N, k)
Entrada: Número de muestras de clase minoritaria T; Cantidad de SMOTE N%;
Número de k vecinos más cercanos
Salida: (N/100)* Muestras sintéticas de clase minoritaria T

      1. (* si N es menor que 100%, aleatorizar las muestras de clase
      minoritaria ya que solo se aplicará SMOTE en un porcentaje aleatorio de
      éstas. *)
      2. if N < 100
      3. then Aleatorizar las muestras de la clase minoritaria T
      4. T = (N/100) * T
      5. N = 100
      6. endif
      7. N = (int)(N/100)(* Asumir que la cantidad de SMOTE está en integrales
      múltiples de 100. *)
      8. k = Número de vecinos más cercanos
      9. numattrs = Número de atributos
      10. Sample[ ][ ]:  matriz para muestras originales de clase minoritaria
      11. newindex: Mantiene un conteo del número de muestras sintéticas
      generadas, asignadas con un valor inicial de 0
      12. Synthetic[ ][ ]: matriz para muestras sintéticas
      (* Calcular los k vecinos más cercanos sólo para cada muestra de clase
      minoritaria. *)
      13. for i ← 1 to T
      14. Calcular los k vecinos más cercanos para i, y guardar los índices en
      el nnarray
      15. Populate(N, i, nnarray)
      16. endfor
      Populate(N, i, nnarray) (* Función para generar las muestras sintéticas.
      *)
      17. while N 0
      18. Escoger un número aleatorio entre 1 y k, nombrarlo nn, Este paso
      escoge uno de los k vecinos más cercanos de i.
      19. for attr ← 1 to numattrs
      20. Calcular: dif = Sample[nnarray[nn]][attr] – Sample[i][attr]
      21. Calcular: gap = número aleatorio entre 0 y 1
      22. Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
      23. endfor
      24. newindex++
      25. N = N – 1
      26. endwhile
      27. return (* Final de Populate. *)
      Final del pseudocódigo.
```

Figura 13: Algoritmo SMOTE con variables numéricas

FUENTE: SMOTE, Synthetic Minority Over-sampling Technique. Nitesh V. Chawla (2002).

- **Aplicación para variables discretas**

  Por lo general, SMOTE es una técnica poderosa que se propuso para conjuntos de datos desbalanceadas con variables numéricas originalmente. Sin embargo, las variables categóricas (o discreta, nominal) se presentan en una variedad de aplicaciones de conjuntos de datos que también se encuentran desbalanceados.

  Como aleatoriamente se elige un vecino tras calcular k vecinos más cercanos, podríamos asignar el valor nominal de la mayoría, es decir la moda, de los vecinos a la nueva instancia sintético de la muestra. Esta es la manera más fácil de tratar con variables categóricas (Tianxiang Gao 2015).

  Cuando se tenga una base de datos mixtas es decir variables cualitativas y cuantitativas, para hallar los vecinos más cercanos se usa el coeficiente de Gower, ya antes descrito, dado que nos permite trabajar con los tipos de variables en a la vez (Figura 14).

```
Algoritmo SMOTE (T, N, k)
Entrada: Número de muestras de clase minoritaria T; Cantidad de SMOTE N%; Número de k vecinos más cercanos
Salida: (N/100)*  Muestras sintéticas de clase minoritaria T

      1. (* si N es menor que 100%, aleatorizar las muestras de clase minoritaria ya que solo se aplicará SMOTE en un porcentaje mayor a 100 *)
      2. if N < 100
      3. then Aleatorizar las muestras de la clase minoritaria T
      4. T = (N/100) * T
      5. N = 100
      6. endif
      7. N = (int)(N/100)(* Asumir que la cantidad de SMOTE está en integrales múltiples de 100. *)
      8. k = Número de vecinos más cercanos
      9. numattrs = Número de atributos
      10. Sample[ ][ ]:  matriz para muestras originales de clase minoritaria
      11. newindex: Mantiene un conteo del número de muestras sintéticas generadas, asignadas con un valor inicial de 0
      12. Synthetic[ ][ ]: matriz para muestras sintéticas
      (* Calcular los k vecinos más cercanos sólo para cada muestra de clase minoritaria. *)
      13. for i ← 1 to T
      14. Calcular los k vecinos más cercanos para i, y guardar los índices en el nnarray
      15. Populate(N, i, nnarray)
      16. endfor
      Populate(N, i, nnarray) (* Función para generar las muestras sintéticas. *)
      17. while N 0
```

```
18. Escoger un número aleatorio entre 1 y k, nombrarlo nn, Este paso
escoge uno de los k vecinos más cercanos de i.
19. for attr ← 1 to numattrs
20.    if(variable j is categorical) then   /*Usar los valores nominales
en  caso de V.Categóricas*/
       Synthetic[newindex][attr]= Sample[nnarray[nn]][attr]
       else
21. Calcular: dif = Sample[nnarray[nn]][attr] – Sample[i][attr]
22. Calcular: gap = número aleatorio entre 0 y 1
23. Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
24. endfor
25. newindex++
26. N = N – 1
27. endwhile
28. return (* Final de Populate. *)
Final del pseudocódigo.
```

Figura 14: Algoritmo SMOTE incluyendo variables categóricas

FUENTE: Hybrid classification approach of SMOTE and instance selection for imbalanced datasets. Tianxiang Gao (2015).

- **Descripción del algoritmo SMOTE**

  El algoritmo de SMOTE realiza los siguientes pasos:

  - Recibe como parámetro el porcentaje de ejemplos a sobre-muestrear.
  - Calcula el número de ejemplos que tiene que generar.

    Para crear las instancias artificiales o sintéticas realiza una interpolación de un objeto con sus vecinos más cercanos como se muestra en la Figura 15.

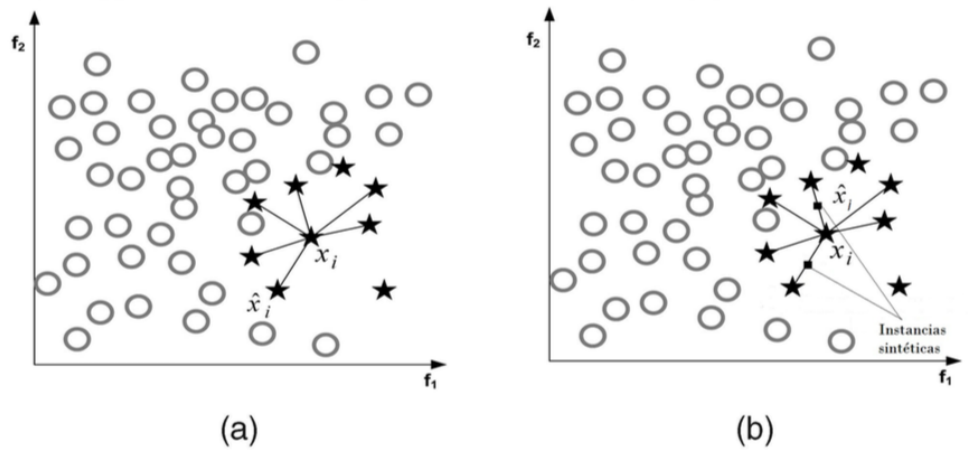


Figura 15: (a) Ejemplo de K-vecinos más cercanos de xi, considerando k = 6. (b) Generación de instancias sintéticas, dos instancias sintéticas entre la línea x_i y $\hat{x}_i$.

FUENTE: Nicandro, C. R,et al. 2013.

  - Calcula los k vecinos más cercanos de los ejemplos de la clase minoritaria.

Para determinar los k vecinos más cercanos, se crean N distancias entre la instancia original y los vecinos.

- Se usa la medida de distancia o similaridad según la situación como se muestra en el siguiente cuadro (Cuadro 5):

Cuadro 5: Medidas a usar según el tipo de variables

Tipo de Variable	Medida de distancia a usar
Variables Numéricas	D.Euclidiana/D.Euclidiana normalizada/D.Mahalanobis
Variables Categóricas	D. Euclidiana Binaria
Variables Mixtas: Nominales, Jerárquicas, escala, dicotómicas	Distancia de Gower

FUENTE: Elaboración propia.

- Para cada atributo del ejemplo a sobre-muestrear, calcula la diferencia entre el vector de atributos muestra y el vecino elegido.
- En el caso la variable sea categórica, se escoge la moda de los vecinos más cercanos.
- Multiplica esta diferencia por un número aleatorio entre 0 y 1.
- Suma este último valor al valor original de la muestra.
- Devuelve el conjunto de ejemplos sintéticos

- **Ejemplo del algoritmo SMOTE**

En el cuadro 6 se presenta un ejemplo con 10 individuos y 4 variables predictoras que constituyen la clase minoritaria.

Cuadro 6: Muestra de 10 individuos fugados en una entidad bancaria

	EDAD	SUELDO	SEXO	REGION
S1	24	4.830	0	0
S2	36	2.813	1	1
S3	46	2.155	1	0
S4	21	2.718	1	1
S5	20	1.285	0	1
S6	56	4.712	1	0
S7	23	4.728	0	0
S8	60	4.883	0	1
S9	55	4.668	1	1

S10	25	3.134	1	0

SEXO: 1=Mujer, 0=Hombre
REGION: 1=Lima, 0=Fuera de Lima

FUENTE: Clientes de una entidad financiera nacional.

Se desea aumentar en 300% los datos mostrados (N=300). Teniendo la instancia original (k =5) donde k es el número de vecinos más cercanos. Sea:

$\mathbf{X}^{V}$ = Matriz que contiene los k vecinos más cercanos.

$\mathbf{X}^{C}$ = Matriz característico.

$\theta \approx U(0,1)$ valor aleatorio entre 0 y 1.

X= Instancia de la clase minoritaria.

NS=Nueva instancia sintética o artificial.

Se tiene un mix de variables por ello se utiliza al coeficiente de similitud de Gower Como ejemplo se calculará la distancia de $d^2{}_{S1,S3}$ y $d^2{}_{S6,S9}$ con la fórmula de similitud Gower (1).

$$s(i,i') = \frac{\sum_{k=1}^{K}\left(1-\frac{|x_{ik}-x_{i'k}|}{Gh}\right)+a+\alpha}{p_1+(p_2-d)+p_3} \quad (1)$$

donde:

$\boldsymbol{p_1}$ = el número de variables cuantitativas continuas.

$\boldsymbol{p_2}$ = el número de variables binarias.

$\boldsymbol{p_3}$ = el número de variables cualitativas (no binarias).

$\boldsymbol{a}$ = el número de coincidencias (1, 1) en las variables binarias.

$\boldsymbol{d}$ = el número de coincidencias (0, 0) en las variables binarias.

$\boldsymbol{\alpha}$ = es el número de coincidencias en las variables cualitativas (no binarias).

$\boldsymbol{Gh}$= es el rango (o recorrido) de la h-ésima variable cuantitativa.

Se empieza a calcular los parámetros necesarios como se muestra en el Cuadro 7

$\boldsymbol{Gh}_{\text{edad}} = |60-20| = 40$

$\boldsymbol{Gh}_{\text{sueldo}} = |4{,}88-1{,}29| = 3{,}60$

Cuadro 7: Cálculo de los parámetros para el cálculo de Gower

	S1 S3 *Edad*	***S1 S3*** *Sueldo*	***S6 S10*** *Edad*	***S6 S10*** *Sueldo*
Gh	40	3,60	40	3,60
$\lvert x_{ik} - x_{i'k} \rvert$	22,00	2,68	31,00	1,58
p_1	2		2	
p_2	2		2	
d	1		1	
p_3	0		0	
a	0		1	
α	0		0	

FUENTE: Elaboración propia.

Reemplazando en (1):

$$S(S1,S3) = \frac{\left(1-\frac{|24-46|}{20-60}\right)+\left(1-\frac{|4,83-2,16|}{4,88-1,29}\right)+0+0}{2+(2-1)+0} = 0,23551$$

$$d^2_{S1,S3} = 1 - 0,23551 = 0,76449$$

$$S(S6,S10) = \frac{\left(1-\frac{|56-25|}{20-60}\right)+\left(1-\frac{|4,71-3,13|}{4,88-1,29}\right)+1+0}{2+(2-0)+0} = 0,59547$$

$$d^2_{S1,S10} = 1 - 0,59547 = 0,40453$$

De esa manera se calcula el resto de distancias obteniendo el cuadro 8.

Cuadro 8: Matriz de distancias cuadradas de Gower

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**
S2	0,715								
S3	0,764	0,358							
S4	0,665	0,100	0,445						
S5	0,695	0,456	0,723	0,356					
S6	0,611	0,507	0,320	0,607	0,963				
S7	0,027	0,714	0,763	0,652	0,677	0,610			
S8	0,638	0,544	0,777	0,644	0,667	0,537	0,656		
S9	0,705	0,248	0,481	0,348	0,704	0,259	0,704	0,296	
S10	0,499	0,341	0,266	0,304	0,660	0,405	0,498	0,840	0,544

FUENTE: Elaboración propia en R-Project, Paquete DMwR.

Ahora que se tienen las distancias para todos, la atención se centra en los vecinos más cercanos para S1, se obtiene el vector-matriz con las distancias cuadradas de Gower, se escogen los k=5 vecinos más cercanos con respecto a S1; son S4, S6, S7, S8 y S10.

$\mathbf{X}^{V}$=Vector que contiene los k vecinos más cercanos escogidos: Dado que N=300%, se escoge 3 vecinos cercano aleatoriamente; S4, S8 y S10. θ se escogió aleatoriamente entre 0 y 1, resulto 0.0174.

$$\mathbf{X}^{V} = \begin{bmatrix} 21 & 2.718 & 1 & 1 \\ 60 & 4.883 & 0 & 1 \\ 25 & 3.134 & 1 & 0 \end{bmatrix}$$

Nuevas instancias (NS):

Variables cuantitativas:

$\boldsymbol{NS}=\mathbf{X}^{C} + \theta\ (\mathbf{X}^{V} - \mathbf{X})$

$$\boldsymbol{NS} = [24 \quad 4.830] + 0.0174 \times \left(\begin{bmatrix} 21 & 2.718 \\ 60 & 4.883 \\ 25 & 3.134 \end{bmatrix} - [24 \quad 4.830] \right)$$

$$\boldsymbol{NS} = \begin{bmatrix} 23.9 & 4.793 \\ 24.6 & 4.831 \\ 24.0 & 4.800 \end{bmatrix}$$

Variables categóricas:

Para variables categóricas se asigna el valor nominal del vector característico que contiene los valores de las variables categóricas.

$$\boldsymbol{NS} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix})$$

Este proceso se realizó para obtener las instancias sintéticas para el primer individuo. De manera similar se procede para obtener los individuos sintéticos de los 9 individuos restantes. En el siguiente cuadro 9 se muestra las instancias completas para todos los individuos.

Cuadro 9: Datos con las 3 instancias sintéticas creadas

	EDAD	SUELDO	SEXO	REGION	
S1	24	4.83	0	0	
S2	36	2.813	1	1	
S3	46	2.155	1	0	
S4	21	2.718	1	1	
S5	20	1.285	0	1	
S6	56	4.712	1	0	
S7	23	4.728	0	0	
S8	60	4.883	0	1	
S9	55	4.668	1	1	
S10	25	3.134	1	0	
Si1,1	24	4.793	1	1	Instancia sinteticas para el S1.
Si1,2	25	4.831	0	1	
Si1,3	24	4.8	1	0	
.	.	.	.	.	
:	:	:	:	:	
Si10,1	45	3.123	0	0	Instancia sinteticas para el S10.
Si10,2	35	2.741	1	1	
Si10,3	28	4.672	1	0	

FUENTE: Elaboración propia.

III. MATERIALES Y MÉTODOS

3.1 MATERIALES Y EQUIPO

Los materiales y equipos de los cuales se hizo uso en la presente tesis son los siguientes:

a) Una computadora laptop marca DELL, con un procesador Intel® Core (TM) i5-6410M CPU @ 3.40GHz, 2501 Mhz, 4 procesadores principales, 8 procesadores lógicos y con una memoria RAM de 8 GB y un sistema operativo Windows 10 Pro de 64 bits.
b) El programa R y su versión 3.4.3 con sus librerías SMOTE, MASS y LMG; el software SPSS Versión 24. En el anexo se presentan las sentencias en R para la aplicación y resultados.

3.2 TIPO DE INVESTIGACIÓN

La investigación que se realizará es de tipo causal. Así mismo, es una investigación no experimental dado que no existe ningún manejo y efecto sobre las variables.

3.3 FORMULACIÓN DE LAS HIPÓTESIS

El modelo de Regresión Logística con datos desbalanceados con aplicación del algoritmo SMOTE clasifica mejor la fuga de clientes en una entidad financiera con respecto a modelos con sub-muestro aleatorio y sin la realización de algún método de muestreo.

3.4 POBLACIÓN

La aplicación de la investigación se realizará sobre datos históricos de 12 meses, es decir el evolutivo mensual de un año del producto CTS (compensación por tiempo de servicios) en todas las agencias de una entidad financiera del País. El periodo contempla desde el mes de enero del 2015 hasta diciembre del 2015. Se tomará en cuenta todas las cuentas de CTS que existen en la entidad financiera, donde las cuentas con saldo cerradas voluntariamente por los clientes serán tomadas como **fugas.** Mientras que las cuentas CTS aun activas con saldo

mayor o igual a S/350 nuevos soles, serán tomadas como **no fugas,** los montos de corte fueron definidos por los especialistas del producto**.** La distribución y proporción de los datos es la siguiente (Cuadro 10).

Cuadro 10: Distribución de los datos a analizar

Situación	clientes	%
Clientes con CTS activa	6,099	90.94%
Clientes con CTS canceladas	610	9.06%
Total de clientes	**6,709**	**100.00%**

FUENTE: Elaboración Propia.

3.5 VARIABLES

Se cuenta con información de los clientes. Las variables que se encuentran disponibles por cliente, ya que fueron entregadas por la empresa, se pueden observar en el cuadro 11.

Cuadro 11: Variables

VARIABLES PREDICTORAS	DESCRIPCIÓN	TIPO DE VARIABLE
Tasa	Tasa de interés de la cuenta CTS	Numérica
Ant_Banco	Tiempo de antigüedad del cliente en la institución financiera en meses.	Numérica
Ant_Cts	Tiempo de antigüedad de la cuenta CTS en la institución financiera, en meses.	Numérica
Edad	Edad del cliente en años.	Numérica
Sexo	Sexo del cliente: 0=Femenino 1=Masculino	Categórica
Saldo_soles	Monto de Saldo de la cuenta CTS, en Soles.	Numérica
EstadoCivil	Estado Civil del Cliente : Div.Sol.Viu = Divorciado, Soltero y Viudo y Cas.Conv = Casado, Conviviente	Categórica
CrossSell	Número de productos vigentes con el banco, tanto pasivos o activos.	Numérica
Flag_Bancarizado	Flag si el cliente tiene deudas en el sistema financiero. 0: No bancarizado 1: Bancarizado	Categórica
Región	Zona a la que pertenece el cliente: NORTE.SUR,ORIENTE,CENTRO o LIMA_CALLAO	Categórica
VARIABLE A PREDECIR	**DESCRIPCIÓN**	**TIPO DE VARIABLE**
Flag_Fuga	0= Cliente no fugado 1= Cliente fugado	Categórica

FUENTE: Elaboración propia

3.6 METODOLOGÍA APLICADA

1. **Pre procesamiento de datos**
   - Detección de valores perdidos o nulos.
   - Detección de valores extremos.
   - Detección de multicolinealidad, selección y transformaciones de variables.

2. **Aplicación de técnica de clasificación: Regresión Logística**
   - Selección de muestra para construcción del modelo: Training y Testing.
   - Regresión Logística con datos sin balancear.
   - Regresión Logística con sub-muestreo aleatorio.
   - Regresión Logística con balanceo mediante el algoritmo SMOTE.

3. **Evaluación y comparación de clasificadores**
   - Matriz de confusión: Clasificación Global
   - Sensibilidad y especifidad.
   - Curva ROC.

4. **Interpretación de resultados**
   - Interpretación del modelo del mejor modelo.
   - Variables que influyen en la fuga del cliente con CTS.
   - Predicción de nuevos individuos.

IV. RESULTADOS Y DISCUSIÓN

4.1 PRE PROCESAMIENTO DE DATOS

- **Exploración, transformación de variables y preparación de datos**

  Antes de realizar cualquier análisis, se procede a la detección de variables nulas, vacías o outliers, si fuera el caso, en resumen cualquier dato que pudiera interferir o alterar el modelamiento con la finalidad de obtener los datos más fiables para la elaboración de nuestros análisis predictivos.

- **Detección de valores perdidos o nulos**

  Se realizó la detección de valores perdidos y/o nulos, se encontraron 3 valores perdidos en las variables; Edad y 1 en Saldo_Soles. Dado que esto llega a representar solo el 0.014% en el peor de los casos (en Edad), se procedió a depuración de los mismos.

- **Detección de valores extremos.**

  Se realizó la detección de valores outliers moderados y extremos, con ayuda de cuartiles y con diagramas de Boxplot para visualizarlos, esto con el fin de mostrar de manera ágil y sobre todo sencilla a las personas encargadas del proyecto que no tienen conocimientos estadísticos. Se encontró presencia de outliers en las variables Edad y Saldo_sol. En la variable Edad se eliminó edad menores a 19 años y mayores a 65 años, estos fueron identificados como outliers extremos, además no cumplen con los rangos de edad estipulados según las políticas de la entidad bancaria y por sugerencias de los especialistas del producto. En Saldo_sol, que es la variable que contiene el monto total de la CTS, se retiró 2 individuos con montos muy por encima de los S/. 150,000.

- **Detección de multicolinealidad, selección y transformaciones de variables**

  Para detectar linealidad entre las variables cuantitativas, correlacionadas entre ellas, se realizó un análisis de correlación de Pearson y correlación de clusters como se ve en el anexo 3.

  Dado que se tiene dos variables de la misma naturaleza como lo son ant.cliente.meses y ant.cts.meses, y por ende refleja linealidad, se decidió fusionar ambas variables sin perder su información convirtiéndola en un ratio, es decir ant.cts.meses/ant.cliente.meses [0,1], donde uno refleja que tiene el mismo tiempo como cliente en el banco como la posesión de una cuenta CTS, y un ratio muy cercano a 0, significa que tiene muy poco tiempo con la CTS. Luego de esta transformación se tienen las variables no correlacionadas entre sí, de esa manera se solucionó el problema de multicolinealidad.

  Y para las variables categóricas se realizó el test de bondad de ajuste de Chi-cuadrado con respecto a Y: fuga, para ver el tema de la dependencia. Se descartaron algunas variables categóricas que no influyen sobre Y.

  Al final se quedó con las siguientes variables (Cuadro 12):

Cuadro 12: Variables finales para el modelamiento

Variables	Tipo
TasaInteres	Numérica
saldo_sol	Numérica
Edad	Numérica
EstadoCivil	Categórica
REGION	Categórica
CrossSell	Numérica
FUGA	Categórica
ratio.ant	Numérica

FUENTE: Elaboración Propia.

4.2 APLICACIÓN DE TÉCNICA DE CLASIFICACIÓN: REGRESIÓN LOGÍSTICA

- **Selección de muestras para la construcción del modelo: Training y Testing**

  Antes de la construcción de los modelos se debe seleccionar una muestra de training y testing, 80% y 20% respectivamente de la data original. La construcción y entrenamiento de los modelos se realizan con el 80%, cualquier técnica de sub-muestreo o sobre-muestreo se realizará con la muestra de testing (figura 16). Así como todos los modelos serán testeadas con este 20%. Algo importante es mantener la proporción original de desbalance de los datos de aproximadamente 91% vs 9% en la variable respuesta en ambas muestras tanto trainnig y testing. (Ripley, 1996). Luego del preprocesamiento de datos se tiene la distribución que se presenta en la figura 16.

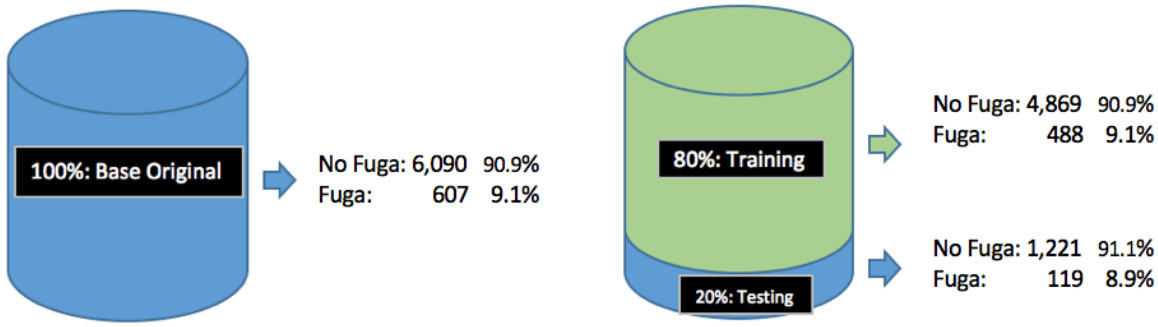


Figura 16: Selección de muestras: training y testing

FUENTE: Elaboración propia.

- **Regresión logística con datos sin balancear**

  Realizar un modelo logístico (Cuadro 13) sin un previo tratamiento de datos para balancear las categorías, es ignorar un paso importante, dado que la regresión no es capaz de "aprender" por la insuficiente cantidad de datos que posee una categoría, en este caso la categoría minoritaria Fuga=1.

Cuadro 13: Regresión logística con datos sin balancear

Variables	Coeficientes	Pr(>|z|)	Exp(B)
(Intercept)	4.055	0.00 ***	57.701
TasaInteres	-1.143	0.00 ***	0.319
saldo_sol	0.000	0.00 ***	1.000
Edad	-0.041	0.00 ***	0.960
REGIONLIMA_CALLAO	0.932	0.07 .	2.540
REGIONNORTE.SUR	1.311	0.01 *	3.709
REGIONORIENTE	1.329	0.05 .	3.778
ratio.ant	-0.798	0.00 ***	0.450

FUENTE: Elaboración propia con R-project.

El resultado de la matriz de confusión (Cuadro 14) revela que la regresión logística presenta tendencia de clasificación hacia la clase mayoritaria, minimizando de esta manera el error de clasificación y clasificando correctamente individuos de la clase mayoritaria en detrimento de individuos de la clase minoritaria.

Cuadro 14: Matriz de confusión de Regresión Logística sin balancear

	PREDICHO		
REAL	NO FUGA	FUGA	TOTAL
NO FUGA	1,221	-	1,221
FUGA	119	-	119
TOTAL	1,340	-	1,340

FUENTE: Elaboración propia con R-project.

Sin bien es cierto que la clasificación global y el ROC son altos (Cuadro 15), en cuanto a la sensibilidad, que mide la correcta clasificación de la clase minoritaroa deseada a predecir es nulo. Así como su R^2 es bajo, lo que indica que el porcentaje de variación de la variable de respuesta que explica su relación con una o más variables predictoras es bajo.

Cuadro 15: Indicadores de clasificación de una regresión logística sin balancear

	RL
Sensibilidad	0%
Especificidad	100%
ROC	64%
R^2	20%
Precisión / C. Global	91%

FUENTE: Elaboración propia con R-project.

- **Regresión logística con balanceo mediante sub-muestreo aleatorio**

  Se realizó un sub-muestro aleatorio simple a la categoría mayoritaria con el fin de balancear los datos, quedando la proporción de datos que se muestra en el Cuadro 16.

Cuadro 16: Datos balanceados mediante sub-muestreo aleatorio

Categoría	Nro	%
No fuga	610	50.2%
Fuga	604	49.8%
Total	1,214	100.0%

FUENTE: Elaboración propia.

La regresión logística resultante (cuadro 17) es la siguiente:

Cuadro 17: Regresión logística con datos mediante sub-muestreo aleatorio

Variables	Coeficientes	Pr(>|z|)	Exp(B)
(Intercept)	5.2640	0.000 ***	193.252
TasaInteres	-1.0473	0.000 ***	0.351
saldo_sol	0.0000	0.000 ***	1.000
Edad	-0.0399	0.000 ***	0.961
EstadoCivilDiv.Sol.Viu	-0.1443	0.035 .	0.866
REGIONLIMA_CALLAO	1.0070	0.068 .	2.737
REGIONNORTE.SUR	1.4256	0.011 *	4.160
REGIONORIENTE	1.3843	0.094 .	3.992
CrossSell	0.2735	0.001 ***	1.315
ratio.ant	-0.5241	0.014 *	0.592

FUENTE: Elaboración propia con R-project.

El resultado de la matriz de confusión (Cuadro 18) revela mejoras en cuanto a la predicción de la categoría minoritaria; fuga. Así como los indicadores de clasificación muestran una mejora a diferencia del modelo anterior (Cuadro 15).

Cuadro 18: Matriz de confusión de Regresión logística a con sub-muestreo aleatorio

	PREDICHO		
REAL	NO FUGA	FUGA	TOTAL
NO FUGA	711	510	1,221
FUGA	37	82	119
TOTAL	748	592	1,340

FUENTE: Elaboración propia con R-project.

La sensibilidad (Cuadro 19) resulta en un 69%, lo que es bueno, dado que se está prediciendo con éxito en un porcentaje aceptable de la categoría minoritaria. La especificidad logra superar el 50%, con 58%. Algo en tener en cuenta es que se desea predecir al cliente potencial en fugar, tratando de identificar también al que no lo hará, para ahorrar gastos en las campañas de retención. Dado que tratar a un cliente como potencialmente a fugarse cuando no lo es, incurre en gasto, pues la campaña consiste en regalos, elevar la tasa, etc.

Cuadro 19: Indicadores de clasificación de una regresión logística con sub-muestreo aleatorio

	RL SUB
Sensibilidad	69%
Especificidad	58%
ROC	67%
R^2	32%
Precisión / C. Global	59%

FUENTE: Elaboración propia con R-project.

- **Regresión logística con balanceo mediante SMOTE**

  Se realizó el balanceo de los datos mediante; sub-muestreo y sobre-muestreo con el algoritmo de SMOTE (Anexo 9) en R-project, con la librería DMwR, quedando de la siguiente manera la proporción (Cuadro 20):

Cuadro 20: Datos balanceados mediante SMOTE

Categoría	Nro	%
No fuga	1,940	44.4%
Fuga	2,425	55.6%
Total	4,365	100.0%

FUENTE: Elaboración propia con R-project, librería DmwR.

La regresión logística resultante se presenta a continuación en el Cuadro 21.

Cuadro 21: Regresión logística con balanceo mediante SMOTE

VARIABLE	Coeficientes	Pr(>|z|)		Exp(B)
(Intercept)	7.7074	< 2e-16	***	2,224
TasaInteres	-1.3756	< 2e-16	***	0.253
saldo_sol	-0.0101	< 2e-16	***	1.000
Edad	-0.0371	< 2e-16	***	0.964
EstadoCivilDiv.Sol.Viu	-0.9284	< 2e-16	***	0.395
REGIONLIMA_CALLAO	0.3847	0.113		1.469
REGIONNORTE.SUR	1.2207	0.000	***	3.390
REGIONORIENTE	2.2695	0.000	***	9.675
CrossSell	0.0305	0.044	.	1.031
ratio.ant	-0.7076	0.000	***	0.493

FUENTE: Elaboración propia con R-project.

Como resultado de la predicción la tabla de confusión (Cuadro 22), muestra una mayor cantidad de individuos clasificados correctamente como fugados, así como el número de individuos clasificados correctamente como no fugados.

Cuadro 22: Tabla Cruzada de Regresión Logística con SMOTE

	PREDICHO		
REAL	NO FUGA	FUGA	TOTAL
NO FUGA	900	321	1,221
FUGA	31	88	119
TOTAL	931	409	1,340

FUENTE: Elaboración propia con R-project.

Al observar el cuadro de indicadores de clasificación (Cuadro 23), confirmamos que la clasificación de ambas categorías mejora: minoritaria, la sensibilidad es de 74% y mayoritaria, la especificidad es 74%. Así mismo la precisión es aceptable con un 74%, área bajo la curva es del 68% y el R^2 es más alto que las regresiones anteriores con un 41%.

Cuadro 23: Indicadores de clasificación de una regresión logística aplicando SMOTE

	RL SMOTE
Sensibilidad	74%
Especificidad	74%
ROC	68%
R^2	41%
Precisión / C. Global	74%

FUENTE: Elaboración propia con R-project.

4.3 EVALUACIÓN Y COMPARACIÓN DE CLASIFICADORES

Tener una clasificación global alta no garantiza que un modelo sea bueno y que clasifique correctamente las categorías, para este caso la minoritaria, fuga y la mayoritaria, no fuga. Para realmente identificar que el modelo es bueno, se debe acudir al análisis de sus indicadores de clasificación tanto globales, R^2, curva ROC, como parciales; sensibilidad y especificidad.

Ahora bien se contrastó los indicadores de clasificación obtenidos (Cuadro 24) de las 3 regresiones realizadas.

Cuadro 24: Comparación de modelos

	RL	RL SUB	RL SMOTE
Sensibilidad	0%	69%	74%
Especificidad	100%	58%	74%
ROC	64%	67%	68%
R^2	20%	32%	41%
Precisión / C. Global	91%	59%	74%

FUENTE: Elaboración propia con R-project.

Si bien es cierto que la precisión más alta es de la regresión logística sin balancear con un 91%, se descarta al tener poder de predicción solo sobre una categoría. Como se puede observar, la regresión logística con SMOTE, obtiene un mejor índice de sensibilidad, es decir predice mejor la categoría de fuga que la Regresión Logística con sub-muestreo aleatorio y la regresión sin balanceo. En cuanto a la curva ROC (Figura 17), el área bajo la curva, se observa que en algunos tramos la regresión logística con SMOTE tiene mayor sensibilidad en algunos tramos.

Además, también obtiene mejor índice de especificidad y clasificación global. Por ende se puede concluir que la regresión con SMOTE posee mejor desempeño como clasificador.

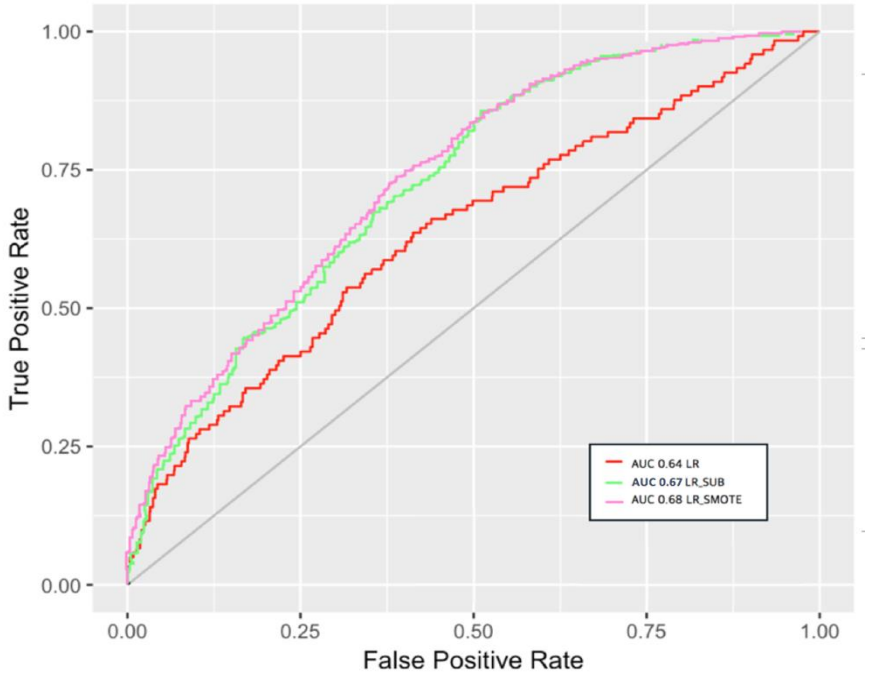


Figura 17: Comparación de áreas bajo la curva (AUC)

FUENTE: Elaboración Propia en R-project.

4.4 INTERPRETACIÓN DE RESULTADOS

- **Interpretación del mejor modelo.**

  Según los indicadores de clasificación obtenidos (Cuadro 24) : La regresión logística con SMOTE es el mejor modelo. Ahora bien se tienen las siguientes interpretaciones de los coeficientes obtenidos:

  - Al aumentar en una unidad la tasa de interés, disminuye la ventaja que el cliente fugue a que no fugue, específicamente en un 74,7%, manteniendo las demás variables constantes.
  - Al aumentar en una unidad la edad de cliente disminuye la ventaja que el cliente fugue a que no fugue, específicamente en un 3,6% manteniendo las demás variables constantes.
  - El ser divorciado, soltero o viudo, disminuye la ventaja que el cliente fugue a que no fugue, específicamente en un 39.5% respecto a ser casado o conviviente, manteniendo las demás variables constantes.
  - Al vivir en Lima o Callao, aumenta la ventaja que el cliente fugue a que no fugue, específicamente en un 46.9% respecto a vivir en el Centro, manteniendo las demás variables constantes.
  - Al vivir en el sur o norte del país, aumenta la ventaja de que el cliente fugue a que no fugue, específicamente en un 239% respecto a vivir en el Centro, manteniendo las demás variables constantes.
  - Al vivir en el oriente del país, aumenta la ventaja que el cliente fugue a que no fugue, específicamente en un 867% respecto a vivir en el Centro, manteniendo las demás variables constantes.
  - Al aumentar en una unidad el CrossSell (nivel de vinculación del cliente), aumenta la ventaja que el cliente fugue a que no fugue, específicamente en un 3,1% manteniendo las demás variables constantes.
  - Al aumentar en una unidad el ratio de antigüedad, aumenta la ventaja de que el cliente fugue a que no fugue, específicamente en un 50.7%, manteniendo las demás variables constantes.

- **Variables que influyen en la fuga del cliente con CTS**

  Se determinó las siguientes variables significativas en la fuga del cliente con CTS: tasa de interés, edad, estado civil, región donde vive, antigüedad del producto y nivel de vinculación con el banco que es medida por el número de productos que posee en la entidad financiera; CrossSell.

- **Prediccion de nuevos individuos.**

  Se realizó el proceso de predicción de nuevos clientes con la regresión logística binaria con sobre-muestreo aplicando el algoritmo de SMOTE. La regresion es la siguiente:

$$y = 1/(1 + \exp(-(7.70 - 1.37(tasaInteres) - 0.01(Saldo.Sol) - 0.04(Edad) - 0.93(E.Civil.Div.Sol.Viu) + 0.38(R.LIMA.CALLAO) + 1.22(R.NORTE.SUR) + 2.27(R.ORIENTE) + 0.03(CrossSell) - 0.71(ratio.ant))$$

  En el cuadro 25 se presenta 5 conjuntos de datos de clientes, con sus respectivas variables predictoras. Con el fin de clasificar a cada cliente como: *Fuga* o *No Fuga*. Se presentan las probabilidades, y la clasificación final.

Cuadro 25: Predicción de nuevos individuos

ID	Tasa Interes	Saldo.Sol	Edad	E.Civil	R.Lima. Callao	R.Norte. Sur	R.Oriente	Cross Sell	Ratio .ant	Probab	Predicción
C1	2.5	1,200	34	0	0	0	1	1	0.01	0.001	*No fuga*
C2	3.1	10,000	35	1	0	1	0	2	0.34	0.000	*No fuga*
C3	4.3	4,500	30	0	1	0	0	2	0.03	0.000	*No fuga*
C4	2.7	450	20	0	0	0	1	3	0.00	0.743	*Fuga*
C5	5.5	250	59	0	0	1	0	3	1.00	0.016	*No fuga*

FUENTE: Elaboración propia.

V. CONCLUSIONES

1. Según los indicadores de clasificacion mostrados; clasificación global, curva ROC sensibilidad y especificidad. La regresión logística binaria aplicando la técnica de balanceo: algoritmo de SMOTE, obtiene mejores resultados en la clasificacion de fuga de clientes en una entidad financiera que la técnica de sub-muestreo simple.

2. El indicador, el área bajo la curva ROC, de la regresión logística sin balanceo, la regresión con sub-muestreo aleatorio y la regresión logística con SMOTE son 68%, 67% y 64% respectivamente, teniendo un aceptable porcentaje de clasificación las dos primeras.

3. Contrastando los indicadores de clasificación parciales: la sensibilidad y especificidad, de la regresión logística sin balanceo, la regresión logística con sub-muestreo aleatorio y regresión logística con SMOTE; se concluye que la regresión logística con SMOTE es mejor, dado que obtiene los indicadores más altos; 74% en ambos casos.

4. El perfil del cliente que fuga se caracteriza por tener una tasa de interés baja, es joven, es soltero, vive en el norte, sur y selva, posee poca antigüedad con la CTS en el banco y tiene un indicador de CrossSell mayor a 1.

VI. RECOMENDACIONES

1. Se recomienda a otros investigadores que tengan interés en este tema, la investigación del algoritmo SMOTE y variantes con otros tipos de modelos, en los últimos años se ha adaptado la técnica en diferentes modelos de Machine Learning como: Random Forest, Adaboosting, Support Vector Machine, etc.

2. Se recomienda la aplicación del algoritmo SMOTE en distintos tipos de datos en otros ámbitos más comerciales como propensión, captación de clientes y la identificación de fraudes.

3. Por parte de la predicción de fuga de clientes en CTS, se recomienda dimensionar los datos y criterios con la Ley 30334 de la constitución peruana, que permite la disposición del excedente de 4 sueldos brutos. Dado que una necesidad de las entidades financieras en conseguir productos pasivos para poder reinvertirlos en productos activos, sería apropiado diseñar un modelo de predicción de retiro de disposición de CTS.

VII. REFERENCIAS BIBLIOGRÁFICAS

Antini, M. (2015): Dealing with imbalanced data: undersampling, oversampling and proper cross-validation. Disponible en http://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation.

Buttle, F. (2008): Customer Relationship Management. Butterworth Heinemann. Butterworth- Heinemann.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16(1): pags 321–357.

De Ullibarri (1998), Curvas ROC. Atención Primaria en la Red, 1998, vol. 5, pags. 229-235.

Debonis, J.N.; Balinski, E.W. y Allen, (2003): Value-based marketing for bottom-line success: 5 steps to creating customer value. An American Marketing Association Title Series. McGraw-Hill, 2003.

DeRouin, E., Brown, J., Fausett, L., & Schneider, M. (1991). Neural Network Training on Unequally Represented Classes. In Intellligent Engineering Systems Through Artificial Neural Networks, pags. 135–141.

Domingos, P. (1999). Metacost: A General Method for Making Classifiers Cost-sensitive. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pags. 155–164.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management., pags. 148–155.

Efron, B. The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association, v. 70, n. 352, pags. 892-898, 1975.

Fawcett,T.(2004), ROC Graphs: Notes and Practical Considerations for Researchers, consultado el 23 set 2016, disponible en http://web.archive.org/web/20151002215629/http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf.

García, J. (2012) : “Calibración Local de Predicciones Numéricas de Viento con Técnicas Estadísticas no Lineales (Downscaling Estadístico)”.
Gower, J. C. 1971. A general coefficient of similarity and some of its properties. En: Biometrics Vol. 27,no. 4; pags.857–871. Diponible en: http://www.scielo.org.co/scielo.php?script=sci_nlinks&ref=000142&pid=S0304-28472007000100004000022&lng=en.
Ha, T. M., & Bunke, H. (1997). Image processing methods for document image analysis. Handbook of Character Recognition and Document Image Analysis, pags 1-47.
Haddena, J; Tiwaria, A; Roya, R y Rutab, D (2007): “Computer assisted customer churn management: State-of-the-art and future trends. “Computers & Operations Research”, pags. 2902–2917.
He, Haibo, Yunqian, Ma. (2013) . Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons.
Hosmer, D. W., & Lemeshow, S. (2000). “Introduction to the logistic regression model. Applied Logistic Regression, Second Edition, pags 1-30.
Japkowicz, N. (2000). Learning from imbalanced data sets: a comparison of various strategies. In AAAI workshop on learning from imbalanced data sets (Vol. 68, pags. 10-15).
Jiang, L., Li, C., & Wang, S. (2014). Cost-sensitive Bayesian network classifiers. Pattern Recognition Letters, 45, pags 211-216.
Johnson, R. A., Wichern D., W. (2007). Applied Multivariate Statistical Analysis.
Klemper, P. (1987): The Competitiveness of Markets with Switching CostsA. The RAND Journal of Economics, pags. 138–150.
Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. En ICML Vol. 97, pp. 179-186.
Lewis, D., & Catlett, J. (1994). Heterogeneous Uncertainity Sampling for Supervised Learn- ing. In Proceedings of the Eleventh International Conference of Machine Learning, pags. 148–156.
Ling, C. X. and Li, C. (1998). Data mining for direct marketing: Problems and solutions. In KDD, volume 98, pags 73–79.
Lomax, S., & Vadera, S. (2013). A survey of cost-sensitive decision tree induction algorithms. ACM Computing Surveys (CSUR), pag16.
Londoño, G. , Lavalett, L., Galido, P., y Afanador, L. (2007) : Uso de métodos multivariantes para la agrupación de aislamientos de Colletotrichum spp con base en

características morfológicas y culturales. Rev. Fac. Nal. Agr. Medellín. 60, : pag 3671-3690. Disponible en http://rev-inv-ope.univ-paris1.fr/files/31310/31310-03.pdf.
Martínez, E. R., Herrera, et al. Edición de Conjuntos de Entrenamiento no Balanceados, haciendo uso de Operadores Genéticos y la Teoría de los Conjuntos Aproximados.
Mladeni, D., & Grobelnik, M. (1999). Feature Selection for Unbalanced Class Distribution and Naive Bayes. In Proceedings of the 16th International Conference on Machine Learning., pp. 258–267. Morgan Kaufmann.
Moreno, J., Rodríguez, D., Sicilia, M. A., Riquelme, J. C., & Ruiz, R. (2009). SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias. Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos. Disponible en : http://www.cc.uah.es/drg/adis2009/articles/adis-09-Moreno-ISMOTE.pdf.
Nelder, J. A., & Baker, R. J. (1972). Generalized linear models. Encyclopedia of statistical sciences.
Nicandro, C. R., Efrén, M. M., María Yaneli, A. A., Enrique, M. D. C. M., Héctor Gabriel, A. M., Nancy, P. C., ... & Rocío Erandi, B. M. (2013). Evaluation of the diagnostic power of thermography in breast cancer using bayesian network classifiers. Computational and mathematical methods in medicine, 2013.
Nieto, S. (2010). Crédito al Consumo: La estadística aplicada a un problema de riesgo crediticio. Tesis Maestra. México. UAM. 96 pp.
Peppers, D. y Rogers, M.(2011): Managing Customer Relationships: A Strategic Framework. John Wiley & Sons.
Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. Machine Learning, 42/3, pags 203–231.
Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. Operations research, 42(4), 589-613.
Solberg, A. H. S., Taxt, T., & Jain, A. K. (1996). A Markov random field model for classification of multisource satellite imagery. IEEE Transactions on Geoscience and Remote Sensing, 34(1), pags 100-113.
Solberg, A. S., & Solberg, R. (1996, May). A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images. Vol. 3, pags. 1484-1486.
Swets, Jhon. A. (1996). Signal detection theory and ROC analysis in psychology and diagnostics. Hillsdale, NJ: Lawrence Erlbaum.
Thompsom, H. (2004): Who stole my customer?: winning strategies for creating and sustaining customer loyalty. Pearson Prentice Hall.

Tianxiang, G (2015), “Hybrid classification approach of SMOTE and instance selection for imbalanced datasets”, pags 2-5.

Van Rijsbergen, C., Harper, D., & Porter, M. (1981). The Selection of Good Search Terms. Information Processing and Management, 17, pags 77–91.

Woods, K., Doss, C., Bowyer, K., Solka, J., Priebe, C., & Kegelmeyer, P. (1993). Compara- tive Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography. International Journal of Pattern Recognition and Artificial Intelligence, pags 1417–1436.

Zadrozny, et al. (2003, November). Cost-sensitive learning by cost-proportionate example weighting. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 435-442). IEEE.

Zhou, Z. H., & Liu, X. Y. (2006). “Training cost-sensitive neural networks with methods addressing the class imbalance problem”. IEEE Transactions on Knowledge and Data Engineering, 18(1), pags 63-77.

VIII. ANEXOS

Anexo 1: Detección de valores perdidos y eliminación de ellos.

```
> columnas_nas<-which(colSums(is.na(cts.f))!=0)
> columnas_nas
saldo_sol      Edad
        2         3
> per.miss.col=100*colSums(is.na(cts.f[,columnas_nas]))/dim(cts.f)[1]
> per.miss.col
 saldo_sol       Edad
0.04476944 0.01492315
> cts.cl=na.omit(cts.f)### se limpia los NA's
> columnas_nas<-which(colSums(is.na(cts.cl))!=0)
> columnas_nas ### se comprueba que no hay mas NA's
named integer(0)
```

Anexo 2: Detección de outlier's.

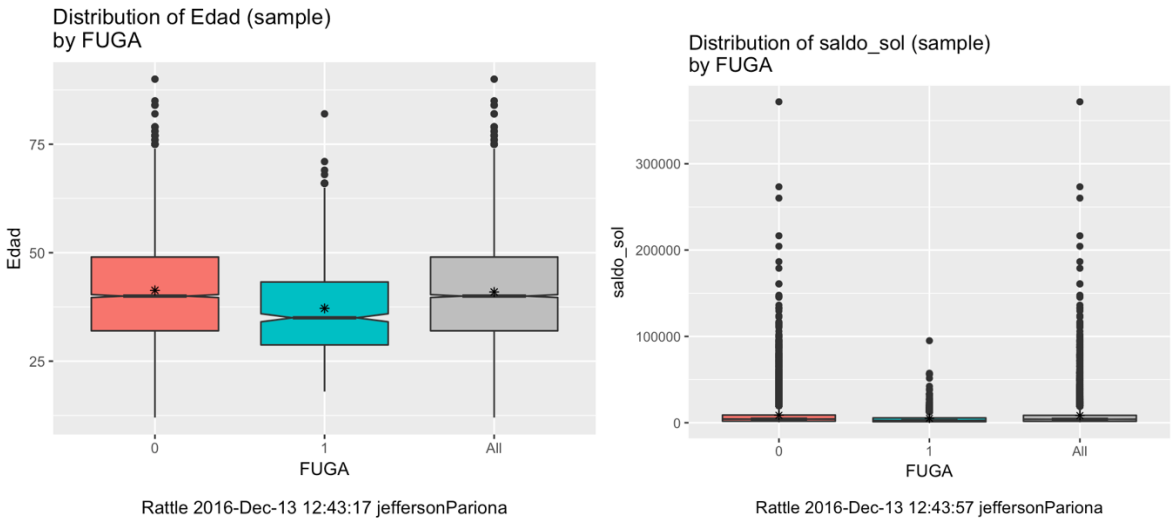
Se detectó outllier's en las variables saldo_sol y Edad

saldo_sol

```
       n   missing distinct     Info     Mean
   6,697         0     6,434        1     8,335
lowest :    500.1808    500.3555    501.0000    501.5106    501.7823
highest: 204423.4282 216585.0773 260199.6368 273310.5879 371783.9580
```

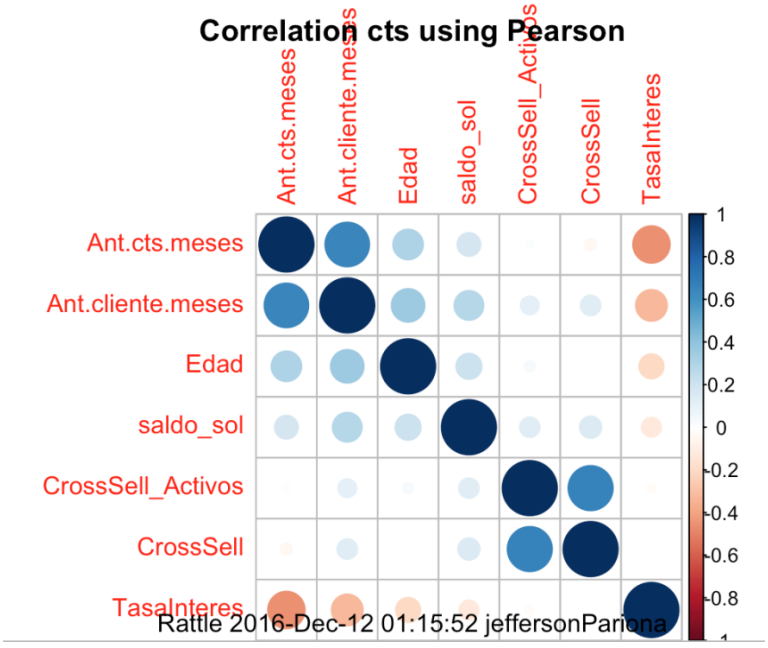
Edad

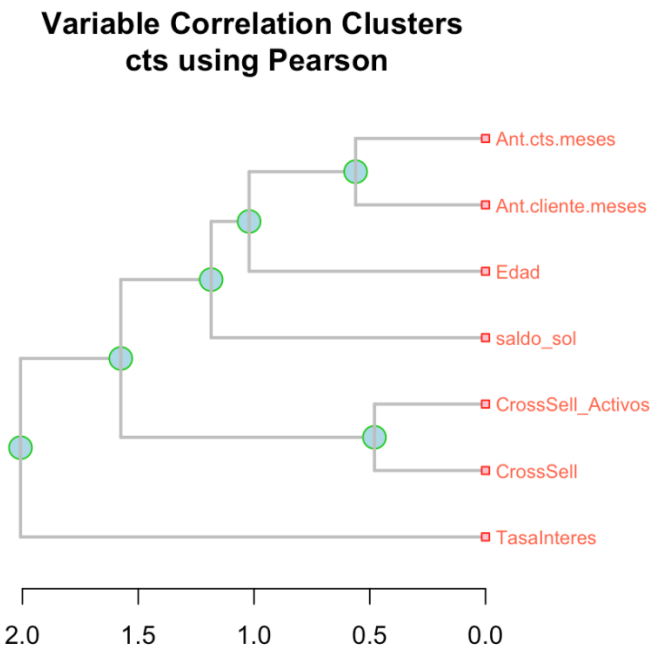
```
       n  missing distinct     Info     Mean
    6697        0       69    0.999    40.94
lowest : 12 15 16 18 19
highest: 79 82 84 85 90
```



Anexo 3: Detección de multicolinealidad (relación entre las variables predictoras).

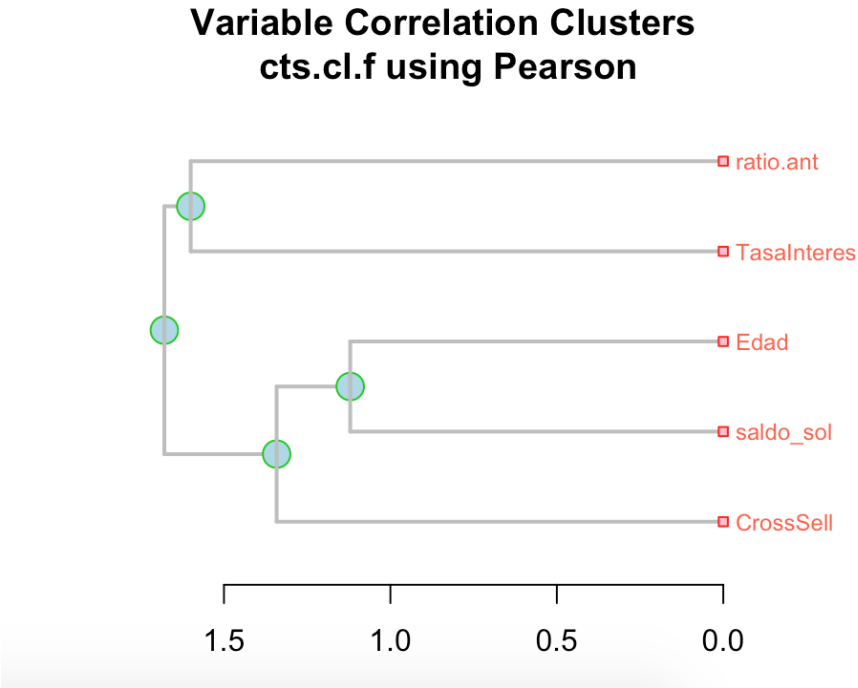
	Ant.cts.meses	Ant.cliente.meses	Edad	saldo_sol	CrossSell_Activos	CrossSell	TasaInteres
Ant.cts.meses	1.000	0.652	0.304	0.182	0.020	-0.044	-0.457
Ant.cliente.meses	0.652	1.000	0.366	0.288	0.112	0.134	-0.328
Edad	0.304	0.366	1.000	0.218	0.037	-0.007	-0.202
saldo_sol	0.182	0.288	0.218	1.000	0.133	0.158	-0.125
CrossSell_Activos	0.020	0.112	0.037	0.133	1.000	0.666	-0.024
CrossSell	-0.044	0.134	-0.007	0.158	0.666	1.000	-0.007
TasaInteres	-0.457	-0.328	-0.202	-0.125	-0.024	-0.007	1.000





Anexo 4: Transformación de variable y solución de multicolinealidad.

```
##transformacion de variables: fusion de 2 variables deudas
cts$ratio.deuda<-cts$Deu_TOTAL_BFP/cts$Deu_TOTAL_SBS
```



Anexo 5: Test de Chi-cuadrado para probar dependencia de las variables categóricas con la variables dependiente. (Influencia sobre Y).

```
> chisq.test(cts.cl$Sexo,cts.cl$FUGA)
      Pearson's Chi-squared test with Yates' continuity correction
data:  cts.cl$Sexo and cts.cl$FUGA
X-squared = 0.25571, df = 1, p-value = 0.6131
> chisq.test(cts.cl$EstadoCivil,cts.cl$FUGA)
      Pearson's Chi-squared test
data:  cts.cl$EstadoCivil and cts.cl$FUGA
X-squared = 7.7178, df = 4, p-value = 0.1025
> chisq.test(cts.cl$REGION,cts.cl$FUGA)
      Pearson's Chi-squared test
data:  cts.cl$REGION and cts.cl$FUGA
X-squared = 17.765, df = 3, p-value = 0.0004918
> chisq.test(cts.cl$FLAG_AHORROS,cts.cl$FUGA)
      Pearson's Chi-squared test with Yates' continuity correction
data:  cts.cl$FLAG_AHORROS and cts.cl$FUGA
X-squared = 6.2355, df = 1, p-value = 0.01252
> chisq.test(cts.cl$FLAG_CS,cts.cl$FUGA)
      Pearson's Chi-squared test with Yates' continuity correction
data:  cts.cl$FLAG_CS and cts.cl$FUGA
X-squared = 18.842, df = 1, p-value = 0.0000142
> chisq.test(cts.cl$Flag.Banc,cts.cl$FUGA)
      Pearson's Chi-squared test with Yates' continuity correction
data:  cts.cl$Flag.Banc and cts.cl$FUGA
X-squared = 0.11856, df = 1, p-value = 0.7306
```

Anexo 6: Selección de muestras: training y testing.

```
>###seleccion de variables de la base de datos final
> cts.cl.f<-cts.cl[,c(1,2,3,4,6,11,12,13)]
>###seleccion aleatoria del tamaño de las muestras con el paquete caret
> library(caret)
> train_size <- floor(0.80*nrow(cts.cl.f))
> set.seed(123)
>   train_cts.cl.f_MAS   <-   sample(seq_len(nrow(cts.cl.f)),   size   =
train_size)
>### se escoge el 80% de la muestra total para training
> train_MAS <- cts.cl.f[train_cts.cl.f_MAS,]
>### se escoge el complemento: 20% para la muestra de testing
> test_MAS <- cts.cl.f[-train_cts.cl.f_MAS,]
> prop.table(table(train_MAS$FUGA))*100

        0         1
90.946425  9.053575
> table(train_MAS$FUGA)

   0    1
4872  485
> prop.table(table(test_MAS$FUGA))*100

        0         1
91.119403  8.880597
> table(test_MAS$FUGA)

   0    1
1221  119
```

Anexo 7: Regresión logística sin balancear datos.

```
> table(train_MAS$FUGA)

   0    1
4872  485
> prop.table(table(train_MAS$FUGA))

   0    1
0.91 0.09

##MODELO DE REGRESION LOGISTICA SIN BALANCEAR

Call:
glm(formula = FUGA ~ ., family = binomial(link = "logit"), data =
crs$dataset[crs$train,
    c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8550  -0.4872  -0.3923  -0.2862   3.3572

Coefficients:
                    Estimate  Std. Error z value Pr(>|z|)
(Intercept)        4.05528125  0.93774890   4.324 1.53e-05 ***
TasaInteres       -1.14290721  0.16197749  -7.056 1.71e-12 ***
saldo_sol         -0.00003710  0.00000796  -4.660 3.16e-06 ***
Edad              -0.04052958  0.00489957  -8.272  < 2e-16 ***
REGIONLIMA_CALLAO  0.93199501  0.51807992   1.799   0.0720 .
REGIONNORTE.SUR    1.31082638  0.52149763   2.514   0.0120 *
REGIONORIENTE      1.32914357  0.68147871   1.950   0.0511 .
ratio.ant         -0.79782165  0.16807590  -4.747 2.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log likelihood: -1533.329 (8 df)
Null/Residual deviance difference: 178.757 (7 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.20363230
```

```
TEST
      Predicted
Actual    0
     0 1221

     1  119
```

```
> prop.table(table(training_mas$FUGA))*100
   0    1
50.3 49.7
> table(training_mas$FUGA)
  0   1
610 604
> table(test_MAS$FUGA)

   0    1
1221  119
```

Anexo 8: Regresión logística con sub-muestreo aleatorio.

```
Call:
glm(formula = FUGA ~ ., family = binomial(link = "logit"), data =
crs$dataset[crs$train,
    c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8350  -1.1001  -0.3716   1.0690   2.4157

Coefficients:
                           Estimate   Std. Error z value Pr(>|z|)
(Intercept)             5.263994681  1.062549486   4.954 7.27e-07 ***
TasaInteres            -1.047255640  0.174122299  -6.014 1.80e-09 ***
saldo_sol              -0.000031667  0.000008273  -3.828 0.000129 ***
Edad                   -0.039946367  0.005859875  -6.817 9.30e-12 ***
EstadoCivilDiv.Sol.Viu -0.144287617  0.153874914  -0.938 0.034840 .
REGIONLIMA_CALLAO       1.007026141  0.551569346   1.826 0.067888 .
REGIONNORTE.SUR         1.425586198  0.558949909   2.550 0.010758 *
REGIONORIENTE           1.384299397  0.825637522   1.677 0.093612 .
CrossSell               0.273482299  0.079215767   3.452 0.000556 ***
ratio.ant              -0.524100586  0.212273425  -2.469 0.013550 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log likelihood: -778.088 (10 df)
Null/Residual deviance difference: 126.756 (9 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.31776231
```

```
TEST
      Predicted
Actual    0  1
     0  711  510

     1   37   82
```

Anexo 9: Regresión logística con la aplicación de SMOTE.

```
> library(DMwR)
> train_S<-SMOTE(FUGA ~. , train_MAS, perc.over = 400, perc.under=100)
> prop.table(table(train_S$FUGA))*100
   0     1
44.4 55.56
> table(train_S$FUGA)
   0    1
1940 2425

##MODELO DE REGRESION LOGISTICA CON SMOTE
Call:
glm(formula = FUGA ~ ., family = binomial(link = "logit"), data =
crs$dataset[crs$train,
    c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0758  -0.9221  -0.5978   1.0447   3.0204

Coefficients:
                           Estimate   Std. Error z value Pr(>|z|)
(Intercept)             7.707387803  0.534081653  14.431  < 2e-16 ***
```

```
TasaInteres            -1.375616742  0.094050805 -14.626  < 2e-16 ***
saldo_sol              -0.010058715  0.000004933 -11.903  < 2e-16 ***
Edad                   -0.037118171  0.002852763 -13.011  < 2e-16 ***
EstadoCivilDiv.Sol.Viu -0.928359225  0.061858555 -15.008  < 2e-16 ***
REGIONLIMA_CALLAO       0.384692955  0.242803198   1.584    0.113
REGIONNORTE.SUR         1.220689972  0.244967525   4.983 6.26e-07 ***
REGIONORIENTE           2.269511294  0.368153333   6.165 7.07e-10 ***
CrossSell               0.030546053  0.039162637   0.780    0.0435.
ratio.ant              -0.707565195  0.104435099  -6.775 1.24e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Log likelihood: -3646.720 (10 df)
Null/Residual deviance difference: 1108.331 (9 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.41195492
```

```
TEST
      Predicted
Actual    0  1
     0  900  321

     1   31   88
```