UNIVERSIDAD NACIONAL
PEDRO RUIZ GALLO

FACULTAD DE CIENCIAS FÍSICAS Y MATEMATICAS

ESCUELA PROFESIONAL DE ESTADÍSTICA

TESIS

"MODELO DE CREDIT SCORING PARA PREDECIR EL OTORGAMIENTO DE CRÉDITO PERSONAL EN UNA COOPERATIVA DE AHORRO Y CRÉDITO"

PARA OPTAR EL TÍTULO PROFESIONAL DE:

LICENCIADO EN ESTADÍSTICA

Autores

Medina Rodríguez, María del Pilar

Ulfe Rentería, Henry Gustavo

Asesor

M.Sc. Alfonso Tesén Arroyo

Lambayeque - Perú

2015

UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

FACULTAD DE CIENCIAS FÍSICAS Y MATEMATICAS

ESCUELA PROFESIONAL DE ESTADÍSTICA

TESIS

"MODELO DE CREDIT SCORING PARA PREDECIR EL OTORGAMIENTO DE CRÉDITO PERSONAL EN UNA COOPERATIVA DE AHORRO Y CRÉDITO"

PARA OPTAR EL TÍTULO PROFESIONAL DE

LICENCIADO EN ESTADÍSTICA

Aprobado por el siguiente jurado:

Lic. Est. HUGO LORGIO SAAVEDRA SAAVEDRA
Presidente

M.Sc. LILIAN ROXANA PAREDES LÓPEZ
Secretario

Lic. Est. LUIS ENRIQUE TUÑOQUE GUTIÉRREZ
Vocal

“MODELO DE CREDIT SCORING PARA PREDECIR EL OTORGAMIENTO DE CRÉDITO PERSONAL EN UNA COOPERATIVA DE AHORRO Y CRÉDITO”

Bach. MARÍA DEL PILAR MEDINA RODRÍGUEZ
Autor

Bach. HENRY GUSTAVO ULFE RENTERÍA
Autor

M.Sc. ALFONSO TESÉN ARROYO
Asesor

Lambayeque, Diciembre del 2015

DEDICATORIA

Dedico esta tesis a Dios por bendecirme, por siempre estar a mi lado ayudándome a cumplir mis objetivos y metas.

A mi familia, en especial a mis padres por todo el apoyo y cariño que me brindan a diario, hoy puedo retribuir en cierta medida tanto esfuerzo y amor.

A mi abuelo José Luis Rodríguez Arrasco^{+}, con esta tesis cumplo mi palabra de que me veas realizada profesionalmente.

Y todos aquellos amigos, compañeros y conocidos que me dieron un voto de confianza y que siempre creyeron en mí.

Pilar Medina.

Dedico esta tesis a Dios porque ha estado conmigo a cada paso que doy, cuidándome y dándome fortaleza para continuar.

A mi madre Rosa Elena Rentería Hernández, a mi padre José Gustavo Ulfe Tepe y mi hermano Franklin Manuel Ulfe Rentería, con esta tesis cumplo mi palabra de que me vean realizado profesionalmente.

A mis amigos, compañeros y a todas aquellas personas que permitieron hacer posible la realización de este trabajo.

Henry Ulfe.

AGRADECIMIENTOS

En primer lugar agradecer a Dios por habernos dado sabiduría y entendimiento para poder concluir nuestra carrera, por bendecirnos y proveernos de lo necesario para salir adelante, gracias por hacer realidad este sueño anhelado.

A nuestra alma mater, la “Universidad Nacional Pedro Ruiz Gallo”, que nos abrió sus puertas para formarnos profesionalmente, gracias por habernos acogido y albergado durante cinco años de estudio.

A todos nuestros maestros que marcaron cada etapa de nuestro camino universitario en especial al Mg. Alfonso Tesén Arroyo por su apoyo como asesor, al Lic. Hugo Saavedra Saavedra, Lic. Luis Enrique Tuñoque Gutiérrez y a la M.Sc. Roxana Paredes López que con sus aportes y experiencias corrigieron algunos puntos tratados en esta tesis que debían mejorarse.

Al M.Sc. Luis Ángel Cajachahua Espinoza por su apoyo y guía durante la elaboración de la tesis, gracias por la disponibilidad y paciencia. Ha sido un privilegio contar con su asesoría y conocimientos.

Por supuesto, el agradecimiento más profundo y sentido a nuestras familias, que sin su apoyo y motivación habría sido imposible llevar a cabo esta tesis. Y por último, gracias a todas esas personas que siempre nos brindaron su ayuda y que, de una manera u otra, han sido claves en nuestra vida profesional y personal.

Con todo nuestro cariño, esta tesis es por ustedes.

Los autores.

RESUMEN

En los últimos años, la Cooperativa de Ahorro y Crédito presenta dificultades en cuanto a la concesión de créditos personales; el retraso de los clientes en las fechas de pago y en algunos casos el incumplimiento de la deuda son causantes de una morosidad variable, que no es más que el reflejo de una ineficiencia en la asignación de créditos y la necesidad de ajustar los criterios de evaluación. Por ello, la siguiente investigación tiene como objetivo la construcción de un modelo scoring que permita predecir el otorgamiento de crédito personal, con la finalidad de clasificar a los clientes a partir de la probabilidad de default. Metodológicamente la investigación es aplicada, con propósito predictivo y explicativo, basada en el proceso CRISP-DM para el desarrollo del proyecto. Las técnicas utilizadas fueron la Regresión Logística, Árboles de Clasificación y Redes Neuronales; la comparación de los modelos se realizó considerando las capacidades de clasificación y predicción, eligiendo como mejor modelo a la Regresión Logística por Agrupación Interactiva (R.L.A.I) por presentar una curva de ROC con 0.792, un GINI de 0.584 y una respuesta capturada del 30.8%. Posteriormente se realizó el scoring de los clientes estableciendo los puntos de corte de acuerdo a los objetivos de la institución, asignando un 30% para rechazo automático, 5% para análisis manual y un 65% para aprobación automática. Finalmente se concluyó que el credit scoring propuesto es una herramienta útil en la evaluación del sujeto a manera de sugerencia de aceptación o rechazo de una futura solicitud de crédito, permitiendo identificar con una mayor eficiencia a qué clientes se les puede otorgar crédito, logrando así la automatización y optimización del proceso crediticio en la institución, previniendo el sobreendeudamiento e incumplimiento.

Palabras clave: Credit Scoring, CRISP-DM, Regresión Logística, Árboles de Decisión, Redes Neuronales

ABSTRACT

In the last years, the credit union has been presenting difficulties in granting personal credits, the customers delays in the payment dates and in some cases the failure to meet debts are causing a default rate variable, that it only reflects the inefficiency in the assignation of the credits and the necessity to adjust the evaluation criteria. Therefore, the objective of this research is to construct a scoring model which allows to predict granting of personal credit in order to classify its customers starting from the probability of default. Methodologically, the investigation is applied, with predictive and explanatory purpose, based on the process CRISP-DM for the development of the project. The techniques used were the Logistics Regression, Decision Trees and Neural Networks, the comparison of the models was performed considering the capacities in classifying and prediction, being chosen the logistic regression by interactive grouping (R.L.A.I.) as the best model by having the ROC curve with 0.792, a GINI coefficient of 0.584 and a cumulative % captured response of 30.8. After that, the customer scoring was performed by setting cut-off points according to the objectives of the institution, allocating 30% to automatic rejection, 5% for manual analysis and 65% for automatic approval. Finally we concluded that the proposed credit scoring is a useful tool in the evaluation of the subject as a suggestion of acceptance or rejection of a future application for credit, allowing greater efficiency to identify which customers can be granted credit, thus achieving the automation and optimization of the credit process in the institution, preventing indebtedness and default.

Keywords: Credit Scoring, CRISP-DM, Logistics Regression, Decision Trees, Neural Networks

INDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

I. INTRODUCCIÓN

La cooperativa de ahorro y crédito enfrenta mes a mes la existencia de un nivel de riesgo, el retraso de sus clientes con las fechas de pagos y en algunos casos la incobrabilidad de las operaciones de crédito son motivos por los cuales presenta un nivel de morosidad variable (El año 2010 alcanzó un 7.94% que disminuyó a un 6.76% al cierre del 2011, a fines del 2012 alcanzó un 9.75% que disminuyó a un 4.79% a junio del 2013); esta situación demuestra una ineficiencia en la asignación de créditos y la necesidad de ajustar los criterios de evaluación por lo que se planteó como pregunta: ¿Cuál es el modelo de credit scoring que permite predecir el otorgamiento de crédito personal en una cooperativa de ahorro y crédito?

La investigación tuvo como objetivo general construir un modelo de credit scoring para la cartera de crédito personal; asimismo como objetivos específicos se buscó: Identificar al menos una variable explicativa que pueda ser considerada en la construcción de un modelo credit scoring, Construir modelos mediante las técnicas de Regresión Logística, Arboles de Clasificación y Redes Neuronales que permitan predecir la probabilidad de incumplimiento del crédito personal, Determinar el mejor modelo basado en indicadores eficiencia y predictibilidad y por último, Estimar el score de cada cliente y determinar los puntos de corte para clasificar los créditos personales en intervalos de aceptación, análisis manual y rechazo.

En términos prácticos, un modelo de credit scoring favorecerá no solamente a la institución sino también a los clientes; entre los beneficios se destaca el poder mejorar la calidad de la cartera crediticia y el servicio de crédito personal que se brinda, ayudar al analista de crédito en la toma de decisiones de forma más rápida y objetiva, y la predicción del

comportamiento de futuros créditos con el fin de aumentar utilidades y reducir costos.

Los modelos de credit scoring han sido extensamente estudiados y usados por un sinfín de instituciones financieras mostrando exitosos resultados en todo el mundo: (Bensic, Sarlija, & Zekic-Susac, 2005) concluyen que las medidas de asociación estadística muestran que el mejor modelo de redes neuronales está mejor relacionado con los datos que los modelos de RL Y CART, (Vigo, 2010) sostiene que no existe diferencias en el porcentaje de error de entrenamiento en los métodos de regresión logística y árbol de clasificación CART utilizados para la clasificación de los clientes que solicitan un préstamo, sin embargo con las redes neuronales se obtuvo un 84.18% de buena clasificación y un 74.32% de buena predicción; y (Soltan & Mohammadi, 2012) concluyen que el modelo de evaluación del riesgo de crédito por redes neuronales NN se comporta razonablemente bien al tener 91.44% de precisión y un 8% de error de tipo I y 0,4% error tipo II, diferencia sustancial en comparación con otros modelos de clasificación.

Dado que los estudios citados mostraron como mejor modelo a las Redes Neuronales, los investigadores decidieron construir distintos tipos de modelos para determinar cuál es el que mejor se ajusta al conjunto de datos. Cabe resaltar que el desarrollo de un credit scoring es pertinente porque permitirá evaluar al cliente mediante un score para clasificarlo como un buen o mal pagador en un futuro crédito personal, contribuyendo con la disminución del índice de morosidad y realización de un adecuado proceso de otorgamiento de crédito.

En la investigación se estableció como hipótesis general que existe al menos un modelo capaz de estimar la probabilidad de que un cliente incumpla con el crédito personal; como hipótesis específicas se planteó que existe al menos una variable explicativa significativa para la

construcción del modelo de predicción de incumplimiento de pago, y que existe al menos un mejor modelo basado en indicadores de eficiencia y predictibilidad.

El presente documento se divide en seis apartados: el capítulo uno abarca problema, objetivos, importancia, justificación e hipótesis que se desean validar. El capítulo dos comprende todo el marco teórico que conforma el estudio. En el capítulo tres se detalla los materiales y métodos, tipo de investigación, población, técnicas de recolección de datos y el análisis estadístico realizado; finalmente en los capítulos cuatro, cinco y seis se muestran los resultados, conclusiones y sugerencias obtenidas durante la investigación.

II. MARCO TEÓRICO

2.1. CRÉDITO , RIESGO y RIESGO DE CRÉDITO

2.1.1. Crédito

En el ámbito bancario, el contrato de crédito es aquel a través del cual la entidad financiera se obliga a poner a disposición del cliente una cantidad de dinero pactada en unas determinadas condiciones y en un cierto plazo. El cliente o acreditado podrá disponer o no de la cantidad estipulada en contrato según sus necesidades financieras. En todo caso, sólo tendrá que pagar intereses por el crédito dispuesto y no por el total disponible. Las operaciones de crédito más comunes son la cuenta de crédito o póliza de crédito y las tarjetas de crédito.

Así como no todas las entidades financieras son iguales (existen bancos, cajas, prestamistas privados, etc.), *no todos los créditos son iguales*. Existen tipos de créditos distintos y acordes a las distintas necesidades de cada persona o empresa.

Según el portal (Emprendedor.pe, 2013), la SBS reconoce 8 tipos distintos de créditos en el Perú:

- **Créditos corporativos:** Otorgados a personas jurídicas cuyas ventas son de al menos s/. 200 millones al año; para que se otorgue este tipo de crédito es indispensable que este monto sea real al menos en los dos últimos años antes de solicitar el crédito.
- **Créditos a grandes empresas:** Otorgados a personas jurídicas con ventas anuales mayores a s/. 20 mil pero menores a s/. 200 millones en los dos últimos años antes de solicitar el crédito.
- **Créditos a medianas empresas:** Dirigidos a personas jurídicas que tengan un endeudamiento de al menos s/. 300 mil en el

Sistema Financiero en los últimos seis meses y que no cumplen con las características para ubicarse entre los corporativos y las grandes empresas; asimismo, este crédito se otorga a las personas naturales con deudas que no sean hipotecarias mayores a s/.300 mil en el SF en los últimos seis meses siempre y cuando parte de este crédito este destinado a pequeñas empresas o microempresas.

- **Créditos a pequeñas empresas:** Otorgados a personas naturales o jurídicas para fines de prestación de servicios, comercialización o producción, cuyo endeudamiento en el SF sea de al menos s/. 20 mil y menor a s/. 300 mil en los últimos seis meses.
- **Créditos a microempresas:** Otorgados a personas jurídicas o naturales para fines de iguales a los de las pequeñas empresas, salvo que en este caso el endeudamiento en el SF debe ser menor de s/. 20 mil.
- **Créditos de consumo revolvente:** El crédito revolvente se refiere a que estos créditos pueden ser pagados por un monto inferior al de la factura, acumulándose la diferencia para posteriores facturas. Se otorgan a personas naturales con la finalidad de pagar servicios, bienes o deudas no empresariales.
- **Créditos de consumo no revolvente:** La diferencia con este crédito está en que en este caso se debe pagar por el mismo monto facturado y no de manera diferida.
- **Créditos hipotecarios:** Estos créditos se otorgan a personas naturales para la compra, construcción, reparación, remodelación, ampliación, etc., de vivienda propia siempre y cuando esos créditos se amparen en hipotecas inscritas.

2.1.2. Riesgo

Desde un punto de vista amplio se define al riesgo como la contingencia, probabilidad o proximidad de un peligro o daño. Desde el punto de vista económico, el riesgo es la pérdida financiera que el inversor debe valorar a la hora de realizar una inversión. Esta pérdida puede deberse a diferentes aspectos, y según éstos, se pueden definir diferentes tipos de riesgo, como por ejemplo: el riesgo de tipo de interés, el riesgo de divisa, el riesgo de volatilidad, el riesgo de precio o el riesgo de crédito, entre otros.

El Comité de Cooperativas Financieras (2011) clasifica los tipos de riesgos en:

a) **Riesgo de Mercado:** Se refiere a la variabilidad en los ingresos generados por la variación de precio de activos intercambiados en los mercados financieros (tasas de interés, tipo de cambio, índices de precios, acciones, etc.) los cuales a su vez inciden en el valor de las posiciones de activos y / o pasivos de la cooperativa. (COFIA, 2011)

b) **Riesgo de Crédito:** Se refiere a la variabilidad en los ingresos generados por el incumplimiento de un acreditado o contraparte. Incluye la variabilidad derivada tanto de las pérdidas por el importe adeudado y no pagado a las cooperativas por los acreditados, como los costos de recuperación incurridos. (COFIA, 2011)

c) **Riesgo de Liquidez**: En el contexto de portafolios, el riesgo de pérdida por diferencias adversas entre el valor de realización y el valor teórico de una posición y por la imposibilidad de enajenar, adquirir o cubrir una posición (COFIA, 2011).

d) **Riesgo Operativo**: Este riesgo considera dos tipos de riesgos cuantificables. El Riesgo de negocio abarca pérdidas por cambios rápidos en el ambiente competitivo o eventos que dañen la franquicia o la forma de operar de un negocio (variación en volumen, precios o costos) y el Riesgo de evento que se da debido a eventos individuales tales como fallas de sistemas, errores y omisiones, fraudes, daños de equipo no asegurados. (COFIA, 2011)

e) **Riesgo Legal:** El riesgo debido al incumplimiento de disposiciones legales o administrativas, a la resolución de disposiciones administrativas y judiciales desfavorables y la aplicación de sanciones por parte de las autoridades. Este riesgo puede generarse como consecuencia de un Riesgo Operativo de Evento o de Negocio. (COFIA, 2011)

2.1.3. Marco internacional de gestión de riesgo (Acuerdo de Basilea)

El marco internacional de gestión de riesgos está dirigido por el Comité de Basilea, también llamado Comité de Supervisión Bancaria de Basilea, fue fundado en 1975 por las máximas autoridades de los bancos centrales del G-10. El objetivo del Comité de Basilea es “emitir recomendaciones y promover la cooperación en materia de supervisión bancaria en el ámbito internacional” (Araujo & Masci, 2007). A pesar de no tener ningún tipo de autoridad ni competencia supranacional, sus recomendaciones y normativas son aplicadas por una gran cantidad de países. Hasta la actualidad el Comité de Basilea ha promulgado varios documentos encaminados a la gestión de riesgos.

Basilea I

Este documento es conocido como el Acuerdo de Capital de Basilea. Fue promulgado en 1988 y debió ser implementado hasta

el año 1992. Las recomendaciones están dirigidas a gestionar el riesgo de crédito, y persiguen dos objetivos principales: el primero objetivo es “El nuevo marco deberá servir para fortalecer la solidez y la estabilidad del sistema bancario internacional”, y el segundo “El marco ha de ser justo y tener un alto grado de coherencia en su aplicación a los bancos en diferentes países con el fin de disminuir una fuente existente de desigualdad competitiva entre los bancos internacionales”. (Comité de Supervisión Bancaria de Basilea, 1988)

El principal aporte de Basilea I es el requerimiento básico de capital de 8% en relación a los activos ponderados por unas medias simples y consensuadas sobre el nivel de riesgo. Con ello, se logró un importante incremento en los niveles de capitalización del sistema bancario internacional, y se emitían recomendaciones para la gestión del riesgo de crédito.

Para 1996, se modificó el Acuerdo de Capital y se logró incorporar la gestión del riesgo de mercado, como aquel derivado de las operaciones en moneda extranjera y de la cartera de negociación. Asimismo, en 1997 se promulgaron los Principios Básicos de Supervisión Bancaria como un gran aporte al monitoreo y control del sistema financiero. Actualmente, Basilea I ha sido implementada por una gran cantidad de países independientemente de su grado de desarrollo.

Basilea II

A pesar del importante avance que constituyó Basilea I, la realidad financiera y bancaria a mediados de la década de los noventa presentaba nuevos desafíos, por lo que era necesaria la definición de una nueva normativa. Así, a finales del año 1998 se inicia el debate y reuniones para definir el nuevo marco regulatorio. Luego de aproximadamente seis años de deliberaciones, a

mediados del año 2004 se promulga Basilea II, llamado Convergencia internacional de medidas y normas de capital. Los tiempos de implementación de Basilea II estaban planeados desde el año 2006 y un año más si son métodos avanzados.

El objetivo primordial de Basilea II ha sido “establecer un marco que fortaleciera en mayor medida la solidez y estabilidad del sistema bancario internacional, manteniendo al mismo tiempo la necesaria consistencia para que la normativa de suficiencia del capital no fuera una fuente de desigualdad competitiva entre los bancos internacionales”. (Comité de Supervisión Bancaria de Basilea, 2004). En Basilea II se emiten recomendaciones para gestionar tres tipos de riesgos: riesgo de crédito, riesgo de mercado y riesgo operacional.

Basilea II consiste fundamentalmente en el desarrollo de tres pilares:

Primer pilar: Requerimientos Mínimos de Capital

Uno de los principales aportes de Basilea II es el permitir a las entidades financieras la generación de modelos internos de medición de riesgo, y con ello, se puede hacer un uso más eficiente de los recursos.

Segundo pilar: El proceso de examen supervisor

Existen cuatro principios básicos, el primero es que los bancos deberán contar con un proceso para evaluar la suficiencia de su capital total en función de su perfil de riesgo y con una estrategia para el mantenimiento de sus niveles de capital. El segundo es que las autoridades supervisoras deberán examinar y evaluar las estrategias y evaluaciones internas de la suficiencia de capital de los bancos, así como la capacidad de éstos para vigilar y garantizar

su cumplimiento de los coeficientes de capital regulador. Las autoridades supervisoras deberán intervenir cuando no queden satisfechas con el resultado de este proceso. El tercer principio es que los supervisores deberán esperar que los bancos operen por encima de los coeficientes mínimos de capital regulador y deberán ser capaces de exigirles que mantengan capital por encima de este mínimo. Por último, el cuarto principio establece que los supervisores traten de intervenir con prontitud a fin de evitar que el capital descienda por debajo de los niveles mínimos requeridos para cubrir las características de riesgo de un banco dado; asimismo deberán exigir la inmediata adopción de medidas correctoras si el capital no se mantiene en el nivel requerido o no se recupera ese nivel.

Tercer pilar: Disciplina de mercado

El tercer pilar busca la transparencia de la información y toma peculiar importancia al existir la posibilidad de crear metodologías de medición y gestión de riesgos para cada banco. Para ello, "intenta fomentar la disciplina de mercado mediante el desarrollo de una serie de requisitos de divulgación que permitirá a los agentes del mercado evaluar información esencial referida al ámbito de aplicación, el capital, las exposiciones al riesgo, los procesos de evaluación del riesgo y, con todo ello, a la suficiencia del capital de la institución".

Basilea III

Es un conjunto integral de reformas para fortalecer la regulación, supervisión y gestión de riesgos del sector bancario; el Comité de Basilea decide crear esta nueva normativa denominada Basilea III o Marco Regulador Internacional para Bancos 2010 a raíz de la crisis financiera internacional iniciada en Estados Unidos por

los créditos hipotecarios. Estas medidas buscan mejorar la capacidad del sector bancario para afrontar perturbaciones ocasionadas por tensiones financieras o económicas de cualquier tipo, mejorar la gestión de riesgos y el bueno gobierno en los bancos y reforzar la transparencia y la divulgación de información de los bancos. (Bank for international settlements, s.f.)

La Superintendencia de Banca y Seguros (s.f.) en Perú, por medio de la SBS, es consciente de las ventajas en seguridad y estabilidad que genera un esquema como el propuesto en Basilea II y no está al margen de esta reforma internacional de la regulación bancaria. Es por ello que el cronograma de implementación se inició en el año 2007 con los estudios de impacto y la emisión de la normativa necesaria para la implementación del NAC, durando esta fase hasta junio del 2009 y a partir de julio entró en vigencia del método estandarizado para riesgo de crédito y riesgo de mercado, y el método básico y estándar alternativo para riesgo operacional. Asimismo, es a partir de esta fecha que las empresas pueden postular para el uso de modelos internos. En cuanto al Basilea III, la SBS actualmente está evaluando la implementación de estos cambios de acuerdo a la realidad peruana. (Superintendencia de Banca y Seguros, s.f.)

2.1.4. Riesgo de crédito

El riesgo de crédito es la posible pérdida que asume un *agente económico* como consecuencia del incumplimiento de las obligaciones contractuales que incumben a las *contrapartes* con las que se relaciona; si bien este concepto está íntimamente ligado con las instituciones financieras y bancos, también afecta a empresas y organismos de otros sectores.

Existe una gran variedad de definiciones sobre riesgo de crédito y entre las principales se encuentran la de los siguientes autores:

Jorión (1999) plantea que “El riesgo de crédito abarca tanto el riesgo de incumplimiento como el de mercado” mientras el primero comprende la evaluación objetiva de la probabilidad de que la contraparte incumpla, el segundo mide la perdida financiera que será experimentada si el cliente incumple (Jorion, 1999). Para Bessis (2002), es “La posibilidad de incurrir en pérdidas por quebrantos que puedan producirse en el desarrollo de la actividad bancaria” (Bessis, 2002). De Lara (2005) manifiesta que el riesgo crediticio es el más antiguo y probablemente el más importante que enfrentan los bancos y también las instituciones de microfinanzas; también se puede definir como la pérdida potencial, producto del incumplimiento de la contraparte en una operación donde se incluye un compromiso de pago. (De Lara, 2005)

El riesgo crediticio está asociado a múltiples factores como por ejemplo: la solidez financiera del cliente y su capacidad de pago, la presencia de una garantía que respalde un préstamo, el entorno económico, etc.; estos factores pueden ser afectados tanto por el entorno macroeconómico como por el que corresponde al sector productivo del cliente. Sólo con una revisión completa del portafolio de créditos se puede evaluar el nivel de riesgo al que está expuesta un entidad financiera; sin embargo, algunos indicadores de los balances pueden iluminar la calidad de la cartera de una institución y el peligro de que ésta no pueda honrar las obligaciones a sus depositantes por falla de sus deudores en el reembolso de los préstamos.

2.1.5. Factores que determinan el riesgo de crédito

2.1.5.1. Políticas de crédito

El blogspot Créditos y cobranzas (2010) considera que “Al ser un crédito un proceso ordenado de pasos y procedimientos interconectados al desenvolvimiento económico y financiero, necesita de políticas que enmarquen las pautas para la consecución de objetivos a los cuales se debe llegar en virtud a una administración efectiva del crédito”. Para poder clasificar las políticas de crédito, son varias las razones las que motivan a los empresarios que venden crédito a orientar sus políticas como liberales o conservadoras, y estas son la competencia, demanda de los clientes, volúmenes de ventas etc. Se dice que una política es liberal cuando las empresas se muestran flexibles al momento de otorgar créditos, tanto en el monto máximo para aprobar como en el grado de riesgo para asumir; y son conservadoras cuando las empresas son estrictas al momento de evaluar un crédito y de determinar el monto máximo por aprobar, así como para definir lo referente al riesgo que asumirán. (Creditos y Cobranzas, 2010)

Sin embargo, autores como Brachfield (2015) consideran a una tercera clasificación llamada políticas normales y las definen como aquellas “Que se sitúan en el término medio, es decir no son ni restrictivas ni liberales (este tipo de políticas lo que busca es el lograr un término medio o equilibrio en el riesgo de clientes, asumiendo riesgos en algunos casos y permitiendo los plazos de pago comunes en la industria)”. (Brachfield, 2015)

2.1.5.2. Centrales de Riesgo

La central de riesgos es un sistema de registro que consolida la información de la situación crediticia de los deudores de las empresas del sistema financiero. Una central de riesgo es un

registro o una base de datos que mantiene información actualizada sobre los deudores, incluyendo datos demográficos, pautas de pago de distintos tipos de obligación de crédito y registros de deudas incobrables y otros.

En el Perú existen dos tipos de centrales de riesgos: la Central de Riesgos de la SBS que es pública y se rige por lo dispuesto en los artículos 158°,159°, y 160° de la Ley N ° 26702 y las Centrales Privadas de Información de Riesgos (CEPIRS) como INFOCORP Y CERTICOM, cada una de ellas tiene una regulación específica, así como procedimientos particulares que permiten a los titulares de información la defensa y exigencia de sus derechos. (Superintendencia de Banca y Seguros, s.f.)

A diferencia de otras centrales de riesgo públicas, donde la cantidad de información es mínima, el ente supervisor peruano sobresale al proveer a los burós de crédito con una amplia cantidad de datos de todos los deudores de las instituciones reguladas. Asimismo, es el único entre países como Bolivia, Brasil, Colombia, El Salvador, Guatemala, México, Nicaragua, Perú, República Dominicana que brinda toda la base de datos a las instituciones reguladas y algunas instituciones públicas. Ello hace de Perú el único país donde hay “alineamiento” de la calificación del deudor (es decir, la calificación del deudor tomando su situación crediticia en todo el sistema financiero). Esta práctica reduce el riesgo sistémico y los costos de las instituciones. En los demás países, las instituciones financieras reguladas solo tienen acceso a consultar individualmente a los deudores desde la página web y, por lo tanto, no pueden realizar este alineamiento. México y Perú son los dos países donde los burós de crédito presentan en los reportes crediticios la información de deuda tributaria, útil en el análisis crediticio. (Rodríguez & Global Methodology, 2012)

2.1.5.3. Los Buró de crédito

En el Perú son sociedades de información crediticia orientadas a integrar información sobre el comportamiento crediticio de personas y empresas, por lo que hoy en día se han convertido en un marco de referencia para el otorgamiento de crédito, no solo en Perú sino en otros países, que ya que cuentan con expedientes crediticios de personas físicas y empresas. Los Burós de Crédito únicamente pueden proporcionar información, en el caso de que el titular del historial crediticio solicite su reporte de crédito especial, o que el otorgante de crédito solicite el reporte de crédito de una persona o empresa previa autorización de esta.

En su página web La Economía (2011) hace mención que los Burós de Crédito no deciden si un crédito debe o no aprobarse, ni tampoco emiten juicio sobre si una persona es sujeta de crédito o no; únicamente proporcionan información sobre los créditos y comportamiento de pago de una Persona o Empresa. Es el otorgante de Crédito, quien en función al análisis que efectúa de un reporte de Crédito y a las políticas que tenga establecidas, decide si otorga o no el crédito. (La Economía, 2011)

Según Rodríguez & Global Methodology (2012) en su informe referente a las centrales públicas de riesgo, buros de crédito y el sector microfinanciero, manifiestan que los burós de crédito en el Perú son: Data Crédito, Equifax, Informa del Perú Sentinel y Xchange Perú; señalan además que estos burós no son regulados, consolidan la información de la central pública de riesgos y la recabada independientemente, de acuerdo a convenios particulares con instituciones financieras no reguladas, comerciales, públicas, etc. (Rodríguez & Global Methodology, 2012)

2.1.5.4. Morosidad Crediticia

Viene a ser la cartera pesada de los clientes que han incumplido su compromiso de pago; la morosidad es consecuencia de una mala calificación del crédito, en cuanto a información, garantías y una pésima administración. Por ello atendiendo a este último factor se debe tener en cuenta la clasificación del deudor o cartera de créditos.

2.2. MODELOS CREDIT SCORING

El credit scoring es una herramienta para valorar una operación en términos de riesgo, siendo usada como herramienta para aprobar o denegar una operación de crédito. Su metodología cuantifica la calidad-riesgo de una operación de un cliente mediante la ponderación de características observables como lo son los datos socioeconómicos, operativos, de negocio, de comportamiento financiero de la misma entidad o de recursos externos.

Los scoring son construidos haciendo uso de modelos estadísticos u otros modelos de inteligencia artificial como las redes neuronales. El output de los mismos permite hacer una predicción del riesgo de la operación para cada cliente y además posicionarlos en un ranking en función de la propensión al riesgo.

Montoya (2010) menciona que “El desarrollo de herramientas de Credit Scoring requiere de personas especializadas que entiendan por un lado los modelos estadísticos y por otro los requerimientos del negocio. La gestión de los scorings tiene como finalidad usar estas herramientas como una solución potente para la compañía. Es indudable el valor que los scorings aportan en la gestión de la relación con los clientes y en la reducción de la temida tasa de mora”. (Montoya, 2010)

2.2.1. Utilidades Específicas del Credit Scoring

En la tesis “Evaluación crediticia aplicando un modelo de Credit scoring en el ámbito microempresarial: Caso CMAC - Paita" (Herrán, 2009) menciona como utilidades específicas a las siguientes:

a) Es posible discriminar entre probables buenos y malos pagadores.

b) Asignar probabilidades de incumplimiento de pago a los clientes para otorgar o no un crédito de acuerdo a estas probabilidades; al establecerse el cut-off score se concederá crédito sólo a aquellos clientes que estén por debajo del punto, rechazando así a aquellos cuya probabilidad esté por encima.

c) Permite que por medio de la PD se identifique aquellas variables que afectan el riesgo crediticio, así como el efecto marginal de las mismas sobre dicha probabilidad.

d) Sirve como mecanismo de ayuda para que las empresas se enfoquen sobre los prestatarios más riesgosos.

e) Permite que las empresas presupuesten provisiones de acuerdo a su riesgo crediticio esperado.

f) Fijar las tasas de interés de acuerdo al riesgo – crédito esperado y a la meta establecida de Ingreso Financiero (o Margen Financiero).

g) Es factible, incluso, establecer tasas de interés diferenciadas según el riesgo – crédito de los clientes.

2.2.2. Tipos de Scoring

Score de Originación

Estima la probabilidad de incumplimiento de pago de un posible cliente otorgando a la empresa la capacidad de decidir si acepta o no a un cliente como posible consumidor de crédito; por medio de

las variables demográficas y las de buró este tipo de score optimiza la tasa de aprobación de las solicitudes de crédito. Además permite a la organización establecer el puntaje mínimo óptimo de aceptación en conjunto con otros departamentos, así como definir productos de crédito personalizados y realizar actividades de mercadeo para aumentar el número de clientes con características deseables y cumplir con las metas corporativas. (Nieto, 2010)

Score de comportamiento

A diferencia del score de originación éste predice la probabilidad de incumplimiento de aquellos que ya se les otorgó crédito en la institución. Por medio de las variables de comportamiento de las cuentas dentro de la propia institución es posible dar seguimiento al comportamiento de los clientes, permitiendo al departamento de cobranzas emplear técnicas para que un cliente siga siendo rentable para la empresa. (Nieto, 2010)

Score de Bureau

Es de naturaleza comportamental, se basa en el historial crediticio de las personas para predecir el futuro. Permite disponer de una primera calificación de riesgo de cualquier persona con referencias sin necesidad de solicitar ningún tipo de información. Es una excelente herramienta para la seleccionar clientes potenciales en una campaña; una de las ventajas es que al combinarlos con modelos internos permite identificar significativamente la calidad de riesgo de las carteras.

Marketing Scores

Se centra en la retención de clientes y en la adquisición de nuevos, por ello, se están desarrollando scorings de retención y

scorings de valor del cliente. Los últimos proporcionan información al gestor de la cuenta y al responsable de Marketing sobre los ingresos potenciales del cliente, permitiendo focalizar las campañas en las cuentas con mayor rentabilidad esperada y gestionar la cartera en función del valor de los clientes. (González & Montoya, 2006)

Utilizados conjuntamente con scoring de riesgos permiten:

- "Mejorar la estimación del éxito de las ofertas en curso y crear nuevas ofertas sin aumentar el riesgo" (González & Montoya, 2006).
- "Realizar campañas de marketing selectivo sobre los clientes de mayor valor y probabilidad de respuesta de una manera efectiva en términos de reducción de costos" (González & Montoya, 2006).
- Tarificar de una manera más completa, por rentabilidad-riesgo. No es lo mismo el valor del cliente considerando rentabilidad, que considerando rentabilidad-riesgo, porque de esta última manera nos aproximamos mucho más al valor real del cliente. (González & Montoya, 2006)
- Centrar esfuerzos en aquellos clientes con mayor valor mediante estrategias de fidelización (González & Montoya, 2006)

2.2.3. Requisitos para la Construcción de un Modelo de Credit Scoring

a) "Se debe definir el grupo objetivo sobre el cual se desea realizar las predicciones cuidando que esté constituido por componentes capaces de explicar el evento a predecir" (AIS Goup, 2011).

Cuando se trabaja con modelos predictivos hay 3 conjuntos de datos fundamentales que todo dataminer debe manejar y son: la

muestra de entrenamiento (TRAINING) con la cual se entrenan los modelos, la muestra de validación (VALIDATION) que selecciona el mejor de los modelos entrenados, y la muestra de prueba (TEST) que entrega el error real cometido con el modelo seleccionado. (Webmining Consultores, 2011)

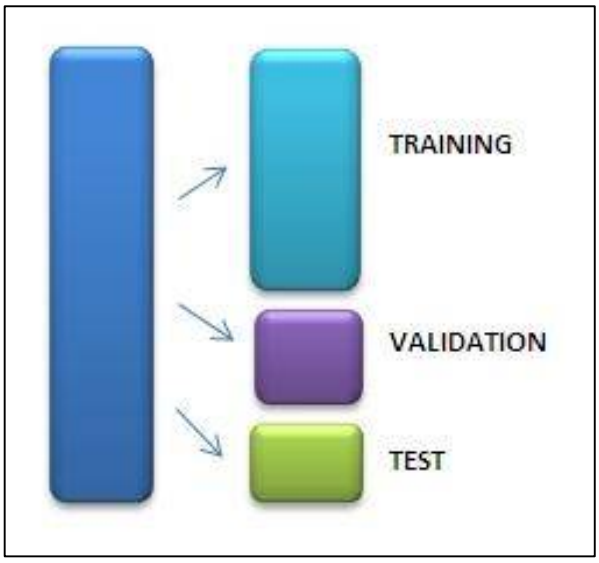


Figura 1. Particionamiento de la base de datos

Hastie, Tibshirani & Friedman como se citó en Webmining consultores (2011) señalan que “Es difícil dar una regla general sobre cuántas observaciones se deben asignar a cada conjunto, aunque indican que una división típica puede ser de 50% para el entrenamiento y 25% para la validación y prueba, respectivamente” (Webmining Consultores, 2011). En situaciones en las que no hay datos suficientes para dividir en tres partes se suele trabajar con dos divisiones, es decir, entrenamiento y validación. Aunque no es algo estricto, en la literatura existente hay valores tales como 50-50, 70-30, 57-43, 67-33, 85-15 y otros más que podrían también probarse. Lo ideal es escoger una muestra 50%-50% con la finalidad de disminuir el sesgo al momento de validar el modelo.

b) “Se deben recolectar todas las características que mejor definan a los componentes del grupo objetivo” (AIS Goup, 2011).

c) “Se deben definir claramente el evento a predecir y su espacio temporal de ocurrencia: *definición de mora”* (AIS Goup, 2011).

d) Se pueden aplicar diversas técnicas para tratar la muestra de datos y obtener un modelo que represente su comportamiento respecto al evento a predecir: análisis discriminante, regresión lineal, regresión logística, modelos probit, logit, métodos no paramétricos de suavizado, métodos de programación matemática, modelos basados en cadenas de Markov, árboles de decisión, algoritmos genéticos, redes neuronales, sistemas expertos entre otros. (AIS Goup, 2011)

e) (AIS Goup, 2011) menciona que se debe aplicar una metodología que contemple:

- “Validar la muestra y asegurar que sea representativa del grupo objetivo a representar a través del modelo”.
- “Conocer detalladamente al grupo objetivo en sí mismo: análisis exploratorio o univariante”.
- “Conocer detalladamente al grupo objetivo respecto al evento a predecir: análisis bivariante”.
- “Recusar los procedimientos de regresión hasta obtener el modelo óptimo a través de los máximos indicadores de calidad”.
    - **Curva de ROC:** Es una gráfica que representa el eje X (1 – especificidad) y el eje Y la sensibilidad, para un sistema clasificador binario según se varía el umbral de discriminación.

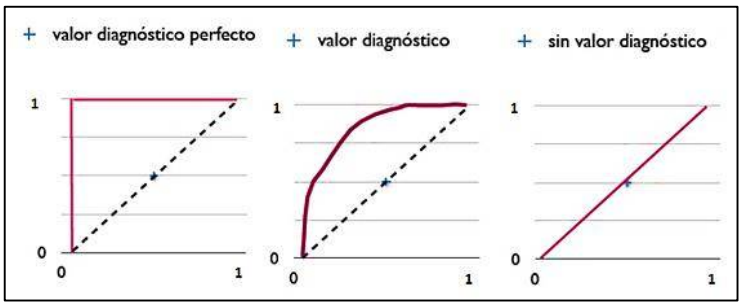


Figura 2. Curva de ROC

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUROC, siendo el mejor modelo aquel que tiene el valor AUROC más alto:

[0.5, 0.6): Test malo.

[0.6, 0.75): Test regular.

[0.75, 0.9): Test Bueno.

[0.9, 0.97): Test muy bueno.

[0.97, 1): Test excelente.

- **Índice de Gini:** Es una medida de desigualdad, donde es un número entre 0 y 1, en donde 0 se corresponde con la perfecta igualdad, siendo el mejor modelo aquel que tiene el valor más alto**:**

$$G = \left| 1 - \sum_{k=1}^{n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k) \right|$$

Se calcula con la fórmula $\boldsymbol{GINI} = 2 * AUROC - 1$.

- **Kolmogórov-Smirnov:** El valor del estadístico KS de un modelo es la mayor diferencia entre las distribuciones acumuladas de los acreditados Buenos y los acreditados Malos. Cuanta más alta sea la diferencia mejor será el poder discriminatorio del modelo, obteniendo un modelo ideal una distancia igual a 1. Un modelo real suele obtener

distancias superiores a 0.5, siendo justificables en algunos casos distancias entre 0.4 y 0.5.

$$KS = \max_{x \in \Re} \left\{ \left| F_{buenos}(x) - F_{Malos}(x) \right| \right\}$$

- **Matriz de confusión:** una herramienta de visualización que se emplea en aprendizaje supervisado.

Valor observado	Resultado del modelo	
	Malo	Bueno
Malo	Verdadero positivo (VP)	Falso negativo (FN)
Bueno	Falso positivo (FP)	Verdadero negativo (VN)

Figura 3. Matriz de Confusión

Indicadores de eficiencia del modelo:

Sensibilidad o Razón de Verdaderos positivos: probabilidad de clasificar correctamente a un sujeto enfermo.

$$Sensibilidad = \frac{VP}{VP + FN}$$

Especificidad o Razón de Verdaderos negativos: probabilidad de clasificar correctamente a un sujeto sano.

$$Especificidad = \frac{VN}{VN + FP}$$

Valor Predictivo Positivo (VPP): probabilidad de que un sujeto enfermo de positivo en la prueba.

$$VPP = \frac{VP}{VP + FP}$$

Valor Predictivo Negativo (VPN): probabilidad de que un sujeto sano de negativo en la prueba.

$$VPN = \frac{VN}{VN + FN}$$

Exactitud: probabilidad de resultados correctos de la prueba.

$$Exactidud = \frac{VP + VN}{VP + FP + VN + FN}$$

- "Validar el modelo con contra muestras y el universo de datos supuestamente representado por la muestra: máximos indicadores de calidad" (AIS Goup, 2011).
- Punto de corte:

  En el curso "Credit scoring, validación de modelos y stress testing nivel I", considera los siguientes pasos para la realización de los puntos de corte:

  - **Selección de puntos de corte**

    "Para seleccionar los puntos de corte de debe establecer el nivel mínimo de score en el cual el aplicante es denegado o aceptado. Umbrales por rentabilidad y riesgo" (Fermac Risk, 2014).

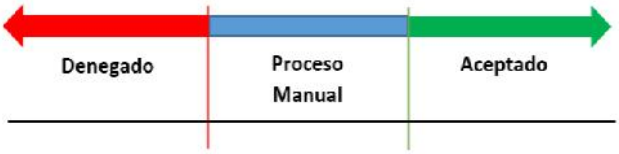


Figura 4. Puntos de corte para calificación de scoring

  - **Análisis de la puntuación del scoring**

    "Se utiliza para la definición de puntos de corte para su implementación en la estrategia de admisión" (Fermac Risk, 2014).

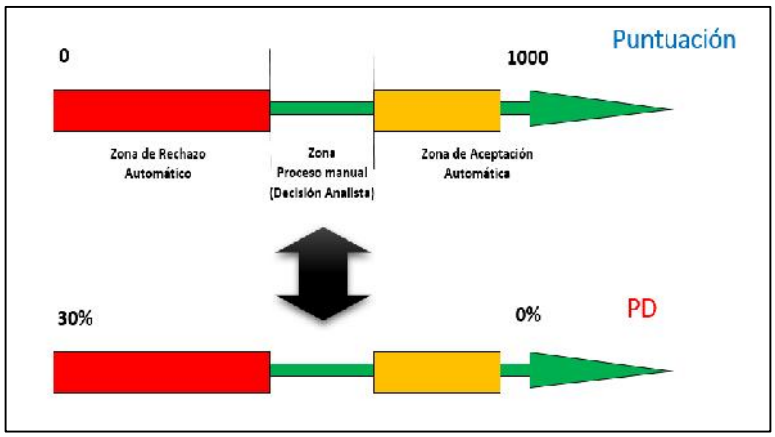


Figura 5. Puntuación de scoring

"Se decide en función de los resultados del modelo, del riesgo asociado a la puntación y del conocimiento de negocio. La decisión de los puntos de corte dependerá de la propensión al riesgo de la entidad" (Fermac Risk, 2014).

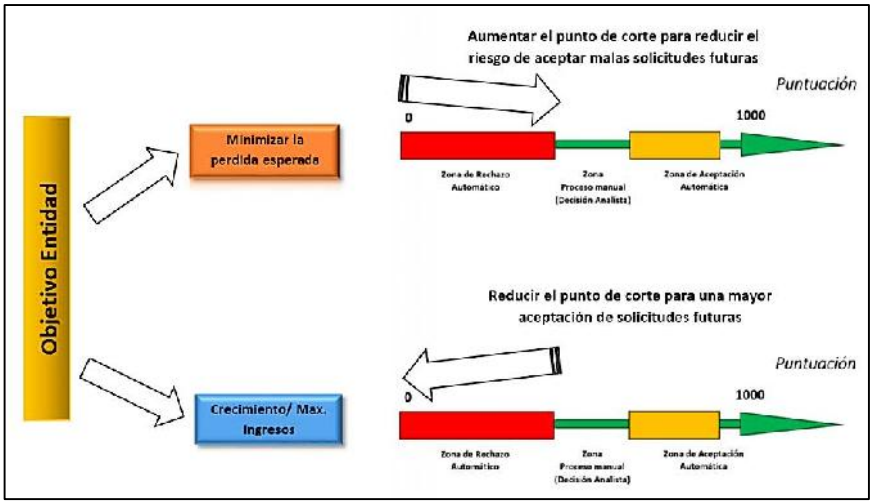


Figura 6. Propensión de riesgo en una entidad financiera

2.2.4. Ventajas y desventajas del Credit Scoring

Schreiner (2002) En su investigación denominada “Ventajas y desventajas del scoring estadístico para las microfinanzas” define como beneficios y dificultades del scoring a las siguientes:

a) Beneficios del Credit Scoring

La evaluación del crédito que realiza el score es objetiva y consistente en el tiempo de riesgo, tiene un proceso más eficiente en tiempo por lo que se traduce en menores costos; además, al tener cada crédito una puntuación desde el inicio permite un control estadístico del portafolio. Por último menciona que esta herramienta permite calificar a cada uno de los clientes. (Schreiner, 2002)

b) Dificultades o limitaciones del Credit Scoring

Entre las dificultades que consideran están que son procesos que pueden tomar bastante tiempo, son modelos de predictibilidad limitada por lo que se deteriora con el tiempo, nunca nos podrá separar completamente por puntaje, se hace necesario contar con un área especial, etc. (Schreiner, 2002)

2.3. TÉCNICAS APLICADAS EN LOS MODELOS CREDIT SCORING

2.3.1. Análisis de Regresión Logística

El análisis de regresión logística es una técnica para el estudio de la relación entre una o más variables independientes y una variable dependiente de tipo dicotómica, representa la ocurrencia o no de un suceso, por ejemplo: un paciente muere o no antes del alta, una persona deja o no de fumar después de un tratamiento, una persona vota a favor o no de un candidato, una persona compra artículos o no de una determinada marca, etc.

Un modelo de regresión logística permite estimar o predecir la probabilidad de que un individuo posea una característica en función de una determinada o unas determinadas características individuales. La regresión logística forma parte de los modelos lineales generalizados, donde la función de enlace es la función logit. Este modelo comúnmente presenta una forma de “S”, limitada en el eje de las ordenadas entre los valores 0 y 1. El modelo antes descrito se denomina “Función Logística”.

En 1937, Bartlett utilizó la transformación $log[y/(1-y)]$ para analizar proporciones. Fisher y Yates sugirieron en 1938 el uso de esa transformación para analizar datos binarios. El término logit fue introducido por Joseph Berkson en 1944 para designar esta transformación y sus trabajos incentivaron la utilización de la regresión logística. Jerome Cornfield utilizó este método para el cálculo de ODDS RATIO como valores aproximados del riesgo relativo en estudios de casos y controles. (Molinero, 2004)

El objetivo fundamental de la regresión logística es determinar si hay relación entre una variable predicha o variable respuesta y un conjunto de predictores o variables regresoras. La regresión

logística permite modelar cómo influye en la probabilidad de aparición de un suceso (generalmente de una variable dicotómica) la presencia o no de diversos factores. ”En este análisis es conveniente tener en cuenta de que las variables categóricas deben ser codificadas de forma apropiada” (Peña, 2002).

2.3.1.1. Definición

Sean $X_1, X_2, \ldots, X_k$ un conjunto de variables regresoras, y sea Y una variable dependiente dicotómica, que toma como valor 0 y 1. Se busca determinar $p = P\,Y = 1/\,X_1, X_2, \ldots, X_k$, donde p es la probabilidad de éxito.

Se construye un modelo de la forma:

$$P\,Y = \frac{1}{X1}, X2, \ldots., Xk \;=\; p\;\; X_1, X_2, \ldots, X_k;\; \beta \ldots (1)$$

Donde $p\,X1, X2, \ldots., Xk;\; \beta$, es una función que recibe el nombre de función de enlace (función de probabilidad) cuyo valor depende de un vector de parámetros $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$.

2.3.1.2. Modelo de Regresión Logística

Es de interés estudiar la relación entre una o más variables independientes o explicativas: $X_1, X_2, \ldots, X_k$ y la variable *Y,* donde Y toma valor 1 si ocurre el suceso y valor 0 si no ocurre. El modelo logístico establece la siguiente relación entre ***la probabilidad de que ocurra el suceso***, dado que el individuo presenta los valores $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$.

La ecuación de regresión tiene que ser diferente de la que se emplea en regresión múltiple $(Y = \beta' X + \varepsilon)$. En este modelo lo que se predice no es directamente la variable sino *la probabilidad* de que la variable adopte cierto valor. Para predecir una probabilidad puede utilizarse diferentes funciones, una es la logística, que es la

base del cálculo de la probabilidad de *p* que queremos predecir.

Si llamamos X_i a los predictores, sea:

$p = p\,X_1, X_2, \ldots, X_k;\ \beta = G \quad \beta + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k\, X_k = GX'\beta \ldots (2)$

Donde la función de densidad acumulada de la función logística, queda denotada como:

$$log\left(\frac{p\,X_1, X_2, \ldots, X_k;\ \beta}{1-p\ \ X_1, X_2, \ldots, X_k;\ \beta}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k\, X_k \quad \ldots (3)$$

(Modelo logit). Esto indica que existe una relación lineal entre el cociente de probabilidades (la probabilidad de pertenecer a un grupo –éxito- dividido por la probabilidad de pertenecer al otro -fracaso-) y los predictores.

Si se aplica la exponencial a la expresión (3), la ecuación queda expresada de la siguiente manera:

$$p = \frac{e^{x\prime\beta}}{1+e^{x\prime\beta}} = \frac{1}{1+e^{-x\prime\beta}} \quad \ldots (4)$$

Tomando la forma de la ecuación de regresión múltiple, donde β_0 es la constante y los β_i son los coeficientes de los predictores X_i correspondientes.

Supuestos del modelo:

Tres supuestos: ausencia de colinealidad entre las variables regresoras, los errores tienen distribución binomial y la no linealidad de la variable de respuesta.

El modelo de regresión logística es robusto con respecto al incumplimiento del supuesto de igualdad de las matrices de covarianza entre grupos.

Supuestos de la regresión logística:

Independencia entre las observaciones sucesivas y la existencia de una relación lineal entre logit (x) y los predictores $X_1, X_2, \ldots, X_k$.

2.3.1.3. Estimación de los coeficientes del Modelo

Si bien existen muchos métodos, el que más se utiliza es el de máxima verosimilitud, que consiste en maximizar la función de verosimilitud de la muestra.

Para ello se considerará una muestra aleatoria simple de tamaño n dada de la siguiente forma $X_i', X_i; i = 1,2,3, \dots, n$ donde: X_i son los valores de las variables independientes del i-ésimo individuo de la muestra, $Y_i = 0,1$ es el valor observado de la variable dependiente del $i - ésimo$ individuo de la muestra.

Además:

$Y/\ X_1, X_2, \dots, X_k \sim Binomial\ (1, p)\ Y = 1/\ X_1, X_2, \dots, X_k;\ \beta,$ y el número de éxito en n repeticiones tiene una distribución binomial $B(n, p)$. La función de verosimilitud es:

$$\partial(\beta) = L\beta / X_1', y_1, \dots, X_n', y_n = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \dots \quad (5)$$

Donde: $p_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}; i = 1,2, \dots, n$

Aplicando logaritmo neperiano a ambas expresiones, se tiene:

$$\mathbf{L = \ln\ \ell(\beta) = \sum_{i=1}^{n}\left(y_i x_i \beta + \ln\left(\frac{1}{1+e^{x_i\beta}}\right)\right)} \qquad \dots(6)$$

El vector de parámetros β se estima por máxima verosimilitud; una forma de calcularla es bajo la estimación incondicional, que maximiza la función de verosimilitud anterior, derivando e igualando al valor cero:

Por β_0:

$$\frac{\partial}{\partial \beta_0} L = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \frac{e^{x_i\bar{\beta}}}{1+e^{x_i\bar{\beta}}}\right) = 0 \qquad \dots(7)$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \hat{p}_i = 0$$

Por β_i:

$$\frac{\partial}{\partial \beta_j} L = 0$$
$$\Rightarrow \sum_{i=1}^{n} x_{ij}\ y_i - \hat{p}_i = 0 \qquad \ldots(8)$$

Las $k+1$ ecuaciones (7 y 8) se resuelven mediante métodos iterativos. Este procedimiento es matemáticamente complejo, se dan a los coeficientes unos valores arbitrarios (en su mayoría el valor 0, aunque no necesariamente). La solución final no depende de estos valores pero sí del tiempo de cálculo.

2.3.1.4. Evaluación del Modelo

a) Evaluación de la Significancia del Modelo

Para determinar si las variables independientes son significativas o no, se plantea las siguientes hipótesis:

$H_0 = \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

$H_1 = Por\ lo\ menos\ un\ \beta_1 \neq 0\ \forall\ i = 1,2, \ldots, k$

*Con un nivel de α de significación.

Estadístico de Prueba: El estadístico que se plantea para la evaluación de la significancia del modelo es la diferencia del valor de la desvianza del modelo sólo con la constante y del modelo incluyendo las variables independientes, este estadístico sigue una distribución Chi – Cuadrada con k grados de libertad:

$$G = D(modelo\ sin\ variables) - D(modelo\ con\ variables)$$

$$G = -2\ln\left(\frac{verosimilitud\ del\ modelo\ sin\ variables}{\text{verosimilitud del modelo con variables}}\right)$$

$$G = -2\left\{ n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n) - \sum_{i=1}^{n} y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) \right\} \sim \chi_k^2$$

b) Evaluación de la Bondad de Ajuste del Modelo

i. Desvianza del Modelo (D):

Para la evaluar la bondad de ajuste del modelo, es decir determinar si el modelo ajustado es el adecuado o no, se plantea las siguientes hipótesis:

$$H_0 = El\ modelo\ ajustado\ es\ significativo.$$

$$H_1 = El\ modelo\ ajustado\ no\ es\ significativo.$$

*Con un nivel de α de significación.

Estadístico de Prueba:

La desvianza es la medida del grado de diferencia entre las frecuencias predichas y las observadas del modelo, el mejor modelo será aquel que tenga menor desvianza, este estadístico sigue una distribución Chi – Cuadrada con n – (k+1) grados de libertad [3], se tiene:

$$D = -2\ln\left(\frac{verosimilitud\ del\ modelo\ sin\ variables}{verosimilitud\ del\ modelo\ con\ variables}\right) \quad \ldots(9)$$

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{p}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{p}_i}{1-y_i}\right)\right] \sim \chi^2_{n-(k+1)}$$

ii. Test de Hosmer y Lemeshow (Hosmer & Lemeshow, 1989)

La prueba de Hosmer-Lemeshow evalúa un aspecto de la validez del modelo: la calibración[1].

Hipótesis para evaluar la bondad de ajuste del modelo:

H_0: No existe diferencia entre los valores observados y los valores estimados a partir del modelo de regresión logística.

H_1: Existe diferencia entre los valores observados y los valores estimados a partir del modelo de regresión logística.

[1] Calibración: grado en que la probabilidad predicha coincide con la observada.

Estadístico de Prueba: La prueba de Hosmer - Lemeshow agrupa los sujetos en patrones según criterios estadísticos. Primero se calculan los 9 deciles $(D1,..,D9)$ de las probabilidades esperadas o estimadas; $\hat{p}_i; i = 1,2,\ldots n$ y se dividen los datos observados en 10 categorías dadas por:

$$A_j = \hat{p}_i \in \left[D_{j-1}, D_j \right) / i \in 1,2,\ldots,n \;\; ; j = 1,2,\ldots,10$$

* Donde $D0 = 0, D10 = 1.$

Se diseña una tabla de contingencia de 10 x 2, basada en las frecuencias observadas y esperadas se construye el estadístico Chi – Cuadrado de Pearson con distribución X^2 de 8 grados de libertad.

$$T = \sum_{j=1}^{10} \frac{o_j - n_j \bar{p}_j}{n_j \bar{p}_j \; 1 - \bar{p}_j}^2 \sim \chi_8^2$$

*Donde: $n_j = Número\ de\ casos$ en A_j ; $j = 1,\ldots,10$

$$o_j = \sum_{i \in A_j} y_i \;\; ; j = 1,2,\ldots,10 ; \;\; \bar{p}_j = \sum_{i \in A_j} \frac{m_i \hat{p}_i}{n_j} \;\; ; j = 1,2,\ldots,10$$

iii. Prueba de Hipótesis sobre la significancia de los coeficientes

Se plantea las siguientes hipótesis:

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

*Con un nivel α de significación.

Estadísticos de prueba:

$$Z = \frac{\hat{\beta}_j}{\hat{EE}(\hat{\beta}_j)} \sim N(0,1) \quad \text{ó} \quad w = \frac{\hat{\beta}_j^2}{\widehat{var}(\hat{\beta}_j)} \sim \chi_1^2$$

Criterio de Decisión:

Se rechaza H_0, si:

$$|Z| > Z_{\alpha/2} \quad \text{ó} \quad w > \chi^2_{1,\alpha}$$

* W = Estadístico de Wald.

2.3.2. Árboles de clasificación

Los árboles de clasificación también llamados de identificación o de decisión, es una técnica de segmentación diseñada para dividir a una población en dos o más grupos basándose en sus atributos, como género, edad, antecedentes familiares, etc.; en general este método es utilizado en diagnóstico médico, análisis de riesgo en la concesión de créditos, elaboración de horarios, etc.

"El modelo presenta una estructura en forma de árbol, en donde las ramas representan conjuntos de decisiones, estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos de datos disjuntos y exhaustivos. Las ramificaciones se generan de forma recursiva hasta que se cumplan ciertos criterios de parada". (Fernández, 2002)

"La construcción automática de árboles de decisión parte de los trabajos en ciencias sociales de Morgan y Sonquist (1963) y Morgan y Messenger (1973). Breiman (1984) influyo tanto en despertar el interés de los estadísticos como en el de proponer nuevos algoritmos para la construcción de árboles. Sobre esa misma época

la inducción mediante árboles de decisión comenzó a usarse en el "Aprendizaje de las Máquinas" (Inteligencia Artificial), especialmente por Quinlan (1979, 1983, 1986, 1993). Sin embargo no hubo una amplia literatura al respecto, estando dispersas la mayoría de las contribuciones estadísticas. Clark y Pregibon (1992) describen los modelos basados en árboles y los implementan en lenguaje S, haciendo que estos métodos sean mucho más asequibles. Sus métodos son muy flexibles y están fundamentalmente orientados al análisis exploratorio de datos". (Gámez & García, 2000)

Los árboles de decisión son empleados para clasificar y pronosticar, es decir identificar el resultado categórico atendiendo a una serie de criterios dados y pronosticar el resultado según una futura serie de criterios o variables independientes. El objetivo de este método es obtener individuos u objetos más homogéneos con respecto a la variable discriminadora dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas a partir de las cuales se va a realizar la discriminación de la población en subgrupos.

2.3.2.1. Definición

Los análisis de clasificación basados en árboles de decisión son técnicas de explotación de datos que consisten en estudiar grandes masas de datos co

n el fin de descubrir patrones.

Un árbol de decisión tiene como entrada un conjunto de atributos $X = X1, X2, \dots, Xk$ donde a partir de los cuales devuelve una respuesta Y. Los valores que pueden tomar las entradas y las

salidas pueden ser valores discretos o continuos, cuando se utiliza valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utiliza los continuos se denomina regresión; aunque en su mayoría se utilizan los valores discretos por simplicidad.

Un árbol de decisión lleva a cabo un test a medida que este se recorre hacia las hojas para alcanzar así una decisión, por eso suele contener distintos tipos de nodos. Uno de ellos es el nodo interno que contiene un test sobre algún valor de una de las propiedades, el nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, nodo hoja representa el valor que devolverá el árbol de decisión, y para finalizar las ramas que brindan los posibles caminos que se tiene de acuerdo a la decisión tomada. (Leung, 2008)

2.3.2.2. Algoritmos

En su mayoría, los árboles de decisión son construidos ayudándose de un algoritmo que divide los registros en grupos, donde la probabilidad del resultado es diferente en cada grupo atendiendo a los valores de las variables independientes.

Algunos de los algoritmos de árboles de decisión:

- **Árboles de Clasificación y Regresión (C&RT o CART):** Algoritmo binario completo que hace particiones de datos y produce subconjuntos homogéneos precisos. Fue diseñado por L. Brieman (Acuña, 2004).

- **CHAID:** Algoritmo estadístico multidireccional y rápido que explora de forma eficiente los datos, construye segmentos y diseña perfiles en función de la variable de respuesta establecida. Fue introducido en 1980 por Kass y es un derivado

del “THAID” (Morgan & Messenger, 1973). El criterio para su partición está basado en χ^2 y para terminar el proceso se requiere definir de antemano un umbral.

- **C4.5:** Es una extensión del algoritmo ID3 desarrollado por Ross Quinlan. “Los árboles de decisión generador por C4.5 pueden ser usados para clasificación, y por esta razón está casi siempre referido como un clasificador estadístico” (Quinlan, 1993).

2.3.2.3. Estructura

Partiendo de una Base de Datos con una variable respuesta $\boldsymbol{Y}$ a discriminar, y un conjunto finito de variables explicativas $X1, X2, \ldots, Xk$, se tratará de seleccionar entre las variables explicativas aquellas que discriminen mejor a la variable $\boldsymbol{Y}$; obteniéndose una partición de la población de forma que se encuentren dos o más subgrupos lo más heterogéneos posibles entre sí con respecto a la variable respuesta $\boldsymbol{Y}$, y lo más homogéneos posibles dentro. Esta discriminación se continúa para los nuevos nodos generados y se aplica un criterio de parada, obteniendo el árbol de clasificación o regresión.

Todo árbol de clasificación comienza con un nodo raíz, el resto de nodos se dividen en nodos intermedios y nodos hojas (también llamados nodos no terminales y nodos terminales).

- **Nodos intermedios:** llamados también segmentos intermedios, se generan dos o más segmentos descendientes inmediatos (dependiendo del método empleado).

- **Nodos terminales:** También denominado segmento terminal, es un nodo que no se puede dividir más.

- **Rama de un nodo t:** Consta de todos los segmentos descendientes de t, excluyendo t.

- **Árbol de decisión completo (A_{max}):** Árbol en el cual cada nodo terminal no se puede ramificar.

- **Sub-árbol:** Se obtiene de la poda de una o más ramas del árbol A_{max}.

A pesar de los diversos tipos de árboles de clasificación y regresión que existen, la forma de actuar en todos ellos es similar. Primero se debe tener un conjunto de datos con una variable respuesta (categórica o continua) y un conjunto de variables explicativas, todas ellas categóricas o continuas que han sido previamente categorizadas. Todos los registros de la base de datos son examinados para encontrar la mejor regla de clasificación de la variable respuesta. La secuencia de particiones define el árbol, donde cada partición se realiza para optimizar la clasificación del subconjunto de datos. El proceso de división es recursivo y finaliza la ramificación cuando se verifica un criterio de parada que ha debido ser definido previamente. (Puerta, 2002)

La construcción de un árbol de decisión se basa en cuatro elementos:

- Un conjunto de preguntas binarias Q de la forma $\{x \in A?\}$, donde A es un subconjunto del espacio muestral.
- El método usado para particionar los nodos.
- La estrategia requerida para parar el crecimiento del árbol.
- La asignación de cada nodo terminal a un valor de la variable de respuesta (regresión) o a una clase (clasificación).

“Las diferencias principales entre los algoritmos para construir árboles está en la regla para particionar los nodos, la

estrategia para podar los árboles, y el tratamiento de valores perdidos". (Acuña, 2004)

a) Formación de Nodos

Hay un gran número de posibles formas de efectuar divisiones en función de los valores que tomen las variables explicativas $X1, X2, \ldots, Xk$, y generalmente no se pueden considerar todas ellas; por lo que dependerá del tipo de variable que estemos tratando. (Puerta, 2002)

- **Variable cualitativa nominal:** "La variable toma C valores distintos entre los que no cabe establecer un orden natural. Si tenemos que discriminar con ayuda de una variable nominal los elementos que van a los distintos nodos hijos en el nodo t, podemos formar todos los subgrupos de los C valores que puede tomar Xi y enviar a un nodo los casos que generan la mejor discriminación con respecto a la variable respuesta y los restantes al otro nodo". (Puerta, 2002)

- **Variable cualitativa ordinal:** "Si la variable toma d valores, una vez ordenadas las categorías se consideran como posibles cortes $d - 1$ valores intermedios. Entre los posibles cortes se considerará el que proporcione grupos más homogéneos con respecto a la variable respuesta". (Puerta, 2002)

- **Variable cuantitativa continua:** "Se trabaja de la misma forma que con las variables ordinales, con la particularidad de que en este caso el número de valores de corte a comprobar será elevado debido al caso de no haber repeticiones, $n - 1$ cortes en el caso de ser n el tamaño de la muestra. De este conjunto

se seleccionarán los grupos que mejor discriminen los individuos con respecto a la variable respuesta". (Puerta, 2002)

2.3.2.4. Medidas de Impureza

a) Impureza de un nodo

Para decidir qué variable se va a utilizar para hacer la partición en un nodo, se calcula primero la proporción de observaciones que pasan por el nodo para cada uno de los grupos. Si se denomina a los nodos como: $t = 1,2, \dots, T$ y $p\ g|t$ a las probabilidades de que las observaciones que lleguen al nodo t pertenezcan a cada una de las clases. (Marín, 2009)

Se define la impureza del nodo t:

$$i(t) = \emptyset\ p\ 1\ |t, p\ 2|t, \dots, p\ G|t$$

Dónde: $\emptyset$ es la función de impureza y, p (g| t) puede calcularse empíricamente como la proporción de casos de clase g en el nodo t (Cortijo, 2001). Es decir:

$$p\ g|t = \frac{n_g\ (t)}{n\ (t)}$$

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

un número excesivo de nodos por lo que se hace necesario definir procedimientos de poda o simplificación del mismo. (Marín, 2009)

b) Impureza de un árbol

La impureza del árbol T, denotado por $I(T)$, se define:

$$I(T) = \sum_{t \in T} I(t) \sum_{t \in T} p\ t\ .\ i(t)$$

*Donde $p(t)$ es la probabilidad de que un caso cualquiera esté en el nodo t.

La impureza de un árbol se calcula solo en base al conjunto de nodos terminales. Además la selección continuada de las particiones que maximizan $\Delta\ i(s,t)$ es equivalente a seleccionar las particiones que minimizan la impureza global $I(T)$, lo que significa que la estrategia de selección de la mejor partición en cada nodo conduce a la solución óptima considerando el árbol final. (Cortijo, 2001)

c) Estimación de la tasa de error

La elección de un árbol respecto de otro dependerá en general de una estimación de su tasa de error $R(T)$, y para poder realizar la estimación de dicha tasa existen diversas formas de calcular la estimación:

- **Estimador por re-sustitución o intramuestral:** "Es el estimador más simple. Consiste en dejar caer por el árbol la misma muestra que ha servido para construirlo, pero como los árboles tienen gran flexibilidad para adaptarse a la muestra se puede obtener una estimación sesgada inferiormente de la tasa de error, y por tanto desconocer realmente el error real del árbol". (Puerta, 2002)

- **Estimador por muestra de validación (muestra de contraste):** Es empleado cuando se tiene tamaño de muestra muy grande debido a que no se pierde mucha información al eliminar del estudio una muestra para la estimación del error. Consiste en dejar caer por el árbol una muestra distinta a la empleada para la realización del árbol. Tenemos de esta forma un estimador de $R(T)$ insesgado, sin embargo este tiene el inconveniente de forzar a reservar, para la validación, una parte de la muestra la cual podía haberse empleado en la construcción del árbol; perdiendo información. (Puerta, 2002)

- **Estimación por validación cruzada:** Se deja fuera de la muestra a una fracción m^t del tamaño muestral total para la construcción del árbol. Obteniéndose m^t estimaciones $R^{(1)}(T), \ldots\ldots, R^m(T)$ y promediándolas de la siguiente forma:

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

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

- **Estimador bootstrap:** Recientemente se ha propuesto esta técnica de remuestreo para la estimación de la tasa de error. Ripley (1996). (Puerta, 2002)

d) Reglas de Parada

Distintos criterios de parada pueden provocar la finalización de los algoritmos que realizan árboles de clasificación o regresión. Entre las causas se encuentran:

- “Se ha alcanzado la máxima profundidad del árbol permitida” (Puerta, 2002).
- No pueden realizarse más particiones, porque se ha verificado alguna de las siguientes condiciones: “No hay variables explicativas significativas para realizar la partición del nodo, “El número de elementos en el nodo terminal es inferior al número mínimo de casos permitidos para poder realizar la partición” y que “El nodo no se podrá dividir en el caso en el cual el número de casos en uno o más nodos hijos sea menor que el mínimo número de casos permitidos por nodo”. (Puerta, 2002).

2.3.3. Redes Neuronales

Los primeros teóricos que concibieron los fundamentos de la computación neuronal fueron Beltran Russell, Warren McCulloch y Walter Pitts, quienes en 1943 lanzaron una teoría acerca de la forma de trabajar de las neuronas. A partir de 1986, el panorama fue alentador con respecto a las investigaciones y el desarrollo de las redes neuronales, es así que en 1988 fue formada la Sociedad Internacional de Redes Neuronales. Actualmente, son numerosos los trabajos que se realizan y publican cada año, las aplicaciones nuevas que surgen (sobre todo en el área de control) y las empresas que lanzan al mercado productos nuevos, tanto hardware como software (sobre todo para simulación).

2.3.3.1. Definición

Consiste en un modelo de nodos e interconexiones que recrea el funcionamiento inter neural del cerebro humano. Se define como técnicas de explotación de datos que extraen datos de una información implícita, desconocida y potencialmente útil; comprenden muchos modelos y métodos de aprendizaje. Respecto

al modo interno de trabajo las redes neuronales son modelos matemáticos multivariantes que utilizan procedimientos iterativos, con el objetivo de minimizar una determinada función de error.

Las Redes Neuronales son sistemas definidos por funciones $f(\cdot)$, que en forma única traza un patrón de entrada en un patrón de salida. “Cuando la entrada al sistema es denotada por el vector X y la salida por el vector Y, la relación entrada-salida puede ser escrita de la forma $Y = f(X, W)$, donde W denota los pesos de la red. Los pesos y la estructura de los nodos interconectados en el sistema definen la transformación de entrada-salida desarrollado por la red”. (Kraneau, 2007)

Estadísticamente, las RNA son modelos no lineales, constituyen una nueva técnica no paramétrica de análisis de datos multivariante ya que no requiere de supuestos. Es mucho más flexible y permite formular relaciones más complejas que las tradicionales técnicas estadísticas (Molera & Caballero, 2001). Con las redes neuronales se puede realizar automática y eficientemente múltiples tareas como modelación, optimización, regresión, clasificación, lógica difusa, patrones y rasgos ocultos, memorización, aprendizaje asociativo, control adaptativo, Pronóstico y Predicción de Series de Tiempo, etc.

2.3.3.2. Componente de una Red Neuronal

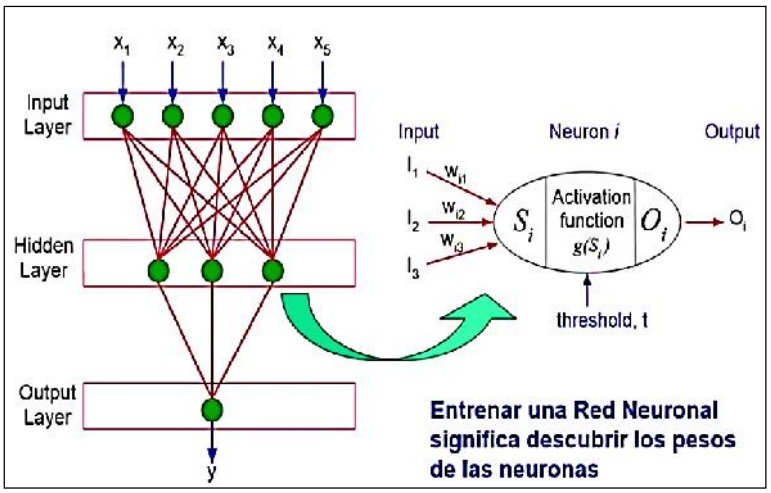


Figura 7. Esquema de una Red Neuronal

La RNA está constituida por neuronas interconectadas y arregladas por capas de entrada, oculta y de salida. Los datos ingresan en la capa de entrada, pasan a través de la o las capas ocultas y salen por la capa de salida. Si no hubiese capa oculta es como si se trabajara con el Perceptron Simple. Al modelo de la neurona se le llama habitualmente nodo o unidad.

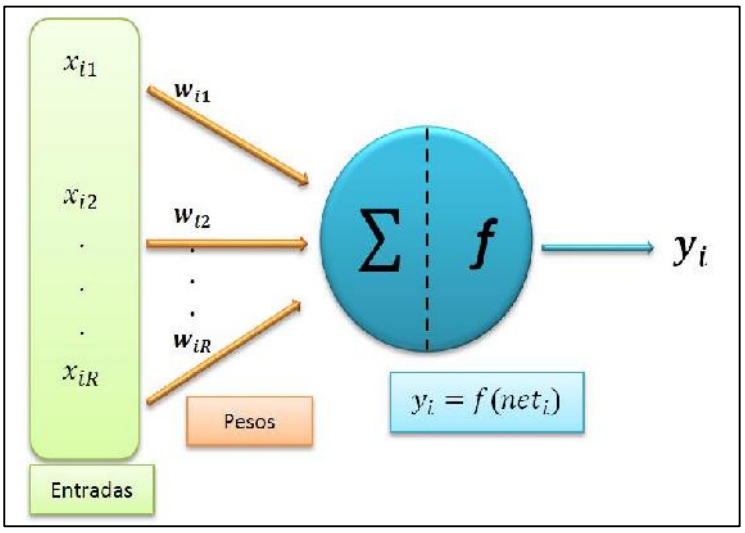


Figura 8. Componentes de una Red Neuronal

a) Entradas

Los patrones de entrada son las variables independientes denotadas como Xj, las cuales ingresan a la neurona j, donde j =1,2,…, R y R es el número de variables independientes; cada registro es denotado como $X_i = (\mathrm{X_{i1}}, \mathrm{X_{i2}}, \ldots, \mathrm{X_{iR}})$ es un vector de orden $1 \times \mathrm{R}$, donde i=1,2,..., n, y n es el número de registros o individuos.

b) Pesos

Generalmente una neurona recibe muchas y múltiples entradas simultáneas. Cada entrada tiene su propio peso relativo $(\mathrm{w_{ij}})$ el cual proporciona la importancia de la entrada dentro de la función de agregación de la neurona. Estos pesos realizan la misma función que realizan las fuerzas sinápticas de las neuronas biológicas. En ambos casos, algunas entradas son más importantes que otras de manera que tienen mayor efecto sobre el procesamiento de la neurona al combinarse para producir la respuesta neuronal. Los pesos son coeficientes que pueden adaptarse dentro de la red que determinan la intensidad de la señal de entrada registrada.

2.3.3.3. Estructura de una Red Neuronal

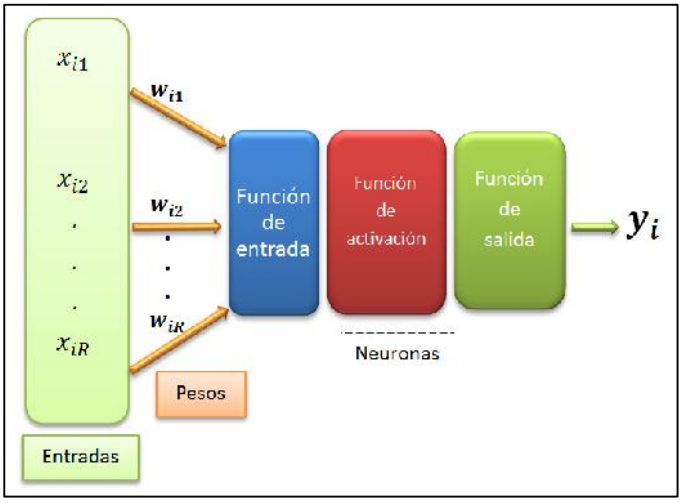


Figura 9. Ejemplo de una neurona con R entradas y 1 salida

a) Función de entrada o de propagación

Esta regla "Permite obtener a partir de las entradas y los pesos, el valor de la entrada neta o potencial post-sináptico "hi" de la neurona: $h_i(t) = \sigma_1(w_{ij}, x_j)$" (Mtz. de Lejarza, 1998)

La función más utilizada es la suma ponderada de todas las entradas, es decir se agrupan las entradas y pesos en dos vectores $(x1, x2, \ldots, xR)$ y $(w_{1j}, w_{2j}, \ldots, w_{Rj})$, con este se calcula la suma realizando el producto escalar sobre los dos vectores denominado también función lineal del tipo hiperplano.

$$h_i(t) = \sum_{j=1}^{R} w_{ij} \, . \, x_j$$

La función de propagación puede ser más compleja que simplemente una suma de productos. Las entradas y los pesos pueden ser combinados de diferentes maneras antes de pasarse el valor a la función de activación. El algoritmo específico para la propagación de las entradas neuronales está determinado por la elección de la arquitectura.

b) Función de activación o transferencia

Esta regla se encuentra en función de la suma ponderada de los registros ingresados y de los sesgos $(bias)$, denotándose de la siguiente forma:

$$a_i(t) = f_i(b_i, h_i(t))$$

En la función de activación el valor de la salida de combinación puede ser comparada con algún valor umbral para determinar la salida de la neurona. Si la suma es mayor que el valor umbral, neurona generará una señal. Si la suma es menor

que el valor umbral, ninguna señal será generada. Normalmente el valor umbral, o valor de la función de transferencia, es normalmente no lineal.

Funciones de transferencia más utilizadas:

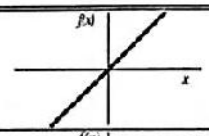
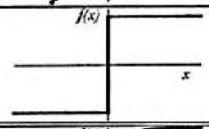
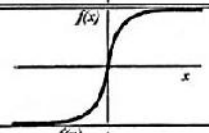
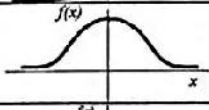
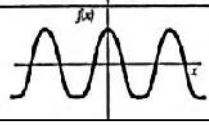
	Función	Rango	Gráfica
Identidad	$y = x$	$[-\infty, +\infty]$	
Escalón	$y = sign(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0, +1\}$	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$ $y = tgh(x)$	$[0, +1]$ $[-1, +1]$	
Gaussiana	$y = Ae^{-Bx^2}$	$[0,+1]$	
Sinusoidal	$y = A\,\mathrm{sen}(\omega x + \varphi)$	$[-1,+1]$	

Figura 10. Funciones de transferencia más utilizadas

Funciones de activación del SAS:

El valor producido por la función de combinación se transforma mediante una función de activación, lo que implica no hay pesas u otros parámetros estimados. Varios tipos generales de funciones de activación se utilizan comúnmente (SAS Enterprise miner 13.1).

- Arc Tan: $H_k = \frac{2}{\pi} * tan^{-1} n_k$
- Elliot: $H_k = \frac{n_k}{1+|n_k|}$
- Hyperbolic Tangent: $H_k = \tanh(n_k)$

- Logistic: $H_k = \frac{1}{1+\exp(n_k)}$
- Gauss: $H_k = \exp(-0.5 * n_k^2)$
- Sine: $H_k = \sin(n_k)$
- Cosine: $H_k = \cos(n_k)$
- Exponencial: $H_k = \exp(n_k)$
- Square: $H_k = n_k^2$
- Reciprocal: $H_k = \frac{1}{n_k}$
- Softmax: $H_k = \frac{1}{\sum_{j=1}^{M} \exp(n_j)}$

c) Función de salida (Competitividad)

Cada elemento de procesamiento tiene permitido una única salida $yi(t)$ que puede estar asociada con un número elevado de otras neuronas. Usualmente, la salida es directamente equivalente al valor resultante de la función de activación.

$$y_i(t) = F_i\big(a_i(t)\big) = a_i(t)$$

Algunas topologías de redes neuronales, sin embargo, modifican el valor de la función de transferencia para incorporar un factor de competitividad entre neuronas que sean vecinas. Las neuronas tienen permitidas competir entre ellas, inhibiendo a otras neuronas a menos que tengan una gran fortaleza.

d) Función de error y el valor propagado hacia atrás

En la mayoría de algoritmos de entrenamiento de redes neuronales necesitamos calcular la diferencia entre la salida actual y la esperada. Esta diferencia es transformada por la función de error correspondiente a la arquitectura particular. El error de la

neurona se propaga normalmente dentro del algoritmo de aprendizaje de otra neurona. Este término de error es algunas veces denominado error actual. El error actual es propagado hacia atrás a la capa anterior, siendo este valor o bien el valor actual de error de esa capa obtenido al escalarlo de alguna manera (lo habitual es usando la derivada de la función de transferencia) o bien es tomado como la salida esperada (esto sucede en algunas topologías). Normalmente el valor que se propaga hacia atrás, una vez escalado por la función de aprendizaje, se multiplica por los pesos de la neurona para modificarlas antes de pasar al ciclo siguiente.

2.3.3.4. Escalamiento y limitación

"El valor de salida de la función de activación puede ser procesado de manera adicional mediante un escalamiento y limitación. El escalamiento simplemente multiplica el valor de la función de transferencia por un factor de escala y después se le suma un desplazamiento". (Piedra, 2007)

El mecanismo de limitación asegura que el resultado del escalamiento no excede una cota superior o inferior. Esta limitación se realiza de manera adicional a los límites que puede imponer la función de transferencia original. Normalmente este tipo de escalamiento y limitación es principalmente usado en topologías para verificar modelos neuronales biológicos. (Piedra, 2007)

2.3.3.5. Función de una Red Neuronal

a) Funcionamiento de una Red Neuronal

Una red neuronal originalmente no dispone de ningún tipo de conocimiento útil almacenado; por ese motivo, para que ejecute

una tarea es preciso entrenarla o lo que en estadística sería como “estimar parámetros”.

b) Procedimiento estadístico

Primero se selecciona un conjunto de datos, o patrones de aprendizaje en términos de RNA, luego se procede a desarrollar la arquitectura neuronal, número de neuronas, tipo de red. Después se debe seleccionar el modelo y los números de variables dependientes e independientes, una vez realizado lo anterior se procede a la fase de aprendizaje o estimación del modelo y a la fase de prueba, con la cual se evaluará el modelo. Finalmente se validan los resultados. (Ver figura 11)

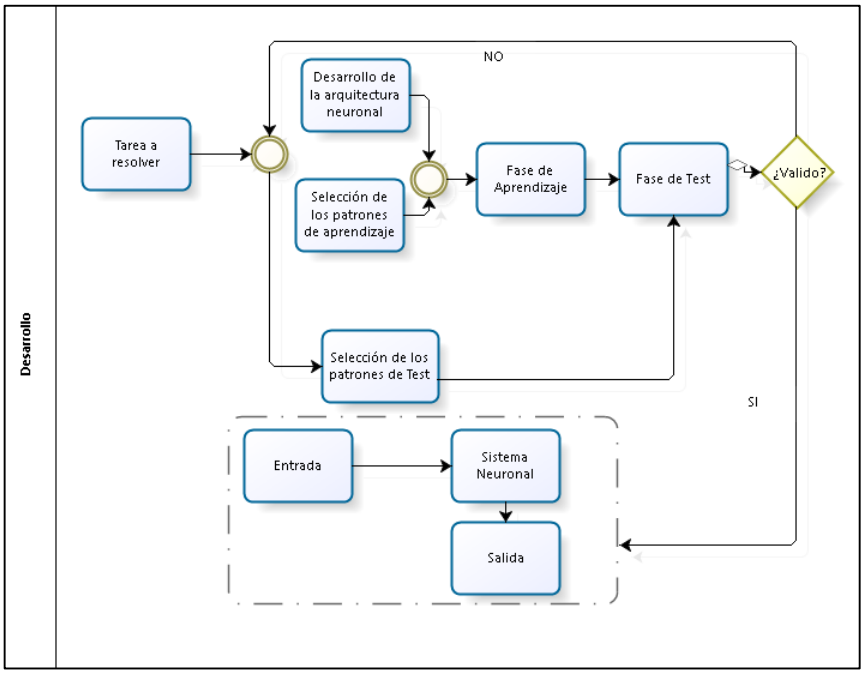


Figura 11. Diagrama de flujo redes neuronales

c) Modelos

Según (Serrano & Martin, 1995) existen diversos modelos que aparecen en estudios académicos y en bibliografía especializada. Entre ellos se encuentran:

- Adaline
- Crecimiento dinámico de células
- Gas Neuronal Creciente
- Mapas Autoorganizados (RNA)
- Máquina de Bolzman
- Máquina de Cauchy
- Memorias asociativas
- Perceptron
- Perceptron multicapa
- Propagación hacia atrás (Backpropagation)
- Red de contrapropagación
- Redes ART (*Adaptative Resonance Theory*)
- Redes de Elman
- Redes de Hopfield
- Redes de neuronas de aprendizaje competitivo
- Redes de neuronas de base radial

Técnicas de Entrenamiento del SAS:

El nodo de red neuronal proporciona una amplia variedad de técnicas de formación, incluyendo tanto las técnicas de optimización convencionales, y técnicas de la literatura de la red neural. Las técnicas de optimización convencionales más populares incluyen los siguientes:

- **Trust-Region:** El método Trust-Región se recomienda para las pequeñas y medianas problemas de optimización con hasta 40 parámetros (SAS Enterprise miner 13.1).

- **Levenberg-Marquardt:** Es muy rápido y fiable para pequeñas redes de mínimos cuadrados, pero requiere una gran cantidad de memoria (cuadrática en el número de hasta 100 parámetros estimados) (SAS Enterprise miner 13.1).

- **Cuasi-Newton techniques:** Son buenos para redes medianas. Ellos requieren memoria cuadrática, pero sólo alrededor de la mitad que el de Levenberg-Marquardt. Técnicas cuasi-Newton por lo general requieren más iteraciones que Levenberg-Marquardt, pero cada iteración requiere menos cálculos de punto flotante (SAS Enterprise miner 13.1).

- **Conjugate gradient techniques:** Son buenos para redes de gran tamaño cuando no hay suficiente memoria para las técnicas anteriores - los requisitos de memoria son solamente lineal. Por lo general requieren más iteraciones que cualquiera de las técnicas anteriores, pero cada iteración requiere menos cálculos de punto flotante. Técnicas de gradiente conjugado puede ser una buena opción si usted tiene una unidad de disco muy rápido, se utiliza la técnica para grandes problemas de minería de datos con más de 500 parámetros (SAS Enterprise miner 13.1).

d) Fases en la modelación con las redes neuronales

Fase de entrenamiento: se usa un conjunto de datos o patrones de entrenamiento para determinar los pesos (parámetros) que definen el modelo de red neuronal (Marín J. M., s.f.). Las redes estarán listas para ser entrenadas, una vez que los pesos y *bias* han sido inicializados; pueden ser entrenadas para aproximar funciones (regresión no lineal), asociar o clasificar patrones…"Durante el entrenamiento los pesos y $bias$ de la red son ajustadas iterativamente para minimizar la función de error de la red; por defecto la función de desempeño para redes con conexiones hacia atrás es el cuadrado medio del error

(MSE), es decir el promedio del al cuadrado entre las salidas de la red *a* y los objetivos *t".* (Aguilar, 2009)

Para determinar cómo ajustar los pesos, todos los algoritmos de entrenamiento utilizan la función del gradiente de desempeño, que a su vez se determina mediante la técnica de retro propagación, que involucra la realización de hacer cálculos hacia atrás de la red. (Aguilar, 2009)

Fase de Prueba: Durante el entrenamiento, puede que el modelo se ajuste demasiado a las particularidades presentes en los patrones, perdiendo su habilidad de generalizar su aprendizaje a casos nuevos (sobreajuste). Para evitar el sobreajuste, lo ideal es utilizar un grupo complementario al de entrenamiento (grupo de validación) que permita controlar el proceso de aprendizaje. (Marín J. M., s.f.)

"Generalmente, los pesos óptimos se obtienen optimizando (minimizando) alguna función de energía. Por ejemplo, un criterio muy utilizado en el llamado entrenamiento supervisado, es minimizar el error cuadrático medio entre el valor de salida y el valor real esperado". (Marín J. M., s.f.)

2.4. METODOLOGIA CRISP-DM PARA MODELADO DE PROCESOS DE DATA MINING

Rodríguez (2010) Menciona que son diversos lo modelados de proceso que han sido propuestos para el desarrollo de proyectos de Data Mining, entre ellos figuran: DMAMC (*Definir, Medir, Analizar, Mejorar, Controlar)* usado por Seis Sigma, SEMMA (siglas en ingles de *Muestreo, Exploración, Modificación, Modelado y Evaluación*) desarrollada por SAS Enterprise Miner, y CRISP-DM *(Cross Industry Standard Process for Data Mining)* incorporado por IBM SPSS Modeler. La presente investigación utilizó la metodología CRISP-DM para el modelamiento de los datos (Rodríguez O. , 2010).

Se puede describir a CRISP-DM como un proceso jerárquico dividido en cuatro niveles de abstracción, que a su vez están conformados por un conjunto de tareas que parten de un nivel general hacia lo específico (ver figura 12) y que organiza el desarrollo de un proyecto de Data Mining en una serie de seis fases (ver figura 13).

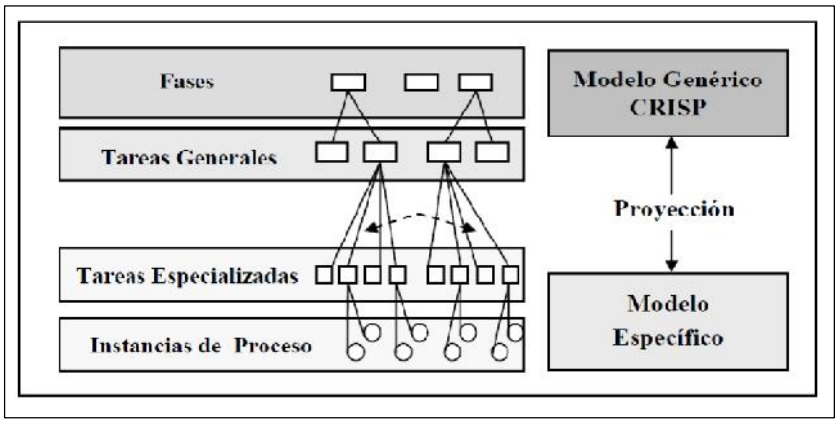


Figura 12. Esquema de los 4 niveles de CRISP-DM (CRISP-DM consortium, 2000)

Una de las ventajas de este proceso es que la sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas

específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

2.4.1. Fase de Comprensión del Negocio

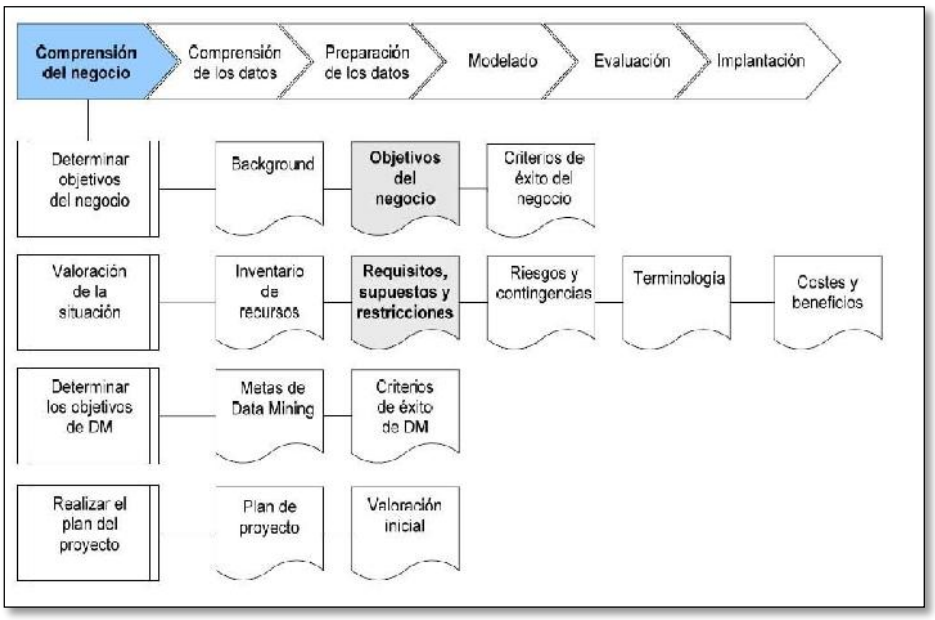


Figura 13. Fase de comprensión del negocio CRISP-DM (CRISP-DM consortium, 2000)

Para Rodríguez (2010) Esta fase inicial llamada también como fase de comprensión del problema, es quizás la más importante porque se enfoca en la comprensión de los objetivos y requisitos del proyecto desde un plano empresarial o institucional, con el fin de poder convertirlos en un problema de Data Mining y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. Para poder realizar un adecuado modelamiento es necesario que se entienda cual es el problema a resolver, ya que si no se es capaz de comprender dichos objetivos no se podrá obtener resultados fiables así se cuente con algoritmos muy sofisticados (Rodríguez O. , 2010).

Las principales tareas que comprenden son:

Determinar los objetivos del negocio: "Esta es la primera tarea a desarrollar y tiene como metas, determinar el problema que se pretende resolver, porqué la necesidad de utilizar minería de datos y definir los criterios de éxito" (Rodríguez O. , 2010). Todo analista de datos tiene como principal objetivo comprender lo que el cliente quiere lograr; "sino determinamos correctamente los objetivos del negocio se estaría haciendo un mal direccionamiento, que traería consigo gastar un gran esfuerzo produciendo respuestas correctas a preguntas incorrectas o erradas" (Dataprix, s.f.).

Evaluación de la situación: Implica evaluar recursos, restricciones, presunciones entre otros factores que deberán ser considerados antes de iniciar el proceso de minería de datos. "Aspectos como: ¿Qué conocimiento previo disponemos acerca del tema?, ¿Contamos con la cantidad de datos requerida para resolver el problema?, ¿Cuáles es la relación coste/beneficio de aplicar DM?, etc. ayudarán a definir los requisitos del problema, tanto en términos de negocio como en términos de minería de datos" (Rodríguez O. , 2010).

Determinación de los objetivos de DM: Tiene como finalidad representar los objetivos del negocio en términos de las metas de un proyecto de minería de datos. "Por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será determinar el perfil de los clientes según su capacidad de endeudamiento" (Rodríguez O. , 2010).

Producción de un plan del proyecto*:* La última tarea de esta fase tiene como meta describir el plan intencionado para alcanzar

los objetivos de minería de datos y así alcanzar los objetivos del negocio. En esta fase se debe especificar los pasos a ser realizados durante el resto del proyecto, incluyendo la selección herramientas y técnica en cada paso (Rodríguez O. , 2010).

2.4.2. Fase de comprensión de los datos

Para Rodríguez (2010) Esta fase "Comprende la recolección inicial de datos con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis" (ver figura 14).

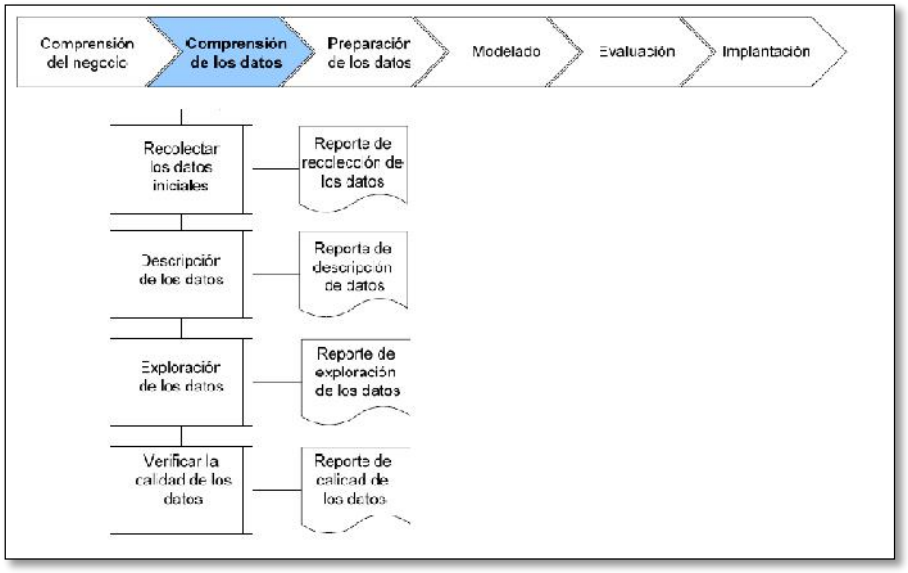


Figura 14. Fase de comprensión de los datos CRISP-DM (CRISP-DM consortium, 2000)

La comprensión de los datos junto con la preparación de los mismos y el modelado, son las que demandan un mayor esfuerzo y tiempo en un proyecto de DM. En el caso de que la organización cuente con una base de datos corporativa, es preferible crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes accesos a la base de datos con el fin de realizar consultas y de ser

necesario, algunas modificaciones, lo cual podría generar muchos problemas (Rodríguez O. , 2010).

Las principales tareas a desarrollar son:

Recolección de datos iniciales: Tiene como objetivo primordial la recolección de datos iniciales y su adecuación para un procesamiento futuro; para ello se debe elaborar informes en donde se liste los datos adquiridos, localización, las técnicas que se utilizaron en su recolección y los problemas y soluciones inherentes a este proceso (Rodríguez O. , 2010).

Descripción de los datos: "Involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial" (Rodríguez O. , 2010).

Exploración de datos: Tiene como finalidad encontrar una estructura general para los datos. La salida de esta tarea es un informe de exploración de datos que involucre la aplicación de análisis estadísticos simples, que revelen propiedades en los datos recién adquiridos, la creación tablas de frecuencia y construcción de gráficos de distribución (Rodríguez O. , 2010).

Verificación de la calidad de los datos: Una forma de evitar posibles problemas a futuro es mediante la realización de un análisis de la calidad de los datos disponibles antes de proceder al modelado. "En esta tarea, se realiza verificaciones de los datos y para ello se debe examinar la calidad de estos dirigiendo preguntas como: ¿Están los datos completos?, ¿Son correctos, o contienen errores?, ¿Hay valores omitidos en los datos?...etc." (Dataprix, s.f.).

2.4.3. Fase de preparación de datos

Rodríguez (2010) menciona que “Una vez realizada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos”. Esta fase incluye las tareas generales de selección de datos, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (ver figura 15). La preparación de los datos guarda relación con la fase de modelado al punto de interactuar permanentemente, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas (Rodríguez O. , 2010).

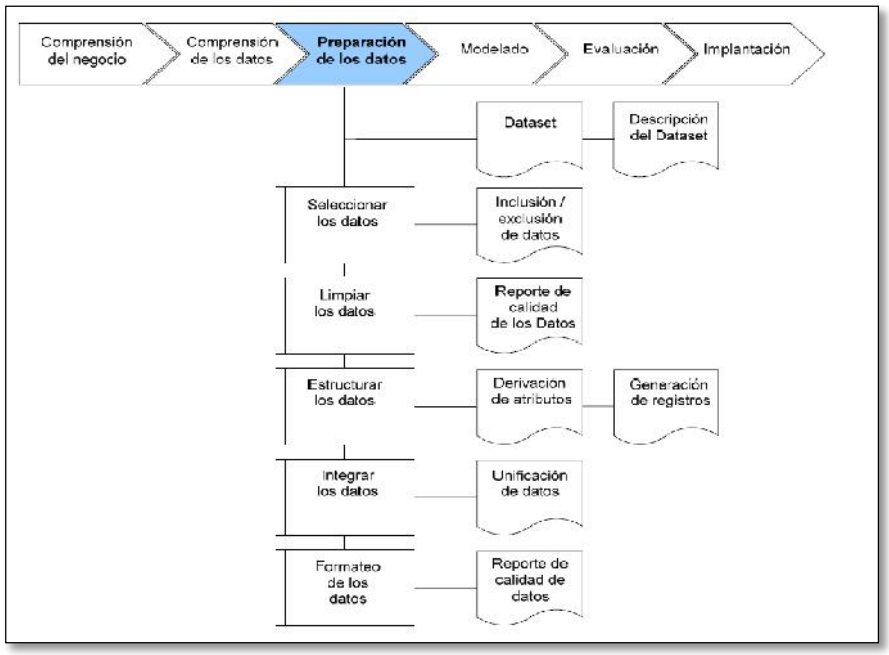


Figura 15. Fase de preparación de los datos CRISP-DM (CRISP-DM consortium, 2000)

Las principales tareas a desarrollar en esta fase del proceso son:

- ***Selección de datos.*** En esta etapa se debe decidir qué datos serán usados para el análisis, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas.

- ***Limpieza de los datos****:* Es una de las tareas que más tiempo y esfuerzo consume, debido a la diversidad de técnicas (normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.) que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación (Rodríguez O. , 2010).

- ***Estructuración de los datos:*** "Incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes" (Rodríguez O. , 2010).

- ***Integración de los datos:*** Esta tarea involucra crear de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, creación de nuevos registros, generación de nuevos campos a partir de otros existentes, fusión de tablas o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen (Rodríguez O. , 2010).

- ***Formateo de los datos:*** Consiste en realizar transformaciones sintácticas de los datos pero sin cambiar su significado, con la

finalidad de poder emplear alguna técnica de DM en particular, como “la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.” (Rodríguez O. , 2010).

2.4.4. Fase de Modelado

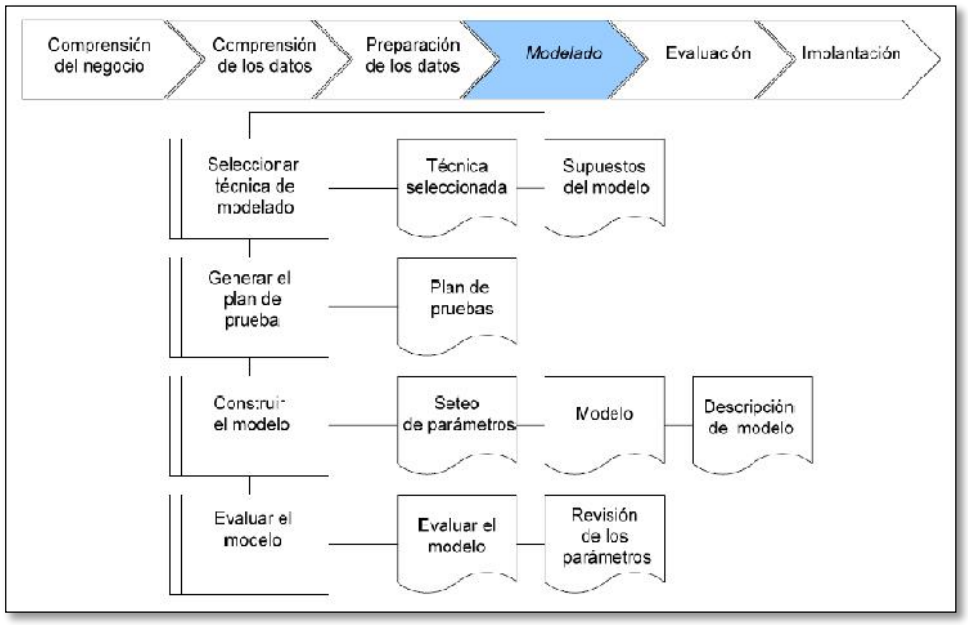


Figura 16. Fase de modelado CRISP-DM (CRISP-DM consortium, 2000)

Rodríguez (2010) manifiesta que en esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico, y que estas “se eligen en función a criterios como: ser apropiada al problema, disponer de datos adecuados, cumplir los requisitos del problema, tiempo adecuado para obtener un modelo, conocimiento de la técnica”. Además añade que, previo al modelado de los datos, se debe establecer un método de evaluación de los modelos que permita establecer el grado de bondad de ellos; después de concluir estas tareas, se procede a la generación y evaluación del modelo (ver

figura 16).

Las principales tareas a desarrollar en esta fase son:

- ***Selección de la técnica de modelado:*** Consiste en seleccionar la técnica de DM más apropiada al tipo de problema a resolver. “Para ello se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes” (Rodríguez O. , 2010).

- ***Generación del plan de prueba:*** *“*Antes de construir un modelo, se debe generar un procedimiento para probar la calidad y validez del modelo”. En tareas de minería de datos supervisados como la clasificación, es común usar tasas de errores como medida de calidad, y para ello se separan los datos en dos conjuntos; es decir, se construye un modelo basado en el conjunto de entrenamiento para después medir la calidad del modelo mediante el conjunto de validación (Rodríguez O. , 2010).

- ***Construcción del Modelo****:* Una vez seleccionada la técnica, se ejecuta sobre los datos que ya han sido tratados para generar uno o más modelos. “La selección de los mejores parámetros es un proceso iterativo y se basa en los resultados generados. Estos deben ser interpretados y su rendimiento justificado” (Rodríguez O. , 2010).

- ***Evaluación del modelo:*** Una vez definido el conjunto de modelos iniciales, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente y a los criterios de éxito preestablecidos (Rodríguez O. , 2010).

2.4.5. Fase de Evaluación

Rodríguez (2010) Menciona que en esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema; considera además, que la fiabilidad calculada para el modelo se aplica solamente para los datos con los que se realizó el análisis. De acuerdo a los resultados obtenidos, es preciso revisar el proceso para poder repetir algún paso anterior por si se haya cometido algún error (Ver figura 17). “Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo” (Rodríguez O. , 2010).

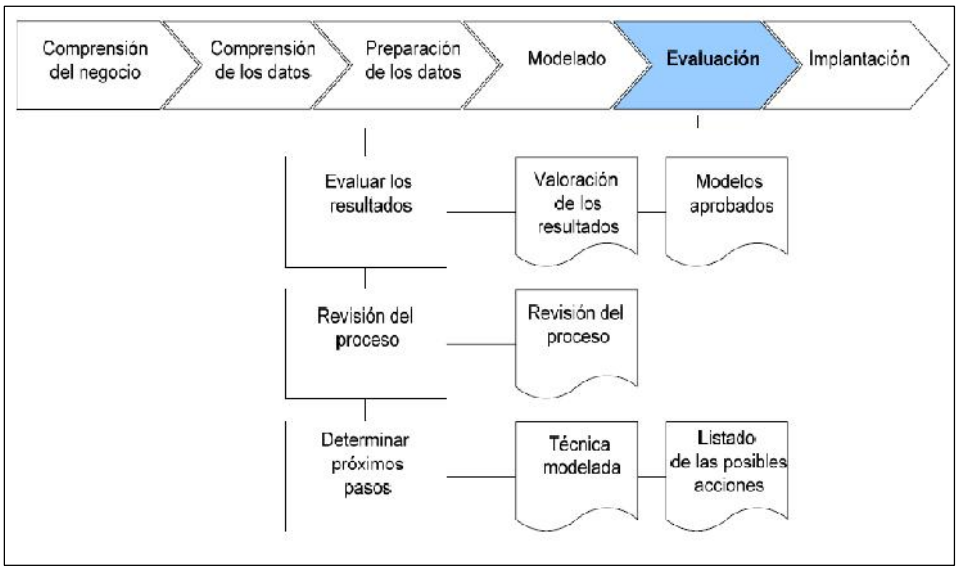


Figura 17. Fase de evaluación CRISP-DM (CRISP-DM consortium, 2000)

Las tareas que conforman esta fase son las siguientes:

- ***Evaluación de los resultados:*** En esta etapa, se formalizará su evaluación en función de si los resultados del proyecto cumplen con los criterios del rendimiento comercial y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo en un problema real si el tiempo y restricciones lo permiten.

Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional. (Dataprix, s.f.)

- ***Proceso de revisión:*** *"*En esta tarea, los modelos resultantes pasan a ser satisfactorios y a satisfacer las necesidades de negocio. Ahora es apropiado hacer una revisión más cuidadosa de los compromisos de la minería de datos para determinar si alguna tarea ha sido pasada por alto" (Rodríguez O. , 2010). Esta revisión también cubre cuestiones de calidad -por ejemplo: ¿Construimos correctamente el modelo? ¿Usamos sólo los atributos que nos permitieron usar y que están disponibles para análisis futuros?

- ***Determinación de futuras fases:*** Está en función de los resultados de la evaluación y la revisión de proceso. En esta tarea, el equipo de proyecto decidirá si debe terminar el proyecto y si es apropiado tomar medidas sobre el desarrollo, tanto iniciar más iteraciones, o comenzar nuevos proyectos de DM; también incluye los análisis de recursos restantes y del presupuesto, que puede influir en las decisiones (Dataprix, s.f.).

2.4.6. Fase de Implementación

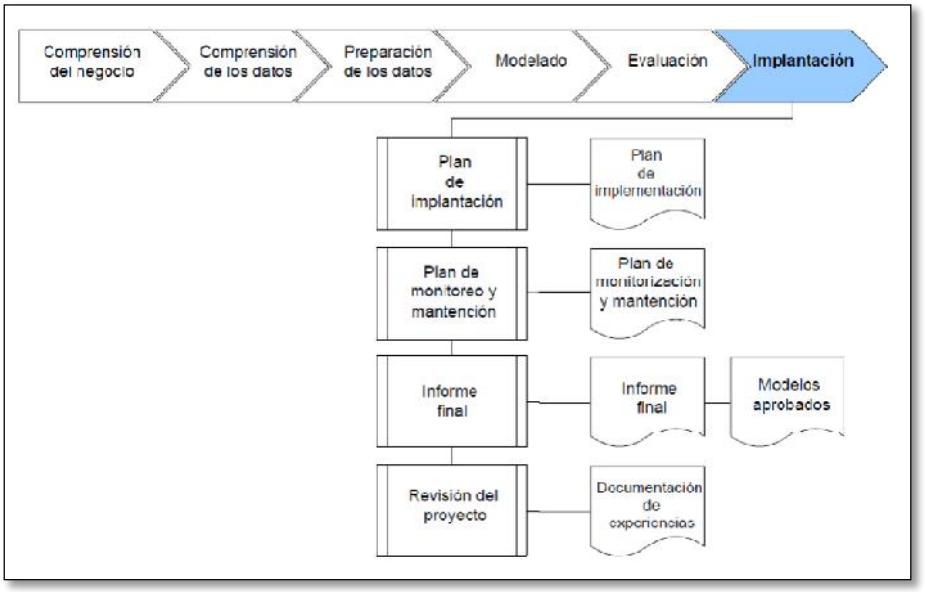


Figura 18. Fase de implementación CRISP-DM (CRISP-DM consortium, 2000)

Rodríguez (2010) hace mención que una vez que el modelo ha sido validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio y esto se logra cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados, o también cuando se aplica el modelo en diferentes conjuntos de datos. Un claro ejemplo son casos de análisis de riesgo crediticio, detección de fraudes, entre otros. Explica además que “La mayoría de veces un proyecto de Data Mining no concluye en la implantación del modelo, pues se debe documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento”. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la probable difusión de resultados (Rodríguez O. , 2010).

Las tareas que se ejecutan en esta fase son las siguientes:

- ***Plan de implementación:*** "Esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación" (Rodríguez O. , 2010).

- ***Monitorización y Mantenimiento:*** Una preparación detallada de una estrategia de mantenimiento previene a que se pase largos periodos innecesarios de uso incorrecto de resultados. "Para supervisar el desarrollo de los resultados de la minería de datos, el proyecto necesita un plan detallado de proceso de supervisión" (Dataprix, s.f.).

- ***Informe Final:*** Es la conclusión del proyecto de DM realizado, este informe puede ser una presentación final que incluya y explique los resultados logrados con el proyecto o puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda (Rodríguez O. , 2010).

- ***Revisión del proyecto:*** Es el paso final del método CRISP-DM y ofrece la oportunidad de formular sus impresiones finales e incorporar los conocimientos adquiridos durante el proceso de minería de datos. En este punto se evalúa qué fue correcto e incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar (Rodríguez O. , 2010).

III. MATERIALES Y MÉTODOS

3.1 Tipo de investigación

La presente investigación es aplicada con propósito predictivo y valor explicativo.

3.2 Población

La población de estudio está constituida por 3000 clientes que forman parte de la cartera de crédito personal de la Cooperativa de Ahorro y Crédito a julio del 2013.

3.3 Técnicas e instrumentos de recolección de los datos

En la presente investigación la principal fuente de información fue conformada por el historial crediticio de los clientes pertenecientes a la cartera de crédito personal, obtenida de la base de datos de la Cooperativa de Ahorro y Crédito objeto de estudio. Cabe resaltar que la información brindada a los investigadores no incumple o infringe la ley del secreto bancario, pues no se muestra nombres, apellidos o cualquier otro dato que identifique los clientes.

Entrevista personal al Administrador de cooperativa de ahorro y crédito objeto de estudio.

3.4 Análisis estadístico de datos

En la presente investigación la base de datos estuvo constituida por el historial crediticio de 3000 clientes que forman parte de la cartera de crédito personal de la Cooperativa de Ahorro y Crédito, contiene en total 30 variables clasificadas en 27 variables cuantitativas y 3 variables cualitativas. Por criterio de los investigadores, el análisis estadístico se llevó a cabo considerando todas las variables en estudio.

Al inicio fue necesario evaluar la relación entre la variable dependiente y las variables independientes con el objeto de encontrar al menos una variable explicativa que permita la construcción del modelo; así como probar la normalidad de las variables.

Antes de modelizar, se particionó la población para trabajar el entrenamiento y la validación del modelo, de esta manera se planteó que un 50% de los casos fueran destinados al conjunto de entrenamiento (construcción del modelo) y el otro 50% al conjunto de validación (ajuste y comparación de los modelos); no se consideró asignar un porcentaje al conjunto de testeo por la limitación de no contar con una mayor cantidad de datos. En la modelización predictiva del conjunto de datos de entrenamiento se generaron fácilmente modelos que predijeran el valor del TARGET (variable Respuesta) a partir de un conjunto de valores de entrada; sin embargo, estas predicciones solo fueron precisas para el propio conjunto de entrenamiento. El intento de generalizar las predicciones de este conjunto de datos a un conjunto independiente pero con una distribución similar puede producir resultados con un deterioro significativo de la precisión, con el fin de evitar este problema se utilizó el conjunto de validación como forma de evaluar independientemente la performance de un modelo.

Para tratar los valores faltantes que se presentaron en el modelo se trabajó con el método de valores faltantes para variables de clase: moda, constante, distribución, árbol de clasificación (con o sin reglas subrogantes).

En el modelamiento de datos se utilizaron las técnicas de Regresión Logística, Árboles de Clasificación y Redes Neuronales para la estimación de la probabilidad de default (PD), una vez hallado los modelos predictivos se procedió a comparar y evaluar utilizando estadísticos como: Matriz de confusión (sensibilidad, especificidad, valor predictivo positivo, precisión o

error de clasificación), la curva ROC, índice de GINI, el estadístico de Kolmogórov-Smirnov (KS) y la Curva de Porcentaje Acumulado de Respuesta Capturada (CAPC) para poder determinar el mejor modelo predictivo en eficiencia y predictibilidad para el otorgamiento del crédito personal en la cooperativa de ahorro y crédito.

El análisis se realizó mediante el software estadístico SAS versión 9.4, SAS Enterprise Miner Workstation 13.1, R Project 3.2.0 y Excel.

IV. RESULTADOS Y DISCUSIONES

4.1. Comprensión del negocio

a) Determinar los objetivos del negocio

- Maximizar la rentabilidad
- Optimizar la asignación de crédito.
- Automatizar el sistema de aprobación de crédito
- Minimizar el riesgo de incumplimiento de pago del crédito.

b) Determinación de los objetivos de minería de datos

Objetivo General

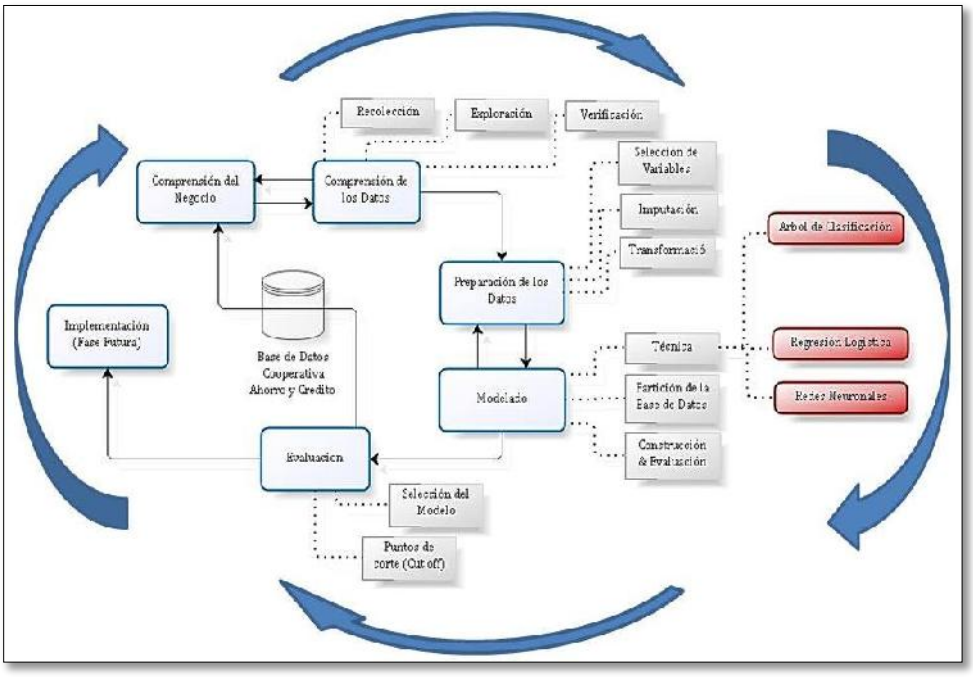
Construir un modelo de Credit scoring que permita predecir el otorgamiento de crédito personal en la cooperativa de ahorro y crédito.

Objetivos Específicos

- Identificar al menos una variable explicativa que pueda ser considerada en la construcción de un modelo de Credit scoring.
- Construir modelos mediante las técnicas de Arboles de Clasificación, Redes Neuronales y Regresión Logística que permitan predecir la probabilidad de incumplimiento del crédito personal.
- Determinar el mejor modelo basado en indicadores eficiencia y predictibilidad.

c) Producción de un plan del proyecto

Figura 19: Plan del proceso de Minería de Datos



Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

d) Herramientas y Técnicas

Se requiere una estación de trabajo que cuente con algunas características técnicas aceptables como: memoria RAM de por lo menos 4Gb, 1Gb de espacio en el disco duro, acceso al Data Warehouse de la organización. Para la presente investigación se utilizó el SAS 9.4 y el SAS Enterprise Miner 13.1, que cuentan con herramientas de conectividad a Excel y a sistemas de gestión de bases de datos.

Para el modelamiento de datos se utilizó las técnicas de Regresión Logística, Árboles de Clasificación y Redes Neuronales.

4.2. Comprensión de los datos

a) Recolección de datos iniciales

Los datos fueron obtenidos de la cartera de crédito personal de la Cooperativa de ahorro y crédito con fecha de corte a Julio del 2013.

b) Descripción de los datos

El número de registros que conforma la base de datos es de 3000 clientes, contiene en total 30 variables clasificadas en 27 variables cuantitativas, 3 variables cualitativas.

Variables Cualitativas

- Target: Indicador de Calificación de Default (Variable Dependiente)
- IndBancarrota: Indicador de Bancarrota.
- ID: Código de identificación del cliente.

Variables Cuantitativas

- NumCob: Número de Cobranzas
- NumRepNeg: Número de Reportes Negativos
- NumVer06: Número de Verificaciones en 6 los últimos meses.
- NumVerFinac24: Número de Verificaciones Financieras en los últimos 24 meses.
- TTransUltVer: Tiempo transcurrido desde la última Verificación.
- LCred50: Número de Líneas de Crédito utilizadas al 50%.
- LCred75: Número de Líneas de Crédito utilizadas al 75%.
- LCredCM24: Número de Líneas de Crédito con calificación mala en los últimos 24 meses.
- LCredRepNeg: Número de Líneas de Crédito con calificación mala o Reportes Negativos.
- PLCredSal: Porcentaje de Líneas de Crédito con Saldo.
- TLCredAb: Número Total de Líneas de Crédito abiertas.

- TLCredAb03: Número Total de Líneas de Crédito abiertas en los últimos 3 meses.
- TLCredAb12: Número Total de Líneas de Crédito abiertas en los últimos 12 meses.
- TLCredAb24: Número Total de Líneas de Crédito abiertas en los últimos 24 meses.
- TLCredMor306024: Número Total de Líneas de Crédito con morosidad de 30 o 60 días en los últimos 24 meses.
- TLCredMor60: Número Total de Líneas de Crédito con morosidad de hasta 60 días.
- TLCredMor6024: Número Total de Líneas de Crédito con morosidad de 60 días o más en los últimos 24 meses.
- TLCredMor60M: Número Total de Líneas de Crédito con morosidad de 60 días o más.
- TLCredMor9024: Número Total de Líneas de Crédito con morosidad de 90 días o más en los últimos 24 meses.
- STLCred: Suma Total de todas las Líneas de Crédito.
- PLCredAb24: Porcentaje de Líneas de Crédito abiertas en los últimos 24 meses.
- PLCredAb: Porcentaje de Líneas de Crédito abiertas
- TLCredSat: Número Total de Líneas de Crédito Satisfactorias.
- PLCredSat: Porcentaje de Líneas de Crédito Satisfactorias.
- SumTLCred: Suma Total de Saldo en todas las Líneas de Crédito.
- TTransPLCred: Tiempo transcurrido desde la apertura de la Primera Línea de Crédito.
- TTransULCred: Tiempo transcurrido desde la apertura de la Última Línea de Crédito.

c) Exploración de los datos

Tabla 1: Análisis descriptivo de la cartera de crédito personal en la Cooperativa Ahorro y Crédito

Variable	Missing	Media	Desviación estándar	Mínimo	Mediana	Máximo	Asimetría	Curtosis
NumCob	0	0.857	2.161	0	0	50	7.557	111.837
NumRepNeg	0	1.43	2.731	0	0	51	5.045	50.938
NumVer06	0	3.108	3.479	0	2	40	2.58	12.821
NumVerFinac24	0	3.555	4.478	0	2	48	2.807	13.051
TTransUltVer	188	3.108	4.638	0	1	24	2.387	5.627
LCred50	99	4.078	3.108	0	3	23	1.443	3.351
LCred75	99	3.122	2.605	0	3	20	1.508	3.687
LCredCM24	0	0.567	1.324	0	0	16	4.377	28.583
LCredRepNeg	0	1.409	2.46	0	0	47	4.58	48.243
PLCredSal	41	0.648	0.266	0	0.696	3.361	-0.181	4.016
TLCredAb	3	7.88	5.422	0	7	40	1.236	2.195
TLCredAb03	0	0.275	0.582	0	0	7	2.806	12.668
TLCredAb12	0	1.821	1.925	0	1	15	1.624	3.685
TLCredAb24	0	3.882	3.397	0	3	28	1.608	4.380
TLCredMor306024	0	0.726	1.164	0	0	8	1.382	1.409
TLCredMor60	0	1.522	2.81	0	0	38	3.308	17.762
TLCredMor6024	0	1.068	1.806	0	0	20	3.08	14.350
TLCredMor60M	0	2.522	3.407	0	1	45	2.564	12.701
TLCredMor9024	0	0.815	1.61	0	0	19	3.624	19.701
STLCred	40	31205.9	29092.91	0	24187	271036	2.061	8.093
PLCredAb24	3	0.564	0.48	0	0.5	6	2.779	18.533
PLCredAb	3	0.496	0.207	0	0.5	1	0.379	-0.019
TLCredSat	4	13.512	8.932	0	12	57	0.851	0.690
PLCredSat	4	0.518	0.235	0	0.526	1	-0.124	-0.484
SumTLCred	40	20151.1	19682.09	0	15546	210612	2.277	10.964
TTransPLCred	0	170.114	92.814	6	151	933	1.031	2.860
TTransULCred	0	11.874	16.321	0	7	342	6.448	80.310

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013.

d) Verificación de calidad de datos

- Existen variables con valores missing las cuales son:

  TTransUltVer: 188 valores missing

  LCred50: 99 valores missing

LCred75 99 valores missing

PLCredSal: 41 valores missing

TLCredAb: 3 valores missing

STLCred: 40 valores missing

PLCredAb24: 3 valores missing

PLCredAb: 3 valores missing

TLCredSat: 4 valores missing

PLCredSat: 4 valores missing

SumTLCred: 40 valores missing

- Se identificó que la mayoría de variables son Asimétricas con excepción de: PLCredAb, PLCredSat, que tienen una distribución Normal (ver anexo 3).
- Se identificó que la mayoría de variables cuantitativas presentaron datos outliers con la excepción de: SumTLCred, PLCredSat, STLCred (ver Anexo 2).

4.3. Preparación de los datos

a) Selección de datos

Se seleccionó 28 variables excluyendo la TARGET y ID

b) Limpieza de datos

Tabla 2: Imputación de las variables con datos Missing

Variables	Imputar Método	Nombre Variable Imputada	Variable Imputada	Missing
TTransUltVer	CONSTANT	IMP_TTransUltVer	0	188
LCred50	DISTRIBUTION	IMP_LCred50	.	99
LCred75	DISTRIBUTION	IMP_LCred75	.	99
PLCredSal	DISTRIBUTION	IMP_PLCredSal	.	41
TLCredAb	DISTRIBUTION	IMP_TLCredAb	.	3
STLCred	CONSTANT	IMP_STLCred	0	40
PLCredAb24	CONSTANT	IMP_PLCredAb24	0	3
PLCredAb	DISTRIBUTION	IMP_PLCredAb	.	3
TLCredSat	DISTRIBUTION	IMP_TLCredSat	.	4
PLCredSat	CONSTANT	IMP_PLCredSat	0	4
SumTLCred	CONSTANT	IMP_SumTLCred	0	40

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

c) Estructuración de los datos

Se realizaron 4 métodos de transformación para la técnica de Regresión Logística:

Transformaciones simples

- Log: La variable se transforma al tomar el logaritmo natural de la variable.
- Square Root: La variable se transforma por la raíz cuadrada de la variable.
- Inverse: La variable se transforma mediante el uso de la inversa de la variable.
- Square: La variable se transforma por el cuadrado de la variable.
- Exponential: La variable se transforma utilizando el logaritmo exponencial de la variable.
- Standardize: La variable ha sido estandarizada restando la media y dividiendo por la desviación estándar.

Transformaciones por Agrupación

- Bucket: Los cubos se crean dividiendo los valores de datos en intervalos igualmente espaciados sobre la base de la diferencia entre los valores máximo y mínimo.
- Quantile: Los datos se dividen en grupos que tienen aproximadamente la misma frecuencia en cada grupo.
- Optimal Binning for Relationship to Target Transformation: Los datos se agrupan con el fin de optimizar la relación a la variable respuesta.

4.4. Modelado

a) Selección de la técnica de modelado

Se utilizaron las técnicas de Regresión Logística, Arboles de Clasificación y Redes neuronales para el modelado predictivo.

b) Generación del plan de Prueba

Se particionó la población planteándose que un 50% de los datos fueran destinados al conjunto de entrenamiento y el otro 50% al conjunto de validación; no se consideró asignar un porcentaje al conjunto de testeo por la limitación de no contar con una mayor cantidad de datos.

c) Construcción y Evaluación del modelo

Arboles de Clasificación

Se planteó construir tres tipos de Arboles de Clasificación (CHAID, CART y C4.5), la comparación de los mismos puede verse a continuación:

Tabla 3: Indicadores de los modelos de Árboles de Clasificación

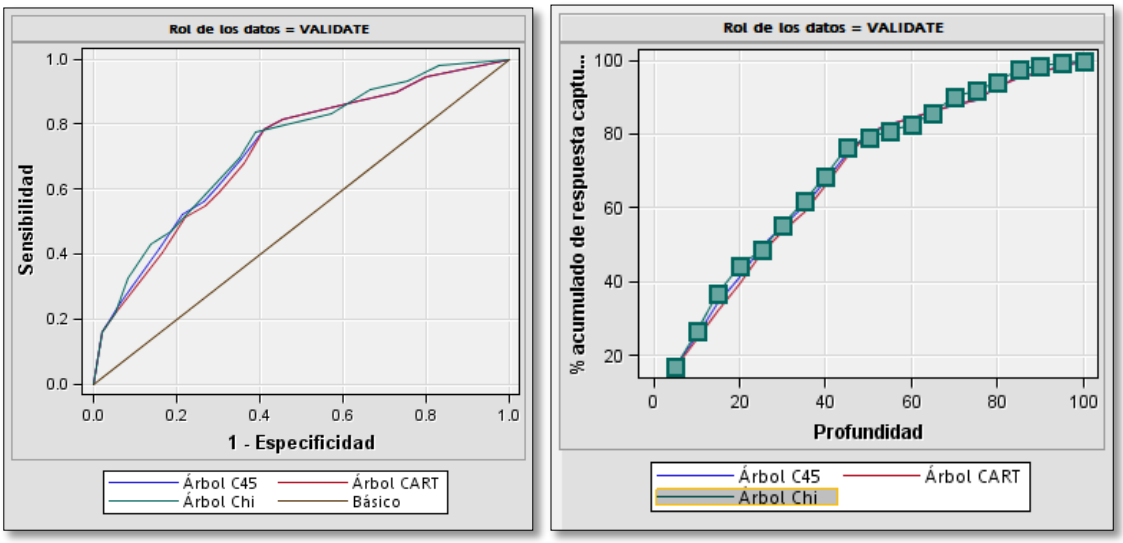
Modelos	ECM	RECM	ROC	GINI	KS	CAPC	SENS	ESP	VPP	VPN	Exactitud	Tasa de Error
Árbol CHAID	0.124	0.352	0.731	0.463	0.384	26.597	0.160	0.978	0.588	0.853	0.841	0.159
Árbol CART	0.127	0.357	0.722	0.444	0.371	24.667	0.160	0.978	0.588	0.853	0.841	0.159
Árbol C4.5	0.127	0.357	0.715	0.430	0.371	25.545	0.160	0.978	0.588	0.853	0.841	0.159

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Leyenda:

ECM: Error cuadrático medio
RECM: Raíz cuadrada del Error cuadrático medio
ROC: Curva ROC
GINI: Índice Gini
VPP: Valor predictivo positivo
KS: Estadístico Kolmogorov-Smirnov
CAPC: Capacidad de respuesta Capturada
SENS: Sensibilidad
E5SP: Especificidad
VPN: Valor predictivo negativo

Figura 20: Curvas ROC y CAPC para los modelos de Árboles de Clasificación



Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

En la *Figura 20*, se puede observar que el árbol CHAID (construido a partir de la base de datos original) es el que tiene el indicador de captura más alto; sin embargo los indicadores no son tan buenos como podría esperarse.

Regresión Logística

Se generaron cinco modelos de Regresión Logística, cuatro modelos se construyeron con la base original usando las transformaciones Normal, Bucket, Cuantil y Máximo Normal, y el último modelo se construyó con agrupación interactiva para seleccionar las variables de acuerdo al valor de la información.

Tabla 4: Indicadores de los modelos de Regresión Logística

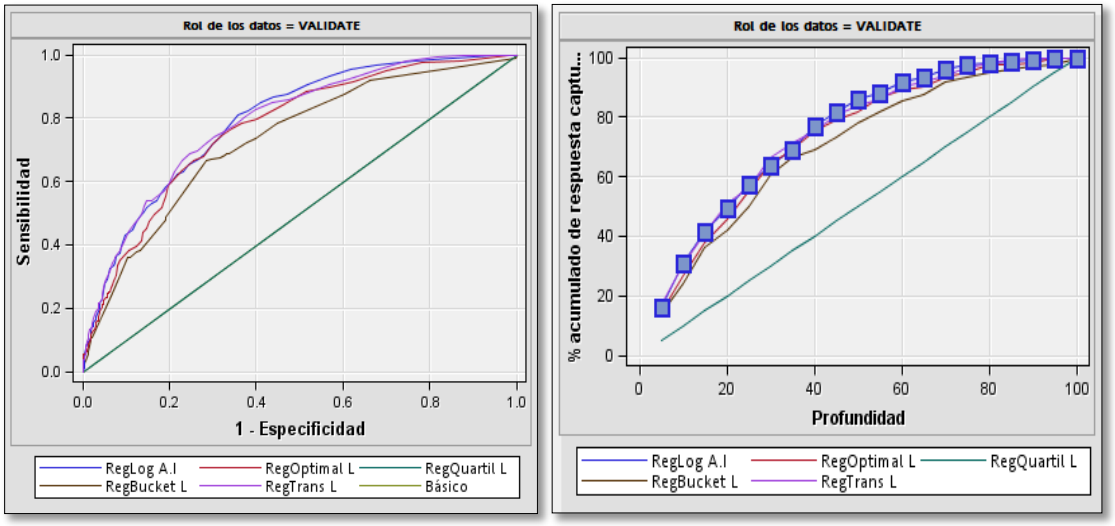
Modelo	ECM	RECM	ROC	GINI	KS	CAPC	SENS	ESP	VPP	VPN	Exactitud	Tasa de Error
Reg Log Trans	0.116	0.341	0.461	0.199	0.722	30.400	0.168	0.975	0.575	0.854	0.841	0.159
Reg Log Bucket	0.125	0.354	0.459	0.199	0.715	24.000	0.108	0.980	0.519	0.846	0.835	0.165
Reg Log Quartil	0.139	0.373	0.499	0.200	0.747	10.000	0.000	1.000	-	0.833	0.833	0.167
Reg Log Optimal	0.122	0.349	0.444	0.184	0.789	26.800	0.192	0.959	0.485	0.856	0.831	0.169
Reg Log A.I.	0.117	0.342	0.792	0.584	0.451	30.800	0.160	0.974	0.548	0.853	0.838	0.162

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Leyenda:

ECM: Error cuadrático medio
RECM: Raíz cuadrada del Error cuadrático medio
ROC: Curva ROC
GINI: Índice Gini
VPP: Valor predictivo positivo
KS: Estadístico Kolmogorov-Smirnov
CAPC: Capacidad de respuesta Capturada
SENS: Sensibilidad
ESP: Especificidad
VPN: Valor predictivo negativo

Figura 21: Curvas ROC y CAPC para los modelos de Regresión Logística



Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

En la *Figura 21,* se puede observar que hay dos modelos con indicadores aceptables; uno de ellos ha sido formado usando la variable transformada, mientras que el otro ha recurrido a las variables resultantes de la agrupación interactiva.

Redes Neuronales

Los modelos de Redes Neuronales que se compararon inicialmente se construyeron sobre las muestras original y agrupación interactiva. La muestra original consideró en la selección de las variables utilizar la Regresión Logística y Random Forest (donde se utilizó dos tipos de técnicas de optimización Levenberg-Marquardt y Gradiente Conjugado con la función de activación Softmax con tres hidden layers.)

Tabla 5: Indicadores de los modelos de Redes Neuronales

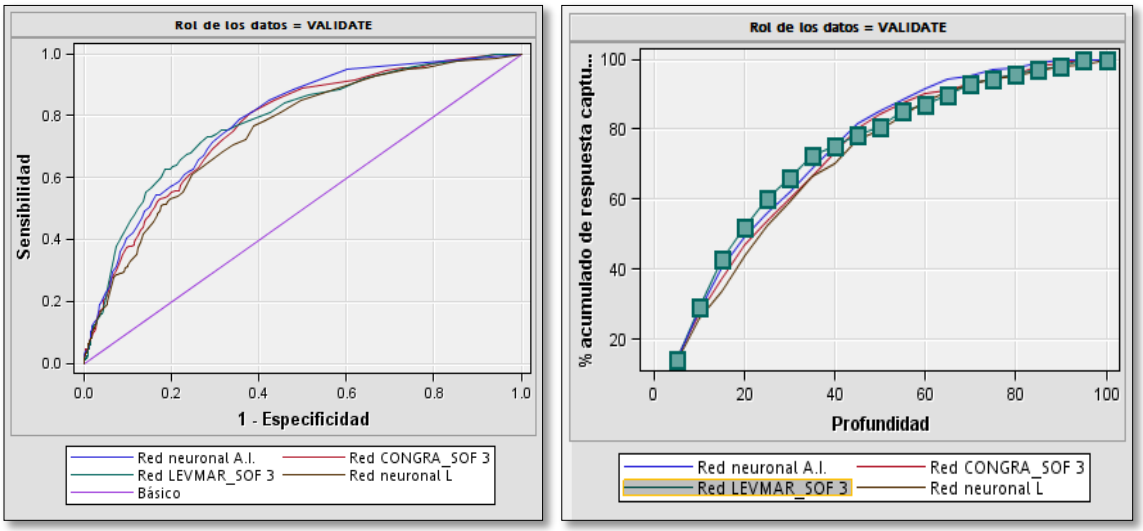
Modelos	ECM	RECM	ROC	GINI	KS	CAPC	SENS	ESP	VPP	VPN	Exactitud	Tasa de Error
Red N. A.I.	0.119	0.346	0.785	0.570	0.431	28.400	0.104	0.984	0.565	0.846	0.837	0.163
Red N. Con. Gra. Sof 3	0.125	0.353	0.768	0.535	0.424	27.600	0.260	0.940	0.464	0.864	0.827	0.173
Red N. Lev. Mar. Sof 3	0.120	0.346	0.779	0.558	0.450	29.200	0.124	0.975	0.500	0.848	0.833	0.167
Red Neuronal L.	0.128	0.358	0.747	0.493	0.381	26.000	0.156	0.964	0.464	0.851	0.829	0.171

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Leyenda:

ECM: Error cuadrático medio
RECM: Raíz cuadrada del Error cuadrático medio
ROC: Curva ROC
GINI: Índice Gini
VPP: Valor predictivo positivo
KS: Estadístico Kolmogorov-Smirnov
CAPC: Capacidad de respuesta Capturada
SENS: Sensibilidad
ESP: Especificidad
VPN: Valor predictivo negativo

Figura 22: Curvas ROC y CAPC para los modelos de Redes Neuronales



Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

En la Figura 22, se puede apreciar que hay dos modelos con indicadores aceptables, uno ha sido formado usando las variables por selección de Random Forest con la técnica de optimización Levenberg-Marquardt y la función de activación Softmax con tres hidden layers; mientras que el otro modelo ha recurrido a las variables resultantes de la agrupación interactiva.

4.5. Evaluación

4.5.1. Evaluación de resultados

Luego de haber identificado el o los mejores modelos de cada técnica, se procedió a determinar el mejor modelo basado en los indicadores de eficiencia y predictibilidad.

Tabla 6: Indicadores de los mejores modelos seleccionados

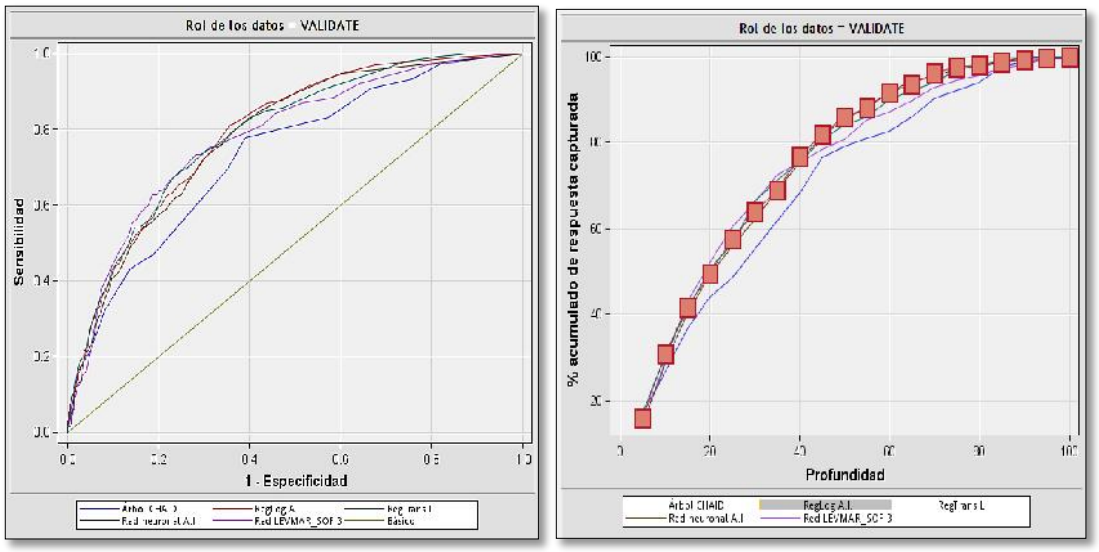
Variables	Árbol CHAID	Reg. Log Trans.	Reg. Log A.I	Red N. Lev. Mar. Sof. 3	Red N. A.I.
Target	X	X	X	X	X
IndBancarrota		X			
NumCob				X	X
NumRepNeg					X
NumVer06				X	X
NumVerFinac24	X	X	X	X	X
TTransUltVer				X	
LCred50					
LCred75		X		X	
LCredCM24					X
LCredRepNeg				X	X
PLCredSal	X	X	X	X	X
TLCredAb					
TLCredAb03		X			
TLCredAb12					
TLCredAb24					
TLCredMor306024	X	X	X	X	X
TLCredMor60				X	X
TLCredMor6024	X	X	X		X
TLCredMor60M				X	X
TLCredMor9024					X
STLCred				X	
PLCredAb24				X	
PLCredAb					
TLCredSat		X		X	
PLCredSat	X	X	X	X	X
SumTLCred					
TTransPLCred	X	X	X	X	X
TTransULCred				X	
ECM	0.124	0.116	0.117	0.120	0.119
RECM	0.352	0.341	0.342	0.346	0.346
ROC	0.731	0.789	0.792	0.779	0.785
GINI	0.463	0.579	0.584	0.558	0.570
KS	0.384	0.441	0.451	0.450	0.431
CAPC	26.597	30.400	30.800	29.200	28.400
SENS	0.160	0.168	0.160	0.124	0.104
ESP	0.978	0.975	0.974	0.975	0.984
VPP	0.588	0.575	0.548	0.500	0.565
VPN	0.853	0.854	0.853	0.848	0.846
Exactitud	0.841	0.841	0.838	0.833	0.837
Tasa de Error	0.159	0.159	0.162	0.167	0.163

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Leyenda:

ECM: Error cuadrático medio
RECM: Raíz cuadrada del Error cuadrático medio
ROC: Curva ROC
GINI: Índice Gini
VPP: Valor predictivo positivo
KS: Estadístico Kolmogorov-Smirnov
CAPC: Capacidad de respuesta Capturada
SENS: Sensibilidad
ESP: Especificidad
VPN: Valor predictivo negativo

Figura 23: Curvas ROC y CAPC de los mejores modelos seleccionados.



Fuente: Elaborado en base a datos de la cooperativa de ahorro y crédito, Julio del 2013

En la Figura 23 se puede observar las variables predictoras y los estadísticos de resumen de los mejores modelos, para el estudio realizado se afirma que el modelo óptimo para identificar la probabilidad de clientes buenos de los malos es la regresión logística por agrupación interactiva (Reg. Log. A.I.) ya que presenta un error cuadrático medio de 0.117, curva de ROC con 0.792, índice de Gini de 0.584 y una respuesta capturada del 30.8%.

a) Análisis del modelo optimo

El modelo seleccionado, basado en tasa de error de clasificación para los datos de validación es el modelo entrenado con los siguientes parámetros:

- NumVerFinac24: Número de Verificaciones Financieras en los últimos 24 meses
- PLCredSal: Porcentaje de Líneas de Crédito con Saldo
- TLCredMor306024: Número Total de Líneas de Crédito con morosidad de 30 o 60 días en los últimos 24 meses
- TLCredMor6024: Número Total de Líneas de Crédito con morosidad de 60 días o más en los últimos 24 meses
- TLCredSat: Porcentaje de Líneas de Crédito Satisfactorias
- TTransPLCred: Tiempo transcurrido desde la apertura de la Primera Línea de Crédito

b) Prueba de Desvianza para los coeficientes de los parámetros

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

Tabla 7: Desvianza de los coeficientes

Sólo términos independientes	Términos independientes & covariables	Chi-Square	DF	Pr>ChiSq
1.351,68	1.099,61	252,07	6	<.0001

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, *Julio del 2013*

Según la *Tabla 7,* se muestra que el nivel de significancia 0.05> 0.0001, por lo tanto existe prueba suficiente para afirmar que los coeficientes (B_i) de los parámetros son diferentes de cero.

c) Análisis de Efectos por parámetro

H_0:	H_1:
$\beta_{\text{NumVerFinac24}} = 0$	$\beta_{\text{NumVerFinac24}} \neq 0$
$\beta_{\text{PLCredSal}} = 0$	$\beta_{\text{PLCredSal}} \neq 0$
$\beta_{\text{TLCredMor306024}} = 0$	$\beta_{\text{TLCredMor306024}} \neq 0$
$\beta_{\text{TLCredMor6024}} = 0$	$\beta_{\text{TLCredMor6024}} \neq 0$
$\beta_{\text{TLCredSat}} = 0$	$\beta_{\text{TLCredSat}} \neq 0$
$\beta_{\text{TTransPLCred}} \neq 0$	$\beta_{\text{TTransPLCred}} \neq 0$

Tabla 8: Análisis de Efectos de cada coeficiente (β_i) de los parámetros

Efecto	DF	Wald Chi-Square	Pr>ChiSq
WOE_NumVerFinac24	1	17.0510	<.0001
WOE_PLCredSal	1	50.1412	<.0001
WOE_TLCredMor306024	1	26.8937	<.0001
WOE_TLCredMor6024	1	20.6274	<.0001
WOE_TLCredSat	1	8.7961	0.0030
WOE_TTransPLCred	1	12.0731	0.0005

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, *Julio del 2013*

Según la *Tabla 8*, el estadístico de Wald contrasta la hipótesis de que un coeficiente aislado es distinto de 0 y sigue una distribución con media 0 y varianza 1, por lo cual se observa que los coeficientes son diferentes de 0 ya que el p-valor es menor a 0.05 por tanto son significativos

d) Estimación de parámetros

Tabla 9: Análisis de estimaciones de Máxima Verosimilitud

Parámetro	Estimate	Standard Error	Wald Chi-Square	Estimador Pr>ChiSq	Estandarizado	Exp(Est)
Intercept	-1.6379	0.0814	404.52	<.0001		0.194
WOE_NumVerFinac24	-0.7952	0.1926	17.05	<.0001	-0.1707	0.452
WOE_PLCredSal	-1.0395	0.1468	50.14	<.0001	-0.2967	0.354
WOE_TLCredMor306024	-0.6983	0.1346	26.89	<.0001	-0.2139	0.497
WOE_TLCredMor6024	-0.5232	0.1152	20.63	<.0001	-0.2313	0.593
WOE_TLCredSat	-0.3773	0.1272	8.80	0.0030	-0.1443	0.686
WOE_TTransPLCred	-0.8455	0.2433	12.07	0.0005	-0.1651	0.429

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, *Julio del 2013*

En la *Tabla 9*, se muestran los coeficientes estimados β, el estadístico Wald que se distribuye de acuerdo a una X^2; por tanto todos los coeficientes son significativos por tener un $X^2 > 4$.

Tabla 10: Estimaciones del ratio Odds

Efecto	Point Estimate
WOE_NumVerFinac24	0.452
WOE_PLCredSal	0.354
WOE_TLCredMor306024	0.497
WOE_TLCredMor6024	0.593
WOE_TLCredSat	0.686
WOE_TTransPLCred	0.429

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, *Julio del 2013*

La oportunidad relativa de cada variable del modelo es:

El cociente entre la oportunidad de los que incumplen el crédito expuestos al Número de Verificaciones Financieras en los últimos 24 meses es de 0.452, por tanto la oportunidad de que ocurra el evento es menor para lo que no tienen verificaciones.

El cociente entre la oportunidad de los que incumplen el crédito expuestos al Porcentaje de Líneas de Crédito con Saldo es de 0.354, por tanto la oportunidad de que ocurra el evento es menor para lo que no tienen líneas de crédito con saldo.

El cociente entre la oportunidad de los que incumplen el crédito expuestos al Número Total de Líneas de Crédito con morosidad de 30 o 60 días en los últimos 24 meses es de 0.497, por tanto la oportunidad de que ocurra el evento es menor para lo que no tienen líneas de crédito con morosidad de 30 o 60 días en los últimos 24 meses.

Es el cociente entre la oportunidad de los que incumplen el crédito expuestos al Número Total de Líneas de Crédito con morosidad de 60 días o más en los últimos 24 meses es de 0.593, por tanto la oportunidad de que ocurra el evento es menor para lo que no tienen líneas de crédito con morosidad de 60 días o más.

Es el cociente entre la oportunidad de los que incumplen el crédito expuestos al Porcentaje de Líneas de Crédito Satisfactorias es de 0.686, por tanto la oportunidad de que ocurra el evento es menor para lo que no tienen líneas de crédito satisfactorias.

Es el cociente entre la oportunidad de los que incumplen el crédito expuestos al Tiempo transcurrido desde la apertura de la Primera Línea de Crédito es 0.429, por tanto la oportunidad de que ocurra el evento es menor para lo que no tienen apertura de línea de crédito.

e) Interpretación del Modelo

$$logit\,(p_i) = log\frac{p_i}{1-p_i}$$

$$\begin{aligned} logit\,(p_i) \;=\; & -1.6379 \;- 0.7952 * NumVerFinac24 - 1.0395 * PLCredSal \\ & - 0.6983 * TLCredMor306024 \;- 0.5232 * TLCredMor6024 \\ & - 0.3773 * TLCredSat \;- 0.8455 * TTransPLCred \end{aligned}$$

Para la variable NumVerFinac24, manteniendo constante el resto de los predictores, un cambio de 0 a una verificación financiera en los últimos 24 meses produce una disminución de 0.7952 unidad logit en la variable TARGET.

Para la variable PLCredSal, manteniendo contante el resto de los predictores, un cambio de una unidad de porcentaje de líneas de crédito con saldo produce una disminución de 1.0395 unidades logit en la variable TARGET.

Para la variable TLCredSat manteniendo constante el resto de los predictores, un cambio de una unidad en el porcentaje de líneas de crédito Satisfactorias produce una disminución de 0.3773 unidad logit en la variable TARGET.

Para la variable TLCredMor6024 manteniendo constante el resto de los predictores, un cambio de una unidad en el número de líneas de crédito con morosidad de 60 días en los últimos 24 meses produce una disminución de 0.5232 unidad logit en la variable TARGET.

Para la variable TTransPLCred manteniendo constante el resto de los predictores, un cambio de un día en el tiempo transcurrido desde la apertura de la línea de crédito produce una disminución de 0.8455 unidad logit en la variable TARGET.

Para la variable TLCredMor306024 manteniendo constante el resto de los predictores, un cambio de una unidad en el número de líneas de crédito morosas de 30 a 60 días en los últimos 24 meses produce una disminución de 0.6983 unidad logit en la variable TARGET.

f) Matriz de Confusión

Tabla 11: Matriz de Confusión del modelo de Regresión Logística por Agrupación interactiva

		Clase Predicha			
		1	0		
Clase Real	1	40	210	250	Negativo
	0	33	1217	1250	Positivo
		73	1427	1500	
		Negativo	Positivo		

Donde
1: incumplimiento
0: cumplimiento

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, *Julio del 2013*

Sensibilidad= 0.160
Especificidad= 0.9736
Exactitud= 0.838

De la *Tabla 11* se obtuvo que el número total de predicciones correctas fue del 83.8%, asimismo la proporción de clientes que cayeron en default fue 16% y los clientes que cumplieron con el crédito otorgado fue del 97.36%.

4.5.2. Puntos de Corte (Cut - Off)

Una vez identificado el modelo con mejores indicadores se realizó el scoring de los clientes que validaron el modelo de acuerdo a la Probabilidad de Default, luego se estableció 20 percentiles los cuales permitieron clasificar el otorgamiento de crédito en intervalos de aprobación, revisión y aceptación.

Por ser el objetivo de la cooperativa de ahorro y crédito maximizar la rentabilidad, se estableció los puntos de corte asignando un 30% para rechazo automático, 5% para decisión del analista de crédito y un 65% para aprobación automática.

Tabla 12: Puntuación para calificación de scoring

Intervalo	Rango	Frecuencia	%	Decisión
1	45-87	75	5%	Rechazar
2	88-98	75	5%	Rechazar
3	99-107	75	5%	Rechazar
4	108-114	75	5%	Rechazar
5	115-121	75	5%	Rechazar
6	122-127	75	5%	Rechazar
7	128-133	75	5%	Revisar
8	134-139	75	5%	Aprobar
9	140-144	75	5%	Aprobar
10	145-150	75	5%	Aprobar
11	151-155	75	5%	Aprobar
12	156-159	75	5%	Aprobar
13	160-164	75	5%	Aprobar
14	165-169	75	5%	Aprobar
15	170-174	75	5%	Aprobar
16	175-179	75	5%	Aprobar
17	180-185	75	5%	Aprobar
18	186-191	75	5%	Aprobar
19	192-200	75	5%	Aprobar
20	201-228	75	5%	Aprobar

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

A partir de la *tabla 12* se obtuvo que 972 clientes serian aptos para el otorgamiento de crédito, 80 clientes se derivarían al área de análisis crediticio y que 448 clientes se les denegaría el crédito por no cumplir con los requisitos del modelo.

Tabla 13: Matriz de Confusión del scoring

		Clase Predicha			
		1	0		
Clase Real	1	173	77	250	Negativo
	0	355	895	1250	Positivo
		528	972	1500	
		Negativo	Positivo		

Donde
1: incumplimiento
0: cumplimiento

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, *Julio del 2013*

Sensibilidad= 0.692
Especificidad= 0.716
Valor predictivo positivo (VP+)= 0.328
Valor predictivo negativo (VP-)= 0.921
Exactitud= 0.712
Tasa de Error= 0.288

De la *Tabla 13* se obtuvo que la proporción de clientes que cayeron en default fue del 69.2% a diferencia de lo encontrado en la matriz de confusión del modelo de regresión logística que predice que el 16% de clientes no son aptos para otorgarles crédito, por lo que se afirma que el score realizado con la probabilidad de default de acuerdo a los objetivos de la cooperativa es más eficiente para predecir el otorgamiento del crédito.

V. CONCLUSIONES

- Se identificó más de una variable explicativa durante la construcción de los modelos scoring; sin embargo, fueron seis las variables (NumVerFinac24, PLCredSal, TLCredMor306024, TLCredMor6024, TLCredSat y TTransPLCred) que permitieron determinar la probabilidad de incumplimiento del crédito personal.

- Se construyó un total de doce modelos utilizando las técnicas de Árboles de Clasificación, Regresión Logística y Redes Neuronales donde se detalla explícitamente en el *capítulo IV. Resultados y Discusiones - Modelado (pág. 95).*

- Se determinó que el mejor modelo basado en indicadores de eficiencia y predictibilidad fue el modelo de Regresión Logística por Agrupación Interactiva obtenido con los datos sin transformaciones (Reg. Log A.I), ya que presentó un CAPC del 30,8%, un GINI del 0.584 y un ROC del 0.792.

- Se estableció los puntos de corte en 20 percentiles con frecuencia uniforme, asignando un 30% para rechazo automático, 5% para decisión del analista de crédito y un 65% para aprobación automática.

VI. SUGERENCIAS

- Se sugiere enfocarse en la fase de la comprensión del negocio; si no se logra desarrollar esta capacidad, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables.

- A fin de lograr una herramienta dinámica y aceptable, el modelo credit scoring construido se deberá calibrar al menos cada tres meses para que se mantenga actualizado, acorde con las necesidades de la institución y resulte efectivo en sus operaciones en el transcurso del tiempo.

- Se sugiere que el scoring diseñado para la cartera de crédito personal sea implementado mediante un software para su aplicación a fin de que las operaciones crediticias sean más ágiles y eficientes; al automatizar la evaluación de los clientes se mejorará el proceso de concesión de créditos, fundamental en toda institución financiera.

- Si bien esta investigación concluyó que la regresión logística por agrupación interactiva es el mejor modelo capaz de predecir eficientemente, cabe aclarar que existe una gran variedad de técnicas utilizadas para la construcción de modelos credit scoring y que dejan por tanto numerosas vías de investigación por donde seguir avanzando.

BIBLIOGRAFÍA

Acuña, E. (2004). *Clasificación Usando Árboles de Decisión.* Obtenido de http://math.uprm.edu/~edgar/clasifall9.pdf

Aguilar, J. (2009). *Laboratorio: Redes con conexiones hacia adelante.* Obtenido de http://www.flacsoandes.edu.ec/comunicacion/aaa/imagenes/publicaciones/pub_30.pdf

AIS Goup. (03 de Noviembre de 2011). *Herramientas idóneas para el control de riesgos y la evolución de la calidad de las carteras.* Obtenido de http://www.ais-int.com/wp-content/uploads/2011/11/04-Herramientas-Idoneas-Control-Riesgo.pdf

Araujo, R., & Masci, P. (Octubre de 2007). *BASILEA II en América Latina.*

Bank for international settlements. (s.f.). *Marco regulador internacional para bancos (Basilea III)*. Obtenido de http://www.bis.org/bcbs/basel3_es.htm

Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). *Small Business Credit Scoring: A Comparison of Logistic Regression.* Obtenido de http://bib.irb.hr/datoteka/182478.zekic-sarlija-bensicITI2004_revised.pdf

Bessis, J. (2002). *Risk management in banking.* Chichester: John Wiley & Sons.

Brachfield, P. (diciembre de 2015). *Políticas de crédito normales, restrictivas o flexibles*. Obtenido de http://www.perebrachfield.com/el-blog-de-morosologia/riesgo-de-credito/politicas-de-credito-normales-restrictivas-o-flexibles

COFIA. (2 de Marzo de 2011). *El riesgo en la organización cooperativa de ahorro y crédito.* Obtenido de http://www.aciamericas.coop/IMG/pdf/riesgoeorganizacioncooperativaayc.pdf

Comité de Supervisión Bancaria de Basilea. (1988). International Convergence of Capital Measurement and Capital Standards (pp.1). Basilea.

Comité de Supervisión Bancaria de Basilea. (2004). Convergencia internacional de medidas y normas de capital (pp. 1). Basilea.

Cortijo, F. (2001). *Técnicas supervisadas II: Aproximación no paramétrica.* Obtenido de http://iie.fing.edu.uy/ense/asign/recpat/material/tema3_00-01/node1.html

Creditos y Cobranzas. (Marzo de 2010). Obtenido de http://creditosycobranzasdinero.blogspot.pe/2010/01/politicasdecredito.html

CRISP-DM consortium. (2000). *CRISP-DM 1.0.* Obtenido de https://www.the-modeling-agency.com/crisp-dm.pdf

Dataprix. (s.f.). Obtenido de http://www.dataprix.com/comprensi%C3%B3n-del-negocio

De Lara, A. (2005). *Medición y control de riesgos financieros.* Mexico: Limusa.

Emprendedor.pe. (6 de Junio de 2013). Obtenido de http://emprendedor.pe/finanzas/878-tipos-de-creditos-en-peru.html/

Fermac Risk. (2014). *Credit Scoring, Validación de Modelos y Stress Testing Nivel I.* Obtenido de http://www.fermacrisk.com/credit-scoring-nivel-1

Fernández, G. (2002). *Data Mining Using SAS Applications.* Chapman & Hall / CRC.

Gámez, M., & García, N. (2000). *Rating de pequeñas y medianas empresas mediante Árboles de Clasificación.* Obtenido de http://www.uclm.es/ab/fcee/D_trabajos/2-2000-2.pdf

González, F., & Montoya, J. (Setiembre de 2006). *Marketing y Riesgo. Llega el Marketing Score.* Obtenido de http://pdfs.wke.es/9/4/7/5/pd0000019475.pdf

Herrán, L. (24 de Julio de 2009). *Repositorio institucional PIRHUA- Universidad de Piura.* Obtenido de Evaluación crediticia aplicando un modelo de Credit Scoring en el ámbito microempresarial : caso CMAC Paita: https://pirhua.udep.edu.pe/bitstream/handle/123456789/1325/ECO_030.pdf?sequence=1&isAllowed=y

Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression.* Ed. John Wiley. New York.

Jorion, P. (1999). *Valor en Riesgo.* México: Limusa S.A de C.V.

Kraneau, E. (2007). *Fundamentos Teóricos de las Redes Neuronales.*

La Economía. (25 de julio de 2011). *Buros de Crédito en el Perú*. Obtenido de http://laeconomia.pe/buro-de-credito.html

Leung, S. (2008). *Análisis Comparativo entre Árboles de Clasificación.* Obtenido de http://techi322.wordpress.com/2008/04/16/analisis-comparativo-entre-arboles-de-clasificacion-2/

Marín. (2009). *Análisis de Clúster y Árboles de Clasificación.* Obtenido de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf

Marín, J. M. (s.f.). *Introducción a las redes neuronales aplicadas.* Obtenido de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema3dm.pdf

McCulloch, W., & Pitts, W. (1943). *Un cálculo lógico de la inminente idea de la actividad nerviosa.* boletín de matemática biofísica.

Molera, L., & Caballero, M. (2001). *Predicción del Éxito en Estudios Universitarios mediante Redes Neuronales.* Obtenido de http://www.pagina-aede.org/Murcia/E07.pdf

Molinero, L. (2004). *Historia del Razonamiento Estadístico.* Obtenido de http://www.seh-lelha.org/pdf/historiastat.pdf

Montoya, J. (2010). *SAS*. Obtenido de http://www.sas.com/es_pe/home.html

Morgan, J., & Messenger, R. (1973). *THAID: A Sequential Search Program for the Analysis of Nominal Scale Dependent Variables, Technical report.* Institute for Social Research. University of Michigan.

Mtz. de Lejarza, I. (1998). *Elementos Básicos de una Red Neuronal.* Obtenido de http://www.uv.es/~mlejarza/redes.htm

Nieto, S. (18 de Mayo de 2010). Obtenido de http://mat.izt.uam.mx/mcmai/documentos/tesis/Gen.07-O/Nieto-S-Tesis.pdf

Peña, D. (2002). *Análisis de Datos Multivariantes.* Mc Graw Hill / Interamericana de España S.A.

Piedra, N. (2007). *Elemento Básicos de una Red Neuronal II.* (E. Publicado por Advanced Tech Computing Group UTPL. Loja, Ed.) Obtenido de http://advancedtech.wordpress.com/2007/09/26/elementos-baiscos-de-una-red-neuron al-artificialparte-ii/

Puerta, A. (Junio de 2002). *Imputación basada en arboles de clasificación. Publicado por EUSTAT.* Obtenido de http://www.eustat.es/document/datos/ct_04_c.pdf

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Elsevier, Burlington, Canada: Morgan Kaufmann Publishers.

Rodríguez, M., & Global Methodology, M. (Noviembre de 2012). *"Centrales públicas de riesgo, burós de crédito y el sector microfinanciero en América Latina".* Obtenido de http://www.microrate.com/media/downloads/2013/06/MicroRate-Centrales-P%C3%BAblicas-de-Riesgo-Bur%C3%B3s-de-Credito-y-el-Sector-Microfinanciero-en-Am%C3%A9rica-Latina-v2.pdf

Rodríguez, O. (2010). *Metodología para el desarrollo de proyectos en minería de datos CRISP-DM.* Obtenido de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

SAS Enterprise miner 13.1. (s.f.). SAS Enterprise miner 13.1 references help.

Schreiner, M. (11 de setiembre de 2002). *Ventajas y desventajas del scoring estadístico para las microfinanzas.* Obtenido de http://www.microfinance.com/Castellano/Documentos/Scoring_Ventajas_Desventajas.pdf

Serrano, C., & Martin, B. (1995). *Fundamentos de las redes neuronales artificiales: Hardware y Software.* Obtenido de http://es.wikipedia.org/wiki/Red_neuronal_artificial

Soltan, A., & Mohammadi, M. (30 de Abril de 2012). *A hybrid model using decision tree and neural network for credit scoring problem.* Obtenido de http://www.growingscience.com/msl/Vol2/msl_2012_88.pdf

Superintendencia de Banca y Seguros. (s.f.). *Basilea II y Basilea III*. Obtenido de http://www.sbs.gob.pe/principal/categoria/basilea-ii-y-basilea-iii/1075/c-1075

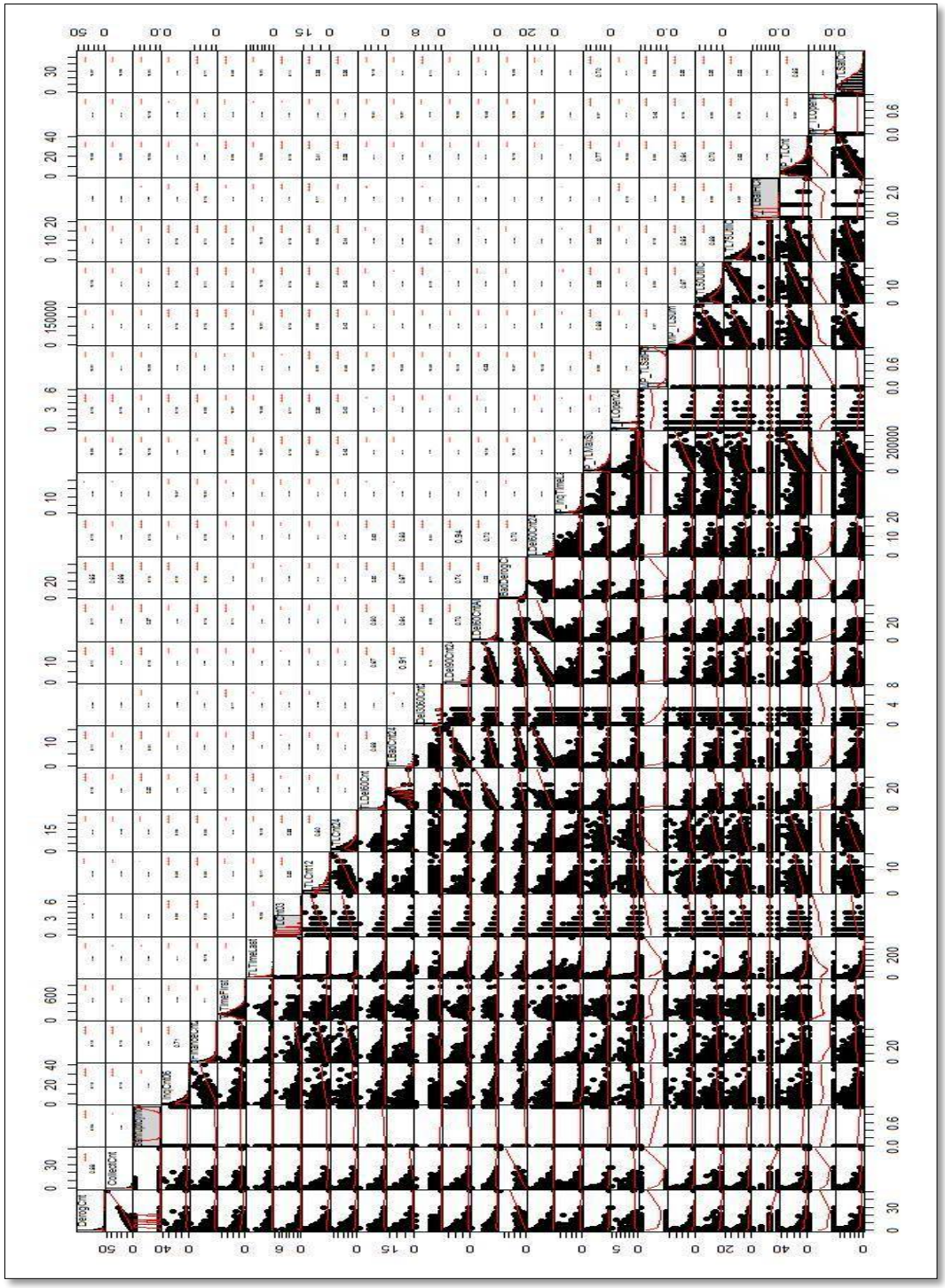
Superintendencia de Banca y Seguros. (s.f.). *Central de riesgos de la SBS: Conoce tus derechos*.

Vigo, G. J. (2010). *Método de clasificación para evaluar el riesgo crediticio: una comparación.* Obtenido de http://cybertesis.unmsm.edu.pe/bitstream/cybertesis/3327/1/Vigo_cg.pdf

Webmining Consultores. (6 de Julio de 2011). Obtenido de http://www.webmining.cl/2011/07/entrenamiento-validacion-y-prueba/

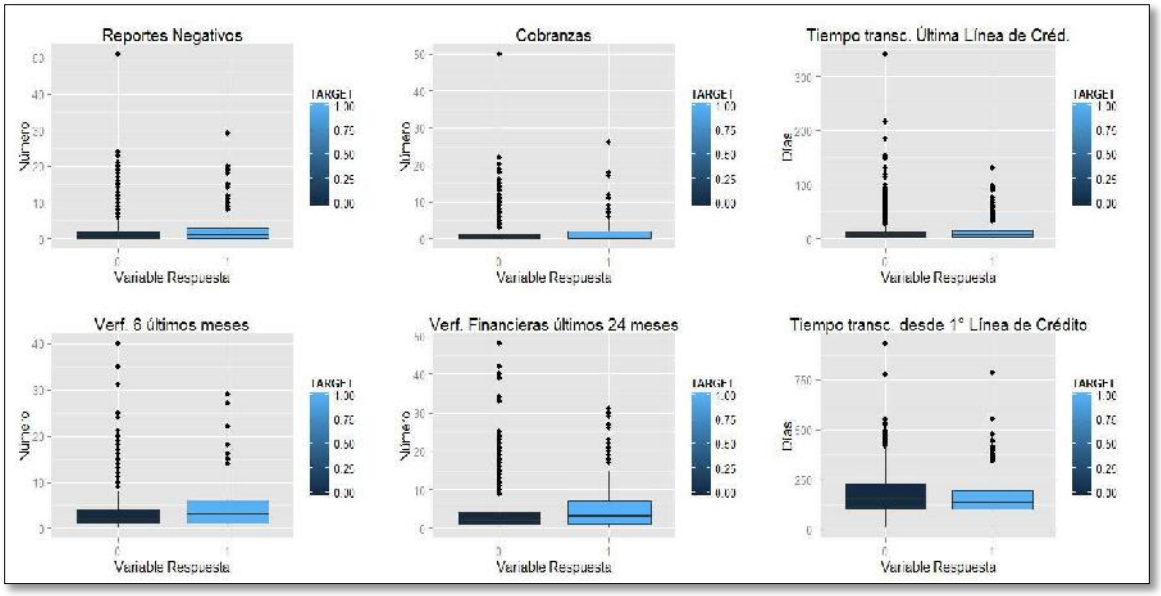
ANEXOS

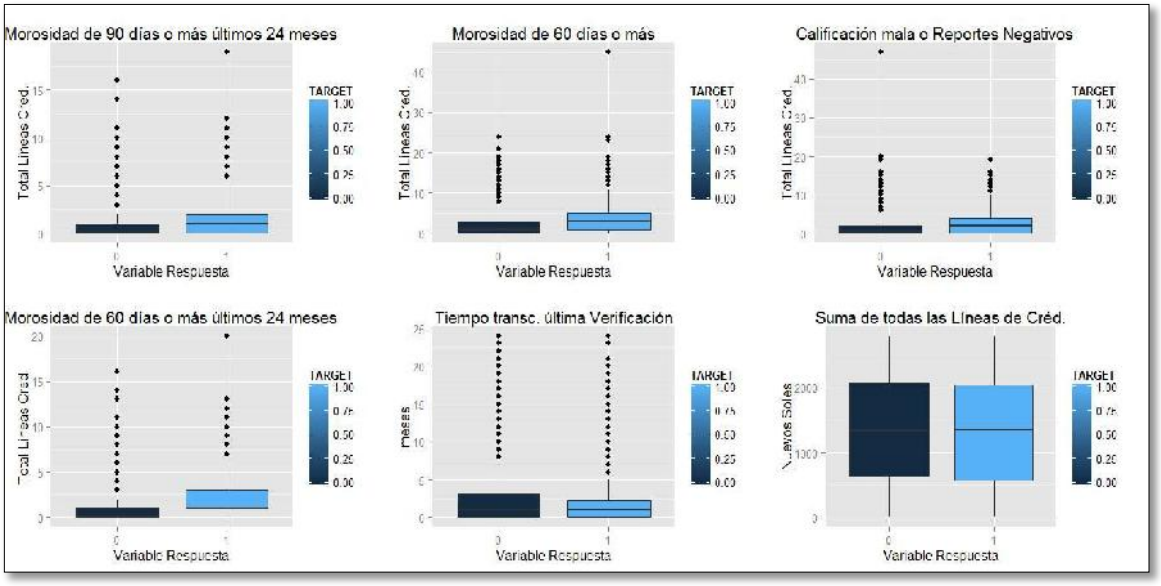
Anexo 1: Diagrama de Correlación

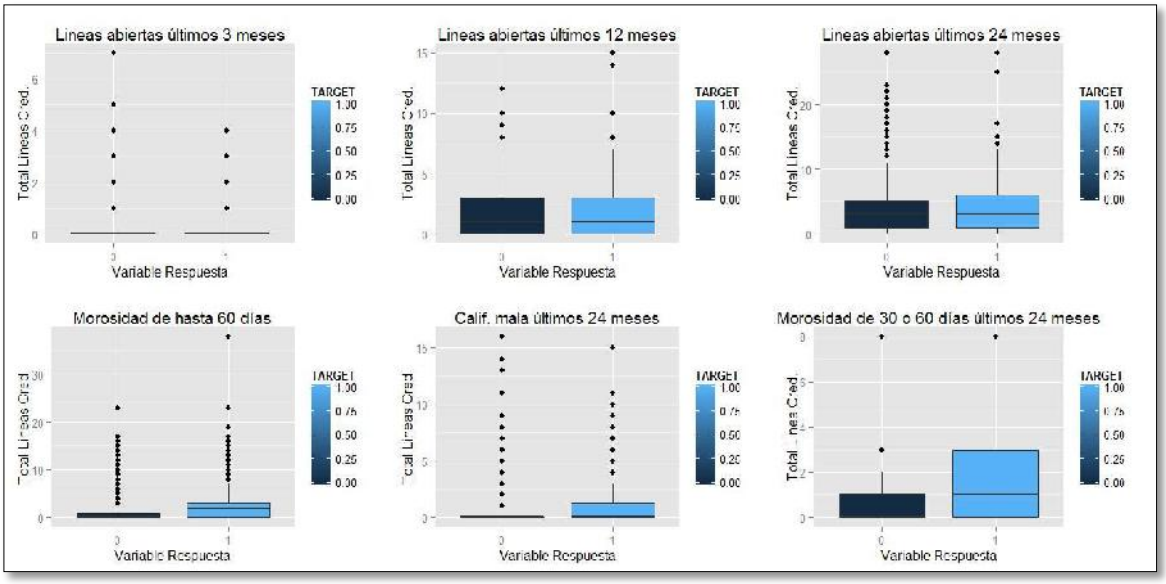


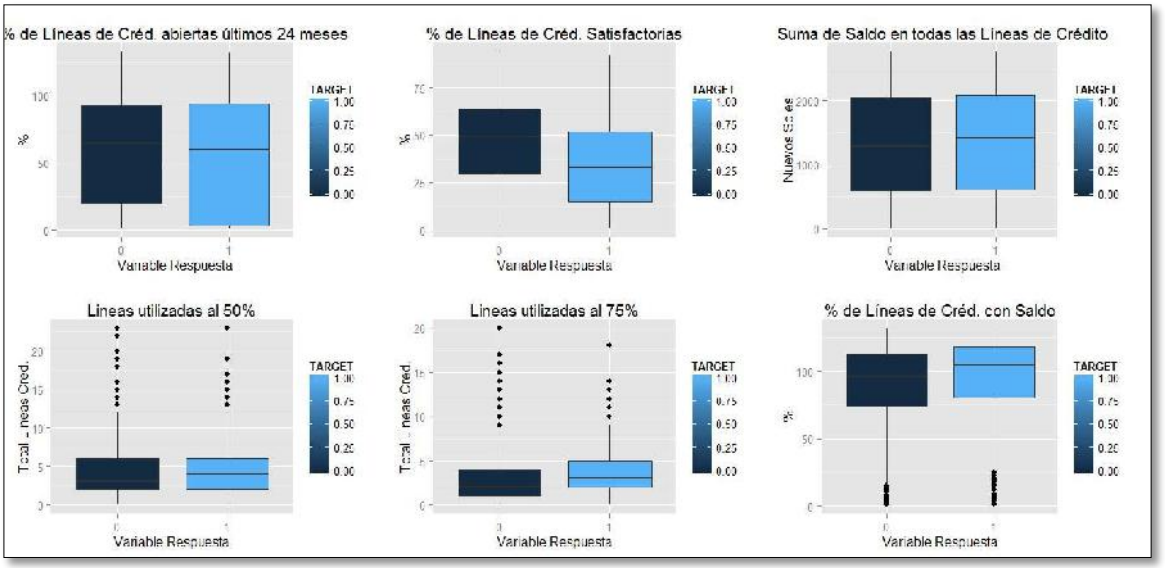
Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

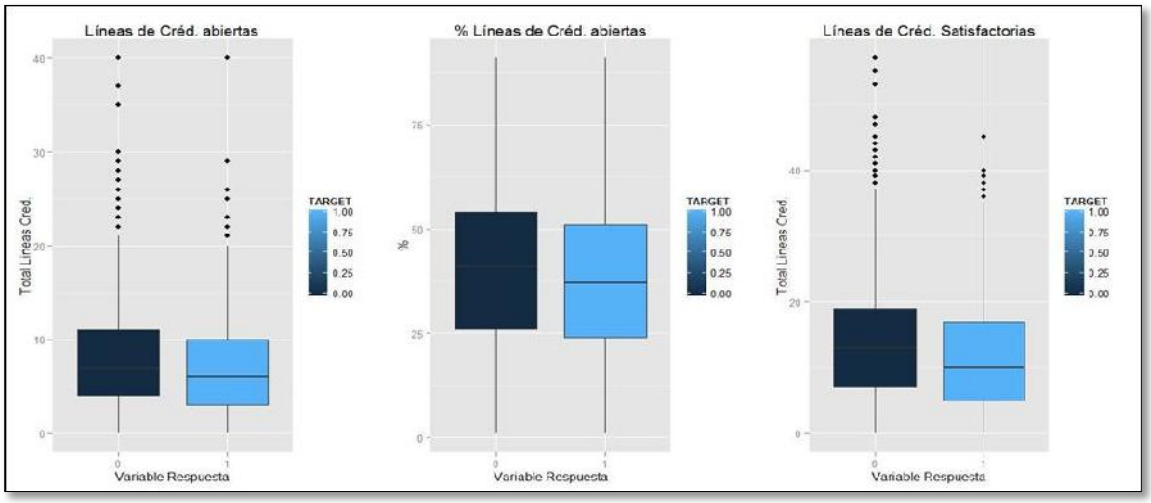
Anexo 2: Diagrama de Cajas de las Variables en Estudio



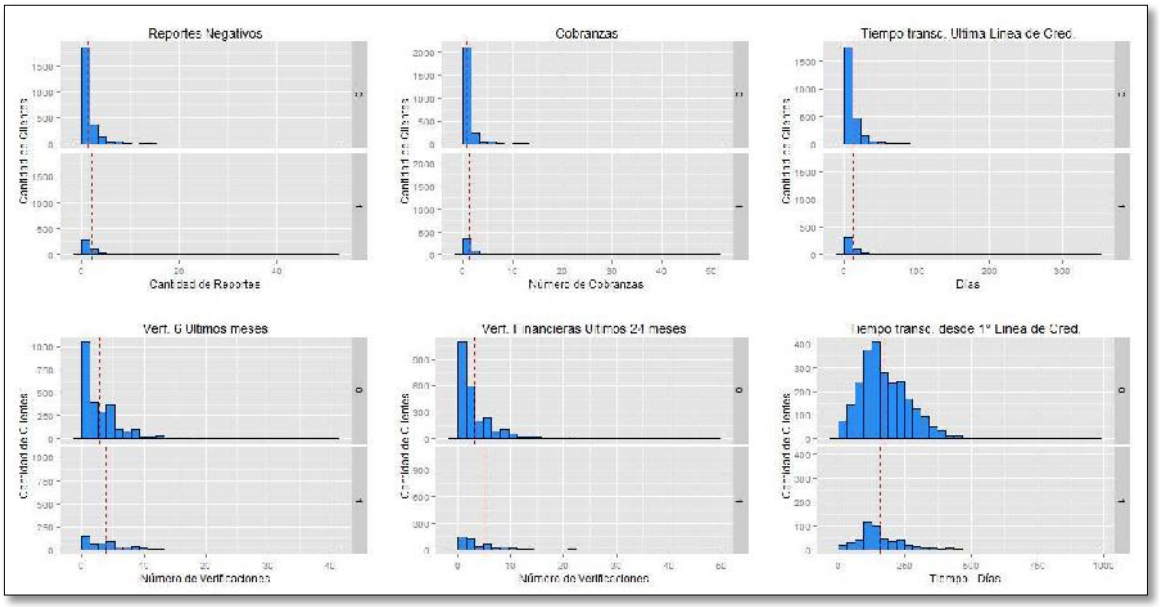


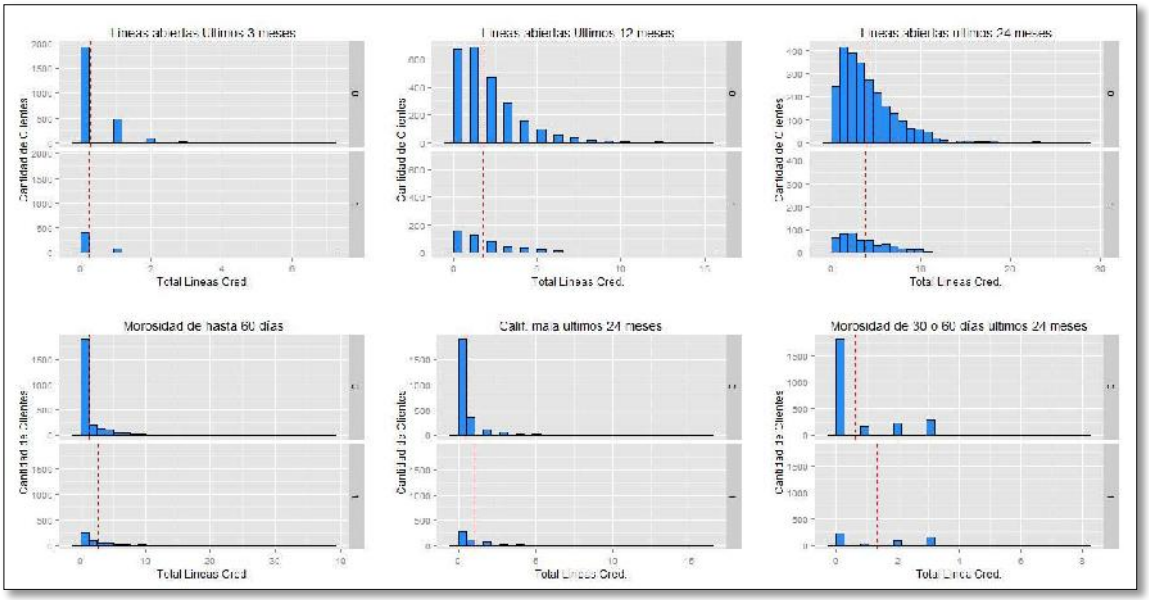

Lineas abiertas últimos 3 meses
Lineas abiertas últimos 12 meses
Lineas abiertas últimos 24 meses
Morosidad de hasta 60 días
Calif. mala últimos 24 meses
Morosidad de 30 o 60 días últimos 24 meses
Total Líneas Créd.
Variable Respuesta
TARGET

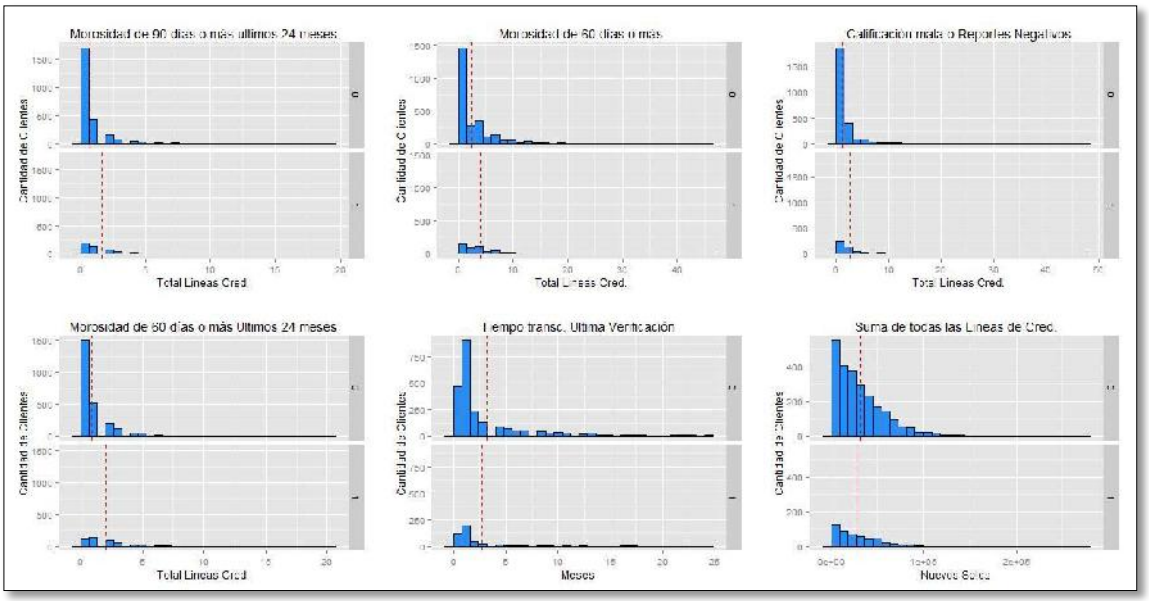

% de Líneas de Créd. abiertas últimos 24 meses
% de Líneas de Créd. Satisfactorias
Suma de Saldo en todas las Líneas de Crédito
Líneas utilizadas al 50%
Líneas utilizadas al 75%
% de Líneas de Créd. con Saldo
Variable Respuesta
TARGET

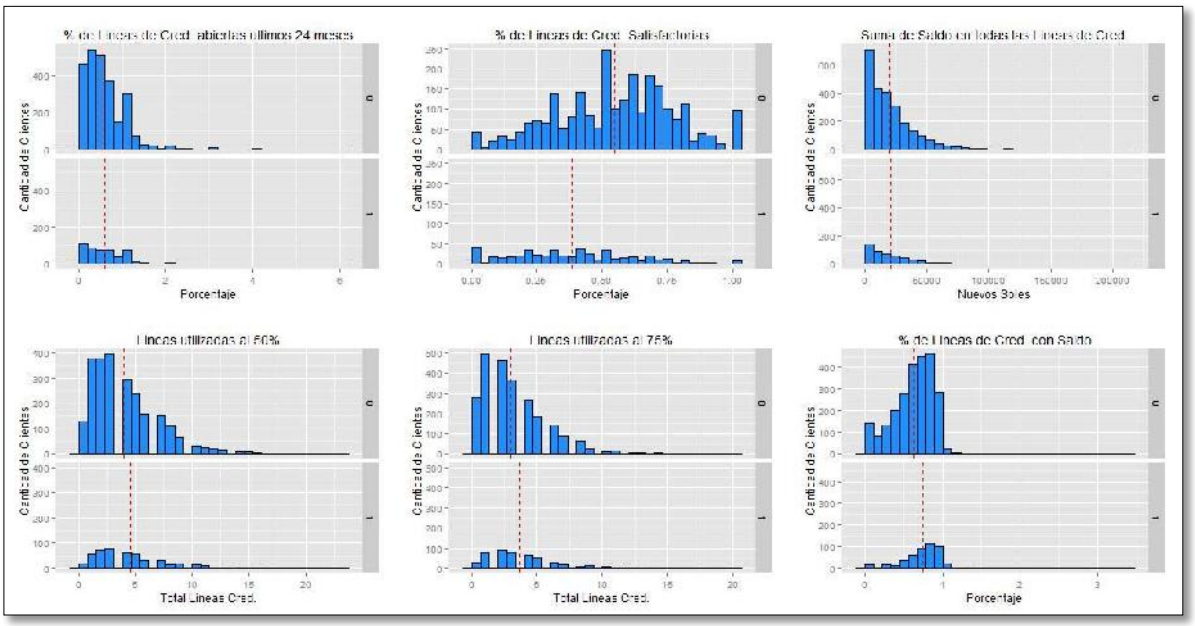


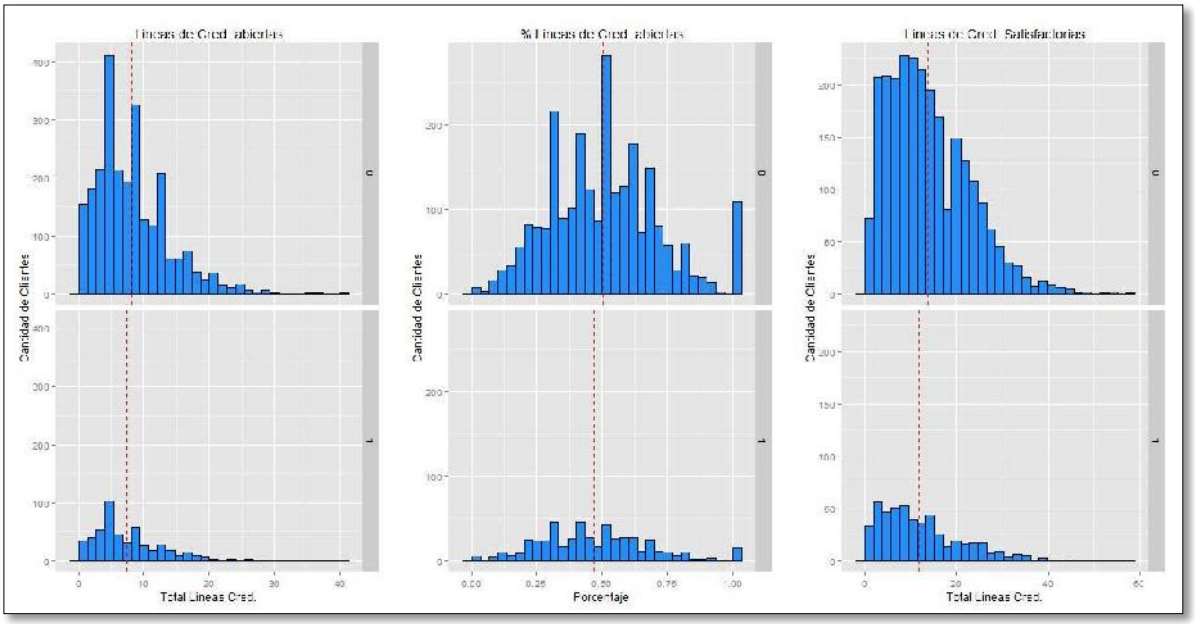
Anexo 3: Histogramas de frecuencia de las Variables en Estudio



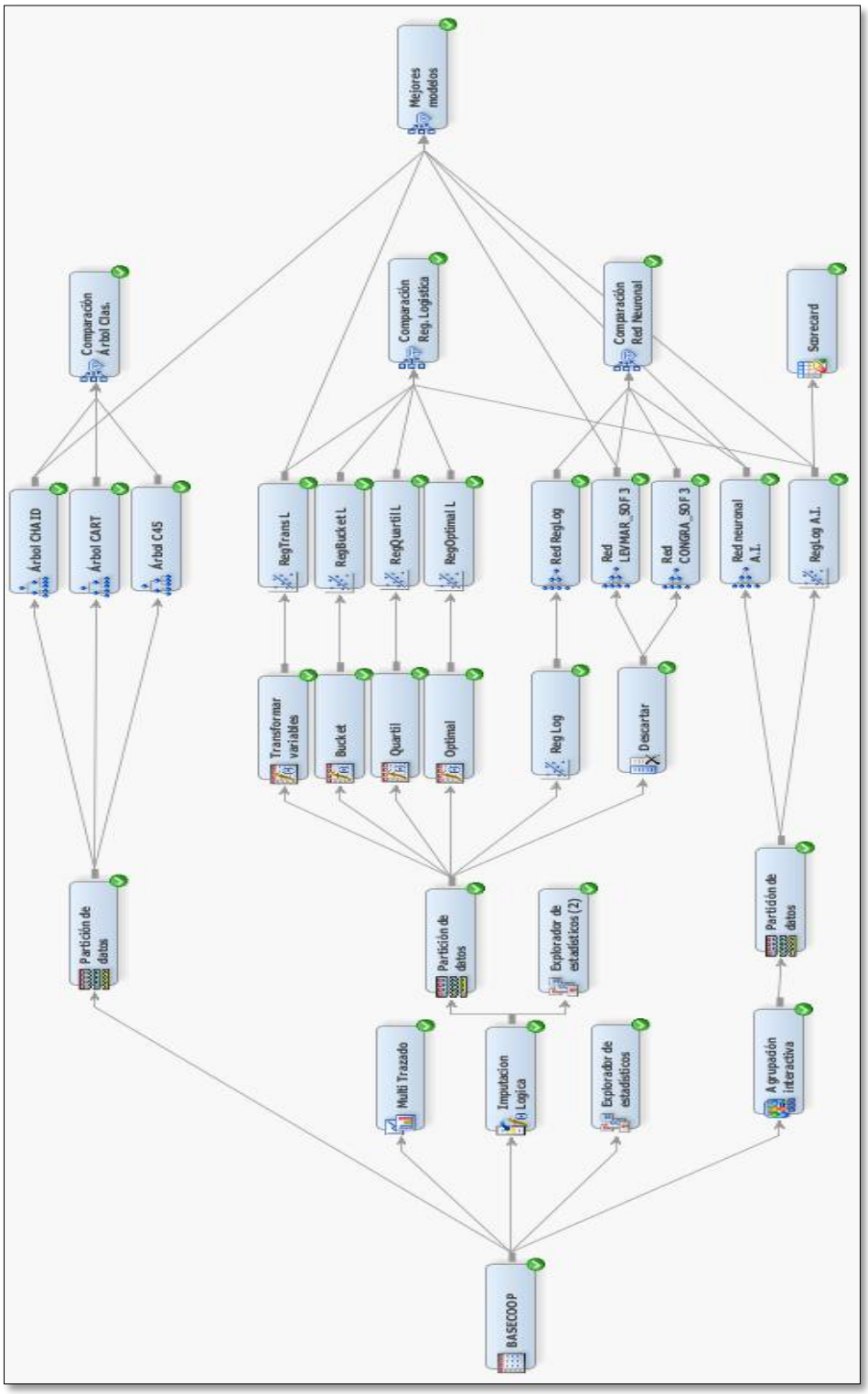

Lineas abiertas Ultimos 3 meses
Lineas abiertas Ultimos 12 meses
Lineas abiertas ultimos 24 meses
Morosidad de hasta 60 dias
Calif. mala ultimos 24 meses
Morosidad de 30 o 60 dias ultimos 24 meses
Cantidad de Clientes
Total Lineas Cred.


Morosidad de 90 dias o más ultimos 24 meses
Morosidad de 60 dias o más
Calificacion mala o Reportes Negativos
Morosidad de 60 dias o más Ultimos 24 meses
Tiempo transc. Ultima Verificación
Suma de todas las Lineas de Cred.
Cantidad de Clientes
Total Lineas Cred.
Meses
Nuevos Soles


% de Líneas de Créd. Abiertas últimos 24 meses
% de Líneas de Créd. Satisfactorias
Suma de Saldo en todas las Líneas de Créd.
Líneas Utilizadas al 50%
Líneas Utilizadas al 75%
% de Líneas de Créd. con Saldo
Cantidad de Clientes
Porcentaje
Nuevos Soles
Total Líneas Créd.


Líneas de Créd. Abiertas
% Líneas de Créd. Abiertas
Líneas de Créd. Satisfactorias
Cantidad de Clientes
Total Líneas Créd.
Porcentaje

Anexo 4: Diagrama del Proyecto en SAS Enterprise Miner



Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Anexo 5: Indicadores y variables de los modelos de Árboles de Clasificación

Tabla 14: Indicadores y variables de los modelos de Árboles de Clasificación

Variables	Arbol CHAID	Arbol CART	Arbol C4.5
Target	X	X	X
IndBancarrota			
NumCob			
NumRepNeg			
NumVer06			
NumVerFinac24	X	X	X
TTransUltVer			
LCred50			
LCred75			
LCredCM24			
LCredRepNeg			
PLCredSal	X	X	X
TLCredAb			
TLCredAb03			
TLCredAb12			
TLCredAb24			
TLCredMor306024	X	X	X
TLCredMor60			
TLCredMor6024	X	X	X
TLCredMor60M			
TLCredMor9024			
STLCred		X	X
PLCredAb24			
PLCredAb			
TLCredSat		X	X
PLCredSat	X	X	X
SumTLCred		X	
TTransPLCred	X	X	X
TTransULCred			
Error cuadrático medio	0.124	0.127	0.127
Raíz del error cuadrático medio	0.352	0.357	0.357
Indice Roc	0.731	0.722	0.715
Coeficiente de Gini	0.463	0.444	0.430
Estadístico Kolmogorov-Smirnov	0.384	0.371	0.371
Sensibilidad (recuerdo)	0.160	0.160	0.160
Especificidad	0.978	0.978	0.978
Valor predictivo positivo (VP+)	0.588	0.588	0.588
Valor predictivo negativo (VP-)	0.853	0.853	0.853
Exactitud	0.841	0.841	0.841
Tasa de Error	0.159	0.159	0.159

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Tabla 15: Indicadores y variables de los modelos de Regresión Logística

Variables	Reg. Logística Trans.	Reg. Logística Bucket.	Reg. Logística Quartil.	Reg. Logística Optimal.	Reg. Logística 10
Target	X	X	X	X	X
IndBancarrota	X		X	X	
NumCob					
NumRepNeg					
NumVer06					
NumVerFinac24	X	X	X	X	X
TTransUltVer					
LCred50					
LCred75	X	X	X	X	
LCredCM24					
LCredRepNeg					
PLCredSal	X	X	X	X	X
TLCredAb					
TLCredAb03	X				
TLCredAb12					
TLCredAb24					
TLCredMor306024	X	X	X	X	X
TLCredMor60				X	
TLCredMor6024	X		X	X	X
TLCredMor60M		X			
TLCredMor9024					
STLCred			X		
PLCredAb24				X	
PLCredAb				X	
TLCredSat	X		X		
PLCredSat	X	X	X	X	X
SumTLCred				X	
TTransPLCred	X	X	X	X	X
TTransULCred				X	
Error cuadrático medio	0.116	0.125	0.139	0.122	0.117
Raíz del error cuadrático medio	0.341	0.354	0.373	0.349	0.342
Indice Roc	0.789	0.733	0.500	0.769	0.792
Coeficiente de Gini	0.579	0.466	0.000	0.538	0.584
Estadístico Kolmogorov-Smirnov	0.441	0.378	0.000	0.426	0.451
Sensibilidad (recuerdo)	0.168	0.108	0.000	0.192	0.160
Especificidad	0.975	0.980	1.000	0.959	0.974
Valor predictivo positivo (VP+)	0.575	0.519	-	0.485	0.548
Valor predictivo negativo (VP-)	0.854	0.846	0.833	0.856	0.853
Exactitud	0.841	0.835	0.833	0.831	0.838
Tasa de Error	0.159	0.165	0.167	0.169	0.162

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Tabla 16: Indicadores y variables de los modelos de Redes Neuronales

Variables	Red Neuronal Log	Red Neuronal A.I	Red Neuronal Lev. Mar. Sof 3	Red Neuronal Con. Gra. Sof 3
Target	X	X	X	X
IndBancarrota				
NumCob		X	X	X
NumRepNeg		X		
NumVer06		X	X	X
NumVerFinac24	X	X	X	X
TTransUltVer			X	X
LCred50				X
LCred75	X		X	
LCredCM24		X		
LCredRepNeg		X	X	X
PLCredSal	X	X	X	X
TLCredAb				
TLCredAb03	X			
TLCredAb12				
TLCredAb24				
TLCredMor306024	X	X	X	X
TLCredMor60		X	X	X
TLCredMor6024	X	X		
TLCredMor60M		X	X	X
TLCredMor9024		X		
STLCred			X	X
PLCredAb24	X		X	X
PLCredAb	X			
TLCredSat	X		X	X
PLCredSat		X	X	X
SumTLCred				
TTransPLCred	X	X	X	X
TTransULCred			X	X
Error cuadrático medio	0.128	0.119	0.120	0.125
Raíz del error cuadrático medio	0.358	0.346	0.346	0.353
Indice Roc	0.747	0.785	0.779	0.768
Coeficiente de Gini	0.493	0.570	0.558	0.535
Estadístico K-S	0.381	0.431	0.450	0.424
Sensibilidad	0.156	0.104	0.124	0.260
Especificidad	0.964	0.984	0.975	0.940
Valor predictivo positivo (VP+)	0.464	0.565	0.500	0.464
Valor predictivo negativo (VP-)	0.851	0.846	0.848	0.864
Exactitud	0.829	0.837	0.833	0.827
Tasa de Error	0.171	0.163	0.167	0.173

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013

Tabla 17: Clasificación de Crédito de los Clientes

Dataobs	ID	TARGET	Reg.TARGET	P_TARGET1	SCORE	DECISION
2	116	0	0	0.04725	174	Aprobar
3	124	0	0	0.08736	155	Aprobar
6	291	0	0	0.08955	155	Aprobar
8	364	0	0	0.03945	179	Aprobar
9	388	0	0	0.02967	188	Aprobar
10	436	0	0	0.07785	158	Aprobar
16	611	0	0	0.18197	130	Analizar
22	777	0	0	0.08983	155	Aprobar
24	911	0	0	0.05774	168	Aprobar
25	1039	0	0	0.37064	103	Rechazar
29	1258	0	0	0.41944	97	Rechazar
31	1402	0	0	0.15850	135	Aprobar
33	1437	0	0	0.02274	196	Aprobar
34	1482	0	0	0.04725	174	Aprobar
39	1548	0	0	0.04385	176	Aprobar
41	1559	0	0	0.04974	172	Aprobar
42	1827	0	0	0.40645	99	Rechazar
43	1864	0	0	0.12416	143	Aprobar
46	1927	0	0	0.07500	159	Aprobar
47	1954	0	0	0.26019	119	Rechazar
48	2033	0	0	0.03229	185	Aprobar
50	2078	0	0	0.14255	140	Aprobar
52	2137	0	0	0.12208	144	Aprobar
54	2156	0	0	0.04725	175	Aprobar
61	2491	0	0	0.15249	136	Aprobar
62	2534	0	0	0.05197	171	Aprobar
64	2596	1	0	0.39844	99	Rechazar
66	2647	0	0	0.04872	173	Aprobar
67	2678	0	0	0.07018	162	Aprobar
68	2687	0	0	0.10672	149	Aprobar
69	2755	0	0	0.06582	165	Aprobar
70	2767	1	0	0.06187	166	Aprobar
71	2782	0	0	0.07407	160	Aprobar
73	2828	0	0	0.14968	138	Aprobar
74	2831	0	0	0.13072	142	Aprobar
75	2875	1	1	0.51116	86	Rechazar
76	2884	1	0	0.37699	103	Rechazar
77	2885	0	0	0.03008	187	Aprobar
78	2940	1	0	0.33035	108	Rechazar

79	2986	0	1	0.54190	82	Rechazar
…	…	…	…	…	…	…
2944	124532	0	0	0.08845	155	Aprobar
2946	124596	0	0	0.04154	178	Aprobar
2949	124730	1	0	0.32744	109	Rechazar
2952	124817	1	0	0.40963	99	Rechazar
2953	124832	0	0	0.18669	130	Analizar
2954	124837	0	0	0.44036	94	Rechazar
2955	124866	0	0	0.14676	138	Aprobar
2957	124922	0	0	0.26182	116	Rechazar
2958	124974	0	0	0.03312	184	Aprobar
2960	124989	0	0	0.03131	187	Aprobar
2963	125118	0	0	0.04397	176	Aprobar
2965	125245	0	0	0.06135	166	Aprobar
2966	125251	0	0	0.06864	161	Aprobar
2967	125260	0	0	0.18238	132	Analizar
2969	125368	0	0	0.16608	134	Aprobar
2971	125399	0	0	0.09674	151	Aprobar
2974	125501	0	0	0.30279	112	Rechazar
2978	125638	0	0	0.14478	139	Aprobar
2980	125666	0	0	0.02907	189	Aprobar
2981	125727	0	0	0.38549	102	Rechazar
2983	125799	1	0	0.43723	96	Rechazar
2985	125930	0	0	0.05655	169	Aprobar
2986	125936	0	0	0.07021	161	Aprobar
2987	125989	0	0	0.03397	184	Aprobar
2988	126065	1	0	0.46016	93	Rechazar
2989	126190	0	0	0.14899	138	Aprobar
2990	126234	0	0	0.16819	134	Aprobar
2991	126271	1	0	0.06429	164	Aprobar
2993	126327	1	0	0.30486	111	Rechazar
2994	126332	1	0	0.41208	97	Rechazar
2995	126402	0	0	0.09268	154	Aprobar
2997	126448	0	0	0.01499	207	Aprobar
2998	126465	0	0	0.11541	147	Aprobar
2999	126487	0	0	0.16406	134	Aprobar

Fuente: Elaborado en base a datos de la Cooperativa de Ahorro y Crédito, Julio del 2013