UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA

PROYECTO DE TÍTULO PARA OPTAR AL GRADO DE
INGENIERÍA ESTADÍSTICA

COMPARACIÓN DE MÉTODOS ESTADÍSTICOS PARA LA DETECCIÓN DE FRAUDE EN CANALES NO PRESENCIALES APLICADOS AL ÁREA BANCARIA

Docente Patrocinante	:	Dr. Guillermo Ferreira	Firma	.....
Profesional Co-guía	:	Sr. Enrique Díaz Peñaloza	Firma	.....
Docente Colaborador:	:	Sr. Sebastian Niklitschek	Firma	.....
Docente Consejero:	:	Dra. Nora Serdyukova	Firma	.....
Nombre Memorante	:	Srta. Pilar Chávez Sandoval	Firma	.....

Concepción, Chile 2019

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Uno de los principales objetivos que persiguen las entidades empresariales de tipo comercial o financiera, es la maximización de utilidades y reducción de costos. Es por esto, que cuando se producen ciertos períodos de crisis, el enfoque primordial se orienta a las posibles causas y por ende, a los efectos que desencadena aquella crisis.

En concreto, a pesar que cualquier tipo de pérdida es importante para la entidad empresarial, la más significativa es la referente a los costos, porque es la que puede llevar a consecuencias desastrosas en términos de las utilidades de la empresa, atacando como tal, su principal objetivo.

Algunas de las causas que explican lo anterior, son los factores en términos sociales, económicos, políticos y tecnológicos. Este último, lo podemos encontrar en los sectores bancarios de las entidades financieras, las cuales, si bien la tecnología aportan a la mejora y al desarrollo en la entrega de bienes o servicios para el cliente, también proporcionan formas adicionales para cometer actos ilícitos. Un ejemplo de ello, deriva en lo que se conoce como **fraude**, el cual puede adoptar una variedad ilimitada de diferentes formas. Actualmente en el comercio electrónico, basta con tener información sobre la tarjeta para perpetruar un fraude.

Las consecuencias de aquellos actos ilícitos, se desenvuelven en pérdidas financieras que no solo afectan a los bancos, sino que también, a sus clientes, ya que si el banco pierde dinero, son los clientes quienes eventualmente pagan a través de intereses más altos, tarifas de membresías más altas, etc.. Adicionalmente, la entidad bancaria tiende a perder credibilidad afectando directa-

mente a su reputación como empresa.

Las acciones tomadas contra el fraude se pueden dividir en **prevención del fraude**, que es donde se intenta bloquear las transacciones fraudulentas; y la **detección del fraude**, que corresponde a las transacciones exitosas de fraude identificadas a posteriori.

Algunas de las tecnologías utilizadas para prevenir aquellos fraudes son los sistemas de verificación de dirección (Address Verification System (AVS)), donde se corrobora el código postal del cliente; verificación de tarjeta (Cardholder Verification Method, (CVM)) y el número de identificación personal (Personal Identification Number, (PIN)), donde ambas implican la comprobación del código numérico que es ingresado por el cliente (Dal, 2015).

Un sistema de detección de fraude debe ser tanto rentable y eficiente, es decir, el costo invertido en este sistema no debe superar las pérdidas por fraudes, por ende, es importante considerar la utilización de modelos y reglas estadísticas para minimizar aquellos costos.

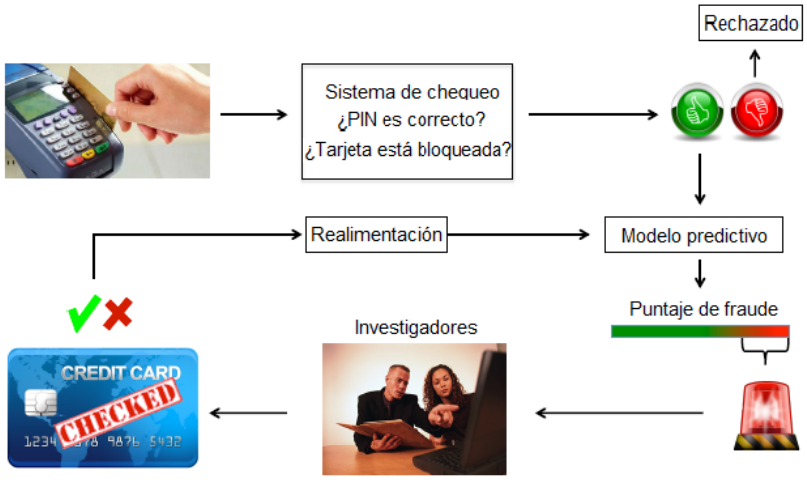

Figura 1.1: Proceso en la detección de fraude con tarjeta de crédito
(Fuente: Dal, 2015)

La Figura 1.1, representa el proceso en la detección de fraude con tarjeta de crédito:

1. Inicialmente, el usuario hace uso de la tarjeta de crédito por medio de un datáfono, dispositivo compacto que permite cobrar por red telefónica o IP vía GSM, GPRS, Wi-Fi, etc. y que a su vez, utiliza un sistema de chequeo corroborando por ejemplo la contraseña (PIN), la disponibilidad de la tarjeta (bloqueada o no), si cuenta con saldo suficiente, entre otras.

2. Al verificar estas condiciones, el sistema decide si rechazar o no la transacción. En el caso de ser aceptada, aquella transacción se filtra y luego se califica mediante un modelo predictivo; en caso contrario, queda rechazada inmediatamente.

3. El modelo predictivo califica cada transacción con alto o bajo riesgo de fraude, entregando alertas cuando se presente un nivel alto de riesgo de fraude.

4. Posteriormente, los investigadores o analistas se encargan de verificar aquellas alertas catalogándolas si son verdaderos positivos (efectivamente son fraudes) o falsos positivos (las transacciones son normales).

Los investigadores clasifican a las transacciones realizadas por el usuario en 4 escenarios que por temas de confidencialidad las designaremos como A, B ,C y D, dando sus respectivas descripciones en los siguientes párrafos.

La primera, hace referencia a las transacciones realizadas por el usuario de manera presencial en comercios correspondientes al país de origen al que pertenezca el cliente; por ende, los fraudes que se pueden dar son mediante: el hurto de la tarjeta, la clonación de la tarjeta ó bajo la modalidad de lo que se denomina el "cuento de tío".

Cabe destacar que el tipo de fraude mayormente presentado en este cuadrante entre los usuarios, es el del hurto de la tarjeta, donde el estafador intenta gastar lo más rápido la mayor cantidad posible de recursos incorporados en ella; por tanto, la detección se basa en el uso inesperado de la tarjeta de crédito comparado con su práctica común.

El segundo punto, es referente a las transacciones que se realizan de manera no presencial en comercios pertenecientes al país de

origen del usuario, es decir, aquellas transacciones hechas a través de internet; lo cual, invita a realizar fraudes vía estafas telefónicas ó phishing. Este último término se define por Valle (2013) como: "una técnica de ingeniería social utilizada por los delincuentes para obtener información confidencial como nombres de usuario, contraseñas y detalles de tarjetas de crédito haciéndose pasar por una comunicación confiable y legítima"(p. 30).

Un ejemplo de esto, se asocia a la capacidad de duplicar páginas web, a modo de llevar a cabo el engaño a través de correos electrónicos que generalmente contienen enlaces de sitios web falsos (correos electrónicos maliciosos) con apariencia casi idénticas a los sitios legítimos.

El tercer punto, infiere a las transacciones realizadas de forma presencial en comercios que se ubican fuera del país de origen del usuario, cumpliendo los mismo posibles escenarios de fraude que en el cuadrante A.

El cuarto punto indica sobre aquellas transacciones realizadas por el usuario de manera no presencial en comercios fuera de su país de origen, lo que hace más probable cometer fraudes mediante pruebas masivas robotizadas respecto a los números de serie de las tarjetas u otros.

Por ende, en cualquiera de los cuadrantes mencionados anteriormente el fraude transaccional, es la acción que en consecuencia atrae a quienes acceden a cierto canal con el propósito de obtener información personal para cometer un hecho ilícito, siendo de vital importancia su rápida detección pues el usuario generalmente no se encuentra en conocimiento del robo y uso de sus datos.

Dentro del contexto de la detección de fraude, el uso de técnicas de aprendizaje permiten descubrir patrones en flujos de datos de alta dimensión, considera que las transacciones fraudulentas generalmente se correlacionan tanto en el tiempo como en el espacio y además, pueden detectar y modelar estrategias fraudulentas existentes, identificando nuevas estrategias asociadas al comportamiento inusual de las tarjetas de crédito (Dal, 2015).

Una encuesta realizada en Colombia, aplicada a 144 directivos de empresas perteneciente el 14 % a la Industria de Bancos y Servicios Financieros y el 8 % a Recursos Naturales y Energía,

operantes en dicho país; afirmaron haber experimentado algún tipo de fraude durante el año 2014 y 2015. En particular, la encuesta cataloga una transacción fraudulenta como cibercrimen y la tipifica en dos aspectos: de acceso no autorizado y de piratería, los cuales corresponden a una cifra indicadora del 4 %, con respecto a las demás categorías de fraude (KMPG,2017).

Dicho esto, existen dos formas de automatizar los inconvenientes en un sistema de detección de fraude y que son:

- Métodos controlados basados en los datos y con un tipo de aprendizaje supervisado o no supervisado.
- Métodos controlados por expertos.

A pesar que su combinación y funcionamiento en paralelo entregarían la "mejor" solución, generalmente hoy en día se utilizan los métodos basados en los datos, ya que responden a la obtención de alertas precisas (reducción de falsos positivos), mientras que los métodos basados en expertos, sólo garantizan que se detecte la totalidad de los fraudes (reducción de falsos negativos), a cambio de obtener pocas alertas falsas (Dal, 2015).

En efecto, para un sistema de detección de fraude basada en los datos se presentan algunos de los siguientes desafíos:

1. El número de fraudes representa una pequeña fracción de todas las transacciones diarias (Dal et al., 2014), es decir, existe un desequilibrio en este tipo de datos.
2. La distribución de fraude evoluciona a través del tiempo debido a la estacionalidad de los datos y a los diseños de nuevas estrategias de ataque (Dal et al., 2014, julio).
3. La verdadera clase a la que pertenecen la mayoría de las transacciones se da a conocer al tiempo después de efectuar la transacción, dado que los analistas tienden a verificar de manera oportuna pocas transacciones (Dal et al., 2015, julio).
4. La carga y el tiempo computacional que se requiere para el procesamiento de los datos.

En consecuencia, como la tarea de detección de fraude no es fácil de resolver, la finalidad de este proyecto de título es proponer

herramientas estadísticas que apuntan inicialmente a la utilización de la técnica de Random Forest y un modelo de Regresión Logística, para realizar una comparación en términos de su funcionalidad y mejora en la reducción de los falsos positivos de transacciones fraudulentas internacionales no presenciales mediante tarjetas de crédito aplicados a datos reales.

La estructura de este documento, inicia con una breve introducción que involucra la contextualización del tema, seguido por el marco teórico en el cual se desarrollan los conceptos claves y las técnicas estadísticas a utilizar. Luego se realiza el planteamiento del problema, para posteriormente definir los objetivos específicos y general del estudio, terminando con los respectivos resultados.

Capítulo 2

Marco Teórico

Un tema recurrente y preocupante que afecta en cierto grado a todos los países, tanto a nivel usuario como empresarial, son los llamados fraudes.

Entonces, como primer punto y para el entendimiento del concepto, según Estupiñan (2006) la definición del término **fraude** es: "Acto intencional, por parte de uno o más individuos del área de administración, personal o terceros, que produce una distorsión en los estados financieros, éste puede involucrar: la manipulación, falsificación o manipulación de registros o documentos, el uso indebido de recursos, entre otras" (p. 324).

De acuerdo a lo anterior, en esta investigación nos interesa el concepto que apunte a los *delitos económicos* definidos por la consultora KPMG (2013) como: "Aquellas actividades ilícitas y de carácter patrimonial que se realizan en perjuicio de una compañía afectando o dañando sus activos, capital social o cualquier otro derecho o bien del que sea propietario" (p. 13).

En consecuencia, existen distintas clasificaciones de los principales comportamientos ilícitos tipificados dentro del fraude siendo:

- Malversación de activos
- Fraude financiero
- Corrupción
- Fraude transaccional

Generalmente, cuando ocurre un fraude en la actividad financiera del tipo transaccional, de inmediato se habla de la participación

de *tarjetas de crédito bancarias*, la cual define Sandoval (1991) como:

> Aquellas en las que un banco o institución financiera asume el rol de emisor y concede el crédito al usuario. Entre el banco y el usuario existe una línea de crédito rotatorio en cuanto a que, una vez utilizado el abono parcial o total que se efectúe origina una nueva disponibilidad en favor del titular de la tarjeta.(p. 17)

Las tarjetas de crédito, han sido parte importante para el crecimiento comercial global de los países con respecto a sus economías, pues ha permitido que la tecnología aporte por ejemplo, en la aplicación de sistemas de transferencia de dinero en línea (transacciones), incorporando un mayor número de consumidores en compra y venta de productos o servicios, constribuyendo en la expansión del comercio electrónico.

Actualmente, a pesar de los beneficios que entrega este medio de pago alternativo, también presenta desventajas en cuanto al manejo de la información debido al abuso tecnológico que se realiza. Una de ellas, es la usurpación mediante correos maliciosos (phishing), donde un programa se encarga de extraer la información personal y corporativa para obtener claves a los accesos de cuentas bancarias y/o también, por el mal manejo de la información personal, desencadenando lo que se denomina **fraude transaccional**.

Una vez, identificada la existencia de fraude transaccional, quienes se ven perjudicadas son las entidades bancarias, ya que en ellas, recae la responsabilidad de dar solución a sus clientes, lo que se traduce en pérdidas importantes de dinero. Así y de manera natural, es evidente que las entidades financieras deseen reducir sus pérdidas provenientes de fraudes, destinando esta misión a los analistas de riesgo, quienes son profesionales especializados en reducir las pérdidas por fraude y los encargados de desalentar a los defraudadores (Rodríguez et al.,2006).

A nivel mundial, la rápida evolución de este tema a generado la utilización de la ciencia de la estadística mediante herramientas de la minería de datos y del machine learning; y también del uso del método de regresión logística, entre otras.

2.1. Detección de fraude vía tarjetas de crédito

La detección de fraude con tarjetas de crédito se basa en el análisis de transacciones registradas, donde los datos de la transacción se componen principalmente de una serie de atributos como la identificación de una tarjeta, la fecha de transacción, el destinatario, el monto de la transacción, entre otras.

En un sistema de detección de fraude se presentan los métodos controlados basados en los datos (**Data Driven**) y los métodos controlados por expertos (**Expert Driven**) (Dal, 2015).

2.1.1. Data Driven

Los métodos controlados mediante los datos, configuran un sistema de detección de fraude con la finalidad de descubrir qué patrones se encuentran más relacionados con un comportamiento fraudulento.

Dentro de las ventajas destacan:

- Aprender de configuraciones fraudulentas complejas.
- Manejo de un gran volumen de datos.
- Predecir nuevos tipos de fraude.
- Adaptación al cambio en la distribución cuando el fraude evoluciona.

Respecto a sus desventajas, el primer método requiere de una cierta cantidad suficiente de muestras y a veces considera modelos de caja negra, el cual lo vuelve poco interpretable.

2.1.2. Expert Driven

Este método, utiliza el conocimiento de dominio de los investigadores de fraude para definir reglas que se realizan en la predicción de la probabilidad que una nueva transacción sea fraudulenta.

Además, se puede definir un conjunto de reglas expertas para diferentes escenarios, a modo de distinguir entre reglas de puntuación y reglas de bloqueo. El primero, asigna una puntuación

a una transacción en función del riesgo que los investigadores asocian a un determinado patrón, donde este último bloquea la transacción si el riesgo de fraude es demasiado alto.

Las ventajas del método son:

- Fácil de desarrollar y comprender.
- Explica la generación de las alertas.
- Explota el conocimiento de dominio del investigador.

Y sus desventajas son:

- Subjetivo.
- Detecta sólo correlaciones sencillas entre las variables y los fraudes.
- Detecta sólo estrategias fraudulentas conocidas.
- Requiere de supervisión humana constante.
- Se vuelven obsoletas rápidamente.

Además de lo anterior, es importante destacar que existen varios inconvenientes asociados a la detección de fraude transaccional presentado mayormente en los Data Driven, tales como:

- Clases desbalanceadas, donde la clase mayor está compuesta por transacciones normales siendo superior en número con respecto a la clase menor (fraude).
- Patrones de fraude cambiantes en el tiempo, es decir, se presenta heterogeneidad en el comportamiento a medida que transcurre el tiempo.
- Escasez de datos reales.
- Falencias en los métodos estadísticos cuando se cuenta con un desequilibrio en los datos.
- Tiempo y carga computacional para el procesamientos de las bases de datos.

2.2. Clasificación de datos no equilibrados

2.2.1. Clasificación

Hoy en día, se sabe que todo tipo de persona es propensa a cometer errores al querer establecer posibles relaciones entre múltiples características o al realizar cualquier tipo de análisis que involucre la toma de decisiones, haciendo difícil encontrar soluciones a los problemas abordados. Es por ello, que el aprendizaje automático a menudo se puede aplicar con éxito a estos problemas, mejorando la eficiencia de los sistemas y los diseños de máquinas (Kotsiantis et al., 2007).

Estos algoritmos de aprendizaje automático utilizan cualquier conjunto de datos o características, de las cuales pueden ser continuas, categóricas o binarias. Si aquél conjunto es de etiqueta conocida, el aprendizaje es de carácter *supervisado*; en caso contrario, si aquellas no se encuentran etiquetadas, el aprendizaje es del tipo *no supervisado*.

Respecto al aprendizaje supervisado, el objetivo es crear una función que pueda predecir la salida correspondiente a cualquier entrada válida después de haber sido sometido a una serie de datos ejemplos (datos de entrenamiento), es decir, el sistema generaliza con base en los datos que no ha visualizado. Este tipo de aprendizaje soluciona principalmente tareas de regresión donde la salida son valores numéricos, y de clasificación donde la salida es una etiqueta de clase (Keider, 2019).

El problema que se presenta para el aprendizaje supervisado de clasificación, es el de poder definir una regla para identificar a cúal clase corresponde una nueva observación, basado en el conjunto de características de entrenamiento cuyas categorías son conocidas. Este tipo de algoritmo permite representar la información adecuadamente para la toma de decisiones.

2.2.2. Desequilibrio de clases

El desbalanceo de datos es un problema frecuente y relativamente nuevo en muchas aplicaciones de aprendizaje automático y cuyos efectos sobre el desempeño de clasificadores estándar son notables.

Esta clase de problema también es precoz en la literatura de aprendizaje automático y minería de datos, sin embargo, es un tema de creciente interés debido a sus efectos sobre los resultados obtenidos y el número de aplicaciones en donde se puede encontrar esta situación.

Se puede definir un conjunto de datos desbalanceados, como aquellos que presentan una desproporción notable en el número de instancias pertenecientes a cada clase. Ello provoca un sesgo en el desempeño de los clasificadores estándares hacia el reconocimiento de las clases más numerosas, en perjuicio de las más raras (Kotsiantis et al., 2006).

Por lo general, sin considerar el problema del desbalance de clases, un algoritmo de clasificación tenderá a predecir que las muestras desconocidas pertenecen a la clase mayoritaria e ignorará completamente la clase minoritaria. Sin embargo, en muchas de las aplicaciones, la clase minoritaria es de vital importancia.
En consecuencia, en muchos algoritmos estándar de aprendizaje, se encuentran clasificadores que proveen un grado severamente desequilibrado de clases, donde la clase mayoritaria bordea casi el 90 % y la minoritaria oscila del 0 % al 10 % (Keider, 2019).

A continuación la Figura 2.1, entregaría un ejemplo intuitivo relacionado con la detección de fraude según lo mencionado anteriormente:

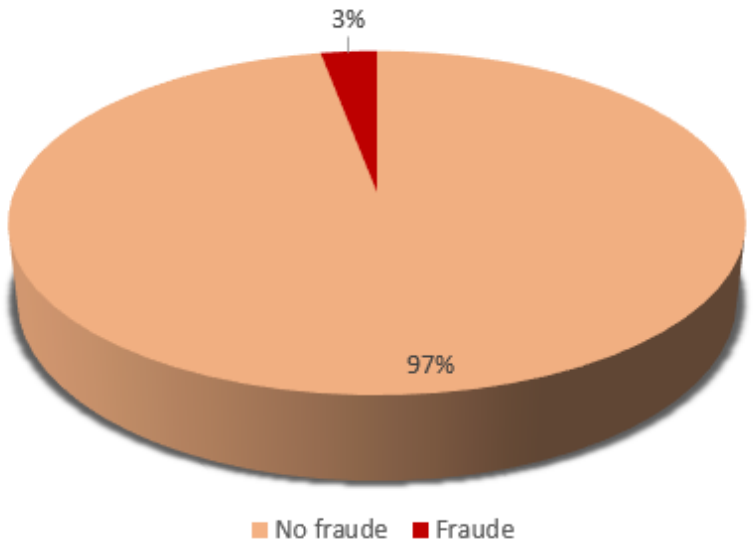

Figura 2.1: Desequilibrio de datos en la detección de fraude (Fuente: elaboración propia)

Otros ejemplos, donde se puede observar la prevalencia de datos desbalanceados son en aplicaciones tales como: el manejo de riesgo, la clasificación de texto, la detección de fallas en procesos industriales, el diagnóstico y monitoreo médico, entre otras (Chawla et al., 2004).

2.2.3. Tácticas para enfrentar el desequilibrio de clases

Como se tiene conocimiento que el desequilibrio de clases afecta de manera considerable a diferentes clasificadores, es que se han desarrollado hoy en día diferentes técnicas o metodologías que tratan este problema tales como:

1. Método de ensamblaje
2. Técnicas de re-muestreo y muestras sintéticas
3. Clustering

2.2.3.1. Método de ensamblaje

A diferencia que los enfoques de aprendizaje ordinarios intentan construir sólo un método de aprendizaje a partir de los datos de entrenamiento, los métodos de conjunto o ensamblajes intentan

construir una agrupación de métodos de aprendizajes y combinarlos (Zhou, 2012).

Por tanto, el Método de ensamblaje o en inglés *Ensemble methods* se utilizan comúnmente para aumentar la precisión predictiva, ya que combina las predicciones de múltiples modelos de aprendizaje automático.

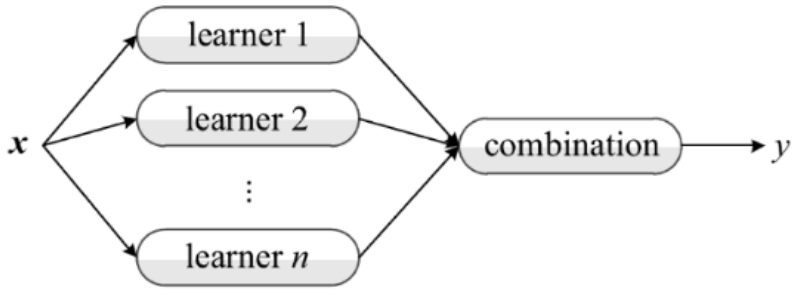

Figura 2.2: Arquitectura de agrupamiento común
(Fuente: Zhou, 2012)

La Figura 2.2 muestra la arquitectura de agrupamiento común, es decir, se tiene un conjunto contenido por una cantidad de aprendices llamados principiantes básicos que generalmente son generados a partir de datos de entrenamiento por un algoritmo de aprendizaje base que puede derivar de un árbol de decisión, red neuronal, random forest u otros tipos de algoritmos de aprendizaje (Zhou, 2012).

Algunas técnicas de modelado de agrupamiento básicas se encuentra el *promedio ponderado*, el cual consiste en aplicar ponderaciones en las predicciones y luego tomar el promedio de las predicciones de los modelos base haciendo que contribuyan de manera diferente a la predicción promediada, dependiendo de su desempeño. Otro método es el de *mayoría de votos*, que consiste en considerar la predicción con el máximo de votos de todos los algoritmos base, aplicándose directamente al problema de clasificación. Y por último, el método de *apilamiento*, que apunta a construir varias capas de modelos, donde la salida de estos en las capas inferiores se usa como entrada para otra capa superior, tomando esta última la decisión final (Barrero, 2018, p.3).

2.2.3.2. Técnicas de remuestreo y muestras sintéticas

Según Keider (2019) dice que: "Las técnicas de remuestreo se utilizan para equilibrar el espacio muestral para un conjunto de datos desequilibrado con el fin de aliviar el efecto de la distribución sesgada de clase en el proceso de aprendizaje"(p.10). Los métodos de remuestreo son más versátiles porque son independientes del clasificador seleccionado (López et al., 2013).

Es por ello, que estas técnicas se basan iniciamente en dos aspectos fundamentales:

- Agregar patrones
- Quitar patrones

de los cuales se destacan 3 grupos dependiendo del método utilizado para el equilibrio de clases, siendo el:

1. Método de sobremuestreo (oversampling)

   Consiste en la creación de nuevas muestras relacionadas con la clase minoritaria, es decir, se realiza un sobremuestreo que deriva en la incorporación de patrones para igualar la muestra.

2. Método de submuestreo (undersampling)

   Al contrario del oversampling, este método consiste en descartar muestras de la clase mayoritaria respecto a algún criterio o regla definida, implicando la eliminación de patrones para emparejar la muestra.

3. Método Híbrido

   Este método involucra la combinación entre los métodos 1 y 2.

En relación al lenguaje de programación, uno de los software que proporciona un amplio abanico de herramientas estadísticas y gráficas es el denominado **R-project**, el cual para efectos del manejo de datos desbalanceados cuenta con un paquete que lleva por nombre `unbalanced` o paquete desbalanceado, que incorpora las técnicas mencionadas anteriormente.

Algunas de las funciones que engloba este paquete son por ejemplo: el submuestreo aleatorio (`ubUnder`), el sobremuestreo (`ubOver`)

(Drummond et al., 2003); basados en la **distancia**, el OSS (`ubOSS`) (Kubat y Matwin, 1997), el ENN (`ubENN`) (Wilson, 1972), el CNN (`ubCNN`) (Hart, 1968), el NCL (`ubNCL`) (Laurikkala, 2001) y el Tomek Link (`ubTomek`) (Tomek, 1976).

Cabe destacar, que estos métodos pueden ser llamados por una función contenedora `ubBalance`, el cual permite probar todas estas estrategias cambiando solo el tipo de argumento (Dal et al., 2015).

2.2.3.3. Ejemplo

A continuación, se presenta un ejemplo para ilustrar lo anterior. Para ello se considera el conjunto de datos denominado *ubIonosphere*, la cual corresponde a una modificación del conjunto de datos *Ionosphere*, ambos disponibles en el paquete `mlbench` del software R-project.

La base en cuestión considera solo variables del tipo numéricas. Además de la variable dependiente llamada **Clase**, la que originalmente correspondía a una variable dicotómica que tomaba valores malos y buenos, pero que fue recodificada con 1 y 0. De esta manera, 1 representa la clase minoritaria (malos) y 0 a la clase mayoritaria (buenos). Esta variable se ubica al final de la base de datos.

Aplicación del sobremuestreo

Con el objetivo de equilibrar el conjunto de datos de *Ionosphere*, aplicamos la técnica del sobremuestreo.

	Variable dependiente	
	Original	Resultado
0	225	225
1	126	225

Cuadro 2.1: Comparación muestral
(Fuente: elaboración propia)

En el Cuadro 2.1, se observa un sobremuestreo a partir de los datos originales, es decir, se presenta un aumento de los datos pertenecientes a la clase minoritaria (transacciones fraudulentas)

a tal punto de alcanzar el número total de transacciones normales.

Aplicación del submuestreo

Por otro lado, si aplicamos la técnica del submuestreo nos entrega lo siguiente:

	Variable dependiente	
	Original	Resultado
0	225	126
1	126	126

Cuadro 2.2: Comparación muestral
(Fuente: elaboración propia)

A diferencia del caso observado de la aplicación anterior, el Cuadro 2.2 presenta un submuestreo, es decir, el método elimina observaciones pertenecientes a la clase mayoritaria (transacciones normales) para alcanzar el número total de transacciones fraudulentas.

2.2.3.4. Clustering

Es una de las técnicas del aprendizaje no supervisado en la que no se requiere una clasificación predefinida. El objetivo es poder particionar los datos obteniendo el conocimiento de acuerdo a las características de los mismos. Generalmente, las clases de los datos no se presentan en el conjunto de datos y los objetos se agrupan basándose en el principio de maximización de similitud dentro de los clusters y minimización de similitud entre clusters diferentes (Tan et al., 2006).

El análisis de Clustering parte de un conjunto de objetos que se caracterizan a través de varias variables con la finalidad de obtener grupos que sean homogéneos entre sí y heterogéneos entre ellos, es decir, según Hernández (2006): "en términos de variabilidad hablaríamos de minimizar la variabilidad dentro de los grupos para al mismo tiempo maximizar la variabilidad entre los distintos grupos"(p.29).
Dentro de las técnicas de Clustering, existen algoritmos que pueden ser clasificados en función del tipo de dato que manejan (numérico, categórico y/o mixto), el criterio utilizado para medir

la similitud entre los puntos y los conceptos y técnicas de clustering empleadas se puede encontrar en lógica difusa, estadísticas, entre otras (Hernández, 2006).

A continuación, se dará un pequeño resumen de los algoritmos asociados en relación al tipo de dato que se tiene, siendo para:

- Datos numéricos: K-means
- Datos categóricos: K-mode
- Datos mixtos: K-prototypes

Dicho lo anterior, para enfrentar el desbalance de clases se propondrá realizar un análisis de clustering al grupo de la clase mayoritaria con la finalidad de obtener aquellos clusters más representativos, para luego realizar un muestreo de aquellos clusters significativos obtenidos, a modo de obtener una muestra de igual número con respecto a la clase minoritaria, realizando de esta manera un equilibrio entre ambas clases.

2.2.4. Comportamiento del cliente

La distribución de los fraudes van evolucionando a medida que transcurre el tiempo, puesto a los cambios en las actividades fraudulentas y principalmente, al **comportamiento del cliente**.

El término comportamiento del cliente según Schiffman y Kanuk (2005) se define como: "el comportamiento que los consumidores muestran al buscar, comprar, utilizar, evaluar y desechar los productos y servicios que, consideran, satisfacerán sus necesidades"(p.8). Por ejemplo, la revolución digital se ha encargado de provocar varios cambios significativos en el rubro de los negocios, de los cuales, destacan a un consumidor con un mayor empoderamiento y una mayor accesibilidad a la información, comparada con hace tiempo atrás.

Es por ello, que el principal responsable de la no estacionariedad en el flujo de transacciones se debe al comportamiento variable que presentan los clientes en ciertos periodos de tiempo, siendo un tema relevante en cuanto a las actualizaciones constantes a los que deben ser sometidos los sistemas de detección de fraude. Las estrategias de aquellos sistemas que no se actualizan o revisan con frecuencia, a menudo pierden su precisión predictiva a largo plazo (Dal et al., 2014).

2.3. Técnicas estadísticas

2.3.1. Regresión logística

El análisis de regresión es una técnica estadística que permite indagar en las relaciones funcionales entre variables, con el objetivo de predecir o estimar el valor de una variable a partir del valor dado de otra variable.

De acuerdo al número y a la naturaleza de las variables en juego , existen distintos tipos de análisis de regresión, las cuales por ejemplo, si la relación involucra una variable independiente (X), se dice que la regresión es simple; y si hay varias, sería de carácter múltiple (Silva, 2004).

Entonces, si se cuenta con una variable respuesta (Y) de 2 niveles y se utilizan una o más variables explicativas para predecir aquella de tipo binaria, inferimos que el tipo de regresión presente es la denominada **Regresión Logística**.

En situaciones como esta, la variable endógena Y refleja la ocurrencia o no de un suceso, es decir, puesto que Y es dicotómica, ella asume los dos siguientes valores:

$$Y = \begin{cases} 0, & \text{el hecho no ocurre} \\ 1, & \text{el hecho ocurre} \end{cases}$$

La situación más simple, es aquella donde se trata de evaluar el efecto de un solo factor, al que se representará mediante la variable exógena X sobre el desenlace de la variable Y.

2.3.1.1. Transformación logística

Para efectos de procurar que se verifique la presentada relación sigmoidal entre el riesgo y los niveles de exposición, se realiza una transformación logística de la probabilidad p_i basada en que un cierto suceso ocurra.

Luego, para múltiples variables predictoras el tipo de modelo lineal generalizado queda expresado como:

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1(x_i)_1 + ... + \beta_p(x_i)_p \qquad (2.1)$$

2.3.1.2. Modelo de regresión logística

El modelo logístico para el caso general toma la siguiente forma:

$$p_i = logit^{-1}(\beta_0+\beta_1(x_i)_1+...+\beta_p(x_i)_p) = \frac{1}{1+e^{-(\beta_0+\beta_1(x_i)_1+...+\beta_p(x_i)_p)}} \tag{2.2}$$

donde p_i representa la probabilidad de que el individuo i desarrolle la característica de interés, x_i definen las variables explicativas independientes, β_0,...,β_p son los parámetros (en principio desconocidos) del modelo. A esta ecuación se le denomina **función logística** y es equivalente a la presentada en la ecuación (2.1).

2.3.1.3. Ejemplo

A continuación, un ejemplo del caso simple descrito anteriormente ocurre cuando es cuando se quiere modelar la probabilidad de fraude por impago (*default*) en función del balance de la cuenta bancaria llamada *balance* de tipo cuantitativa, Amat (2016), planteó su modelo ajustado como:

$$\widehat{default} = -10.65 + 0.005499(\widehat{balance})$$

Luego, se representan los valores de las probabilidades de cometer fraude por impago en función de los valores del balance en un sistema de ejes cartesianos, donde se puede comprobar que la expresión gráfica del modelo cuando la pendiente es positiva ($\beta_1 = 0.005499$) se describe como muestra en la Figura 2.1.

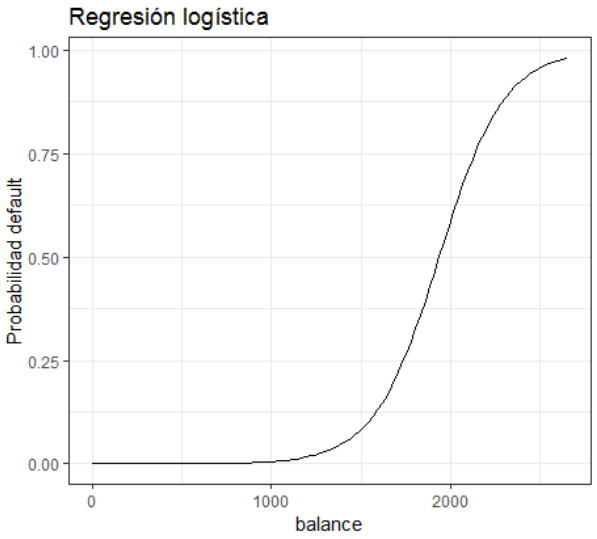

Figura 2.3: Función logística univariada $\beta_1 > 0$
(Fuente: elaboración propia)

2.3.2. Random Forest

La técnica estadística Random Forest (RF), es un algoritmo que corresponde a un tipo de aprendizaje automático supervisado, es decir, el algoritmo es capaz de aprender sobre un conjunto de características etiquetadas con la finalidad de realizar predicciones basadas en patrones aprendidos.

Esta técnica además se basa en los llamados **árboles de decisión**, que vienen siendo otro tipo de aprendizaje automático y que sirven igualmente para la creación de modelos predictivos.

2.3.2.1. Árboles de decisión

Los métodos basados en árboles para la regresión y la clasificación, implican estratificar o segmentar el espacio predictor en varias regiones simples. Entonces, dado que el conjunto de reglas de división utilizadas para la segmentación del espacio predictivo se resume en un árbol, estos tipos de enfoques son conocidos como árboles de decisión (James et al., 2017).

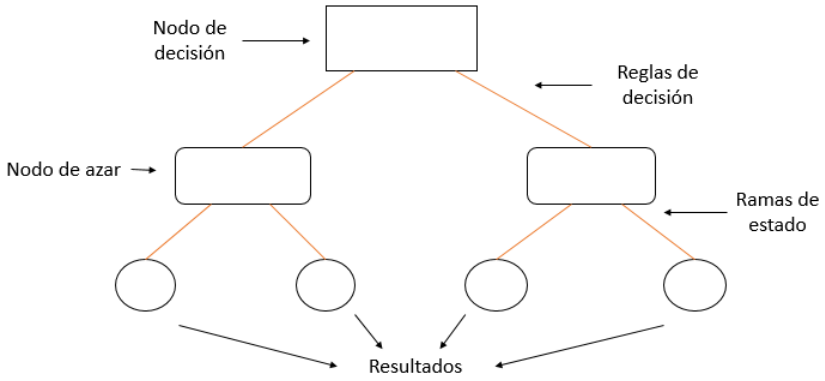

Figura 2.4: Partes de un árbol de decisión
(Fuente: elaboración propia)

La Figura 2.4, indica las partes de un árbol de decisión donde el algoritmo parte de una cierta característica dada por ejemplo la edad (nodo de decisión). Desde ahí, se desprenden las reglas decisivas que considera dividir el nodo principal, obteniendo nuevos nodos definidos por nuevas ramas, repetiéndose lo anterior hasta llegar a las divisiones finales que serían los resultados siendo una clase o un valor predicho del árbol.

Este algoritmo aporta simpleza y utilidad en términos de su inperpretación, pero no es competitivo respecto al enfoque de la precisión en su predicción.

Es por ello, que los algoritmos de aprendizaje supervisado que se enfocan a una mejora en la precisión predictiva son el método bagging, boosting y random forests. Cada uno de ellos, implica la producción de múltiples árboles que luego se combinan para producir una sola predicción de consenso a expensas de la pérdida de interpretación (James et al., 2017).

2.3.2.2. Bagging

También denominada agregación Bootstrap, es un algoritmo que construye cada uno de los clasificadores del conjunto a partir de lo que se conoce como una muestra bootstrap, las cuales se generan tomando del conjunto de entrenamiento tantos elementos con reemplazamiento como este contenga (Ruiz, 2014).

Esta técnica que sirve para reducir la varianza de una función de predicción estimada y funciona bien para procedimientos de alta

varianza y bajo sesgo, como los árboles.

Para la regresión, simplemente se ajusta el mismo árbol de regresión muchas veces a las versiones muestreadas con bootstrap de los datos de entrenamiento, y se promedia el resultado. Y para la clasificación, un comité de árboles emite un voto para la clase prevista (Hastie et al., 2017).

2.3.2.3. Boosting

Al igual que el método bagging, boosting funciona de manera similar excepto que los árboles se cultivan secuencialmente, es decir, cada árbol se arma utilizando información de tres árboles armados anteriormente.

Según Ruiz (2014), el método boosting "es un algoritmo adaptativo en el que cada clasificador se construye en base a los resultados obtenidos en los clasificadores previos mediante la asignación de pesos a cada uno de los ejemplos de entrenamiento"(p.34). De esta forma, los patrones que han sido detectados y clasificados erróneamente por los clasificadores anteriores tendrían más importancia a la hora de construir el nuevo clasificador.

2.3.2.4. Random forest

Este algoritmo de ensamble utiliza el método bagging junto a una selección aleatoria de atributos. En cada nodo de cada árbol del bosque, se selecciona aleatoriamente un subconjunto de los atributos disponibles en ese nodo y se elige el mejor de ellos de acuerdo al criterio de división empleado en el algoritmo base (Breiman, 2001).

Algoritmo empleado para la regresión o la clasificación

1. Para $b = 1$ en B:

   - Obtener una muestra bootstrap Z^* de tamaño N desde la base de entrenamiento.
   - Se construye un árbol de bosque aleatorio T_b a los datos de la muestra bootstrap, repitiendo recursivamente los pasos para el nodo terminal del árbol, hasta alcanzar el tamaño mínimo n_{min} del nodo.

     i) Seleccione m variables al azar de las p variables.
     ii) Elije la mejor variable/punto de división de las m variables.
     iii) Divide el nodo en dos nodos "hijos".

2. Salida del conjunto de árboles $\{T_b\}_1^B$

   Luego se realiza una predicción en un nuevo punto x si es de:

   Regresión: $\hat{f}_{rf}^B(x) = \dfrac{1}{B}\sum_{b=1}^B T_b(x)$

   Clasificación: Si $\hat{C}_b(x)$ es la predicción de la clase de los b árboles de bosques aleatorios entonces $\hat{C}_{rf}^B(x) =$ voto mayoritario $\{\hat{C}_b(x)\}_1^B$.

   Por ende, tal y como lo indica su algoritmo, al hacer crecer un árbol que parte de una muestra bootstrap, este antes de cada división selecciona al azar de las variables de entrada p, un conjunto m de variables tal que $m \leq p$ como candidatos para la división del nodo (Trevor, 2017), obteniendo y dependiendo de la salida (regresora o de clasificación) las predicciones esperadas. Un ejemplo visual lo presenta la Figura 2.5.

   Así, la idea de los bosques aleatorios es mejorar la reducción de la varianza del bagging reduciendo la correlación entre los árboles, sin aumentar abruptamente la varianza, lográndolo por medio del proceso de crecimiento de los árboles mediante la selección aleatoria de las variables de entrada.

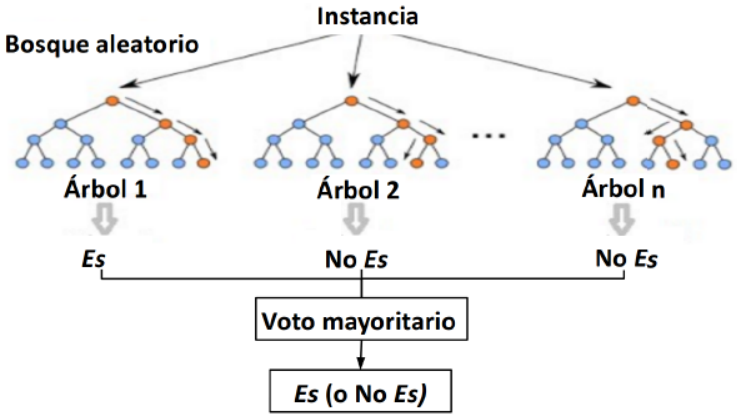

Figura 2.5: Bosques aleatorios de clasificación
(Fuente: Sucuple, 2019)

Capítulo 3

Planteamiento del Problema

Actualmente, existen muchas metodologías estadísticas aparte de las que se proponen para el modelamiento de las transacciones fraudulentas, donde todas tienen como objetivo, encontrar el mejor modelo que prediga de manera correcta cuando se comete fraude, y a su vez, cuál es el que presenta menor error en sus mismas predicciones.

Formalmente, si tenemos la variable dependiente Y que expresa la ocurrencia o no de un cierto suceso, para nuestro caso, esta se definirá como: si existe o no fraude transaccional; donde interesa además, la evaluación del efecto de uno o más antecedentes sobre el hecho de que una situación se produzca o no, es decir, se consideran variables predictoras X_i que sean capaces de explicar la ocurrencia de si existe fraude transaccional o si la transacción es de tipo normal.

Sea n el número de registros de la base de datos con forma (X_i, Y_i), $i = 1, ..., n$; donde X_i es el vector de las variables de entrada (independientes) e Y_i es la variable dependiente codificada como:

$$Y_i = \begin{cases} 0 & \text{transacción normal} \\ 1 & \text{transacción fraudulenta} \end{cases}$$

Luego, el interés se centra en buscar una función $F(X_i)$ que entregue valores entre entre 0 y 1, a modo de comunicar si la transacción es de tipo fraudulenta o no.

3.1. Matriz de confusión

Una matriz de confusión o matriz de error o tabla de contingencia, sirve para evaluar la precisión en la identificación de clases temáticas, es decir, se trata de una matriz bidimensional, en donde las filas y columnas representan las m clases ordenadas (Maass et al., 2006), haciéndola a su vez una matriz simétrica.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Cuadro 3.1: Matriz de confusión
(Fuente: elaboración propia)

Como se observa en el Cuadro 3.1, las columnas corresponden a los resultados arrojados del pronóstico del modelo, mientras que las filas representan la clasificación real de los sujetos. La diagonal de dicha matriz, indica los sujetos que fueron clasificados correctamente por el modelo, es decir, son los que se identifican como los verdaderos positivos (VP) y los verdaderos negativos (VN), a diferencia de las marginales, que indican las asignaciones erradas dadas por los falsos negativos (FN) y los falsos positivos (FP).

Específicamente, las cuatro clases a predecir junto a sus tasas indican lo siguiente:

- Verdaderos Positivos (VP): Número de casos normales que fueron clasificados de manera correcta por el modelo, denominado como éxitos.
- Verdaderos Negativos (VN): Número de casos fraudulentos que fueron clasificados de manera correcta por el modelo, denominado como rechazos correctos.
- Falsos Positivos (FP): Número de casos normales que fueron clasificados de manera incorrecta por el modelo, denominado como falsas alarmas o error de tipo I.
- Falsos Negativos (FN): Número de casos fraudulentos que fueron clasificados de manera incorrecta por el modelo, denominado como error de tipo II.

- Tasa de Verdaderos Positivos (TVP): Es el porcentaje de casos normales correctamente clasificados como pertenecientes a la clase normal.
- Tasa de Verdaderos Negativos (TVN): Es el porcentaje de casos fraudulentos correctamente clasificados como pertenecientes a la clase fraudulenta.
- Tasa de Falsos Positivos (TFP): Es el porcentaje de casos fraudulentos mal clasificados como pertenecientes a la clase normal.
- Tasa de Falsos Negativos (TFN): Es el porcentaje de casos normales mal clasificados como pertenecientes a la clase fraudulenta.

3.2. Métricas de evaluación del modelo predictivo

A continuación, se presentan distintas herramientas indicadoras que se desprenden de la matriz de confusión y que permiten evaluar el rendimiento del modelo como también su calidad.

1. **Exactitud (Accuracy)**

   Mide la calidad de las predicciones realizadas y evalúa la capacidad del modelo de clasificar de manera correcta las categorías de los casos positivos y negativos.

   Matemáticamente se expresa como:

$$\text{Exactitud}_{\text{global}} = \frac{VP + VN}{VP + VN + FP + FN}$$

2. **Precisión**

   Es un indicador de calidad al igual que la exactitud, el cual se define según González (2018) como la "probabilidad promedio de recuperación relevante de información" (p.10), es decir, es el porcentaje de casos positivos detectados por el modelo. Matemáticamente se define como:

$$\text{Precisión} = \frac{VP}{VP + FP}$$

3. **Exhaustividad (Recall)**

   Según González (2018), define la exhaustividad como "la probabilidad promedio de recuperación completa, realizando un promedio de varias consultas de recuperación"(p.10). Matemáticamente se define como:

$$\text{Exhaustividad} = \frac{VP}{VP + FN}$$

4. **Curva ROC**

   El indicador *receiver operating characteristics* (ROC) traducido al español como características operativas del receptor, muestra la compensación entre la tasa de verdaderos positivos (TVP) y la tasa de los falsos negativos (TFN) (Han et al., 2011).

   Matemáticamente se define como:

$$\text{Sensibilidad}_{TVP} = \frac{VP}{VP + FN}$$

$$\text{Sensibilidad}_{TFP} = \frac{VP}{FP + VN}$$

$$\text{Especificidad} = \frac{VN}{FP + VN} = 1 - TFP$$

La Figura 3.1, muestra un gráfico bidimensional donde el eje vertical representa la TVP (sensibilidad), mientras que el eje horizontal representa la TFP (especificidad). La curva ROC, llamada también AUC (área bajo la curva), se encarga de clasificar desde el punto de vista estadístico el mejor modelo obtenido, utilizando las tablas de aprendizaje y entrenamiento. Para calcular el AUC sólo hay que obtener el área bajo la curva como se muestra en la ecuación 3.1.

$$AUC = \frac{1 + TVP - TFP}{2} \tag{3.1}$$

Así mismo, el Cuadro 3.2 indica los intervalos de clasificación para los valores obtenidos del AUC, siendo para el caso de la Figura 3.1 un valor de 0.74, clasificándose como un modelo del tipo regular.

Niveles	Intervalos
Excelente	0.98 - 1.00
Muy Bueno	0.91 - 0.97
Bueno	0.76 - 0.90
Regular	0.61 - 0.75
Malo	0.50 - 0.60

Cuadro 3.2: Intervalos de calificación de valores AUC
(Fuente: elaboración propia)

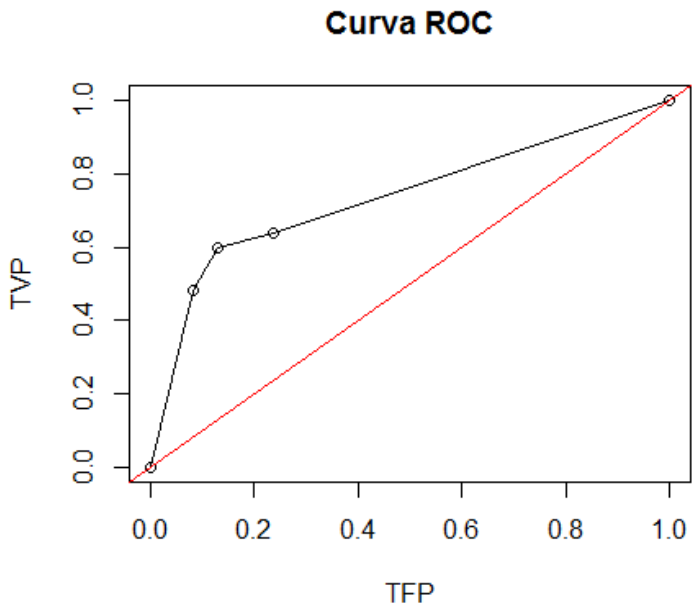

Figura 3.1: Curva ROC
(Fuente: elaboración propia)

Se destaca que la clase de interés para efectos del Proyecto de Título, es la detección del fraude transaccional en el escenario D, ya que al predecir la existencia de una transacción normal cuando en realidad es fraudulenta, provoca pérdidas innecesarias para cualquier entidad financiera.

Por tanto, la finalidad del Proyecto de Título es proponer la mejor metodología estadística que cumpla con lo expuesto anteriormente y a su vez, que entregue simplicidad al momento de replicar su seguimiento.

Capítulo 4

Objetivos

4.1. Objetivo General

Proponer herramientas estadísticas para la detección de fraude transaccional de tarjetas de crédito en canales no presenciales.

4.2. Objetivos Específicos

- Aplicar diferentes técnicas de construcción de variables a partir de variables brutas.
- Implementar los modelos de la metodología estadística de Regresión Logística y Random Forest en datos reales.
- Comparar las técnicas estadísticas para predecir el fraude transaccional de tarjetas de crédito.

Capítulo 5

Resultados

Los datos utilizados en este proyecto provienen de un país externo proporcionados por cierta empresa bancaria, de los cuales contienen información de las transacciones realizadas por los usuarios en el ámbito financiero.

Inicialmente se consideran distintas bases de datos que corresponden a un periodo de tiempo de un total de 16 meses. En aquellas bases, se destacan 4 tipos de escenarios denominados como: A, B, C y D.

Figura 5.1: Tamaños proporcionales por escenario de las transacciones normales
(Fuente: elaboración propia)

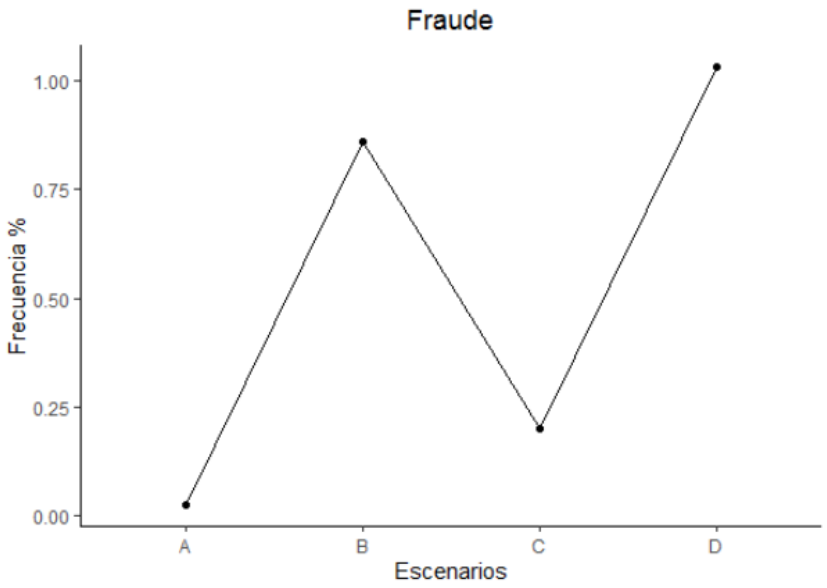

Figura 5.2: Tamaños proporcionales por escenario de las transacciones fraudulentas
(Fuente: elaboración propia)

Las Figuras 5.1 y 5.2 muestran las frecuencias porcentuales por escenario de acuerdo a las transacciones etiquetadas como normales (Figura 5.1) y fraudulentas (Figura 5.2). De ellas, interesa observar la Figura 5.2, donde se percibe que el escenario con mayor proporción de fraude es en el **D**. Por tanto en el presente Proyecto, se trabajará con aquellos datos que pertenezcan sólo a ese escenario.

5.0.1. Análisis descriptivo

Para identificar las principales características de los datos, se decidió realizar un breve análisis descriptivo, visualizando la distribución de las variables relevantes que se presentan frecuentemente en este tipo de estudio, anexando en el Capítulo 6 (sección 6.1) las variables restantes.

- Variable 1 (V1): indica el tipo de transacción realizada por un cliente.
  - 0 : indica una transacción de tipo normal.
  - 1: indica una transacción de tipo fraudulenta.

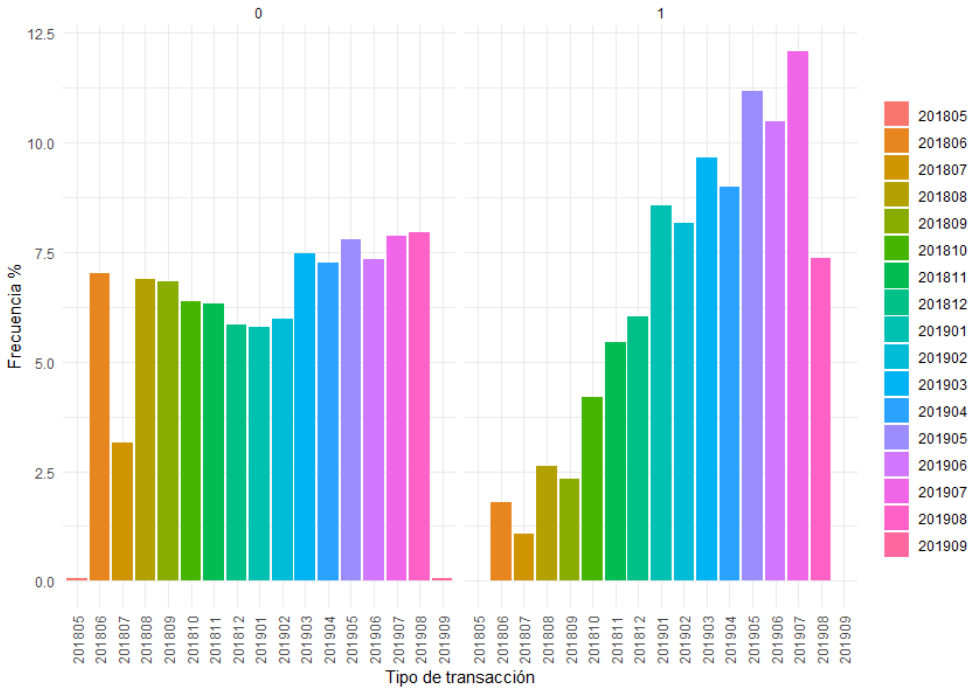

Figura 5.3: Distribución de la V1
(Fuente: elaboración propia)

Se observa en la Figura 5.3 una tendencia positiva en la distribución de la clase fraudulenta (panel derecho de la Figura 5.1) a medida que transcurren los meses respecto a la clase de transacciones de normales (panel izquierdo de la Figura 5.1) , esto es debido al aumento del fraude no presencial en 3 campos identificados por los analistas:

- Fraude amigable: es aquél que involucra las transferencias hechas en la compra de videojuegos, Appstore, Itunes, entre otras; presentándose en mayor proporción respecto a los demás tipos de fraude.
- Fraude duro: es aquél que involucra las transferencias realizadas en comercios como Aliexpress, Amazon, Netflix, Uber, etc.; es decir, en los comercios masivos de distintos países.
- Fraude de prueba: es en donde utilizan la programación para la detección de combinaciones de números de tarjetas de crédito con el fin de encontrar aquella codificación que permita realizar alguna compra vía transferencia.

- Variable 2 (V2): Montos de la transacción realizada por un cliente.

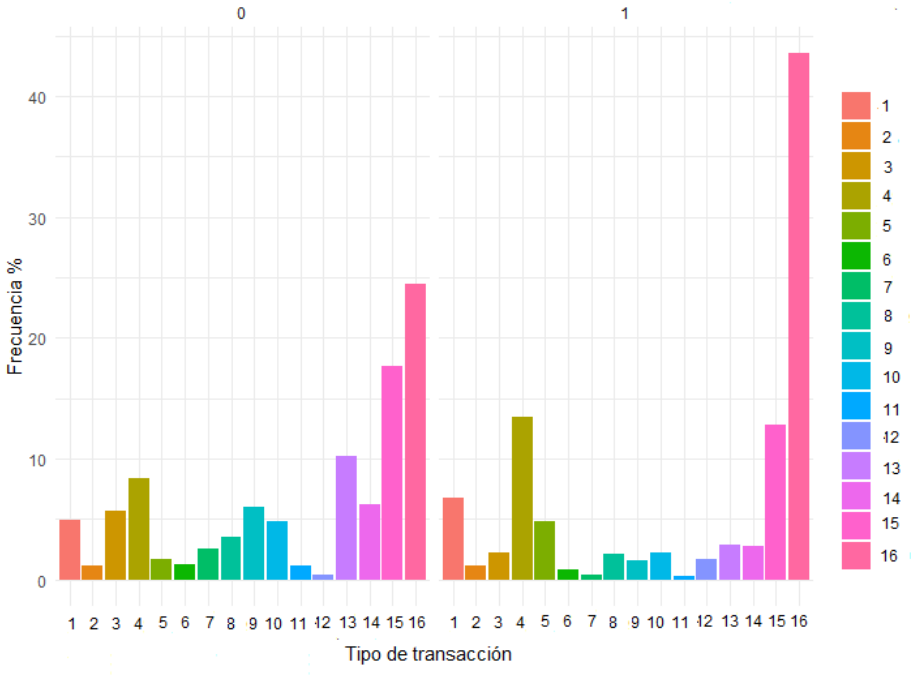

Figura 5.4: Distribución de la V2
(Fuente: elaboración propia)

En el gráfico de la Figura 5.4, se observa una proporción mucho mayor respecto a los demás tramos para los saldos pertenecientes a la categoría 16 en ambas clases, siendo

significativamente más alto en la clase fraudulenta (panel derecho de la Figura 5.2). Este comportamiento se explica debido a que se hicieron compras no presenciales de montos importantes en comercios como lo son por ejemplo las aerolíneas.

- Variable 3 (V3): Indica el código del país en donde se ejecuta la transacción hecha por un cliente.

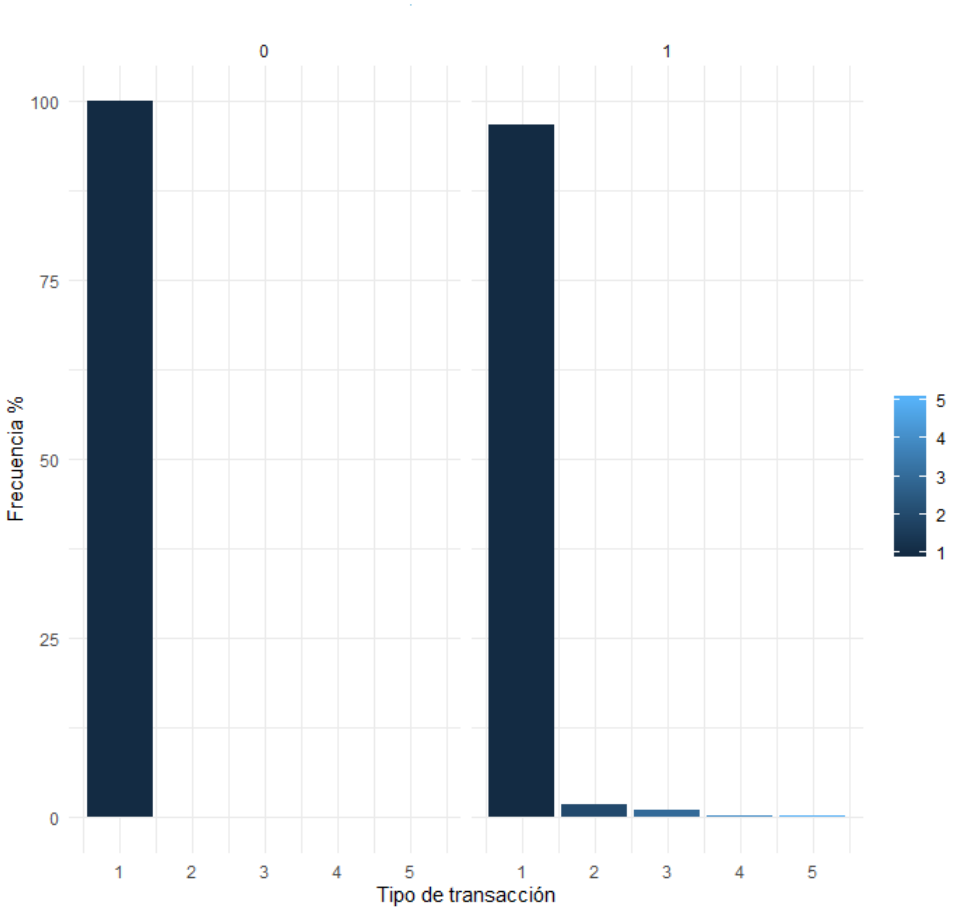

Figura 5.5: Distribución de la V3
(Fuente: elaboración propia)

La Figura 5.5, muestra que la mayoría de las transacciones no presenciales realizadas por los usuarios pertenecientes a la empresa bancaria, son frecuentemente hechas en los mismos países, presentando una proporción similar de aquellas transacciones etiquetadas como fraudulentas y no fraudulentas.

- Variable 4 (V4): Indica el código del comercio en donde se ejecuta la transacción hecha por un cliente.

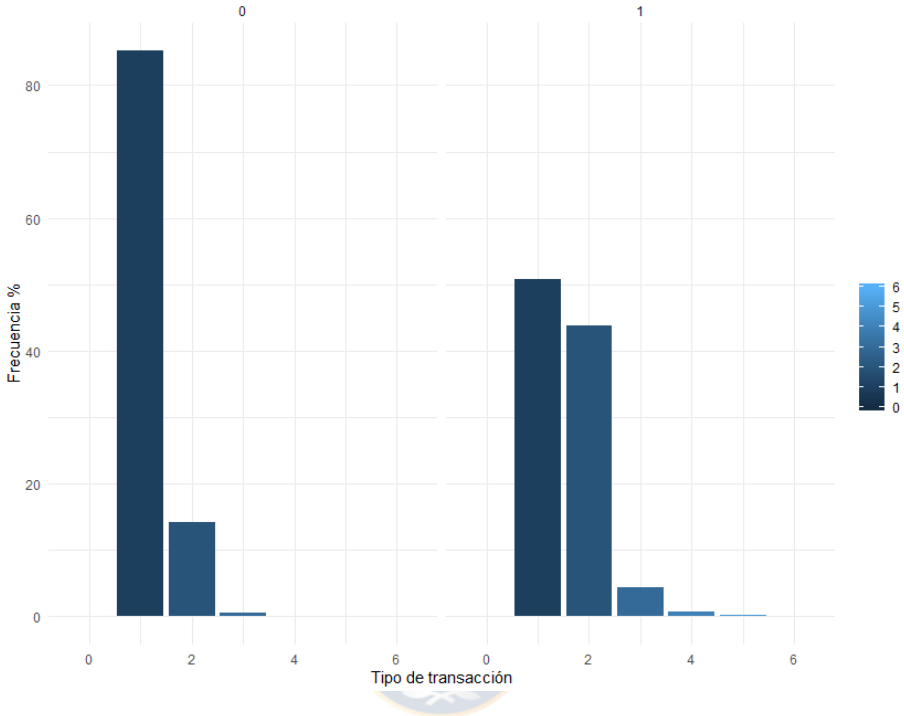

Figura 5.6: Distribución de la V4
(Fuente: elaboración propia)

Para la Figura 5.6, se observa que para las transacciones catalogadas como normales (panel izquierdo de la Figura 5.4) se presenta en 2 conjuntos de comercios (categoría 1 y 2), mientras que para las transacciones etiquetadas como fraudulentas destacan mayores movimientos de fraude en 3 conjuntos de comercios (categorías 1,2 y 3). En particular, observando la categoría 2 de ambas clases, indicaría que del total de transacciones fraudulentas comparado con el total de transacciones normales realizadas en los mismos comercios, gran parte de ellas se detectan como fraudes.

5.1. Tratamiento de los datos

5.1.1. Exclusión de variables

La base del cuadrante de interés, contiene 122 variables con un total de 7.589.782 registros.

Analizando aquellas variables, se realizaron 2 tipos de filtros visualizadas en el Cuadro 5.1, obteniendo un total de 30 variables finales.

Variables	n
Inicial	**122**
Sin datos	54
Criterio experto	14
Mal definidas	14
Rebundante	7
Ilógica	1
Con menos del 5 % de los datos	1
Sensible	1
Final	**30**

Cuadro 5.1: Filtro de variables
(Fuente: elaboración propia)

5.1.2. Exclusión de registros

Respecto al manejo de los datos faltantes se destaca, que a partir del análisis Bi-variado realizado a la base, se visualizan los valores sin información o NA por variable.

Por literatura, existen distintas metodologías que sirven para sobrellevar el tema de los valores faltantes por variable, siendo algunas de ellos los:

- Métodos que eliminan observaciones
- Métodos que utilizan todos los datos disponibles
- Métodos que imputan los datos faltantes

5.1.2.1. Métodos que eliminan observaciones

Estos métodos consisten en eliminar la información de cualquier individuo en el que haya pérdida de datos, por lo menos, en una de las variables de estudio.
Se dicen ser los más sencillos de implementar y los que menos recursos computacionales necesitan. Además, estos métodos reducen el tamaño de la muestra, lo que disminuye la eficiencia y conlleva que los errores estándar sean mayores en los parámetros de interés.

5.1.2.2. Métodos que utilizan todos los datos disponibles

Estos métodos utilizan la información tanto de los casos completos como de los incompletos. Por ejemplo, en regresión multivariante, los casos con datos incompletos en algunas variables pueden dar información sobre la relación entre las variables respuesta y otras variables observadas.

5.1.2.3. Métodos que imputan los datos faltantes

La idea de estos métodos es que la información de los valores que faltan se puede extraer de las variables observadas; de los cuales, se destacan dos clases de métodos: imputación simple e imputación múltiple.

1. Imputación simple

   Los métodos de imputación simple reemplazan los datos faltantes por un único valor, haciendo que en general, para cualquier enfoque de este tipo de métodos, se subestiman los errores estándar de las variables en las que faltan datos.

2. Imputación múltiple

   Consisten en reemplazar cada valor perdido por un conjunto de m valores, obteniéndose así m conjuntos completos de datos, lo que da lugar a m estimaciones con sus respectivas varianzas o errores estándar.

   Así, se combinan las estimaciones que como resultado dan estimaciones e intervalos de confianza que incorporan la incertidumbre causada por la pérdida de datos.

En primer lugar, se analiza una variable importante que llamaremos V5. Esta variable consta con un total de 81.633 registros de datos faltantes, el cual corresponde al 1,07 % del total de la base.

Entonces, para ese 1,07 % de datos faltantes y considerando que todas aquellos registros fueron no-fraudulentas (normales), se decide eliminar esos casos, quedando la base final con un total de 7.508.149 registros.

Respecto a las demás variables y para el *método que elimina observaciones*, se recomienda utilizarlo cuando el porcentaje de valores faltantes es menor igual al 10 % (ver la siguiente tabla).

Variables	**Proporción datos faltantes**	**Prop. Normal**	**Prop. Fraude**
V6	0,66 %	99,54 % (49.684)	0.46 % (230)
V7	21,44 %	99,56 % (1.602.713)	0,44 % (6.984)
V8	4,67 %	97,95 % (343.385)	2,05 % (7.159)
V9	20,08 %	99,46 % (1.499.545)	0,54 % (8.147)
V4	0,051 %	98,79 % (3.753)	1,21 % (46)
V10	48,40 %	99,29 % (3.607.819)	0,71 % (25.779)
V11	4,82 %	99,94 % (361.895)	0,06 % (209)
V12	6,04 %	99,68 % (452.177)	0,32 % (1.449)
V13	6,02 %	99,69 % (450.218)	0,31 % (1.410)
V14	27,51 %	99,59 % (2.056.791)	0,41 % (8.441)
V15	27,51 %	99,59 % (2.056.791)	0,41 % (8.441)
V16	27,51 %	99,59 % (2.056.791)	0,41 % (8.441)
V17	48,40 %	99,29 % (3.607.819)	0,72 % (25.779)
V18	0.00008 %	100 % (6)	0 %
V19	7,10 %	99,74 % (531.851)	0,26 % (1.410)

Cuadro 5.2: Variables con datos faltantes
(Fuente: elaboración propia)

Como se observa en el Cuadro 5.2, existen variables con valores NA en la columna de la variable fraude, por lo tanto, se decidió no eliminar estos registros debido a que estaríamos aumentando el problema del desbalance de nuestros datos. No así para la variable V18, la cual solo tiene 6 registros que según se indica fueron transacciones no fraudulentas.

Para el método de imputación simple y dado que las variables que tenemos son categóricas, una posibilidad es reemplazar aquellos valores faltantes por la moda, sin embargo, este tipo de método conlleva problemas como la subestimación de varianzas y covarianzas. Además, donde tiene mayor influencia es en la modificación de las relaciones entre las variables, lo cual perjudica el tipo de asociación que pudiesen tener. Por lo tanto, este tipo de método se descarta.

Así mismo, el método de imputación múltiple presenta una importante desventaja respecto a los de imputación simple, y es que son más difíciles de implementar y además podría conllevar mayor gasto computacional.

Finalmente, se decide utilizar el *método con todos los datos disponibles* filtrando los registros mencionados previamente, quedando como se visualiza en el Cuadro 5.3.

	Nro. de registros
Inicial	7.589.782
V5	81.633
V18	6
Final	7.508.143

Cuadro 5.3: Filtro de registros finales
(Fuente: elaboración propia)

A su vez se considera además, aquél conjunto de observaciones faltantes como parte de una de las categorías por variable, la cual denominaremos como "-99"(ver Cuadro 5.4).

Variables	**Categoría**	n
V6	-99	49.914
V7	-99	1.609.677
V8	-99	350.257
V9	-99	1.507.599
V4	-99	3.799
V10	-99	3.633.592
V11	-99	280.471
V12	-99	453.621
V13	-99	451.623
V14	-99	2.065.227
V15	-99	2.065.227
V16	-99	2.065.227
V17	-99	3.633.592
V19	-99	451.623

Cuadro 5.4: Lista de variables con valores faltantes
(Fuente: elaboración propia)

5.2. Construcción de variables

En este capítulo, se presentan algunas ideas utilizadas para la creación de variables las cuales se basan en 2 aspectos: las **variables acumuladoras**, que cumplen el rol de identificar distintos tipos de patrones según se presente y las **variables de número índice**, las cuales son descritas a continuación.

5.2.1. Variables acumuladoras

Sea C_s un cierto identificador, $s \in \mathbb{N}$. Sea F_m, $m \in \mathbb{N}$, la variable asociada a ciertas transacciones efectuadas por el C_s y sea T_n, $n \in \mathbb{N}$, la cantidad total de transacciones realizadas en F_m por este identificador C_s.

Así, la lectura de variables creadas con respecto a su movilidad de forma recursiva y para un C_1 por ejemplo quedaría como sigue:

$$L_1 = \{(F_m/T_n), (F_{m-1}/T_{n-1}), ..., (F_1/T_1)\}$$
$$L_2 = \{(F_{m-1}/T_{n-1}), ..., (F_1/T_1)\}$$
$$\vdots$$
$$L_m = \{(F_1/T_1)\}$$

Algunas de las variables creadas se presentan como ejemplo en el Cuadro 5.5.

Variables	**Descripción**
V20	Nro. de transacciones hechas en los últimos 7 dias.
V21	Monto acumulado de los últimos 7 dias.
V22	Monto promedio en los últimos 7 dias.
V23	Monto máximo de los últimos 7 dias.
V24	Nro. de transacciones hechas en los últimos 15 dias.
V25	Monto acumulado de los últimos 15 dias.
V26	Monto promedio en los últimos 15 dias.
V27	Monto máximo de los últimos 15 dias.
V28	Nro. de transacciones hechas en los ultimos 31 dias.
V29	Monto acumulado de los últimos 31 dias.
V30	Monto promedio en los últimos 31 dias.
V31	Monto máximo de los últimos 31 dias.
V32	Nro. de transacciones hechas en los últimos 62 dias.
V33	Monto acumulado de los últimos 62 dias.
V34	Monto promedio en los últimos 62 dias.
V35	Monto máximo de los últimos 62 dias.

Cuadro 5.5: Variables acumuladoras
(Fuente: elaboración propia)

5.2.2. Variables de número índice

5.2.2.1. Número índice

Se le llama **número de índice** o simplemente **índice** a una medida estadística diseñada para poner de relieve cambios en una variable o en un grupo de variables relacionadas con respecto al tiempo, situación geográfica o cualquier otra característica. Por tanto, una colección de números índice para diferentes años, lugares,etc. recibe el nombre de *serie de índices*.

En el caso más sencillo, los números índice sirven para conocer la variación porcentual de una determinada magnitud en el tiempo o en el espacio.
En este caso, los números índice no son otra cosa que el porcentaje de variación de cada valor de la variable con respecto a un valor de referencia llamado **periodo base** o periodo de referencia. Sean x_a e x_b dos valores de una variable X en dos instantes de tiempo a y b. Entonces, el cociente entre x_b y x_a es:

$$x_{\frac{b}{a}} = \frac{x_b}{x_a} \cdot 100\,\%$$

.
esto determina un número índice que representa la relación entre los valores de la variable en esos dos instantes.

5.2.2.2. Propiedades de los números índice

1. Propiedad identidad: El índice de un periodo respecto al mismo periodo es 1, es decir, $x_{\frac{a}{a}} = 1$.

2. Propiedad de inversión temporal: Establece una relación entre los 'indices correspondientes a dos periodos de tiempo.

$$x_{\frac{a}{b}} \cdot x_{\frac{b}{a}} = 1 \Longleftrightarrow x_{\frac{a}{b}} = \frac{1}{x_{\frac{b}{a}}}$$

3. Propiedad cíclica o circular: Establece una relación entre los índices de varios periodos de tiempo encadenados:

$$x_{\frac{a}{b}} \cdot x_{\frac{b}{c}} \cdot x_{\frac{c}{a}} = 1$$

$$x_{\frac{a}{b}} \cdot x_{\frac{b}{c}} \cdot x_{\frac{c}{d}} \cdot x_{\frac{d}{a}} = 1$$

5.2.3. Clasificación de los números índice

En función del número de variables que se quiere relacionar el número índice se divide en 2 tipos:

- Índice simple: considera los elementales y en cadena.
- Índice complejo: considera los sin ponderar y ponderados.

5.2.3.1. Índice simple

Los índices simples son los que hacen referencia a una variable concreta, es decir, a los que dan a conocer la evolución de una única variable comparándola con ella misma al tomar un periodo de tiempo como referencia o base.

5.2.3.2. Índices simples elementales

Los índices elementales son un tipo de índices simples que responderán estrictamente a la definición como cociente de valores de la variable. En este caso se toma un único valor como periodo base o periodo de referencia y es fijo para todos los valores de la variable.

El Cuadro 5.6 muestra un ejemplo de los índices simples elementales.

tiempo t (días)	0	7	14	31	62	92
variable X	x_0	x_7	x_{14}	x_{31}	x_{62}	x_{92}
Indice simple elemental	1	$\frac{x_7}{x_0}$	$\frac{x_{14}}{x_0}$	$\frac{x_{31}}{x_0}$	$\frac{x_{62}}{x_0}$	$\frac{x_{92}}{x_0}$

Cuadro 5.6: Ejemplo de los índices simples elementales
(Fuente: elaboración propia)

5.2.3.3. Índices en cadena

Los índices en cadena son un tipo de índice simple donde el periodo base va a ir cambiando de un valor de la variable a otro. Para calcular el índice de un periodo se toma como base el valor de la variable en el periodo inmediatamente anterior.

tiempo t (días)	0	7	14	31	62	92
variable X	x_0	x_7	x_{14}	x_{31}	x_{62}	x_{92}
Índice simple elemental	1	$\frac{x_0}{x_7}$	$\frac{x_7}{x_{14}}$	$\frac{x_{14}}{x_{31}}$	$\frac{x_{31}}{x_{62}}$	$\frac{x_{62}}{x_{92}}$

Cuadro 5.7: Ejemplo de los índices en cadena
(Fuente: elaboración propia)

Al igual que para las variables acumuladoras, el Cuadro 5.8 muestra un ejemplo de algunas de las variables creadas.

Variables	Descripción
V36	Relación entre el Nro. de transacciones hechas en los últimos 7 días y el Nro. de transacciones hechas en los últimos 15 días.
V37	Relación entre el monto acumulado en los últimos 15 días y el monto acumulado en los últimos 92 días.
V38	Relación entre el monto promedio de los últimos 7 días y el monto promedio de los últimos 15 días.
V39	Relación entre el monto máximo de los últimos 31 días y el monto máximo de los últimos 92 días.
V40	Relación entre el monto acumulado en los últimos 15 días y el monto acumulado en los últimos 62 días.
V41	Relación entre el monto promedio de los últimos 7 días y el monto promedio de los últimos 92 días.

Cuadro 5.8: Ejemplo de variables de número índice (Fuente: elaboración propia)

Con la teoría descrita en este capítulo, las variables en total creadas fueron alrededor de 180 variables del total considerando además, combinaciones entre ellas, donde por motivos de confidencialidad de la empresa no serán expuestas en detalle.

5.3. Selección de variables

En muchas situaciones se dispone de un conjunto grande de posibles variables explicativas, por lo que una posible pregunta sería saber si todas las variables deben entrar en el modelo de regresión o de clasificación según corresponda, y en caso negativo, saber qué variables deben entrar y cuáles no. Los métodos de selección de variables se encargan de abordar el problema de construcción o selección del modelo.

Así, es que existen distintas metodologías para decidir qué variables son las que se considerarían como variables de entrada en la creación de nuestros modelos de clasificación, siendo una de ellas la metodología del *Random Forest*, la cual fue utilizada en el presente Proyecto.

La metodología utilizada entrega la importancia de las variables, que para proceder a una identificación de los predictores más influyentes proporciona los índices Gini por variable. Ella es una medida de desorden, es decir, el "decrecimiento" del índice de Gini, indica que cuanto mayor sea esta medida, más variabilidad aporta a la variable dependiente siendo para nuestro caso la variable llamada V1.

Seguido de lo anterior, para realizar esta selección de variables se consideró una muestra aleatoria representativa manteniendo las mismas proporciones aproximadamente que la de la base global (ver Cuadro 5.9), esto se realizó así para disminuir el tiempo debido a la carga computacional por el tamaño de la base de datos.

	V1		
	0	1	Total
Base global	7.429.943 (98,96 %)	78.200 (1,04 %)	7.508.143
Muestra	371.548 (98,97 %)	3.860 (1,03 %)	375.408

Cuadro 5.9: Proporciones de la base global y su muestra (Fuente: elaboración propia)

El criterio de selección de variables fue considerando aquellas variables que presentaron un índice de Gini (MeanDecreaseGini) mayor o igual al 0.4 (valores multiplicados por 100), obteniendo 79 variables predictoras resultantes como se observan en los Cuadros 5.10, 5.11 y 5.12.

Variable	Indice de Gini	Variable	Indice de Gini	Variable	Indice de Gini
V4	99.57	V25	67.08	V27	61.38
V33	76.73	V47	66.99	V53	61.23
V42	75.02	V48	65.07	V54	60.90
V29	72.68	V35	64.70	V55	60.42
V43	70.70	V49	64.51	V56	60.41
V44	70.43	V31	64.12	V57	60.19
V45	69.97	V50	63.57	V21	59.75
V34	69.64	V51	62.92	V58	59.27
V46	69.35	V6	62.85	V59	58.43
V30	67.14	V52	62.52	V60	58.40

Cuadro 5.10: Variables y sus índices de Gini
(Fuente: elaboración propia)

Variable	Indice de Gini	Variable	Indice de Gini	Variable	Indice de Gini
V61	57.92	V3	55.81	V77	53.40
V22	57.91	V69	55.50	V78	52.87
V62	57.44	V70	54.91	V79	52.81
V63	57.33	V26	54.72	V80	52.74
V64	57.13	V71	54.71	V81	52.35
V65	56.80	V72	54.34	V82	52.31
V23	56.73	V73	54.23	V83	52.17
V66	56.60	V74	53.83	V84	51.99
V67	56.51	V75	53.76	V85	51.84
V68	56.02	V76	53.47	V86	51.80

Cuadro 5.11: Variables y sus índices de Gini
(Fuente: elaboración propia)

Variable	Indice de Gini	Variable	Indice de Gini	Variable	Indice de Gini
V96	51.60	V102	46.32	V92	41.25
V97	50.34	V103	45.29	V93	41.22
V98	50.29	V87	44.88	V94	41.17
V99	49.97	V88	44.00	V95	40.18
V100	48.07	V89	43.91	V32	40.08
V2	47.43	V90	43.86		
V101	47.22	V91	43.05		

Cuadro 5.12: Variables y sus índices de Gini
(Fuente: elaboración propia)

5.4. Partición de la base de datos

Para evaluar la capacidad predictiva de un modelo, se debe comprobar qué tan exactas son sus predicciones a los verdaderos valores de la variable respuesta.

Para poder cuantificar de forma correcta este error, se necesita disponer de un conjunto de observaciones, de las cuales se conozca la variable respuesta, pero que el modelo no haya "conocido", es decir, que no hayan participado en su ajuste.

De esto, se procede a dividir los datos disponibles en una base de entrenamiento y una base de prueba. El tamaño adecuado de las particiones depende en gran medida de la cantidad de datos disponibles y la seguridad que se necesite en la estimación del error, el cual para nuestro caso se decide utilizar la partición 60 % - 40 %.

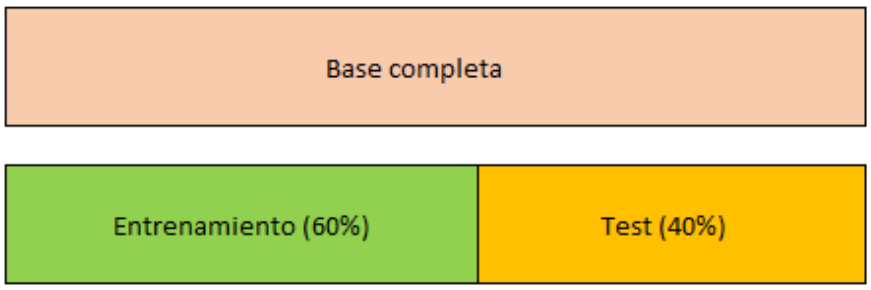

Figura 5.7: División de la base de datos
(Fuente: elaboración propia)

Al igual que para la selección de variables, se procedió a realizar un muestreo (ver Cuadro 5.14 y 5.16) a ambas bases por la demora en su tiempo de procesamiento y carga computacional.

5.4.1. Conjunto entrenamiento

Este conjunto sirve para entrenar el modelo con la técnica estadística propuesta.

	n			
	0		**1**	
Entrenamiento	4.458.028	98,96 %	46.858	1,04 %

Cuadro 5.13: Registros por clases en la base entrenamiento
(Fuente: elaboración propia)

5.4.1.1. Muestra

Se realiza un muestreo aleatorio simple sin reemplazo manteniendo las proporciones de ambas clases pertenecientes a la base de entrenamiento.

	n			
	0		**1**	
Entrenamiento	445.784	98,96 %	4.705	1,04 %

Cuadro 5.14: Registros de la muestra por clases en
la base entrenamiento
(Fuente: elaboración propia)

5.4.2. Conjunto test

Este conjunto sirve para evaluar el modelo con la técnica estadística propuesta.

	n			
	0		**1**	
Test	3.003.257	98,96 %	31.342	1,04 %

Cuadro 5.15: Registros por clases en la base prueba
(Fuente: elaboración propia)

5.4.2.1. Muestra

El muestreo se realizó análogamente al conjunto de la sección 5.4.1.1..

	n			
	0		**1**	
Test	297.181	98,96 %	3.145	1,04 %

Cuadro 5.16: Registros de la muestra por clases en la base prueba
(Fuente: elaboración propia)

5.5. Desbalance de clases

Otro problema clásico en este tipo de datos, es el desbalance de las clases en la variable que queremos predecir y que apriori "conocemos". El cual para nuestro caso, la etiqueta "1" (Fraude), corresponde al 1 % aproximadamente de la base total, versus el 99 % aproximadamente que corresponde a la clase "0"(No-fraude).

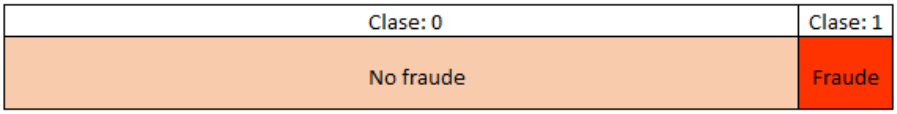

Figura 5.8: Desbalance de clases
(Fuente: elaboración propia)

Para enfrentar esta dificultad problemática, se procedió a separar la base de datos respecto a los 2 tipos de clases existentes para aplicar la metodología clustering a la clase mayoritaria. Posteriormente, se realiza un muestreo aleatorio sin reemplazo a estos 3 grupos representativos de la clase de transacciones no fraudulentas sugeridas por el algoritmo para igualar en tamaño muestral con la clase minoritaria, es decir, con el número de transacciones fraudulentas no presenciales (ver Figura 5.8).

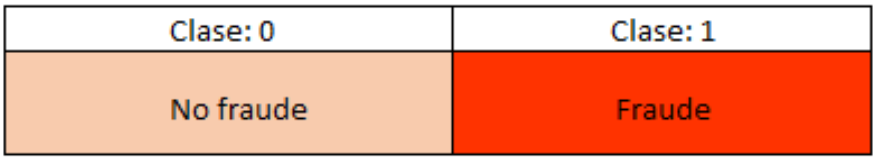

Figura 5.9: Equilibrio de clases
(Fuente: elaboración propia)

En efecto, aplicando lo anterior a cada base se obtiene lo que se visualiza en el Cuadro 5.17 y en el Cuadro 5.18.

	Tamaños		
	0	1	Total
A-priori	4.458.028	46.858	4.504.886
Muestra	445.784	4.705	450.489
Muestra Balanceada	4.705	4.705	9.410

Cuadro 5.17: Registros por clases en la base entrenamiento
(Fuente: elaboración propia)

	Tamaños		
	0	1	Total
A-priori	2.971.915	31.342	3.003.257
Muestra	297.181	3.145	300.326
Muestra Balanceada	3.145	3.145	6.290

Cuadro 5.18: Registros por clases en la base prueba
(Fuente: elaboración propia)

5.6. Aplicación de las técnicas estadísticas

Esta sección, se presenta las matrices de confusión y sus respectivas interpretaciones de ambos modelos aplicados junto a sus métricas, que sirven para medir el rendimiento de aquellos algoritmos.

5.6.1. Regresión logística

En el Cuadro 5.19, se tiene la matriz de confusión para el modelo entrenado de la muestra descrita en la sección 5.4.1.1..

		Predicción		
		0	**1**	**Total**
Observación	**0**	38,01	11,98	50,00
	1	14,55	35,45	50,00
	Total	52,56	47,44	

Cuadro 5.19: Matriz de confusión porcentual
(Fuente: elaboración propia)

La intepretación que deriva de la matriz de confusión son las siguientes:

- Existen 11,98 % de predicciones incorrectas de clase negativa.
- Existen 38,01 % de predicciones correctas de la clase positiva.
- Existen 35,45 % de predicciones correctas de la clase negativa.
- Existen 14,55 % de predicciones incorrectas de clase positiva.

Y la interpretación deribada de la problemática nos dice que: existe un 11,98 % del total de transacciones no presenciales que fueron realizadas por los usuarios usando su tarjeta de crédito de una cierta empresa bancaria, las cuales el modelo etiquetó erróneamente como fraudulentas, cuando ciertamente ellas eran transacciones normales. Por otro lado, existe un 14,55 % del total de transacciones no presenciales que el modelo etiquetó como transacciones normales, cuando realmente eran fraudulentas.

Respecto a la diagonal de la matriz de confusión, existe un total de 38,01 % de transacciones no presenciales realizadas por los usuarios con tarjeta de crédito de cierta empresa bancaria identificadas por el modelo correctamente como transaccines normales. Análogamente, existe un total de 35,45 % de transacciones no presenciales identificadas por el modelo correctamente como fraudulentas.

5.6.1.1. Métricas

- Precisión $= \dfrac{38,01}{38,01+14,55} = 72,31$
- Sensibilidad $= \dfrac{38,01}{38,01+35,45} = 51,74$
- Exactitud $= \dfrac{38,01+35,45}{38,01+14,55+11,98+35,45} = 73,47$
- Especificidad $= \dfrac{35,45}{14,55+35,45} = 70,90$
- Tasa de error $= \dfrac{14,55+11,98}{38,01+11,98+14,55+35,45} = 26,53$
- Valor de predicción negativa (VPN):

$$VPN = \frac{35,45}{47,44} = 74,73$$

- AUC $= \dfrac{100+72,31-29,10}{2} = 71,61$

El modelo de Regresión Logística con clases equilibradas, presenta un 74 % aproximadamente de exactitud, es decir, este porcentaje indica qué tan fiable o exacto es el modelo para predecir la clasificación de ambas clases, indicando ser regularmente bueno. Respecto a su sensibilidad, el modelo indica que el porcentaje de casos clasificados como transacciones normales correctamente fue de un 52 % aprox., al contrario de la especificidad, donde el modelo indicó que el porcentaje de casos clasificados correctamente de transacciones fraudulentas fue de un 71 % aproximadamente.

La precisión del modelo dió un valor del 72 % aprox. con una tasa de error del 27 % aproximadamente, catalogándose como un modelo con resultados aceptables.

5.6.2. Random Forest

El Cuadro 5.20 muestra la matriz de confusión para el modelo entrenado de la muestra descrita en la sección 5.4.1.1..

		Predicción		
		0	**1**	**Total**
Observación	**0**	37,55	12,44	50,00
	1	12,42	37,58	50,00
	Total	49,98	50,02	

Cuadro 5.20: Matriz de confusión porcentual
(Fuente: elaboración propia)

La intepretación teórica que deriva de la matriz de confusión para esta metodología son las siguientes:

- Existen 12,44 % de predicciones incorrectas de la clase negativa.
- Existen 37,55 % de predicciones correctas de la clase positiva.
- Existen 37,58 % de predicciones correctas de la clase negativa.
- Existen 12,42 % de predicciones incorrectas de la clase positiva.

Del mismo modo que la matriz de confusión del modelo de Regresión Logística, la interpretación correspondiente a la problemática de este proyecto de título es análoga para esta metodología.

5.6.2.1. Métricas

- Precisión $= \dfrac{37,55}{37,55+12,42} = 75,15$
- Sensibilidad $= \dfrac{37,55}{37,55+37,58} = 49,98$
- Exactitud $= \dfrac{37,55+37,58}{37,55+12,42+12,44+37,58} = 75,14$

- Especificidad = $\dfrac{37,58}{12,42+37,58} = 75,16$
- Tasa de error = $\dfrac{12,42+12,44}{37,55+12,44+12,42+37,58} = 24,86$
- Valor de predicción negativa (VPN):

$$VPN = \frac{37,58}{50,02} = 75,13$$

- AUC = $\dfrac{100+75,15-24,84}{2} = 75,16$

Para el modelo de Random Forest con clases equilibradas, la interpretación es análoga al modelo anterior, por tanto, el porcentaje respecto a la clasificación de ambas clases predichas correctamente bordea el 75 % aproximadamente de exactitud, indicando ser bueno. Respecto a su sensibilidad, el modelo indicó que el porcentaje de casos clasificados como transacciones normales correctamente fue de un 50 % aprox., al contrario de la especificidad, donde el modelo indicó que el porcentaje de casos clasificados correctamente como transacciones fraudulentas fue de un 75,2 % aproximadamente.

La precisión del modelo además, entregó un valor del 75,2 % aprox. con una tasa de error del 25 % aproximadamente, siendo un modelo con resultados considerablemente buenos.

5.7. Comparación de modelos

Al abordar uno de los desafíos como lo es el desbalance de las clases, los métodos estadísticos aplicados respecto a sus medidas de rendimiento entregan resultados aparentemente favorecedores desde el punto de vista de la empresa bancaria.

Respecto a las medidas de rendimiento finales aplicada a la muestra proveniente de la base de entrenamiento de transacciones financieras, la diferencia entre el mejor y peor valor en términos de precisión global, especificidad, sensibilidad y exactitud es relativamente no significativa, es decir, no se encuentran grandes diferencias en los modelos en cuanto a sus métricas finales según se puede observar en el Cuadro 5.21 y 5.22.

	Presición	**Sensibilidad**	**Exactitud**	**Especificidad**
RL	72,31	51,74	73,47	70,90
RF	75,15	49,98	75,14	75,16

Cuadro 5.21: Comparación de métricas 1
(Fuente: elaboración propia)

	VPN	**Tasa de error (error rate)**	**AUC**
RL	74,73	26,53	71,61
RF	75,13	24,86	75,16

Cuadro 5.22: Comparación de métricas 2
(Fuente: elaboración propia)

Aunque los AUC de ambos modelos según el Cuadro 5.22, indican ser de tipo regulares, el mayor y más cercano a ser catalogado del tipo bueno es el de RF, es decir, existe un 75,15 % de probabilidad de que la detección de fraude realizada a una transacción clasificada como fraudulenta sea más correcta que la detección de fraude realizada a una transacción catalogada como normal. Sobre el VPN, se escoge el valor más alto siendo igualmente el de RF, indicando que de las transacciones detectadas como fraude, el 75,13 % efectivamente eran fraudulentas, mientras que existe una probabilidad del 24,87 % de que las transacciones detectadas como fraude sean transacciones normales.

Por lo tanto, considerando un equilibrio de las clases y con respecto a sus medidas de rendimiento, el mejor modelo en cuanto a precisión, exactitud, especificidad, tasa de error, VPN y AUC es aplicando la metodología Random Forest, es decir, el modelo predice mejor clasificando ambas clases (transacciones normales y fraudulentas) correctamente, con una diferencia del 0.016 % menos en su tasa de error con respecto al modelo RL. En caso contrario, en cuanto sólo a sensibilidad, es mejor aplicando la metodología Regresión Logística.

Capítulo 6

Conclusiones

Uno de los problemas principales en la detección de fraude es el desequilibrio del número de transacciones normales y fraudulentas, ya que se tiene una mayor cantidad de transacciones normales con respecto a la cantidad de transacciones fraudulentas, de las cuales muchas veces estas últimas no superan el 1 %.

Para la construcción de los modelos se utilizaron 16 bases correspondientes a periodos de 16 meses distintos, identificando en cada uno 4 tipo de cuadrantes existentes: A, B, C y D, siendo el enfoque primordial de este Proyecto de Título el escenario D por su "alto"porcentaje de fraude respecto a los demás escenarios, equivalente al 1,03 % respectivamente .

La base de datos extraída a posteriori se encuentra compuesta por 122 variables y 7.589.782 registros en total, que mediante un análisis de datos descriptivo y un tratamiento de valores faltantes de los datos, se reduce a una base final con 30 variables y 7.508.143 registros.

Respecto a la contrucción de variables, ellas se enfocan en dos ámbitos respectivamente, las variables acumuladores y las variables de número índice. Para aquellas variables acumuladoras, se consideraron distintas brechas de tiempo para su construcción, ya que se pensó que este tipo de creación de variables influye en la mejora de resultados respecto al poder predictivo provenientes de los modelos. De esta manera, el enfoque pensado se aplica al segundo tipo de variables mencionada, pues ellas ayudan a percibir las tendencias de los usuarios en los movimientos de sus transacciones. Así del total de variables construidas, se utilizó el indicador de Gini para observar qué variables previamente en-

trarían a los modelos, obteniendo un conjunto de 79 variables escogidas.

Los modelos propuestos para la detección de fraude fueron dos: Random Forest y Regresión Logística. En cuanto a tiempos operacionales y carga computacional se procedió a trabajar sólo con muestras de entrenamiento , obtenidas realizando un equilibrio de las clases de transacciones tanto fraudulentas como normales. Se compararon los modelos con diversas medidas de rendimiento, para capturar sus capacidades predictivas y de clasificación, obteniendo un mejor resultado la metolodía Random Forest.

Desde el punto de vista del negocio, se desarrolló una evidente metodología con las problemáticas claras para la detección de fraude, y aunque el modelo con mejores resultados requiere de la implementación de herramientas más sofisticadas a su sistema de detección; el modelo de Regresión Logística, ofrece no ser el mejor resultado pero si el de tener una implementación más amena. Además, se recomienda al negocio la depuración de las bases de datos, principalmente que las transacciones marcadas como fraudulentas o normales estén de manera correcta y correspondiente al tiempo de efectuada una transacción, para así en un futuro, si se aplica la implementación de modelos, su evaluación y el posterior mantenimiento y/o seguimiento, sea más limpia y rápida para el investigador o analista.

Finalmente y como conclusión general, si se considera el mundo de transacciones fraudulentas completa, es decir, sin fragmentarlas por rubro o tipos de fraude y equilibrando ambos números de transacciones normales con aquellas fraudulentas; los modelos tienden a obtener métricas favorables en la detección de fraude. Esto entrega una idea intuitiva, de que si se examinaran distintos modelos más complejos aplicados a la división de tipos de fraude, estos entregarían mejores resultados que un modelo sencillo como lo es la Regresión Logística.

6.0.1. Trabajos Futuros

Debido al gran volumen y manejo de datos trabajados, es que el tiempo utilizado para realizar este Proyecto de Título se hizo estrecho debido a la carga computacional que se tenía, lo que ocasionó que no se completaran ciertas tareas que complementarían este estudio y que hubieran sido importantes observar para la posterior comparación de resultados. Se recomienda analizar algunos de ellos:

1. Realizar un análisis estadístico completo a las variables construidas y utilizadas en la creación de los modelos, con el fin de observar su poder discriminante.

2. Replicar la aplicación de las metodologías propuestas en este estudio sin balancear las clases, para observar el comportamiento predictivo de los modelos.

3. Aplicar las metodologías propuestas al 40 % de la base de los datos (conjunto prueba) y comparar sus matrices de confusión con el conjunto entrenamiento.

4. Construir variables con brechas de tiempo mayores, es decir, variables que acumulen más historia, a modo de observar como hipótesis si ellas mejoran el poder predictivo de los modelos.

5. Implementar otras técnicas más sofisticadas para predecir la detección del fraude como por ejemplo Redes Neuronales, Gradient Boosting, entre otras.

Capítulo 7

Anexo

7.1. Resultados - Análisis descriptivo

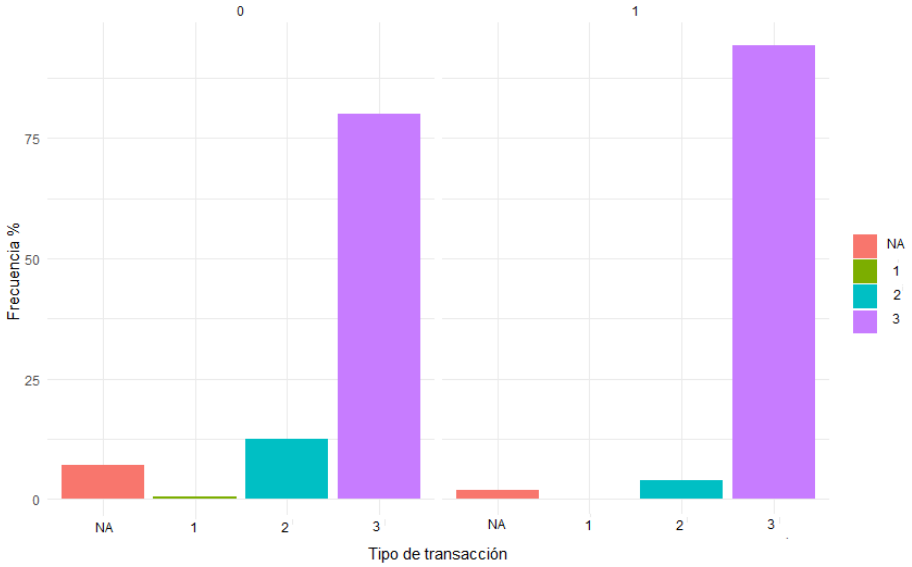

Figura 7.1: Distribución de la Variable V104
(Fuente: elaboración propia)

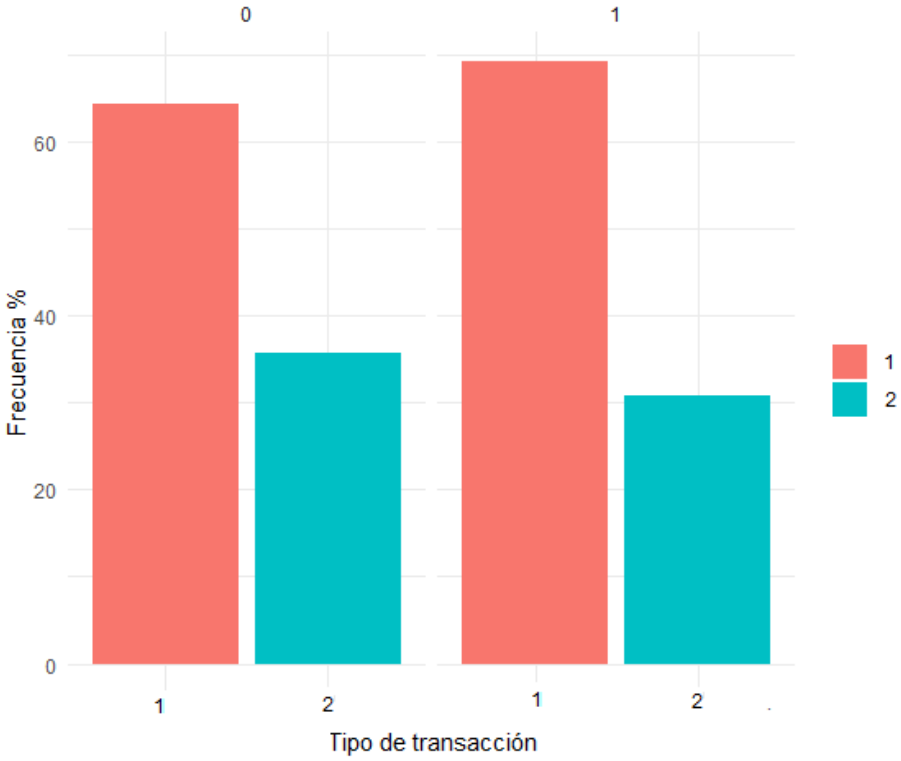

Figura 7.2: Distribución de la Variable V105
(Fuente: elaboración propia)

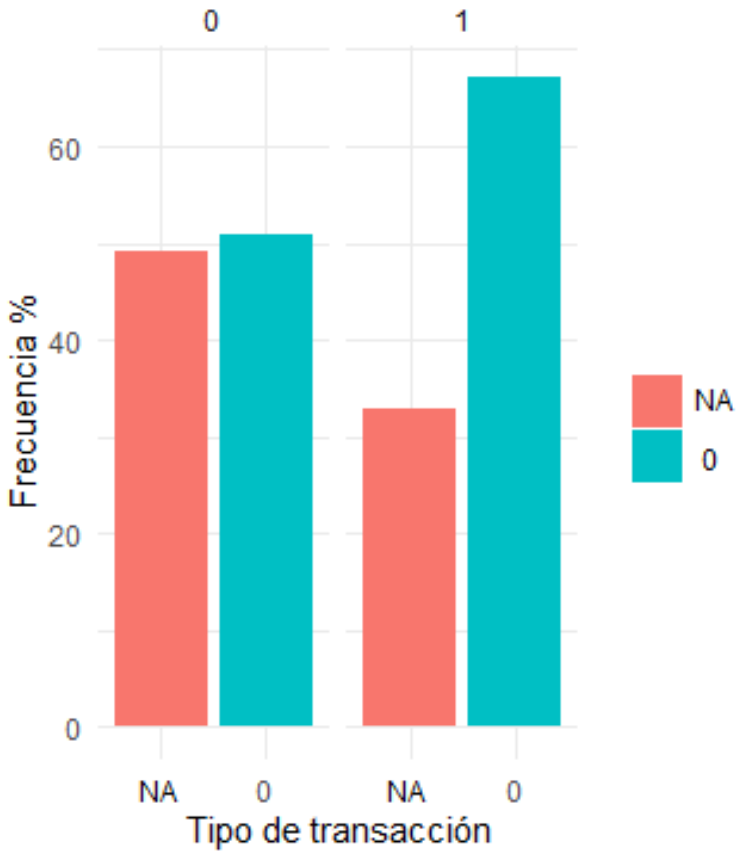

Figura 7.3: Distribución de la Variable V106
(Fuente: elaboración propia)

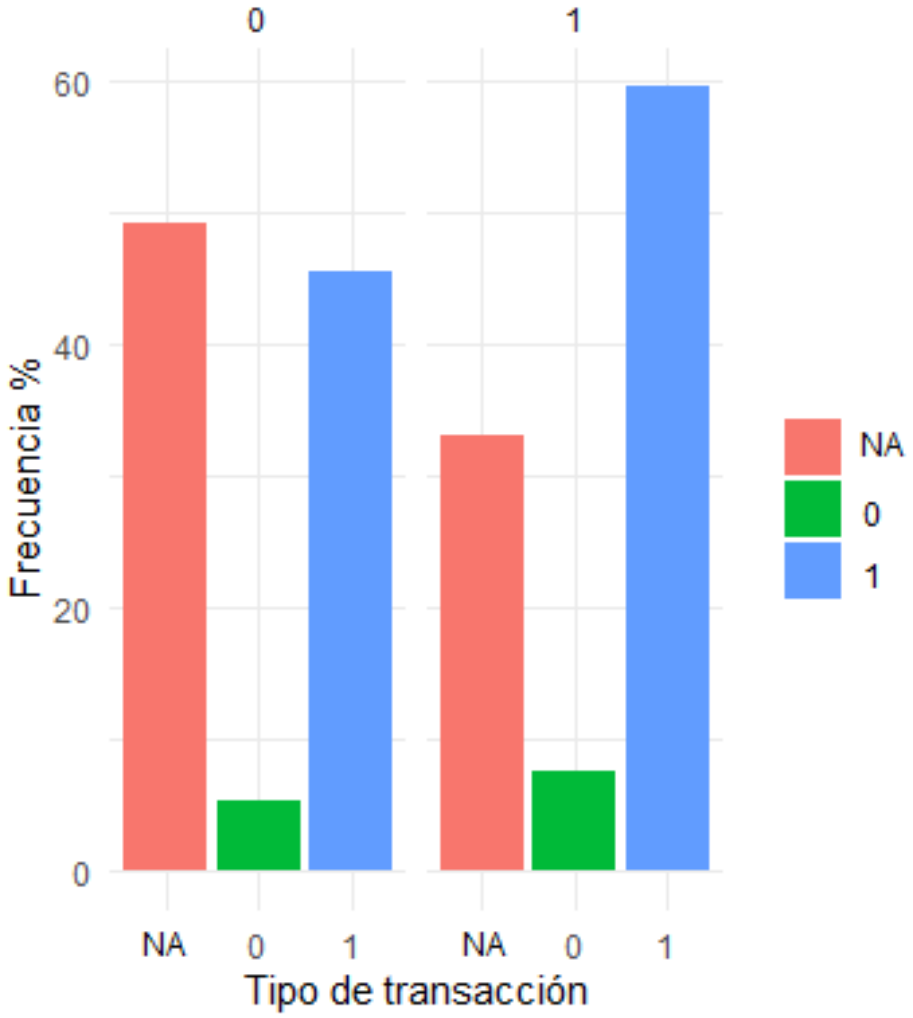

Figura 7.4: Distribución de la Variable V107
(Fuente: elaboración propia)

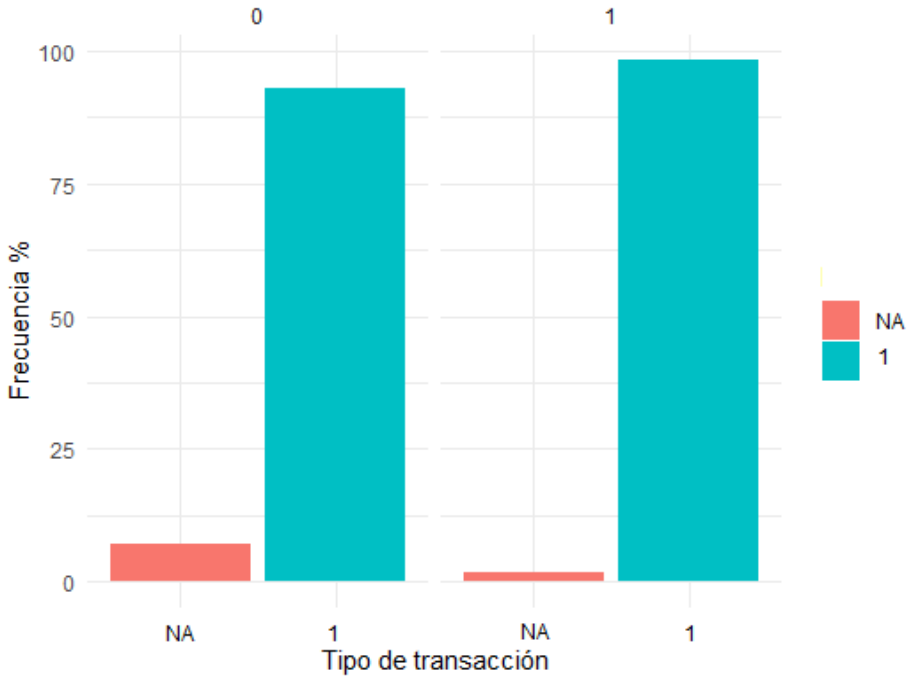

Figura 7.5: Distribución de la Variable V108
(Fuente: elaboración propia)

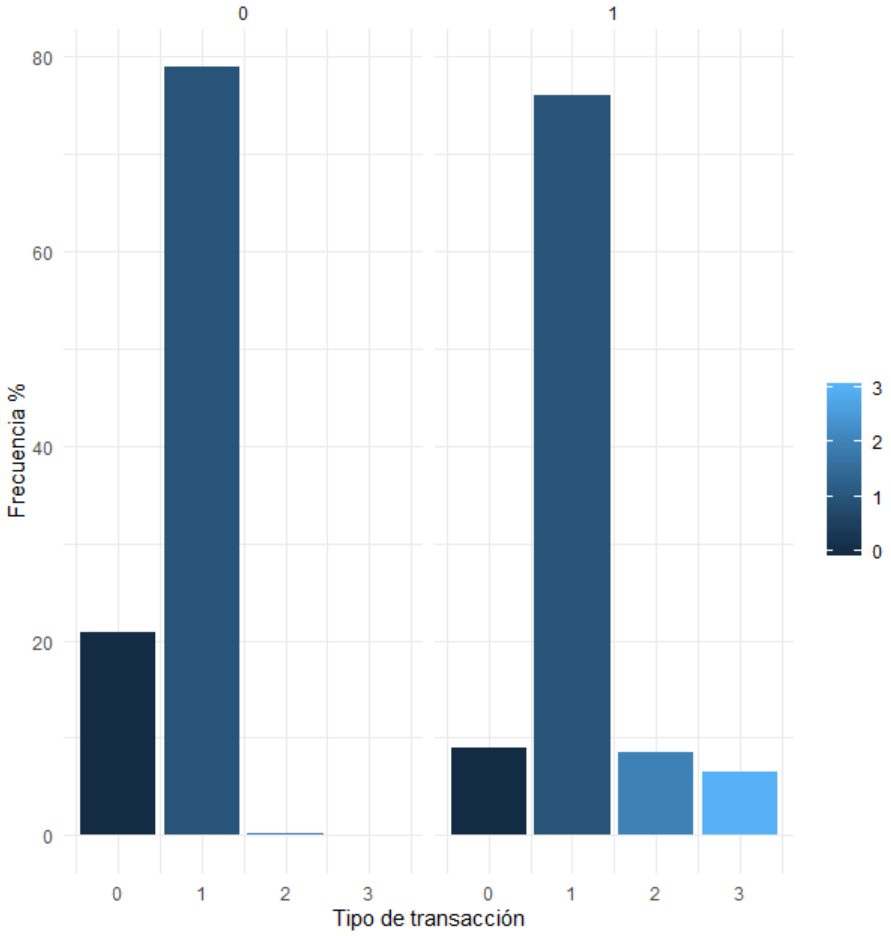

Figura 7.6: Distribución de la Variable V109
(Fuente: elaboración propia)

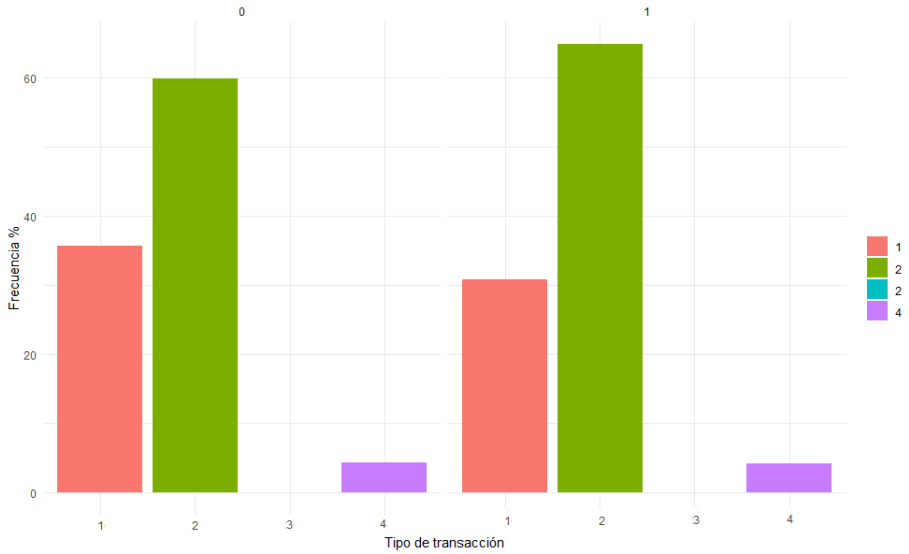

Figura 7.7: Distribución de la Variable V110
(Fuente: elaboración propia)

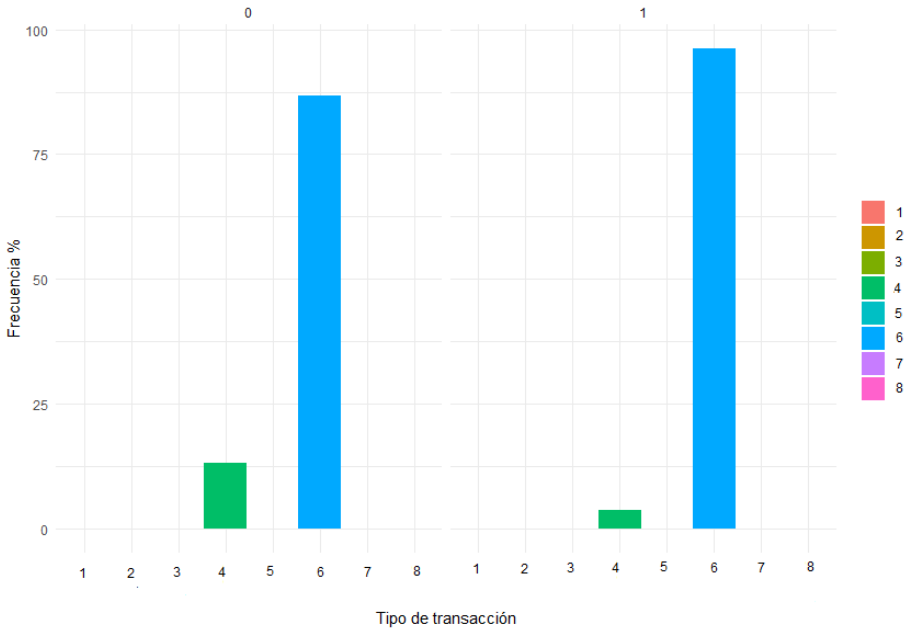

Figura 7.8: Distribución de la Variable V111
(Fuente: elaboración propia)

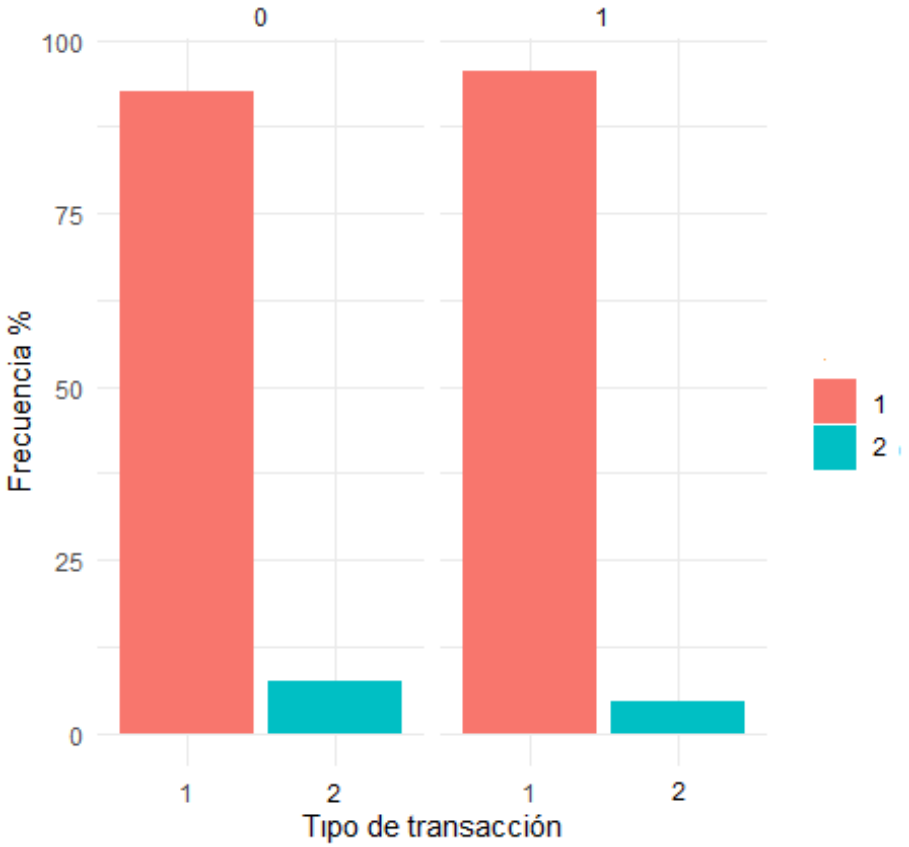

Figura 7.9: Distribución de la Variable V112
(Fuente: elaboración propia)

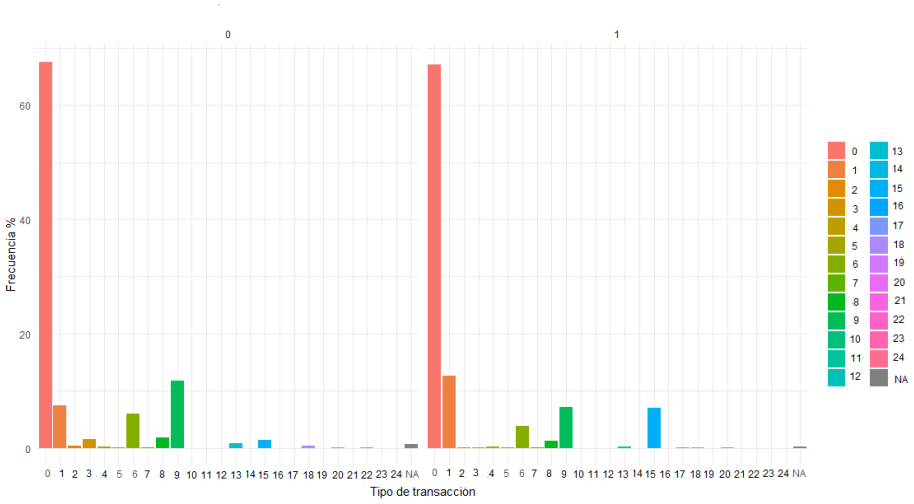

Figura 7.10: Distribución de la Variable V6
(Fuente: elaboración propia)

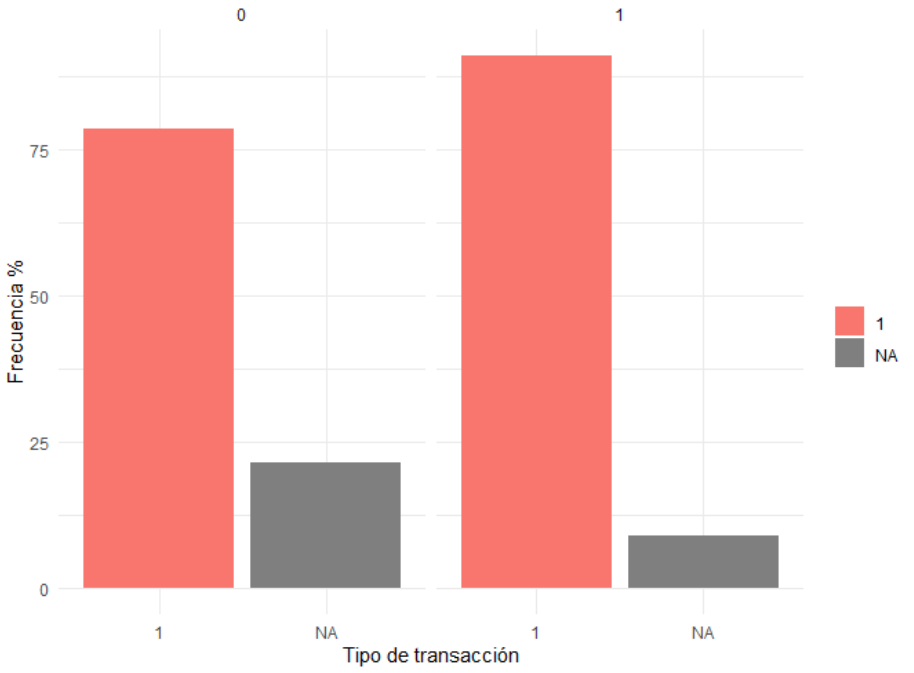

Figura 7.11: Distribución de la Variable V113
(Fuente: elaboración propia)

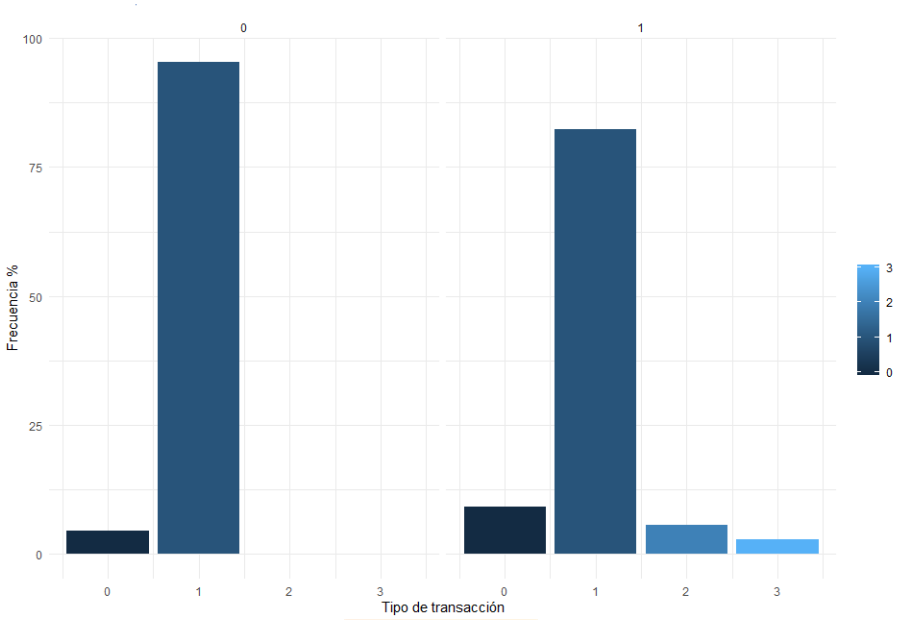

Figura 7.12: Distribución de la Variable V114
(Fuente: elaboración propia)

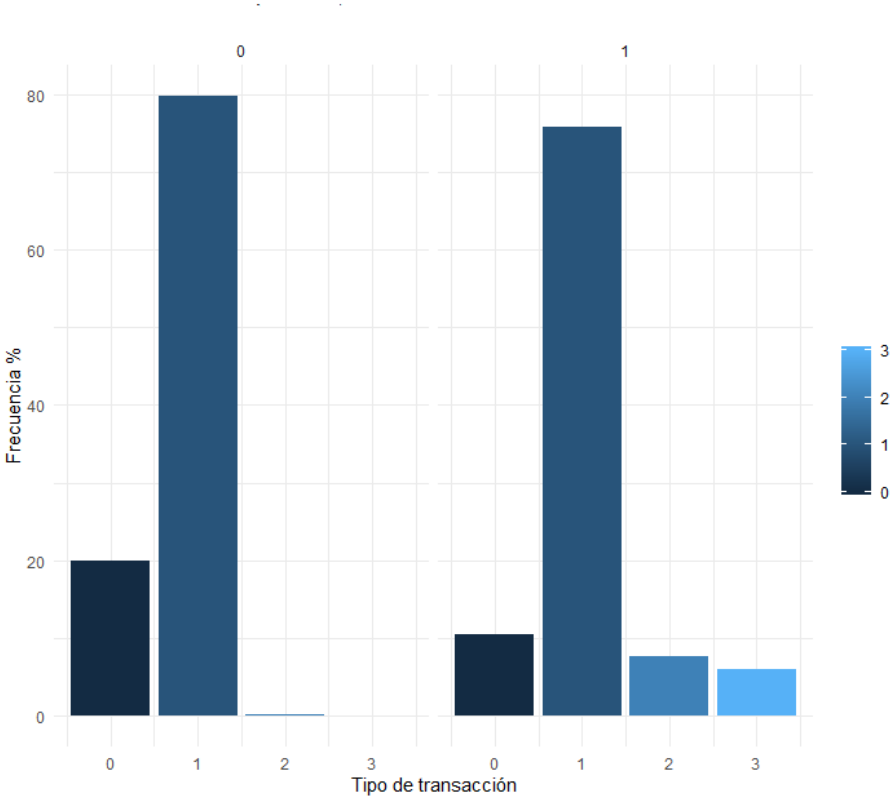

Figura 7.13: Distribución de la Variable V115
(Fuente: elaboración propia)

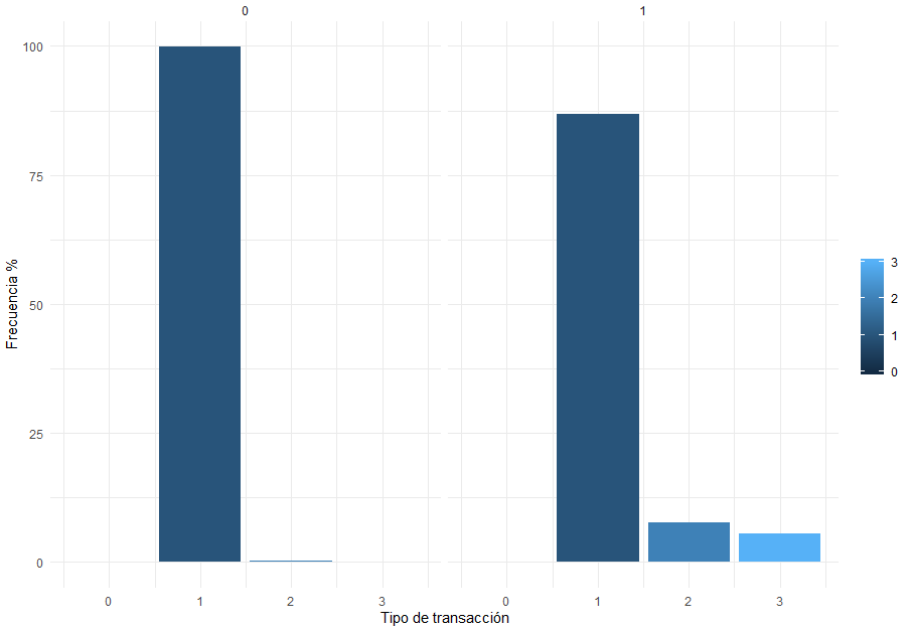

Figura 7.14: Distribución de la Variable V116
(Fuente: elaboración propia)

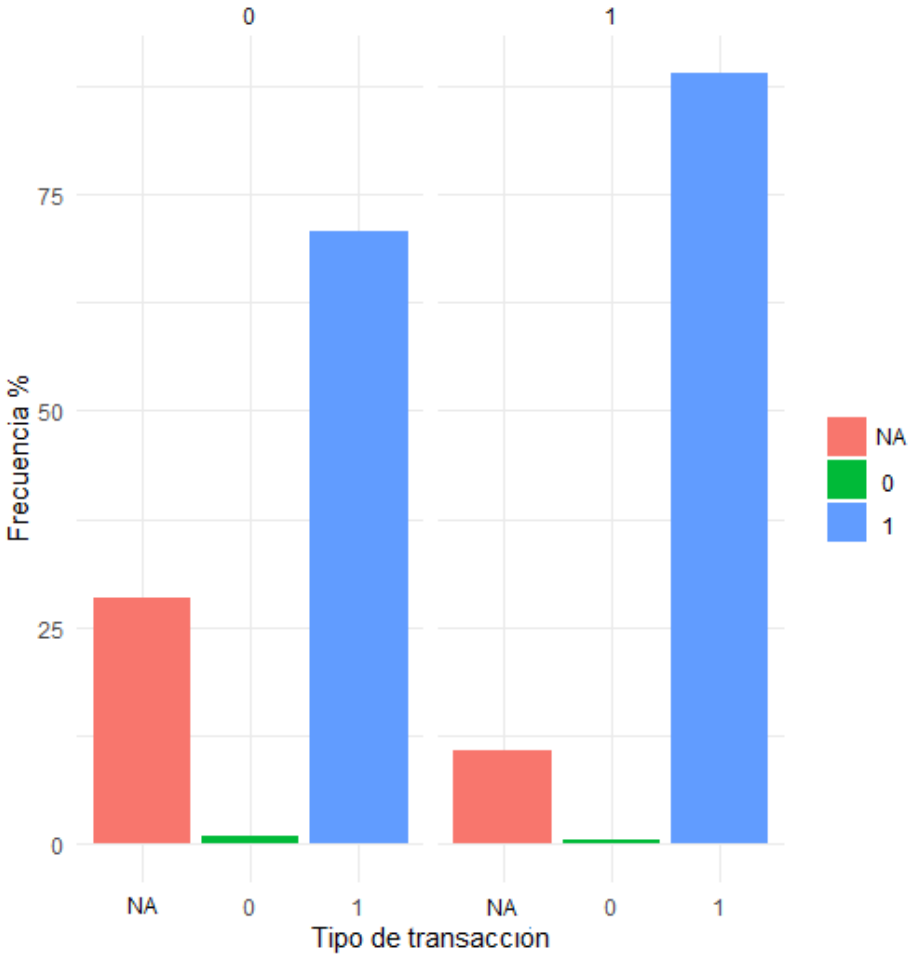

Figura 7.15: Distribución de la Variable V117
(Fuente: elaboración propia)

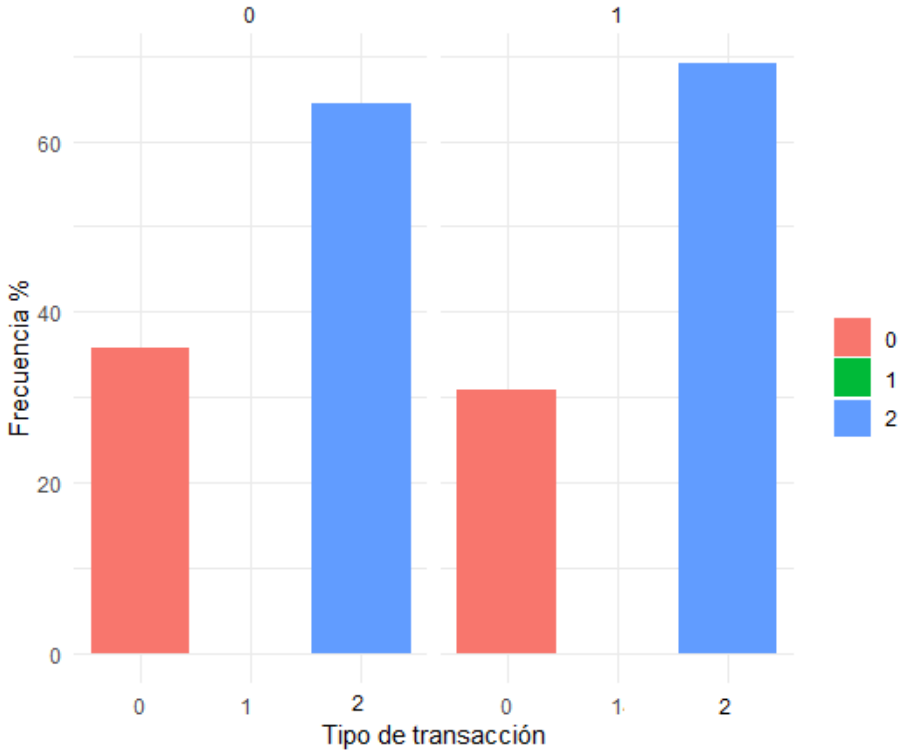

Figura 7.16: Distribución de la Variable V118
(Fuente: elaboración propia)

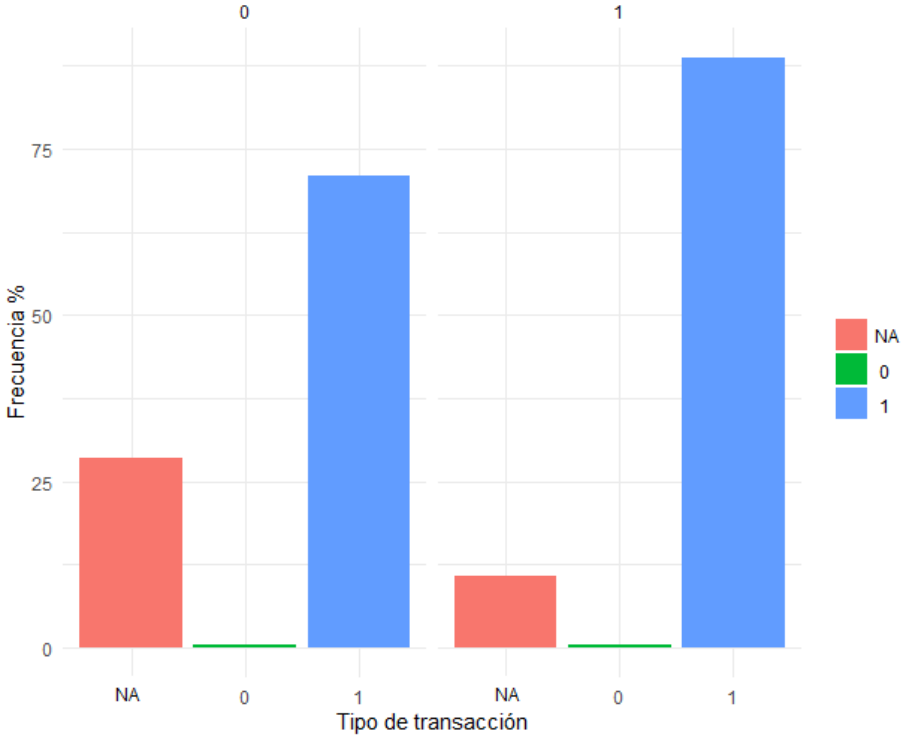

Figura 7.17: Distribución de la Variable V119
(Fuente: elaboración propia)

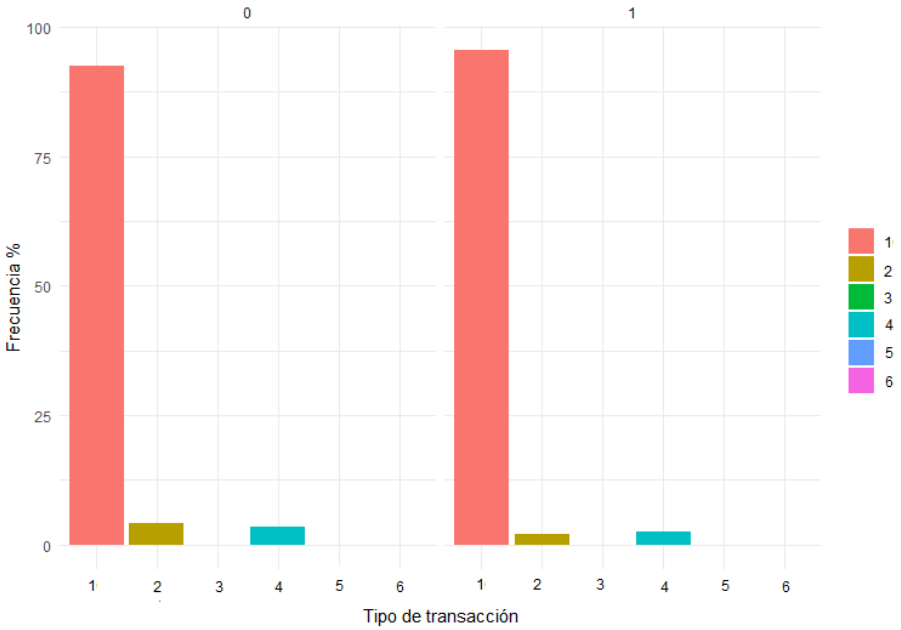

Figura 7.18: Distribución de la Variable V120
(Fuente: elaboración propia)

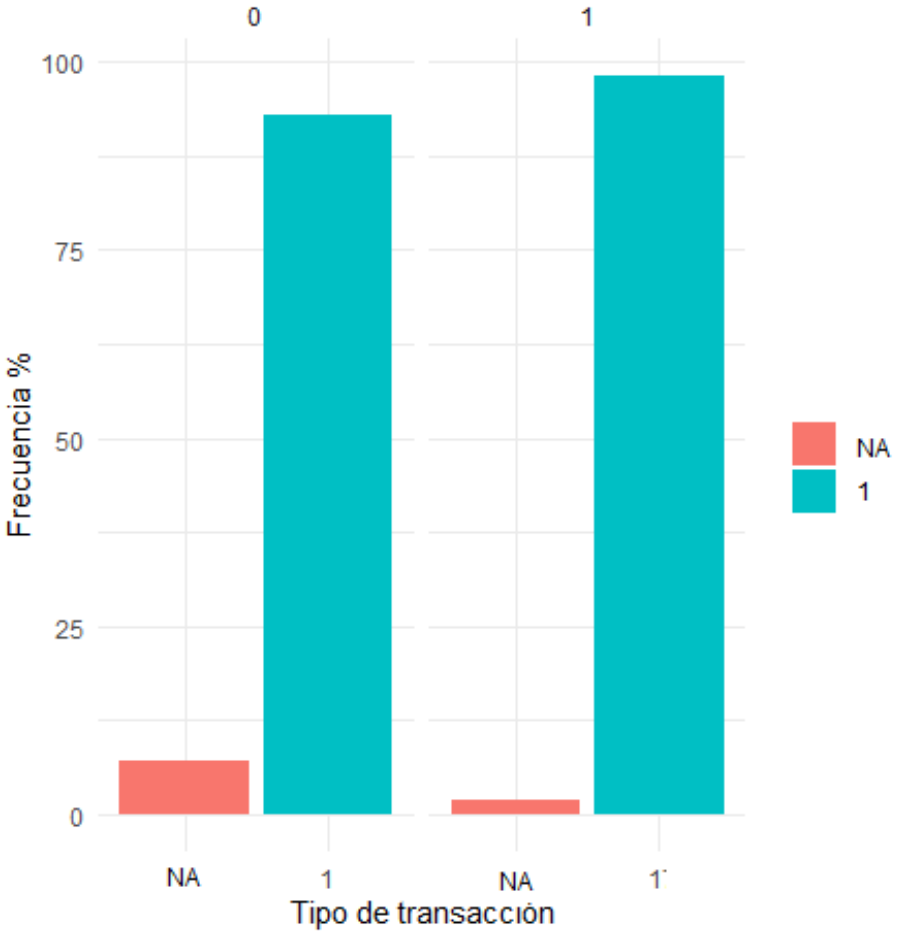

Figura 7.19: Distribución de la Variable V121
(Fuente: elaboración propia)

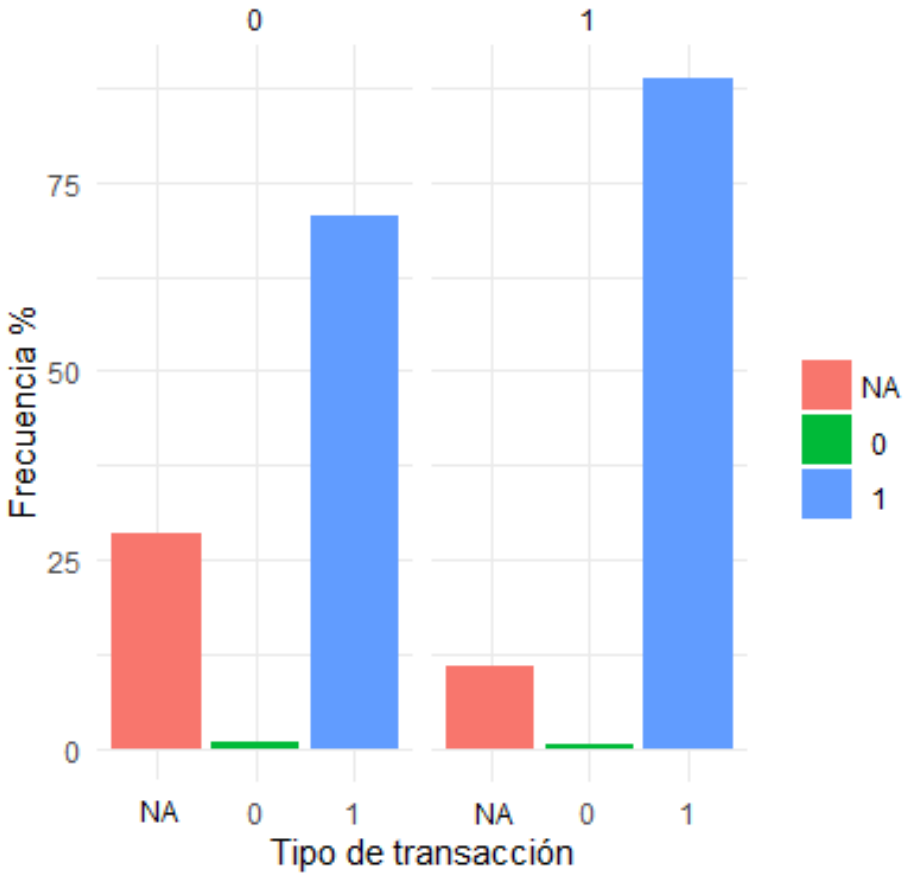

Figura 7.20: Distribución de la Variable V122
(Fuente: elaboración propia)

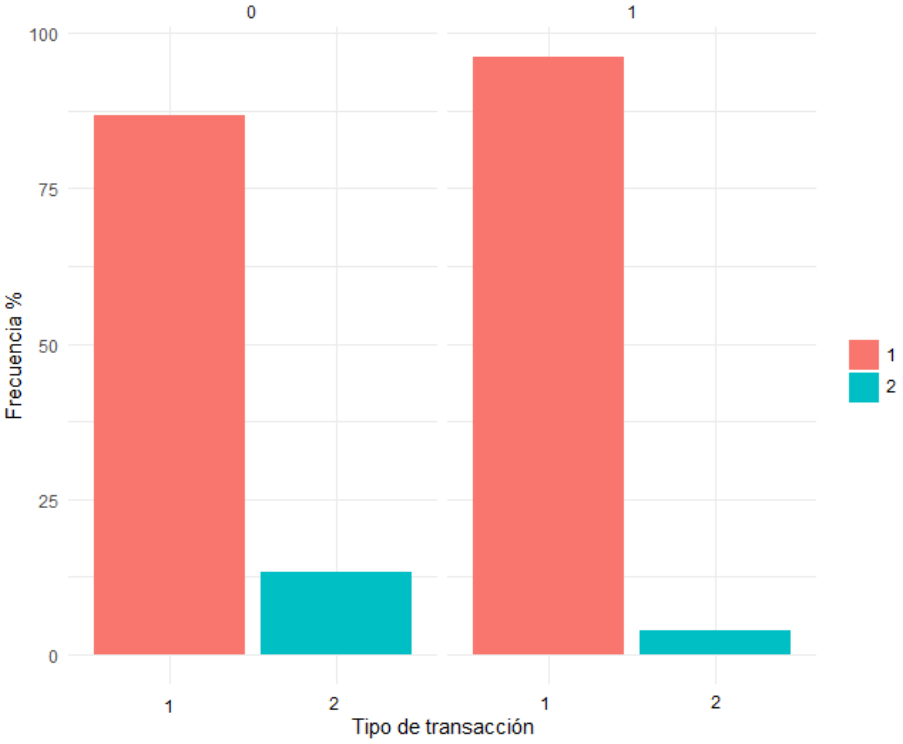

Figura 7.21: Distribución de la Variable V123
(Fuente: elaboración propia)

7.2. Técnicas de remuestreo y muestras sintéticas - Ejemplo

A continuación, se presentan los códigos con los que se obtuvieron los resultados en el ejemplo de la sección 2.2.3.3 para cada método aplicado.

7.2.1. Aplicación del sobremuestreo

```
library(unbalanced)
data(ubIonosphere)
n <- ncol(ubIonosphere)
salida <- ubIonosphere[ ,n]
summary(salida)
entrada <- ubIonosphere[ ,-n]
set.seed(1234)
data_0 <- ubOver(entrada, salida, k=0)
```

7.2.2. Aplicación del submuestreo

```
data <- ubUnder(entrada, salida, perc=50,
method="percPos")
sub_data <- data.frame(data$X,data$Y)
summary(sub_data$data.Y)
```

Bibliografía

[1] Amat, J. (2016). *Regresión logística simple y múltiple.* Recuperado de: `https://rpubs.com/Joaquin_AR/229736`

[2] Barrero, G. (2018). *Uso de la técnica Machine Learning de conjunto o agrupación para la predicción de la calidad del software desarrollado en IBM RPG.* doi: 10.13140/RG.2.2.16603.62246

[3] Breiman, L. (2001). *Random forests.* Machine Learning (vol. 45, pp. 5-32).

[4] Chawla, N., Japkowicz, N. y Kotcz, A. (2004). *Special issue on learning from imbalanced data sets.* ACM Sigkdd Explorations Newsletter, 6(1), 1-6.

[5] Cepeda, D. (2012). *Detección de fraude en tarjetas de crédito* (Memoria para optar al título de Ingeniero Civil Matemático). Universidad de Chile, Chile.

[6] Dal, A., Johnson, R., Caelen, O., Waterschoot, S., Chawla, N. y Bontempi, G. (2014, julio). *Using HDDT to avoid instances propagation in unbalanced and evolving data streams.* In 2014 International Joint Conference on Neural Networks (IJCNN) (pp. 588-594). IEEE.

[7] Dal, A., Caelen, O., Le Borgne, Y., Waterschoot, S. y Bontempi, G. (2014). *Learned lessons in credit card fraud detection from a practitioner perspective*. Expert systems with applications, 41(10), 4915-4928.

[8] Dal, A., Boracchi, G., Caelen, O., Alippi, C. y Bontempi, G. (2015, julio). *Credit card fraud detection and concept-drift adaptation with delayed supervised information.*In 2015 international joint conference on Neural networks (IJCNN) (pp. 1-8). IEEE.

[9] Dal, A. (2015). *Adaptive Machine Learning for Credit Card Fraud Detection* (Memoria para optar a la licenciatura del doctorado en informática). Universidad Libre de Bruselas, Bélgica.

[10] Dal, A., Olivier, C. y Gianluca B. (2015). *unbalanced: Racing For Unbalanced Methods Selection.* Recuperado de: `http://CRAN.R-project.org/package=unbalanced.Rpackageversion2.0.`

[11] Delgado, J. (2006). *Redes neuronales*. Recuperado de: `https://www.monografias.com/trabajos38/redes-neuronales/redes-neuronales2.shtml`

[12] Drummond, C. y Holte, R. (2003). *C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling.*In Workshop on learning from imbalanced datasets II (Vol. 11, pp. 1-8). Washington, DC: Citeseer.

[13] Estupiñan, R. (2006). *Control interno y fraudes: Análisis de informe COSO I, II y III con base en los ciclos transaccionales.* Bogotá: Eco Ediciones Ltda.

[14] Fernández, E., Rodríguez, J. y Merino, H. (2014). *Comportamiento Adaptable de Chatbots Dependiente del Contexto*. Recuperado de: `https://www.researchgate.net/figure`

[15] González, E. (2018). *Detección de Fraude en Tarjetas de Crédito Mediante Técnicas de Minería de Datos* (Trabajo de grado).Universidad Santo Tomas, Colombia.

[16] Han, J., Kamber, M. y Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier Science.

[17] Hart, P. (1968). *The condensed nearest neighbor rule (Corresp.)*. IEEE transactions on information theory, 14(3), 515-516.

[18] Hernández, E. (2006). *Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto* (Tesis de grado). Centro de investigación y de estudios avanzados, México.

[19] James, G., Witten, D., Trevor, H. y Tibshirani, R. (2017). *An Introduction to Statistical Learning*. doi: 10.1007/978-1-4614-7138-7

[20] Keider, J. (2019). *Metodología de clasificación de datos desbalanceados basado en métodos de submuestreo* (Tesis de Magister). Universidad Tecnológica de Pereira, Colombia.

[21] Kotsiantis, S., Kanellopoulos, D. y Pintelas, P. (2006). *Handling imbalanced datasets: A review*.GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

[22] Kotsiantis, S., Zaharakis, I. y Pintelas, P. (2007). *Supervised machine learning: A review of classification techniques.*Emerging artificial intelligence applications in computer engineering, 160, 3-24.

[23] Kubat, M. y Matwin, S. (1997). *Addressing the curse of imbalanced training sets: one-sided selection.*In Icml (Vol. 97, pp. 179-186).

[24] KPMG (2013). *Encuesta de Fraude en Colombia 2013.* Recuperado de: https://docplayer.es/4401402-Encuesta-de-fraude-en-colombia-2013.html

[25] KPMG (2017). *Encuesta de Fraude en Colombia 2017.* Recuperado de: https://public.tableau.com/profile/kpmgco#!/vizhome/EncuestadeFraudeenColombia2017/Historia1

[26] Laurikkala, J. (2001). *Improving identification of difficult small classes by balancing class distribution.*In Conference on Artificial Intelligence in Medicine in Europe (pp. 63-66). Springer, Berlin, Heidelberg.

[27] Maass, S., Regil, H., González, C. y Nava, G. (2006). *Cambio de uso del suelo y vegetación en el Parque Nacional Nevado de Toluca, México, en el periodo 1972-2000.* Investigaciones geográficas, (61), 38-57. Recuperado de: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-46112006000300004

[28] López, V., Fernández, A., García, S., Palade, V. y Herrera, F. (2013). *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics.* Information Sciences (vol. 250, pp. 113-141).

[29] Parra, F. (2017). *Estadística y Machine Learning con R.*
Recuperado de: https://rpubs.com/PacoParra/293405

[30] Robologs (2017). *Red neuronal con varias capas ocultas*. Recuperado de:
https://robologs.net/2017/01/22/tutorial-de-redes-neuronales-con-vrep-c-y-linux/#comments

[31] Rodríguez, A. ,Rivero, D. y Spangenberg, D. (2006). *Análisis y detección de patrones de fraude en medios de pago* (Informe de proyecto de grado).Universidad de la República, Uruguay.

[32] Ruiz, S. (2014). *Random Forests para detección de fraude en medios de pago* (Tesis de Magister). Universidad Autónoma de Madrid, España.

[33] Sandoval, R. (1991). *Tarjeta de crédito bancaria.* Chile: Editorial jurídica de Chile.

[34] Santamaria, W. (2006). *Técnicas de Minería de Datos Aplicadas en la Detección de Fraude: Estado del Arte.* Recuperado de:
https://www.researchgate.net/publication/240724702_Tecnicas_de_Mineria_de_Datos_Aplicadas_en_la_Deteccion_de_FraudeEstado_del_Arte

[35] Schiffman, L. y Kanuk, L. (2005). *Comportamiento del consumidor.* México: Pearson Educación.

[36] Silva, L. (2004). *Regresión Logística.* España: La Muralla, S.A.

[37] Tan, P., Steinbach, M. y Kumar, V. (2006). *Introduction to Data Mining.* Addison Wesley.

[38] Tomek, I. (1976). *Two modifications of CNN.*IEEE Trans. Systems, Man and Cybernetics, 6, 769-772.

[39] Trevor, H., Tibshirani, R. y Friedman, J. (2017). *The elements of Statistical Learning.* doi: 10.1007/b94608

[40] Valle, J. (2013). *El delito informático de Phishing* (Tesis de Magister). Universidad Regional Autónoma de Los Andes.

[41] Wilson, D. (1972). *Asymptotic properties of nearest neighbor rules using edited data.*IEEE Transactions on Systems, Man, and Cybernetics, (3), 408-421.

[42] Zhou, Z. (2012). *Ensemble Methods. Foundations and algorithms.* Estados Unidos.Taylor Francis Group.