VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN EN SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN SISTEMAS DE GESTIÒN DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: "MODELO DE MEDICIÓN DE RIESGO CREDITICIO EN ENTIDADES FINANCIERAS BASADO EN MINERÍA DE DATOS. CASO PRÁCTICO: CACPECO LTDA."

AUTORAS: SALINAS PÉREZ, ADRIANA CUMANDÁ
CHEE TSE, JENNY YEN LIE

DIRECTORA: MBA. DUQUE CRUZ, LORENA GESELLE

SANGOLQUÍ

2019

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN EN SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"MODELO DE MEDICIÓN DE RIESGO CREDITICIO EN ENTIDADES FINANCIERAS BASADO EN MINERÍA DE DATOS. CASO PRÁCTICO: CACPECO LTDA."** fueron realizado por la Ing. **Salinas Pérez, Adriana Cumandá** y la Ing. **Chee Tse, Jenny Yen Lie** el mismo que ha sido revisado en su totalidad y analizado por la herramienta de verificación de similitud de contenido: por lo tanto, cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 4 de diciembre del 2019

MBA. Duque Cruz, Lorena Geselle

C.C.: 1711592525

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN EN SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORÍA DE RESPONSABILIDAD

Nosotras, **Salinas Pérez, Adriana Cumandá**, con cédula de ciudadanía nº 1804255923 y **Chee Tse, Jenny Yen Lie**, con cédula de ciudadanía nº 1710274562 declaramos que el contenido, ideas y criterios del trabajo de titulación: **"MODELO DE MEDICIÓN DE RIESGO CREDITICIO EN ENTIDADES FINANCIERAS BASADO EN MINERÍA DE DATOS. CASO PRÁCTICO: CACPECO LTDA."** son de nuestra autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de la Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 2 de diciembre del 2019

Ing. Adriana Cumandá Salinas Pérez
C.C.: 1804255923

Ing. Jenny Yen Lie Chee Tse
C.C.: 1710274562

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN EN SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORIZACIÓN

Nosotras, **Salinas Pérez, Adriana Cumandá**, con cédula de ciudadanía nº 1804255923 y **Chee Tse, Jenny Yen Lie**, con cédula de ciudadanía nº 1710274562 autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación "**MODELO DE MEDICIÓN DE RIESGO CREDITICIO EN ENTIDADES FINANCIERAS BASADO EN MINERÍA DE DATOS. CASO PRÁCTICO: CACPECO LTDA**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 2 de diciembre del 2019

Ing. Adriana Cumanda Salinas Pérez
C.C.: 1804255923

Ing. Jenny Yen Lie Chee Tse
C.C.: 1710274562

AGRADECIMIENTOS

Agradezco a mi familia por todo el apoyo incondicional que he recibido para poder cumplir con cada uno de mis objetivos.

Salinas Pérez, Adriana Cumandá

AGRADECIMIENTOS

Agradezco a Dios, a mi familia y, a todas aquellas personas que han estado allí de una u otra manera con su demostración de apoyo y guía durante el proceso de realización y culminación de este trabajo.

Chee Tse, Jenny Yen Lie

DEDICATORIA

Dedico el presente trabajo de titulación a mi hijo Tomás por ser mi fuente de motivación, y en quien deposito todo mi amor.

Salinas Pérez, Adriana Cumandá

DEDICATORIA

Dedico el presente trabajo a mi familia por ser un gran soporte en cada una de mi etapa de preparación personal y profesional.

Chee Tse, Jenny Yen Lie

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Antecedentes: La Cooperativa de Ahorro y Crédito CACPECO LTDA. mantiene entre uno de sus objetivos, el potenciar las colocaciones de cartera sin llegar a inflar el indicador de morosidad ni disminuir su liquidez, de tal forma que permitirá fortalecer el crecimiento institucional. **Problema:** La institución posee deficiencias en cuanto a su proceso de selección de potenciales sujetos de crédito, ya que la misma es realizado a través de análisis manuales en herramientas ofimáticas convencionales, sin embargo, no se logran analizar todos los factores deseados que permitan determinar con exactitud cuando un socio es un posible candidato; generando como resultado una deficiente promoción de productos crediticios y, por ende, un bajo porcentaje de colocación de crédito. **Objetivo:** Proveer al área de crédito información estructurada, fácil de comprender y un método útil para el análisis de riesgo del perfil del socio. **Metodología:** La metodología de investigación de la Ciencia del Diseño que en inglés es *Design Science Research* (DSR), fue elegida porque conforman un conjunto de fases o pasos a seguir, las cuales van desde la definición del problema, el desarrollo del artefacto (modelo predictivo) hasta la validación y evaluación del modelo. **Resultados esperados:** Con el desarrollo de este trabajo se pretende obtener un modelo de medición de riesgo crediticio (predictivo) utilizando técnicas de minería de datos que permita promocionar diferentes productos de una cartera en forma más eficientemente a los clientes e incrementar el número de créditos concedidos.

Palabras clave:

- **MODELO DE MINERÍA DE DATOS**
- **RIESGO CREDITICIO**
- **CARTERA DE PRODUCTOS**
- **INVESTIGACIÓN DE LA CIENCIA DEL DISEÑO**

ABSTRACT

Background: The CACPECO Saving and Credit Cooperative Ltd., maintains among one of its objectives, strengthening portfolio placements without inflating the delinquency indicator or reducing its liquidity, in such a way that it will strengthen institutional growth. **Problem**: The institution has deficiencies in the process of selecting potential credit subjects, since it is carried out through manual analysis with conventional office automation tools, however, it is not possible to analyze all the desired factors that allow determining with accuracy when a partner is a possible candidate; generating as a result a poor promotion of credit products and, therefore, a low percentage in the placement of credits portfolio. **Objective**: To provide the credit department with structured information, easy to understand and useful method for the risk analysis of the partner's profile. **Methodology**: The Design Science Research (DSR) methodology was chosen because they form a set of phases or steps to follow, which range from problem definition, artifact development (model predictive) until the validation and evaluation of the model. **Expected results**: The development of this work aims to obtain a credit risk measurement model (predictive) using data mining techniques that allow to promote different products portfolio more efficiently and increase the number of credits granted.

Key words:

- **DATA MINING MODEL**
- **CREDIT RISK**
- **PRODUCTS PORTFOLIO**
- **DESIGN SCIENCE RESEARCH**

CAPITULO I

EL PROBLEMA DE INVESTIGACIÓN

1.1. ANTECEDENTES

En el Ecuador, la cantidad de entidades financieras han crecido notoriamente a través de los años, dando cabida a un alto índice de competencia, de tal forma que el objetivo de estas instituciones es la de posicionarse como líderes en el mercado y, asegurar su estabilidad, enfatizándose principalmente en la adjudicación de préstamos, debido a que dicho producto es la principal fuente de ingresos de las instituciones cooperativas; sin embargo, este proceso debe ser realizado tras un amplio análisis y evaluaciones rigurosas de datos crediticios, ya que el otorgamiento de créditos puede convertirse en un arma de doble filo, dependiendo del retorno del capital, la institución puede posicionarse como líder en el mercado o desfinanciarse y llegar hasta el cierre de la misma. Dentro de este contexto, la cooperativa de Ahorro y Crédito de la pequeña empresa del Cotopaxi (CACPECO LTDA.) tiene como su principal objetivo brindar al micro y pequeño empresario la oportunidad de fortalecer su negocio. La familia CACPECO LTDA. comenzó con su funcionamiento el 14 de marzo de 1988 la cual fue fundada por Luigi Ripalda, con esto le dieron a la provincia de Cotopaxi su primera institución financiera.

El factor de riesgo crediticio en las entidades financieras es un factor primordial de análisis ya que está directamente relacionado con la probabilidad de que el deudor termine de cancelar su operación de crédito en las fechas esperadas, lo cual influye totalmente en el éxito o fracaso de una institución cooperativa. Las técnicas para medir el riesgo de crédito son muy variadas, estas técnicas abarcan procedimientos que van desde simples cálculos, hasta sofisticadas metodologías

con simulaciones dinámicas del futuro más próximo. Bajo un enfoque operacional las campañas bancarias han funcionado en un trabajo conjunto entre las áreas comerciales, riesgos y tecnología para determinar factores determinantes de cada socio como edad, perfil, ingresos, etc. Las cooperativas utilizan provisiones para estimar probabilidades, dichas provisiones son calculadas a partir de pérdidas esperadas, constituyendo estimaciones de las probabilidades de pérdida vinculada a cada cliente individual.

1.2. PLANTEAMIENTO DEL PROBLEMA

CACPECO LTDA. es una cooperativa con una amplia trayectoria a nivel nacional, posterior a realizar un análisis observacional se ha detectado que en el último año ha decaído en porcentaje sobre la cantidad de créditos que se otorgan mensualmente, lo cual se asume es influenciado por el alto número de cooperativas existentes ya que de acuerdo a la SEPS[1] existen más de 800 cooperativas de ahorro y Crédito en el Ecuador[2], además del alto número de solicitudes crediticias negadas, esto ya que en gran número son solicitadas por socios con mal historial crediticio, deficiente calificación de riesgo entre otros ámbitos negativos. En la figura 1 se puede observar la diagramación del problema mediante la técnica “Espina de Pescado”

[1] Superintendencia de Economía Popular y Solidaria

[2] http://www.seps.gob.ec/noticia?ecuador-tiene-un-total-de-887-cooperativas-de-ahorro-y-credito

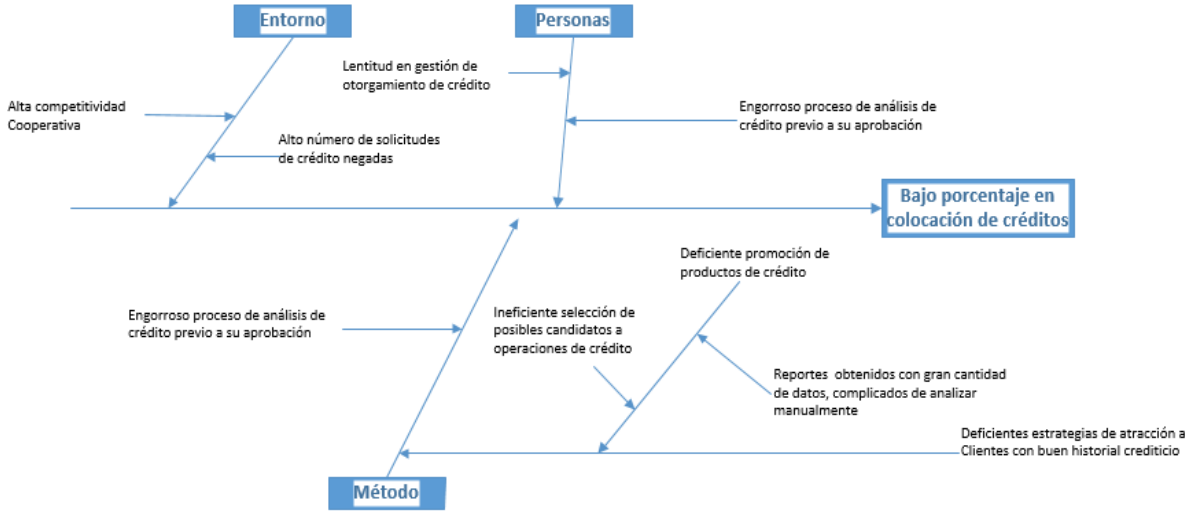


Figura 1. Espina de Pescado

La institución entre sus prácticas para el crecimiento continuo incentiva a socios con altas probabilidades de pago a acceder a créditos para su crecimiento patrimonial. El proceso de selección de los potenciales sujetos de crédito es realizado a través de análisis manuales en herramientas ofimáticas convencionales; de tal forma que, no se logran analizar todos los factores deseados que permitan determinar con exactitud cuando un socio es un posible candidato; llegando a una deficiente promoción de productos crediticios y, por ende, un bajo porcentaje de colocación de crédito.

1.3. OBJETIVOS DE LA INVESTIGACIÓN

1.3.1. OBJETIVO GENERAL

Desarrollar un modelo de medición de riesgo crediticio utilizando técnicas de minería de datos con el fin optimizar los métodos de promoción de créditos e incrementar el porcentaje de su colocación.

1.3.2. OBJETIVOS ESPECÍFICOS

OE1: Analizar la situación actual a través de la entrevista e investigación documental y evaluar las técnicas de minería de datos para medir el riesgo crediticio de una institución financiera con base en la revisión bibliográfica.

OE2: Desarrollar el modelo de medición de riesgo crediticio utilizando técnicas de minería de datos.

OE3: Validar y evaluar el modelo de medición de riesgo crediticio a través del uso de herramientas de minería de datos.

1.4. JUSTIFICACIÓN IMPORTANCIA Y ALCANCE

La colocación de cartera es una de las principales fuentes de ingreso de las entidades financieras, es así que mientras más créditos sean concedidos y cancelados en las fechas previstas, más rentable es la cooperativa y, mayor utilidad es la que percibe; por el contrario, si la institución deja de colocar productos de crédito en el mercado puede llegar a la quiebra.

El presente proyecto se justifica en la cooperativa "CACPECO LTDA." basado en que la institución realiza campañas de socialización y oferta de créditos, para lo cual selecciona sus clientes objetivo únicamente a través de la observación de comportamientos de pago, como resultado se obtienen pocos candidatos, es así que se ve disminuida la cantidad de créditos colocados mensualmente en el mercado, decreciendo el indicador de colocación de cartera.

El beneficio que ofrecerá a la institución es la realización de análisis más detallados a través de modelos predictivos que permitan identificar del total de socios, una mayor cantidad de candidatos con altas posibilidades de retorno de capital, de esta forma se propone hacer campañas

más agresivas en colocación de cartera e incrementar el número de créditos colocados en el mercado, mejorando así las utilidades para la institución.

1.5. PREGUNTAS DE INVESTIGACIÓN

1.6. HIPÓTESIS

El modelo de medición de riesgo crediticio basado en minería de datos permitirá estructurar las campañas de crédito y mejorar el porcentaje de colocación de cartera.

1.6.1. Señalamiento de las variables de la hipótesis

Variable Dependiente: Colocación de Cartera

Variable Independiente: Técnicas de Minería de Datos

CAPITULO II

MARCO REFERENCIAL

2.1. Marco Teórico

En la economía de mercado existen volatilidad en las variables macroeconómicas como el riesgo país, los tipos de cambios, las tasas de interés y demás variables que afectan directamente el costo de capital que ocasionan que las empresas en general registren pérdidas considerables, lo que le motiva el desarrollo de nuevos métodos y técnicas para la gestión de riesgo crediticio con el fin de disminuir las pérdidas por incumplimiento de deudas.

El riesgo crediticio representa la probabilidad de impago de las deudas o crédito contraído con las entidades financiera, por esta razón, es de vital importancia para las entidades financieras predecir el futuro para la toma de decisiones en la colocación de cartera de créditos. Las entidades financieras utilizan diferentes métodos para clasificar a los clientes de acuerdo al riesgo de impago, para lo cual analizan un conjunto de variables personales y la situación financiera del cliente que es sujeto de crédito (Sepúveda Rivillas, Reina Guitiérrez, & Guitierrez Bentacur, 2012).

Realizar un análisis y procesamiento de la información del cliente conlleva mucho tiempo debido a que los datos a analizar no son homogéneos. Las personas solicitan diferentes tipos de crédito como son los créditos comerciales, inmobiliario, vivienda, microcrédito, productivo y de consumo; pero no todos son créditos de sector privado, también existe créditos para la inversión pública. La aprobación de la solicitud de crédito dependerá de que la entidad financiera analice la gran cantidad de variables microeconómicas del sujeto de crédito.

El propósito de esta investigación es de desarrollar un modelo de medición de riesgo crediticio utilizando las diferentes técnicas de minería de datos que permita realizar una búsqueda de patrones en los datos con la finalidad de explorar los atributos que pueden ser relevante para identificar y predecir el comportamiento futuro, y una posterior toma de decisiones objetiva en el otorgamiento de créditos.

El desarrollo del modelo tiene que ser simple como sea posible y con la menor cantidad de características del sujeto de crédito a ser analizado, que permitirá que las decisiones sean tomadas con mayor rapidez y sea a su vez, una motivación para el asesor de crédito de la institución financiera en responder en el menor tiempo posible las solicitudes de crédito logrando en sí una ventaja competitiva con respecto a otras entidades financieras.

La implementación y validación del modelo de medición de riesgo crediticio se aplicará en la base de datos de la Cooperativa de Ahorro y Crédito CACPECO LTDA. que otorga diferentes tipos de créditos en la cual tiene varias agencias distribuida en la provincia de Pichincha, Cotopaxi, Los Ríos, Chimborazo y Tungurahua.

2.1.1. Basilea

El Comité de Basilea de Supervisión Bancaria (organización creada en 1975 y antiguamente estaba formada por los bancos centrales del anterior G10, Luxemburgo y España) que establecen unas garantías mínimas sobre los créditos (PowerData, 2013). Actualmente está formado por representantes de los países del G20 (las economías más industrializadas) a las que se le unen otros países con peso en finanzas, cuyos representantes son de: Alemania, Arabia Saudí, Argentina, Australia, Bélgica, Brasil, Canadá, China, Corea, España, Estados Unidos, Francia,

Hong Kong, India, Indonesia, Italia, Japón, Luxemburgo, México, Países Bajos, Reino Unido, Rusia, Singapur, Sudáfrica, Suecia, Suiza y Turquía (Self Bank, 2017).

El Comité de Basilea realiza recomendaciones, por lo que sus dictámenes y acuerdo no se imponen, pero son aceptados por los representantes de los países miembros. Estas recomendaciones son conocidas con el nombre de Acuerdos de Basilea (Self Bank, 2017).

2.1.1.1. Acuerdos de Basilea

El acuerdo de Basilea son recomendaciones elaboradas por el Comité de Basilea. Fundamentalmente, se encargan de fijar el coeficiente de caja y los niveles de riesgo asumible (PowerData, 2013).

2.1.1.1.1. El acuerdo de Basilea I

El acuerdo fue firmado en 1988, estableció unos principios básicos en los que debía fundamentarse la actividad bancaria. Los más importantes fueron el capital regulatorio, el requisito de permanencia, la capacidad de absorción de pérdidas y la de protección ante quiebra. Este capital debía ser suficiente para hacer frente a los riesgos de crédito, de mercado y de tipo de cambio. El acuerdo establecía también que el capital mínimo de la entidad bancaria debería constituir el 8% del total de los activos de riesgo (crédito, mercado y tipo de cambio sumados) (PowerData, 2013).

2.1.1.1.2. El acuerdo de Basilea II

El acuerdo fue aprobado en 2004, aunque en España no se llegó a aplicar hasta el 2008. Desarrollaba, de manera más extensa, el cálculo de los activos ponderados por riesgo. de esta forma, permitía que las entidades bancarias aplicasen calificaciones de riesgo basadas en sus

modelos internos, siempre que estuviesen previamente aprobadas por el supervisor. Este acuerdo incorporaba, por lo tanto, nuevas tendencias en la medición y el seguimiento de las distintas clases de riesgo. Se hizo énfasis en las metodologías internas, la revisión de la supervisión y la disciplina de mercado (PowerData, 2013).

2.1.1.1.3. El acuerdo de Basilea III

El acuerdo fue aprobado en diciembre de 2010, intentó adaptarse a la magnitud de la crisis económica. Trataba de atender a la exposición de gran parte de los bancos de todo el mundo a los "activos tóxicos" en sus balances y en los derivados que circulaban en el mercado. El temor al efecto dominó que pudiera causar la insolvencia de los bancos, hizo que se establecieron nuevas recomendaciones como (PowerData, 2013):

- ✓ Endurecimiento de los criterios y aumento de la calidad del volumen de capital
- ✓ para asegurar su mayor capacidad para absorber pérdidas.
- ✓ Modificación de los criterios de cálculo de los riesgos para disminuir el nivel de exposición real.
- ✓ Constitución de colchones de capital durante los buenos tiempos que permitan hacer frente el cambio de ciclo económico.
- ✓ Introducción de una nueva ratio de apalancamiento, como medida complementaria a la ratio de solvencia.

2.2. Marco Legal

El presente trabajo de investigación tiene su fundamento Legal en las siguientes leyes:

- ✓ Ley Orgánica de Economía Popular y Solidaria y, en la Superintendencia de Economía Popular y Solidaria

- ✓ Ley Orgánica Del Sistema Cooperativo Y Financiero Popular Y Solidario

En el apartado de anexos se puede encontrar los diferentes artículos determinados las leyes mencionadas.

2.3. Marco Epistemológico

La epistemología, o filosofía de la ciencia, es la rama de la filosofía que estudia la investigación científica y su producto, el conocimiento científico. (Bunge, 1980)

Desde el punto de vista epistemológico, la investigación contará con fuentes de investigación científico teórico que permitirá apoyar las variables independientes y dependientes, obteniéndose de esta manera todo un conjunto de teorías que permitirá sostener el proyecto de investigación.

2.4. Marco Conceptual

En el marco conceptual se orienta a definir lo que es la minería de datos en la cual comprende la descripción de sus métodos, técnicas y herramientas cuyo fin es la exploración de grandes volúmenes de datos con vistas al descubrimiento de la información que servirá de mucha ayuda en el proceso de toma de decisiones para el desarrollo del proyecto de investigación.

2.4.1. Crédito

Se define como crédito a la utilización de un capital ajeno en un periodo de tiempo, a cambio del pago de intereses los cuales son calculados a razón de términos iniciales pactados (Superintendencia de bancos, s.f.).

Crédito bancario: Es un contrato a través del que una entidad financiera otorga a un socio o cliente cierta cantidad de dinero, mismo que deber retornar a la institución con intereses y

comisiones de acuerdo a los plazos establecidos en el contrato inicial. (Superintendencia de bancos, s.f.).

2.4.1.1. Tipos de crédito

El 1 de abril del 2015, la Junta de política y regulación monetaria y financiera expide la resolución No. 043-2015-F, en donde se establece los segmentos de la cartera de crédito que pueden otorgar las entidades del sistema financiero nacional:

2.4.1.1.1. Crédito Productivo

El crédito productivo es aquel que se otorga a personas naturales obligadas a llevar contabilidad o personas jurídicas por un plazo superior a un año con el fin de financiar proyectos productivos cuyo monto, en al menos el 90%, sea destinado en la adquisición de bienes de capital, terrenos, construcción de infraestructura o compra de derechos de propiedad industrial. Exceptuando la adquisición de franquicias, marcas, pagos de regalías, licencias y la compra de vehículos de combustible fósil.

Se incluye en este segmento el crédito directo otorgado a favor de las personas jurídicas no residentes de la economía ecuatoriana para la adquisición de exportaciones de bienes y servicios producidos por residentes. Para el Crédito Productivo se establece los siguientes subsegmentos de crédito: (La Junta de politica y regulacion monetaria, 2015):

a) Productivo Corporativo. - Operaciones de crédito productivo otorgadas a personas jurídicas cuyas ventas anuales superen los $ 5,000,000.00(La Junta de politica y regulacion monetaria, 2015).

b) Productivo Empresarial. - Operaciones de crédito productivo otorgadas a personas jurídicas cuyas ventas anuales se encuentren en un rango de $1,000,000.00 y $5,000,000.00(La Junta de politica y regulacion monetaria, 2015).

c) Productivo PYMES. – Corresponde a operaciones de crédito productivo otorgadas a personas naturales obligadas a llevar contabilidad o a personas jurídicas que registren ventas anuales en un rango entre $100,000.00 y $1,000,000.00(La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.2. Crédito Comercial Ordinario

Es el otorgado a personas naturales obligadas a llevar contabilidad o a personas jurídicas que registren ventas anuales superiores a $ 100,000.00, destinado a la adquisición o comercialización de vehículos livianos, incluyendo los que son para fines productivos y comerciales (La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.3. Crédito Comercial Prioritario

Es el otorgado a personas naturales obligadas a llevar contabilidad o a personas jurídicas que registren ventas anuales superiores a $ 100,000.00 destinado a la adquisición de bienes y servicios para actividades productivas y comerciales, que no estén categorizados en el segmento comercial ordinario. Se incluye en este segmento las operaciones de financiamiento de vehículos pesados y los créditos entre entidades financieras (La Junta de politica y regulacion monetaria, 2015).

Para el Crédito Comercial Prioritario se encuentran establecidos la clasificación siguiente:

a) Comercial Prioritario Corporativo. - Operaciones de crédito comercial prioritario otorgadas a personas naturales obligadas a llevar contabilidad o personas jurídicas que

registren ventas anuales superiores a $5,000,000.00(La Junta de politica y regulacion monetaria, 2015).

b) Comercial Prioritario Empresarial. - Operaciones de crédito comercial prioritario otorgadas a personas naturales obligadas a llevar contabilidad o personas jurídicas que registren ventas anuales en un rango entre $1,000,000.00 y $5,000,000.00(La Junta de politica y regulacion monetaria, 2015).

c) Comercial Prioritario PYMES. - Operaciones de crédito comercial prioritario otorgadas a personas naturales obligadas a llevar contabilidad o personas jurídicas cuyas ventas anuales se encuentren en un rango entre $100,000.00 y $1,000,000.00(La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.4. Crédito de Consumo Ordinario

Otorgado a personas naturales, cuya garantía sea de naturaleza prendaria o fiduciaria, exceptuando los créditos prendarios de joyas. Se incluye los anticipos de efectivo o consumos con tarjetas de crédito corporativas y de personas naturales, cuyo saldo adeudado sea superior a $ 5,000.00; con excepción de los efectuados en los establecimientos médicos y educativos (La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.5. Crédito de Consumo Prioritario

Es el otorgado a personas naturales, destinado a la compra de bienes, servicios o gastos no relacionados con una actividad productiva, comercial y otras compras y gastos no incluidos en el segmento de consumo ordinario, incluidos los créditos prendarios de joyas. Incorpora los anticipos de efectivo o consumos con tarjetas de crédito corporativas y de personas naturales, cuyo saldo adeudado sea hasta $ 5,000.00; con excepción de los efectuados en los

establecimientos educativos. Comprende los consumos efectuados en los establecimientos médicos cuyo saldo adeudado por este concepto sea superior a $ 5,000.00(La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.6. Crédito Educativo

Corresponde a operaciones de crédito otorgadas a personas naturales para su formación y capacitación profesional o técnica y a personas jurídicas para el financiamiento de formación y capacitación profesional o técnica de su talento humano, en ambos casos la formación y capacitación deberá ser debidamente acreditada por los órganos competentes. Se incluye todos los consumos y saldos con tarjetas de crédito en los establecimientos educativos (La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.7. Crédito de Vivienda de Interés Público

Es el otorgado con garantía hipotecaria a personas naturales para la adquisición o construcción de vivienda única y de primer uso, concedido con la finalidad de transferir la cartera generada a un fideicomiso de titularización con participación del Banco Central del Ecuador o el sistema financiero público, cuyo valor comercial sea menor o igual a $ 70,000.00 y cuyo valor por metro cuadrado sea menor o igual a $ 890.00(La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.8. Crédito Inmobiliario

Es el otorgado con garantía hipotecaria a personas naturales para la adquisición de bienes inmuebles destinados a la construcción de vivienda propia no categorizados en el segmento de crédito Vivienda de Interés Público, o para la construcción, reparación, remodelación y mejora de inmuebles propios (La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.9. Microcrédito

Este tipo de crédito es adjudicado a personas naturales o jurídicas con niveles de ventas anuales igual o inferior a $100,000.00, con garantía solidaria, destinado a financiar actividades de producción y/o comercialización en pequeño nivel, su principal fuente de pago está determinada por los ingresos concebidos por las actividades antes mencionadas, y verificados adecuadamente por la entidad financiera.

Para el Microcrédito se establecen los siguientes subsegmentos de crédito (La Junta de politica y regulacion monetaria, 2015):

a) **Microcrédito Minorista**. - Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a las entidades del sistema financiero nacional, sea menor o igual a $ 1,000.00, incluyendo el monto de la operación solicitada (La Junta de politica y regulacion monetaria, 2015).
b) **Microcrédito de Acumulación Simple**. - Operaciones de crédito cuyo saldo adeudado en microcréditos se encuentre en el rango de $ 1,000.00 hasta USD 10,000.00, incluyendo el monto de la operación solicitada (La Junta de politica y regulacion monetaria, 2015).
c) **Microcrédito de Acumulación Ampliada**. – Operaciones de microcrédito cuyo saldo adeudado es superior a $ 10,000.00, incluyendo el monto de la operación solicitada (La Junta de politica y regulacion monetaria, 2015).

2.4.1.1.10. Crédito de Inversión Pública

Es el destinado a financiar programas, proyectos, obras y servicios encaminados a la provisión de servicios públicos, cuya prestación es responsabilidad del Estado, sea directamente o a través de empresas; y, que se cancelan con cargo a los recursos presupuestarios o rentas del deudor

fideicomitidas a favor de la institución financiera pública prestamista. Se incluyen en este segmento a las operaciones otorgadas a los Gobiernos Autónomos Descentralizados y otras entidades del sector público (La Junta de politica y regulacion monetaria, 2015).

2.4.1.2. Riesgo crediticio

Es la posibilidad de perder el valor desembolsado, debido al no pago de obligaciones por parte del deudor o la contraparte, debido a la falta de pago de las obligaciones adquiridas.

2.4.1.2.1. Elementos del riesgo crediticio

El riesgo crediticio puede analizarse en tres elementos (Galicia, 2003) citado en (Saavedra García & Saavedra García, 2010):

- ✓ **Riesgo de incumplimiento**: Es la probabilidad de que se presente el no cumplimiento de una obligación de pago, el rompimiento de un acuerdo en el contrato de crédito o el incumplimiento económico. A este respecto, generalmente las autoridades establecen plazos de gracia antes de poder declarar el incumplimiento de pago (Saavedra García & Saavedra García, 2010).
- ✓ **Exposición:** Es la incertidumbre respecto a los montos futuros en riesgo. El crédito debe amortizarse de acuerdo con fechas establecidas de pago y de esta manera será posible conocer anticipadamente el saldo remanente a una fecha determinada; sin embargo, no todos los créditos cuentan con esta característica de gran importancia para conocer el monto en riesgo. Tal es el caso de los créditos otorgados a través de tarjetas de crédito, líneas de crédito resolventes para capital de trabajo, líneas de crédito por sobregiro, etc., ya que los saldos en estas modalidades de crédito se modifican según las necesidades del cliente, los desembolsos se otorgan sin fecha fija contractual y no se conoce con exactitud

el plazo de liquidación; por ello se dificulta la estimación de los montos en riesgo (Saavedra García & Saavedra García, 2010).

- ✓ **Recuperación**: Se origina por la existencia de un incumplimiento. No se puede predecir, puesto que depende del tipo de garantía que se haya recibido y de su situación al momento del incumplimiento. La existencia de una garantía minimiza el riesgo de crédito siempre y cuando sea de fácil y rápida realización a un valor que cubra el monto adeudado. En el caso de los avales, también existe incertidumbre, ya que no sólo se trata de una transferencia de riesgo en caso del incumplimiento del avalado, sino que podría suceder que el aval incumpliera al mismo tiempo y se tuviera entonces una probabilidad conjunta de incumplimiento (Saavedra García & Saavedra García, 2010).

2.4.1.2.2. Modelo de medición del riesgo crediticio

El objeto de medir el riesgo crediticio es identificar los determinantes del riesgo crediticio que afecta a las entidades financieras, con el propósito de prevenir las pérdidas potenciales en las que podría incurrir. En el análisis del riesgo crediticio se deben tomar en cuenta los criterios de calificación de las carteras crediticias de la entidad financiera, la estructura y la composición de los portafolios crediticios, el impacto de las variables macroeconómicas y las características históricas de las carteras de crédito de cada entidad financiera (Saavedra García & Saavedra García, 2010).

Modelos tradicionales

Los modelos tradicionales de administración del riesgo se basan en un esquema de análisis de ciertos componentes básicos para evaluarlo. Estos modelos se aplican cuando no se cuenta con

herramientas avanzadas o con expertos que puedan aplicarlos o cuando la experiencia del evaluador y el conocimiento acerca del cliente no permiten tomar decisiones de otorgar o no el crédito, sin necesidad de profundizar más en el análisis del riesgo.

a) **Modelo de sistemas expertos:** Los sistemas expertos tratan de captar la intuición de los expertos y sistematizarla aprovechando la tecnología, pues su campo de dominio es la inteligencia artificial, por medio de la cual intentan crear sistemas expertos y redes neuronales. Sin embargo, quedan limitados tan sólo a la etapa de calificación, ya que no pueden establecer un vínculo teórico identificable con la probabilidad de impago y la gravedad de la pérdida, aunque sí les resulta posible establecer una correspondencia entre calificaciones y probabilidades de quiebra (Saavedra García & Saavedra García, 2010).

b) **Modelo 5 Cs:** Esos procedimientos están basados en la opinión subjetiva de los directivos más experimentados, que ponderan, de forma personal, dos tipos de información. La primera de ellas, relacionada con la calidad y posible liquidez del colateral o de las garantías que se han depositado. La segunda relacionada con una serie de elementos característicos del prestatario (por ejemplo, sus ratios financieras) relacionados con su habilidad para generar flujos de caja suficientes para hacer frente a las obligaciones del contrato de préstamo (Valle Carrascal, 2015).

   - ✓ **Capacidad:** Corresponde a evaluar la capacidad de pago del cliente, su trayectoria en los negocios, su gestión y los resultados que ha obtenido. Entre los aspectos que se consideran se encuentran: el crecimiento que ha tenido y antigüedad de la compañía, entre otras. Se busca también conocer el flujo de efectivo que posee el negocio, a fin de poder determinar cómo el cliente pagará el préstamo (Saavedra García & Saavedra García, 2010).

- ✓ **Capital:** Corresponde a un análisis financiero, el cual posibilita conocer en su totalidad, las probabilidades de pago del cliente, sus ingresos, gastos, nivel de endeudamiento, rotación de inventario, liquidez, etc. (Saavedra García & Saavedra García, 2010).
- ✓ **Colateral:** Corresponde a las garantías o apoyos colaterales que el cliente posee para asegurar el pago de la deuda. Este punto se analiza mediante los activos fijos que posee el cliente y su respectivo valor económico y calidad, dado que no se otorgará un crédito sin que se cuente con una fuente de pago secundaria.
- ✓ **Carácter:** Corresponde a analizar los comportamientos de pago pasados y actuales del deudor. Este análisis se debe ejecutar mediante elementos que puedan verificarse y cuantificarse (reporte de buró de crédito, revisión de demandas judiciales, referencias bancarias, etc.) (Saavedra García & Saavedra García, 2010).
- ✓ **Condiciones:** Se analizan aspectos externos que puedan influir en el desempeño del negocio del deudor (situación económica y política del sector o región, etc.), pese a que esto, no son controlados por el deudor (Saavedra García & Saavedra García, 2010).

c) Modelo Credit Scoring: El Credit Scoring, también conocido como calificación del riesgo de impago (default) o morosidad, puede concebirse como un sistema que, mediante predicciones, califica un crédito y mide el riesgo inherente al mismo. La medición del riesgo de crédito implica la utilización de modelos estadísticos apropiados que permiten obtener un valor esperado y su correspondiente variabilidad o volatilidad. Estas técnicas estadísticas permiten conocer el comportamiento financiero de los prestatarios y su morosidad, la relación entre el riesgo y la rentabilidad, y la determinación del coste de la

operación, tendiente a lograr su reducción futura en el marco del proceso de concesión de un crédito (Seijas Giménez, Vivel Búa, Lado Sestayo, & Fernández López, 2017).

2.4.1.2.3. Técnicas paramétricas

Las técnicas paramétricas son aquellas que presentan una función de distribución o clasificación conocidas, al igual que estiman parámetros para explicar un determinado suceso de tal modo que estos se ajusten a las observaciones de una muestra. Dentro de este grupo, es posible distinguir técnicas paramétricas lineales (análisis discriminante y modelos de probabilidad lineal) así como no lineales (modelos logit y probit) (Seijas Giménez, Vivel Búa, Lado Sestayo, & Fernández López, 2017).

a) **Análisis discriminante (lineal):** El análisis discriminante consiste en una técnica multivariante que permite estudiar simultáneamente el comportamiento de un grupo de variables independientes con la intención de clasificar una serie de casos en grupos previamente definidos y excluyentes entre sí (Fisher, 1936) citado por (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).La principal ventaja de esta técnica está en la diferenciación de las características que definen cada grupo, así como las interacciones que existen entre ellas. Se trata de un modelo apropiado para clasificar buenos y malos pagadores a la hora de reembolsar un crédito (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).
b) **Modelos de Probabilidad Lineal (lineal):** Los modelos de probabilidad lineal utilizan un enfoque de regresión por cuadrados mínimos, donde la variable dependiente (variable dummy) toma el valor de uno (1) si un cliente es fallido, o el valor de cero (0) si el cliente

cumple con su obligación de pago. La ecuación de regresión es una función lineal de las variables explicativas (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).

c) **Modelo Logit (no lineales): Los modelos** de regresión logística permiten calcular la probabilidad que tiene un cliente para pertenecer a uno de los grupos establecidos a priori (no pagador o pagador). La clasificación se realiza de acuerdo con el comportamiento de una serie de variables independientes de cada observación o individuo. La principal ventaja del modelo de regresión logística radica en que no es necesario plantear hipótesis de partida, como por ejemplo la normalidad de la distribución de las variables, mejorando el tratamiento de las variables cualitativas o categóricas (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).

d) **Modelo Probit (no lineales):** El modelo Probit es aquel donde la variable dependiente puede tomar únicamente dos valores. Ejemplo: casados o no casados, morosos o no morosos, solventes o no solventes.

   El objetivo de este modelo es determinar la probabilidad de que una observación con características particulares caiga en una categoría específica; también es un tipo de modelo de clasificación binario (Wikipedia, Modelo probit, 2018).

2.4.1.2.4. Técnicas no paramétricas

Las técnicas no paramétricas no se encuentran ligadas a ninguna forma funcional ni distribución concreta de probabilidad. Estas técnicas no tienen por objetivo la búsqueda de parámetros de una función conocida, sino que tratan de obtener formas funcionales que aproximen a la función objetivo. El grupo de las técnicas no paramétricas comprende la

programación lineal, las redes neuronales y los árboles de decisiones (Seijas Giménez, Vivel Búa, Lado Sestayo, & Fernández López, 2017).

a) **Programación lineal:** Los modelos de programación lineal permiten programar plantillas o sistemas de asignación de rating sin perder de vista el criterio de optimización de clientes correctamente clasificados (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).
b) **Redes neuronales:** Las redes neuronales artificiales tratan de imitar al sistema nervioso, de modo que construyen sistemas con cierto grado de inteligencia. La red está formada por una serie de procesadores simples, denominados nodos, que se encuentran interconectados entre sí. Como nodos de entrada consideramos las características o variables de la operación de crédito. El nodo de salida sería la variable respuesta definida como la probabilidad de no pago. La finalidad de cada nodo consiste en dar respuesta a una determinada señal de entrada. El proceso de credit scoring mediante el uso de esta técnica resulta complicado, pues el proceso interno de aprendizaje funciona como una "caja negra" (capa oculta), donde la comprensión de lo que ocurre dentro requiere de conocimientos especializados (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).
c) **Arboles de decisiones:** La principal ventaja de esta metodología es que no está sujeta a supuestos estadísticos referentes a distribuciones o formas funcionales. Aunque conllevan una comprensión interna difícil sobre su funcionamiento, presentan relaciones visuales entre las variables, los grupos de la variable respuesta y el riesgo; por ello, este método es muy usado en el credit scoring. Los algoritmos más comunes para construir los árboles de decisión son el ID3, C4.5 y C5. En cada uno de ellos se persigue la separación óptima en la muestra, de tal modo que los grupos de la variable respuesta ofrecen distintos perfiles de riesgo (Rayo Cantón, Lara Rubio, & Camino Blasco, 2010).

d) **Modelo de Rating:** Un modelo de rating se entiende que al igual que un modelo scoring, es una herramienta que utiliza técnicas estadísticas para predecir si un cliente es mal pagador o buen pagador. La diferencia está en que un modelo scoring tiene como output un resultado objetivo, mientras que el rating es visto por un analista experto en la materia, pero que no quita el sesgo que puede haber a la hora de analizar. Otra gran diferencia es que el modelo de rating es utilizado para grandes y medianas empresas, mientras que el modelo scoring es usado para pequeñas empresas y personas.

e) **Modelos modernos:** Los modelos modernos de administración del riesgo intentan registrar la alta volatilidad a la que están sujetos los valores y emplean técnicas más sofisticadas para su determinación. Estos modelos se aplican cuando se manejan los créditos en un entorno altamente inestable y cuando los montos son relevantes

f) **Modelo CyRCE:** El modelo CyRCE permite evaluar la suficiencia del capital asignado por una institución financiera a una cartera de riesgos crediticios, a través de su comparación con el valor en riesgo de dicha cartera, definido como la máxima pérdida posible con una probabilidad de ocurrencia alta y durante un determinado horizonte temporal. El modelo supone que están dadas las probabilidades de incumplimiento de los créditos y sus covarianzas. Con estas últimas obtiene la forma funcional de la distribución de pérdidas, suponiendo que las mismas pueden ser caracterizadas por dos parámetros: la media y la varianza. De esta forma, el VaR puede establecerse como la pérdida esperada más un cierto múltiplo de la desviación estándar de las pérdidas, que es el valor de la pérdida que acumula el porcentaje de probabilidad impuesto por el intervalo de confianza elegido (Fernandez & Soares Netto).

g) **Modelo KMV:** Este es un modelo de diversificación basado en las correlaciones del mercado de acciones que permite estimar la probabilidad de incumplimiento entre activos y pasivos. El modelo KMV (Kealhofer, McQuown and Vasicek) toma ideas del modelo de Frecuencias de Incumplimiento Esperado (EDF, por su sigla en inglés), además de considerar la diversificación requerida en los portafolios de deuda (Saavedra García & Saavedra García, 2010). El modelo KMV define la probabilidad de incumplimiento como una función de la estructura del capital de la firma, la volatilidad del rendimiento esperado de los activos y su valor actual. Las EDF son específicas de una empresa y pueden ser transformadas hacia cualquier sistema de calificación para derivar la calificación equivalente del acreditado. Así también, las EDF pueden verse como calificaciones cardinales de los acreditados respecto del riesgo de incumplimiento, en lugar de la más convencional calificación ordinal propuesta por las agencias de calificaciones, expresadas en las letras, como AAA, AA, etc (Saavedra García & Saavedra García, 2010).

h) **Modelo CreditMetrics:** El modelo CreditMetrics fue desarrollado por J.P. Morgan en el año de 1997. CreditMetrics es un modelo de medición de riesgo en crédito basado en el uso de matrices de transición y de elementos obtenidos del mercado que permite obtener la distribución del valor de un activo con riesgo de crédito o de una cartera de créditos en un periodo de tiempo T. Los elementos clave del modelo son (Téllez Cabrera, 2010):

   - ✓ Las calificaciones que otorgan las empresas calificadoras de valores, que reflejan la apreciación de éstas sobre la calidad del papel y que permiten obtener las tasas o probabilidades de migración de los papeles con diferentes calificaciones (Téllez Cabrera, 2010).

- ✓ Los diferenciales de tasas entre instrumentos de deuda con diferente calidad, según quedan reflejados en distintos niveles de calificación (Téllez Cabrera, 2010).
- ✓ Las tasas de recuperación de los créditos que caen en cartera vencida, que
- ✓ dependen de la prelación del crédito, por ejemplo, un crédito con garantías tiene
- ✓ una tasa de recuperación mayor al de un crédito sin garantías (Téllez Cabrera, 2010).

i) Modelo Credit Risk Plus (Credit Suisse): El modelo Credit-Risk+ fue propuesto por Credit Suisse Financial Products (CSFP) consiste en obtener la función de distribución de perdida esperada de la cartera de créditos. No obstante, bajo este último planteamiento se asume como hipótesis de partida que la tasa de quiebra es una variable aleatoria continua, cuya volatilidad recoge la incertidumbre acerca de su comportamiento futuro. Se aplican una serie de técnicas de cálculo actuarial para modelizar el riesgo de quiebra (Ruza & Curbera, 2013).

j) Modelo RAROC: El modelo RAROC (Risk- adjusted return on Capital) o el modelo de rentabilidad sobre capital ajustada por el riesgo fue desarrollado por el Bankers Trust. RAROC es una medida de rentabilidad que puede aplicarse a operaciones aisladas o a una cartera de préstamos. Es una medida que también incorpora los riesgos que se asumen al realizar una operación de crédito. Es un indicador para la gestión del negocio bancario que mide la relación entre rentabilidad y la gestión del riesgo. El análisis del RAROC revela la cantidad de capital económico que es requerido por cada línea de negocio, producto o cliente, y cómo estos requerimientos crean la rentabilidad total de capital producido por la empresa (Cortez Cortez, 2011).

RAROC = Utilidad neta / Capital en riesgo

2.4.2. Minería de Datos

La minería de datos o descubrimiento de conocimiento en bases de datos, es una herramienta tecnológica eficaz de grandes dimensiones en la búsqueda y selección de información inédita y potencialmente útil a partir de un gran cumulo de datos de información. La minería de datos, define patrones y las relaciones que existen en los datos analizados, automatizando este proceso y proporciona resultados que pueden ser utilizados en un sistema de apoyo para la toma de decisiones estratégicas de la organización (Cendejas Valdez, Acuña Lopez, & Cortes Morales, 2017).

2.4.2.1. Métodos de Minería de Datos

El proceso de minería de datos está conformado de algoritmos asociadas a diferentes tipos de tareas como análisis predictivo y descriptivo, estas características son descritos en la tabla 1.

Tabla 1.
Algoritmo asociado al tipo de método de minería de datos.

Nombres	**PREDICTIVO**		**DESCRIPTIVO**		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones/ Factorizaciones
Redes neuronales	✓	✓	✓		
Árboles de decisión ID3, C5.0	✓				
Árboles de decisiones CART	✓	✓			
Otros árboles de decisión	✓	✓	✓	✓	
Redes de Kohonen			✓		

CONTINÚA

Regresión lineal y logarítmica			✓			
Regresión logística	✓				✓	
Kmeans				✓		
A priori					✓	
Naive Bayes	✓					
Vecinos más próximos	✓		✓	✓		
Análisis factorial y de componentes principales						✓
Twostep, Cobweb				✓		
Algoritmos genéticos y evolutivos	✓		✓	✓	✓	✓
Máquinas de vectores de soporte	✓		✓	✓		
CN2 rules (cobertura)	✓				✓	
Análisis discriminante multivariante	✓					

Fuente: Elaborada por las autoras basada en (Hernández Orallo, Ramírez Quintana, & Ramírez Ferri, 2004)

2.4.2.1.1. Método Descriptivos o de Aprendizaje no Supervisado

Los Métodos Descriptivos o de aprendizaje no supervisado permiten formar grupos de datos rápidamente, también son conocidos como métodos simétricos, no supervisados o indirectos. Las observaciones son generalmente clasificadas en grupos que no son conocidos con anterioridad, los elementos de las variables pueden estar conectados entre sí de acuerdo a vínculos desconocidos de antemano, de esta manera, todas las variables disponibles son tratados en el mismo nivel y no hay hipótesis de causalidad. (Molina López & García Herrero, 2006)

2.4.2.1.2. Método Predictivo o de Aprendizaje Supervisado

Los Métodos predictivos se basan en entrenar a un modelo o método por medio de diferentes datos para poder predecir una variable partiendo de estos mismos datos. El objetivo de los

modelos predictivos es describir una o más de las variables en relación con todas las demás, son conocidos como métodos asimétricos, supervisados o directos. Se llevan a cabo mediante la búsqueda de normas de clasificación o de predicción basada en los datos, estas normas nos ayudan a predecir o clasificar el resultado futuro de una o más variables de respuesta o de destino en relación a lo que ocurre en la práctica con los motivos que la causan o bien en relación con las variables de entrada. (Molina López & García Herrero, 2006)

2.4.2.2. Técnicas de Minería de Datos

Las técnicas de Minería de Datos basan su clasificación en dos grandes categorías: Supervisadas (Predictivas) y No Supervisadas (Descriptivas). La técnica es un enfoque conceptual que permite extraer información de los datos y, por lo general es implementada por varios algoritmos. En la práctica, cada algoritmo representa una manera de desarrollar una específica técnica paso a paso; por lo que, es indispensable tener una comprensión de alto nivel de los algoritmos para conocer cuál de tantas técnicas son las más apropiadas para cada problema; adicionalmente, es necesario entender las características y los parámetros de los algoritmos para preparar los datos a analizar. (Molina López & García Herrero, 2006), en la Figura 2 se puede visualizar las diferentes tecnicas de mineria de datos.

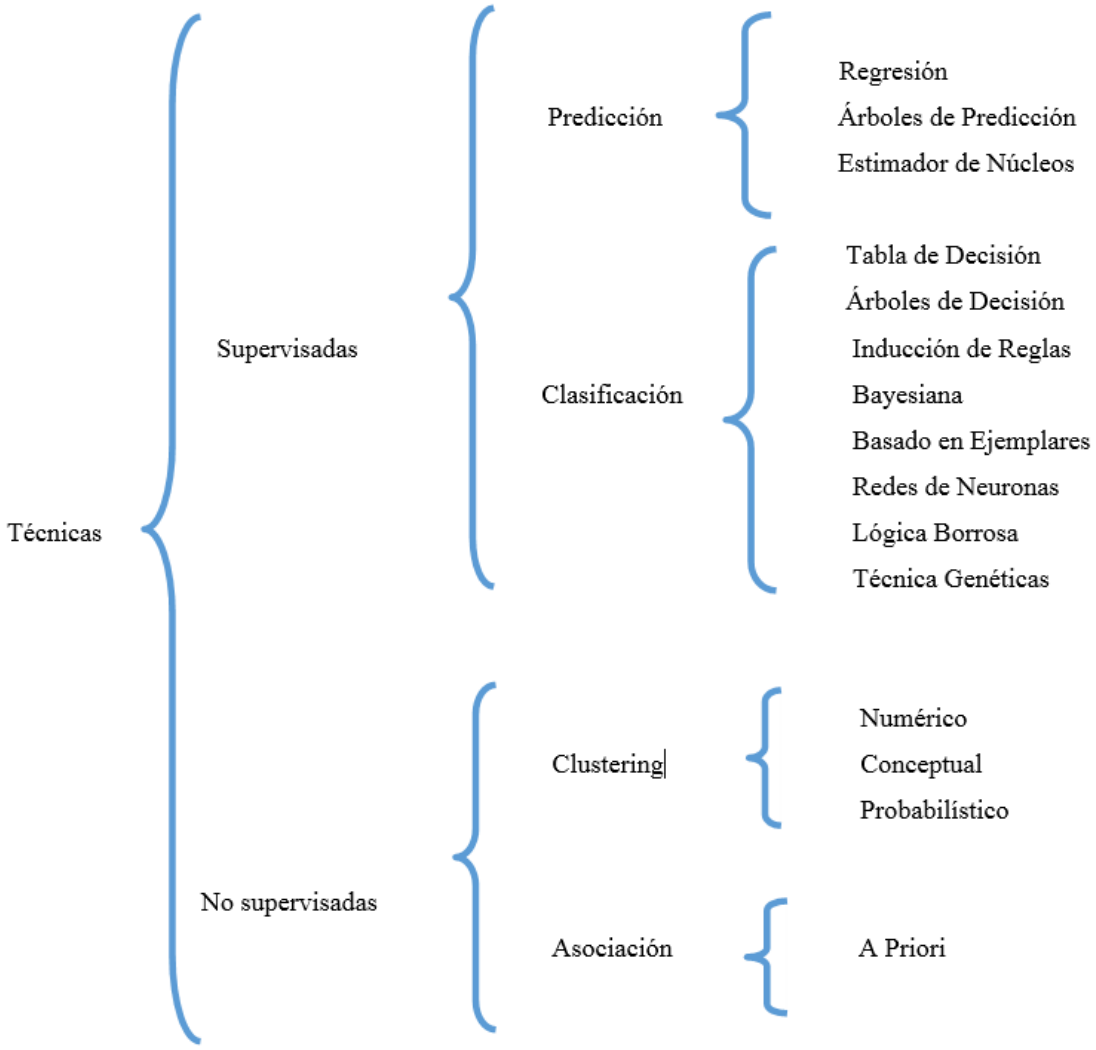


Figura 2. Técnicas de Minería de Datos.
Fuente: Elaborada por las autoras y cuadro sinóptico obtenido de (Molina López & García Herrero, 2006)

2.4.2.2.1. Técnicas Supervisadas

En las técnicas Supervisadas se tiene: la Predicción (Regresión, Arboles de Predicción, Estimador de Núcleos) y la Clasificación (Tabla de Decisión, Arboles de Decisión, Inducción de

Reglas, Bayesiana, Basado en Ejemplares, Redes de Neuronas, Lógica Borrosa, Técnicas Genéticas). (Molina et al., 2006).

Predicción

Predicción es una técnica que trata de determinar los valores de las variables objetivo a partir del análisis y estudio un grupo de datos, tratando así de predecir la ocurrencia de algún suceso. Las técnicas de predicción más comunes se muestran a continuación:

a) **Regresión:** El objetivo de la regresión es predecir el valor numérico de alguna variable a través de un modelo de minería de datos. Está clasificada en diferentes tipos como son:
   - ✓ **Regresión Lineal:** En la regresión lineal es posible hacer predicciones sobre la respuesta en base a los valores de la variable predictora. (Carrasquilla-Batista, Chacon Rodriguez, Núñez Montero, & Gómez Espinoza, 2016)
   - ✓ **Regresión Lineal simple:** La regresión lineal simple nos permite obtener una función lineal de una variable independiente o predictora (X1) a partir de la cual se va a explicar o predecir el valor de una variable dependiente o criterio (Y). (Rodriguez Jaume & Morar Catala, 2001)
   - ✓ **Regresión Lineal Múltiple:** La Regresión Lineal Múltiple nos permite establecer la relación que se produce entre una variable dependiente (Y) y un conjunto de variables independientes (X1, X2, ... XK). El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción. (Rodriguez Jaume & Morar Catala, 2001)

b) Arboles de Predicción: Los árboles de predicción numérica son semejantes con los árboles de decisión con la diferencia en que la clase que se predecirá es continua. La predicción de valores continuos es posible realizarla a través de técnicas estadísticas de regresión lineal, por lo mismo los modelos obtenidos sólo operan con atributos numéricos.

c) Estimador de Núcleos: Los estimadores de densidad de núcleo (Kernel density) son estimados no paramétricos. De entre los que destaca el conocido histograma, por ser uno de los más antiguos y más utilizado, que tiene ciertas deficiencias relacionadas con la continuidad que llevaron a desarrollar otras técnicas. (Molina López & García Herrero, 2006)

Clasificación

Se define como la identificación de características o atributos que hacen que un elemento se vincule a un grupo siguiendo un patrón de datos. Este último se puede utilizar para predecir cómo se comportarán las nuevas instancias. (Valcárcel Asencios, 2004)

a) **Tabla de Decisión**: Se define a la tabla de decisión como la manera más rudimentaria de simbolizar la salida de un algoritmo de aprendizaje, que es figurarlo como la entrada. Estos algoritmos realizan un proceso de selección de subconjuntos de atributos para calcular su precisión, con el fin último de realizar predicciones o clasificaciones, oficinas. Es necesario seleccionar el mejor de los subconjuntos, posterior la tabla de decisión se deberá formar por los atributos seleccionados (más la clase), se insertarán todos los casos de entrenamiento y el subconjunto de atributos seleccionado.

b) **Arboles de Decisión**: Es una técnica de predicción aplicada en el campo de la inteligencia artificial, en donde a partir de una base de datos se crean esquemas lógicos, mismos que

simbolizan y categorizan una serie de condiciones que ocurren de forma repetitiva. Por lo general en esta técnica de construcción de árboles de clasificación, se toma la próxima partición de manera óptima en el conjunto del árbol, esto evita la confusión combinatoria en cuanto número de decisiones futuras a considerar, por eso hay que elegir la medida justa a optimizar en cada corte, para facilitar las próximas divisiones. (Solarte Martínez & Ocampos, 2009)

c) **Inducción de Reglas**: Las técnicas de Inducción de Reglas permiten la elaboración y oposición de árboles de decisión, reglas y patrones a partir de datos de entrada. Los datos de entrada consisten en un conjunto de casos o sucesos donde se ha elaborado una previa clasificación.

d) **Bayesiana**: Un clasificador Bayesiano sencillo es conocido como el clasificador "Naive Bayesiano". Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones. (Aluja, 2001)

e) **Basado en Ejemplares**: El Aprendizaje basado en ejemplares, es aquel que permite el almacenamiento de los ejemplos de entrenamiento, en el momento que se requiera realizar la clasificación de un nuevo objeto, se extraen los objetos más parecidos y se utiliza su clasificación para clasificar al nuevo objeto. Este tipo de aprendizaje también se conoce como lazy learning o memory-based learning donde los datos de entrenamiento se procesan solo hasta que se requiere, la importancia de los datos es medible en función de una medida de distancia. (Gonzalez Bernal, 2011)

f) **Regresión Logística**: La regresión logística en su forma más simple, es decir, con una respuesta binaria, propone que el logaritmo de la "razón de probabilidad" sea entendida como el cociente entre la probabilidad de éxito y la de fracaso en un ensayo de Bernoulli, es igual a una función lineal en los parámetros, denominada usualmente predictora lineal (Ponsot, Sinha, Surendra, & Goitía, 2009). En esta situación el objetivo es estimar y establecer la significancia estadística de los factores frente a una respuesta observada, y al operar con la inversa del logaritmo de la razón de probabilidad en función de la predictora lineal, se predicen las probabilidades de éxito en cada combinación de niveles de los factores. (Ponsot, Sinha, Surendra, & Goitía, 2009)

g) **Redes de Neuronas**: La Red Neural Artificial fue presentada por Warren S. McCulloch y Walter Pitts para clasificación y predicción problemas. El perceptrón multicapa (MLP) es una técnica supervisado con una capa de entrada, una o más capas ocultas y una capa de salida en la que cada capa involucra algunas neuronas, que es una red neuronal de avance (no hay bucle de retroalimentación). El algoritmo de aprendizaje más popular para redes neuronales es propagación de vuelta. Una Red Neuronal tienen salida probabilística. El número de neuronas, la cantidad de capas y conexiones ocultas entre los nodos determinan la topología de una red neuronal. (Ghobadi & Rohani, 2016)

h) **Lógica Borrosa**: La lógica Borrosa, Difusa o Fuzzy (Términos usados indistintamente a lo largo del proyecto), se basa en lo relativo de lo observado. Este tipo de lógica toma dos valores aleatorios, pero contextualizados y referidos entre sí. La lógica difusa se adapta mejor al mundo real, e incluso puede comprender y funcionar con nuestras expresiones, del tipo "hace mucho calor", "no es muy alto", "el ritmo del corazón está un poco acelerado". La clave de esta adaptación al lenguaje, se basa en comprender los

cuantificadores de nuestro lenguaje como, por ejemplo, "mucho", "muy", "un poco". (Carranza Rueda, 2008)

i) **Técnicas Genéticas**: Los algoritmos genéticos, también llamados algoritmos evolutivos, están inspirados en los principios que rigen la evolución de los seres vivos en la naturaleza que aparecen en el libro Origen de las especies de Charles Darwin (1859). Aunque en la década de 1960 surgieron trabajos sobre algoritmos que simulaban estrategias evolutivas, no fue hasta 1975 que John H. Holland definió y concretó en su libro Adaptation in Natural and Artificial Systems (1975) las bases utilizadas hoy en día para los algoritmos genéticos. Estos algoritmos se fundamentan en que los individuos mejor adaptados al entorno en el que viven son los que tendrán más probabilidades de tener descendencia y, consecuentemente, que sus características se combinen con las de otros individuos. Es probable que un individuo bien adaptado al entorno escoja a otro individuo que también esté muy adaptado, ya que ambos poseen características que les hacen sobresalir del resto. La combinación del material genético de los dos producirá una descendencia que tendrá características de ambos individuos, que, combinadas en distinto grado, posiblemente mejoren la adaptación de la descendencia al entorno aún más. Como ejemplo, si una gacela que acierta a detectar depredadores con mucha antelación, tiene descendencia con otra que destaca por su rapidez, su descendencia combinará en cierto grado estas dos características y sus posibilidades de sobrevivir ante el ataque de un depredador posiblemente aumentarán respecto a las de sus progenitores. Es también más probable que sobrevivan más tiempo que la media de la población, al poseer ambas características que lo facilitan, con lo cual es probable que tengan mayor número de descendencia que la media. De esta forma, siendo los mejor adaptados los que tienen más

descendencia, se aumenta la calidad de la población con el paso de las generaciones. (Llorente Lopez , 2012)

2.4.2.2.2. Técnicas No Supervisadas

En las técnicas no Supervisadas se tiene: el Clustering (Numérico, Conceptual, Probabilístico) y la Asociación (A Priori). (Molina et al., 2006).

Clustering

También llamada agrupamiento, permite identificar conjuntos de elementos que mantienen muchas semejanzas entre sí y alto grado de diferencias con los de otros conjuntos. De esta forma es posible segmentar: socios, estudiantes, oficinas, índices financieros, docentes, empleados, sucursales, clientes etc.

El clustering o segmentación ofrece importantes beneficios al permitir el tratamiento de grandes cantidades de datos de forma particular, manejando un punto intermedio entre el tratamiento individualizado y masificado. (Molina López & García Herrero, 2006)

a) **Numérico:** Es un algoritmo particional, es decir, divide los objetos en un número de clústeres pre especificado, sin atender a una estructura jerárquica, puede aplicarse para problemas de "agrupación por similitud" y puede ayudar al investigador a una comprensión cualitativa y cuantitativa de grandes cantidades de datos N-dimensionales. Funciona de forma iterativa, dividiendo óptimamente el conjunto inicial de datos en un número (K) de clústeres, el cual se indica como parámetro. (Hernández Cáceres, 2016)
b) **Conceptual:** Es un algoritmo jerarquico de clustering, utiliza aprendizaje incremental, es decir que realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo

se forma un árbol de clasificación donde las hojas representan los segmentos y el nodo raíz abarca por completo el conjunto de datos de entrada. Al inicio, el árbol mantiene un único nodo raíz, posteriormente las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso, esta actualización consiste en identificar el mejor sitio en donde se pueda incluir la nueva instancia, esta operación puede necesitar de la reestructuración de todo el árbol o simplemente la inclusión de la instancia en un nodo preexistente. La medida denominada "utilidad de categoría" permite medir la calidad de una partición de instancias y de tal manera identificar como y donde actualizar el árbol, La reestructuración que mayor utilidad de categoría proporcione es la que se adopta en ese paso. (Garre, Cuadrado, & Sicilia, s.f.)

c) **Probabilístico:** El clústering probabilístico es conocido también como algoritmo EM (Expectation Maximization). El algoritmo EM permite un acercamiento probabilístico al problema del clústering, solucionando los mencionados problemas. Ahora, en lugar de buscar sujetos parecidos entre sí de manera iterativa, lo que se intenta es buscar el grupo de clústeres más probables dado un conjunto de puntuaciones. El algoritmo se basa en calcular las probabilidades que existen de que un sujeto tenga una puntuación en la variable, si se supiera que el sujeto es miembro de ese clúster. Así, se obtienen k distribuciones de probabilidad, una por cada uno de los k clústeres. Lo que hace el algoritmo EM es adivinar inicialmente los parámetros de las distribuciones para, a continuación, emplear esos parámetros para llevar a cabo el cálculo de las probabilidades de que cada sujeto pertenezca a un clúster. Posteriormente, emplea esas probabilidades para re-estimar los parámetros. Y así hasta llegar al criterio de parada establecido, en base a un valor mínimo de convergencia. (Martínez Abad)

Asociación

Las técnicas de asociación son utilizadas con el propósito de establecer posibles relaciones entre diferentes acontecimientos aparentemente independientes; logrando identificar como la ocurrencia de un suceso puede inducir en la aparición de otros sucesos.

Se utilizan cuando se requiere realizar buscar asociaciones directas o indirectas dentro del conjunto de datos. Las relaciones que se logran identificar pueden ser utilizadas para predecir comportamientos, descubrir correlaciones y ocurrencias de eventos.

Estas técnicas son aplicadas en gran medida en campos como el comercial ya que son útiles para la comprensión de hábitos de compra de los clientes, forman un eje fundamental para la elaboración de ofertas y ventas cruzadas. (Molina López & García Herrero, 2006)

a) **A priori:** Este algoritmo se basa en el conocimiento previo o "a priori" de los conjuntos e datos que aparecen con mayor frecuencia, esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia. El algoritmo Apriori se usa en minería de datos para encontrar reglas de asociación para encontrar correlaciones entre las observaciones de un conjunto de datos. (Velandia Ortega & Hernández Suárez, 2010)

2.4.2.3. Herramientas ETL

La definición de ETL proviene de las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load). El ETL es el proceso que ayuda a una organización a mover los datos desde diferentes fuentes de datos, para realizar el reformateo y limpieza de los datos. Luego cargarlo a otra base datos, data mart o datawarehouse para analizarlo, con el fin de apoyar un proceso de negocio.

2.4.2.3.1. Pentaho Data Integration

Figura 3. Logo de Pentaho
Fuente: Imagen obtenida de (Wikipedia, Pentaho, 2019)

El Pentaho Data Integration conocido como Kettle, es una herramienta perteneciente a la suite de Pentaho que ayuda a realizar el proceso de ETL (Extract-Transform-Load), es decir, es una herramienta que permite realizar la extracción de los datos de una o varias fuentes, para luego proceder a transformar los datos para su posterior carga de datos a otro sitio o destino. El uso de esta herramienta evita el manejo manual de grandes cargas de trabajo constantemente difícil de mantener y de desplegar. (Durán, 2017)

2.4.2.3.2. Knime

Figura 4. Logo de Knime
Fuente: Imagen obtenida de (Wikipedia, Knime, 2019)

Knime conocido con el nombre en inglés (Konstanz Information Miner), es una plataforma analítica que permite manipular los datos, analizarlos, realizar workflows, etc. Está construido sobre la plataforma Eclipse. Knime es una herramienta que incluye ETL y permite realizar minería de datos e integrar con Weka, con Python, con R y definir las clases en java. (jortilles, 2017).

2.4.2.4. Herramientas para Minería de Datos

Se debe realizar un análisis de los sistemas implementados y de los datos para determinar las técnicas de minería de datos que más se adecue y luego, elegir las herramientas que ayude a crear estrategias que contribuyan en el incremento de la eficiencia y estimulación de la innovación con fin de definir argumentos para la toma de decisiones.

La decisión de elegir la herramienta de minería de datos se debe tener en cuenta los siguientes aspectos (García Bermúdez & Acevedo Ramirez, 2010):

- ✓ **Acceso a Datos:** Tener la capacidad de leer diferentes fuentes de datos y formatos de entrada
- ✓ **Selección de Datos:** Debe realizar operaciones con distintos criterios de selección de datos e incluso los filtrados de datos no deseados.
- ✓ **Sensibilidad a la calidad de los datos:** Detectar datos faltantes o incompletos e informar sugerencia que solucione dicho inconveniente.
- ✓ **Visualización de datos:** La presentación de los resultados debe poseer una interfaz gráfica intuitiva y orientado a su excelente usabilidad.
- ✓ **Extensibilidad:** Capaz de integrarse con la administración de almacenamiento de datos y gestión de metadatos.
- ✓ **Rendimientos:** Debe proporcionar en todo momento un rendimiento constante independientemente de la cantidad de datos a extraer, transformar, cargar, procesar o analizar.
- ✓ **Escalabilidad:** Sea capaz de trabajar con grandes cantidades de datos para descubrir patrones significativos, útil y relacionado entre sí.

- ✓ **Apertura:** Debe ser capaz de integrarse con diferentes tipos de herramientas sea a su vez hojas de cálculos, servidor de archivos, servidor de base de datos, compartición en la nube y entorno distribuido durante el procesamiento de análisis de datos.

Las herramientas de minería contienen funcionalidades como lo siguientes (García Bermúdez & Acevedo Ramirez, 2010):

- ✓ **Herramientas Estadísticos:** proporcionan análisis exploratorio para analistas expertos o estadísticos. Al igual que con la programación tradicional, el usuario necesita entender el lenguaje, ya que se necesita una sintaxis específica para poder extraer los datos de interés. Adicionalmente, los usuarios necesitan saber cómo condicionar los datos, así como saber estructurar y administrar las consultas (García Bermúdez & Acevedo Ramirez, 2010).
- ✓ **Herramientas Núcleo de Minería de Datos**: suministran análisis exploratorio a analistas expertos y estadísticos usando un formato gráfico y amigable con el usuario. No se necesita entender un lenguaje o sintaxis. Este tipo de herramientas descubren patrones escondidos, tendencias, relaciones e indicadores predicativos. A pesar de ser muy gráfico el ambiente de trabajo, se requiere saber cómo condicionar los datos y como estructurar y manejar las consultas. Una herramienta núcleo de Minería de Datos, puede proporcionar la capacidad de usar múltiples técnicas, como por ejemplo detección de conglomerados, árboles de decisión y redes neuronales; queda a criterio del minero escoger la técnica que mejor se ajusten a las situaciones del negocio (García Bermúdez & Acevedo Ramirez, 2010).
- ✓ **Herramientas de Consulta**: proporcionan el acceso a los datos detallados, también pueden ser usadas para extraer datos. Estas herramientas son directas, es decir se requiere

que el usuario tenga un muy buen conocimiento de lo que está buscando. Son útiles para desarrollar una idea particular o para probar o invalidar una hipótesis (García Bermúdez & Acevedo Ramirez, 2010).

- ✓ **Herramientas de Visualización de Datos**: muestran los datos gráficamente para mejorar su comprensión, permiten entender grandes cantidades de datos, y datos con complejas relaciones, usando generalmente cubos que muestran jerarquías de dimensiones, aunque no son exactamente herramientas de minería, asisten al minero a visualizar los factores más predictivos en cierta situación, también ayudan a comunicar los resultados que arrojan algoritmos complejos de minería como por ejemplo los de agrupación o clustering, a personas que no tienen conocimientos previos en estadística (García Bermúdez & Acevedo Ramirez, 2010).

2.4.2.4.1. Weka

Figura 5. Logo de Weka
Fuente: Imagen obtenida de (Wikipedia, Pentaho, 2019)

Weka es una herramienta de tipo software para el aprendizaje automático y minería de datos diseñado a base de Java y desarrollado en la universidad de Waikato en Nueva Zelanda en el año 1993, esta herramienta por su nombre en inglés (Waikato Environment for Knowledge Analysis) además es una herramienta de distribución de licencia GNU-GLP o software libre, la cual

dispone de las herramientas necesarias para transformar los datos como por ejemplo: clasificación, regresión, clustering, asociación y visualización. Weka permite añadir nuevas funcionalidades de manera sencilla ya que es una herramienta orientada a la extensibilidad (García Morate, 2011).

2.4.2.4.2. Orange

Figura 6. Logo de Orange
Fuente: Imagen obtenida de (Wikipedia, Orange (Software), 2019)

La herramienta Orange es un programa informático para realizar minería de datos y análisis predictivo, desarrollado en la facultad de informática de la Universidad de Liubliana. Consta de una serie de componentes desarrollados en C++ que implementan algoritmos de minería de datos, así como operaciones de preprocesamiento y representación gráfica de datos. Los componentes de Orange pueden ser manipulados desde programas desarrollados en Python o a través de un entorno gráfico y se distribuye bajo licencia GPL (Licencia Pública General) (Acosta Henríquez, 2015).

2.4.2.4.3. R

Figura 7. Logo de R
Fuente: Imagen obtenida de (Wikipedia, R (programing language), 2019)

R es un entorno estadístico tremendamente potente y completo. Las llamadas a R se realizan en línea de comando, si bien existen algunas interfaces gráficas (Rcommander, etc) que facilitan el uso de este programa. Fue desarrollado inicialmente por el Departamento de Estadística de la Universidad de Auckland, Nueva Zelanda, en 1993. R es un lenguaje de programación y entorno de software de código abierto para computación y gráficos estadísticos. Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásicas, clasificación, agrupación en clústeres, etc (García González, 2013). El entorno de R se caracteriza por su flexibilidad e incluye, entre otros:

- ✓ Un buen gestor de datos.
- ✓ Un conjunto de operadores para cálculos en arrays (vectores de gran tamaño)
- ✓ Un conjunto integrado de herramientas de análisis de datos. 16
- ✓ Funciones gráficas para análisis y visualización de los datos.
- ✓ Un lenguaje de programación simple que incluye condicionales, bucles, funciones recursivas definidas por el usuario y capacidades de entrada y salida. (García Gonzales,2013)
- ✓

2.4.2.4.4. XL Miner

Figura 8. Logo de XL Miner
Fuente: Imagen obtenida de (Provider, n.d.)

XLMiner es un complemento para Excel, con funcionamiento mediante macros, que permite muchos tipos de análisis tanto para datos de tipo corte transversal, como secuencias temporales. Entre las principales características de XLMiner se encuentran (García González, 2013):

- ✓ Manejo de bases de datos, con imputación de datos faltantes.
- ✓ Realización de predicciones.
- ✓ Modelos ARIMA, Holt winters, Polinomiales.
- ✓ Arboles de decisión, análisis clúster.
- ✓ Facilidad para la entrega de informes.
- ✓ Redes neuronales.

2.4.2.4.5. Tanagra

Figura 9. Logo de Tanagra
Fuente: Imagen obtenida de (Wikipedia, Tanagra (machine learning), 2019)

La herramienta Tanagra es un software libre de minería de datos para propósitos académicos y de investigación. Propone varios métodos de minería de datos a partir de análisis exploratorio de

datos, aprendizaje estadístico, aprendizaje automático y base de datos. Provee varios paradigmas de aprendizaje supervisado, agrupamiento, análisis factorial, reglas de asociación, etc (Gambin & Pallotta, 2009).

2.4.2.4.6. Knime

Figura 10. Logo de Knime
Fuente: Imagen obtenida de (Wikipedia, Knime, 2019)

KNIME (Konstanz Information Miner) es una herramienta basada en la colaboración abierta direccionada a integración, procesamiento, análisis y exploración de datos. Esta herramienta fue desarrollada originalmente en el departamento de Bioinformática y Minería de Datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com, radicada en Zúrich, Suiza, continúa su desarrollo, además de prestar servicios de formación y consultoría. KNIME permite a los usuarios crear flujos de datos, ejecutar selectivamente los pasos de análisis, para posteriormente analizar los resultados obtenidos (García González, 2013).

2.4.2.4.7. Rapidminer

Figura 11. Logo de Rapidminer
Fuente: Imagen obtenida de (Rapidminer, 2019)

En el año 2001, en la unidad de Inteligencia Artificial de la Universidad Tecnológica de Dortmund en Alemania, se desarrolló una herramienta de minería de datos denominado “YALE” cuyas siglas significan (Yet Another Learning Envireoment) que luego se cambió el nombre a lo que hoy conocemos como Rapidminer. La herramienta Rapidminer permite diseñar procesos analíticos muy avanzado de aprendizajes automático, minería de datos, minería de texto, análisis predictivo y análisis de negocios (Medina Patrón, Ortiz Servin, Castillo, Montes Tadeo, & Perusquía, 2015). Rapid-i cuenta con dos componentes:

- ✓ RapidMiner: Versión stand-alone para analistas. Implementa todos los operadores de data mining, modelos predictivos, modelos descriptivos, transformación de datos, series de tiempo, etc.
- ✓ RapidAnalytics: Versión Servidor de RapidMiner. Permite trabajo colaborativo, escalable y concurrente de múltiples usuarios, capacidad de delegar en bases de datos (In-Database Mining) y otras mejoras de funcionalidad

RapidMiner ofrece la posibilidad de desarrollar análisis de datos a través del enlace de operadores mediante un entorno gráfico.

Entre las características principales de RapidMiner se encuentran las siguientes:

- ✓ Desarrollado en Java.
- ✓ Multiplataforma.
- ✓ Puede ser utilizado a través de línea de comando, batch.
- ✓ Software de código abierto.
- ✓ Extensible.
- ✓ Incluye gráficos y herramientas de visualización de datos.

- ✓ Ofrece más de 500 operadores para todos los principales procedimientos de máquina de aprendizaje, y también combina esquemas de aprendizaje y evaluadores de atributos del entorno de aprendizaje Weka.
- ✓ Posee de un módulo de integración con la herramienta R. (Garcia Gonzales,2013)

2.4.2.4.8. Tableau

Figura 12. Logo de Tableau
Fuente: Imagen obtenida de (Tableau, 2019)

Tableau Software ofrece productos de inteligencia de negocios y análisis de datos. Fue fundada en 2003 por los pioneros Chris Stolte, Pat Hanrahan y Christian Chabot, con sede en Seatle, Estados Unidos. Tableau es una potente herramienta de cuadros de mandos y análisis, fácil de usar y con la última tecnología. Permite conectarse a los datos y convertirlos en información. Descubrir, analizar e identificar tendencias en segundos, publicar cuadros de mandos y compartirlos dentro de la empresa. (Mora Maqueda & Luque Calvo, 2017)

2.4.3. Validación del modelo predictivo

Para realizar la validación de los modelos predictivos se tomará en consideración las siguientes técnicas.

2.4.3.1. Validación simple

En esta validación se divide el conjunto de datos en dos subconjuntos el uno para entrenamiento y el otro para prueba, el conjunto de prueba no se debe utilizar en el proceso de entrenamiento.

2.4.3.2. Validación cruzada

Esta validación se utiliza cuando se dispone de pocos datos, en este caso el conjunto de datos se divide aleatoriamente en dos subconjuntos del mismo tamaño A y B, el procedimiento es:

- ✓ Entrenar el modelo con los datos del subconjunto A y posterior validarlos con los datos del subconjunto B, calcular el error
- ✓ Entrenar el modelo con los datos del subconjunto B, y validarlos con los datos del subconjunto A, calcular el error
- ✓ Finalmente entrenar el modelo con ambos subconjuntos tanto A como B, y se usa el modelo con el menor error

2.4.3.3. Validación cruzada con n pliegues

Esta validación consiste en la división del conjunto en n subgrupos, utilizando uno para el entrenamiento y n-1 para las pruebas, se calcula el error y se repite el proceso n veces cambiando el conjunto de entrenamiento.

2.4.3.4. Validación bootstrapping

El bootstrapping es un método de remuestreo propuesto por Bradley Efron en 1979. El objetivo del bootstrap es tratar las muestras como si fuera la población y extraer con reposición un gran número de remuestras de tamaño n. El remuestreo con reposición puede incluir datos

originales más de una vez, por lo que cada remuestra será diferente a la muestra original (Miranda Moles, 2003).

2.4.4. Evaluación del modelo predictivo

En la evaluación del modelo predictivo se considerará los siguientes test o pruebas:

2.4.4.1. Test de Back Testing

El objetivo es comprobar el ajuste y la consistencia del modelo. La prueba de Back Testing es modelar con una pequeña cantidad de datos que no han sido utilizado para comprobar si los resultados predichos por el modelo de medición de riesgo crediticio se ajustan a los resultados reales (Ochoa P., Galeano M., & Agudelo V., 2010).

2.4.4.2. Test Kolmogorov-Smirnov

La prueba de Kolmogórov-Smirnov es conocido también como prueba K-S la cual es utilizada para verificar diferencias significativas entre la distribución de la frecuencia observada y la distribución de la frecuencia esperada (Kishinani, 2016).

2.4.4.3. Curva ROC

La curva de ROC (Receiver Operanting Characteristic) mide la relación de la tasa de verdadero positivo (predicciones acertadas) con respecto a la tase de falsos positivos (predicciones erradas). La curva de ROC posee una métrica llamada AUC (Area Under the curve) por lo que se define el Área bajo la curva del ROC (Barrientos & Ríos, 2013).

2.4.4.4. Test de Hosmer-Lemeshow

Es un método para estudiar la bondad de ajuste del modelo de regresión logística, consiste en comparar los valores esperados por el modelo con los valores encontrados. Las distribuciones esperada y observada son contrastadas mediante la prueba de x^2.

La hipótesis nula del test de Hosmer-Lemeshow se basa en que no existe ninguna diferencia entre los valores observados y esperados, en caso de existir diferencias se estaría indicando que el modelo no está ajustado (Universidade de Santiago de Compostela, 2012).

2.4.5. Metodología

La Metodología es la ciencia que nos enseña a dirigir determinado proceso de manera eficiente y eficaz para alcanzar los resultados deseados y tiene como objetivo darnos la estrategia a seguir en el proceso (Cortés Cortés & Iglesias León, 2004).

2.4.5.1. Metodología de investigación

La Metodología de la Investigación (M.I.) o Metodología de la Investigación Científica es aquella ciencia que provee al investigador de una serie de conceptos, principios y leyes que le permiten encauzar de un modo eficiente y tendiente a la excelencia el proceso de la investigación científica. El objeto de estudio de la M.I. Lo podemos definir como el proceso de Investigación Científica, el cual está conformado por toda una serie de pasos lógicamente estructurados y relacionados entre sí. Este estudio se hace sobre la base de un conjunto de características y de sus relaciones y leyes (Cortés Cortés & Iglesias León, 2004).

2.4.5.1.1. Metodología Design Science Research (DSR)

El diseño significa “inventar y hacer realidad”. Al diseñar se trata de crear nuevo artefacto (cosas o procesos que tienen o pueden ser un material existente) que no existe. Si el conocimiento requerido para crear un artefacto de este tipo ya existe, entonces el diseño es rutina, de lo contrario es innovador. El diseño innovador puede requerir la realización de investigaciones (investigación científica del diseño) para llenar los vacíos de conocimiento y puede dar lugar a publicaciones de investigación o patentes (Vaishnavi, Kuechler, & Petter, 2017).

La investigación en ciencias del diseño implica la creación de nuevos conocimientos a través de diseño de artefactos innovadores (cosas o procesos que tienen o pueden ser un material existente) y el análisis de uso y/o rendimiento de tales artefactos junto con la reflexión y la abstracción para mejorar y comprender el comportamiento de los aspectos de los sistemas de información (SI). Los artefactos pueden ser: algoritmos, humanos, computadoras, interfaces y metodologías de diseño de sistemas o lenguajes. Los investigadores de la ciencia del diseño se encuentran en muchas disciplinas y campo, como es el caso de la ingeniería y la informática, en la cual usan variedades de enfoques, métodos y técnicas (Vaishnavi & Kuechler, Design Science rersearch in Information System, 2013).

El propósito de esta metodología es hacer visible y útil para los investigadores en el campo de la ingeniería. El ciclo del diseño comienza con el conocimiento del problema y un análisis del problema que puede ser difícil y complejo. Luego se plantea el desarrollo de un modelo para su posterior evaluación. La evaluación proporciona información y una mejor comprensión del problema para mejorar tanto en la calidad del producto y el proceso de diseño. La evaluación del modelo no es el último paso, ya que el método es iterativo. El proceso iterativo brinda la

oportunidad de refinar los modelos. Por ejemplo, al diseñar un auto, el prototipo no es el último paso, siempre hay una nueva versión del diseño del automóvil desde el momento en que un prototipo está listo para su lanzamiento (Carstensen & Bernhard, 2015).

La metodología de investigación Design Science Research(DSR) es conocida también como Investigación científica basada en el diseño.

El modelo DSR de Peffers et al., fue publicado en el año 2006 y actualizado en el año del 2007 por Peffer, Tuunanen, Rothenberger y Chantterjee (Peffers, Tuunanen, Rothernberger, & Chatterjee, 2007)

El modelo DSR de Peffers que se muestra en la Figura 13. consta de 6 fases que son:

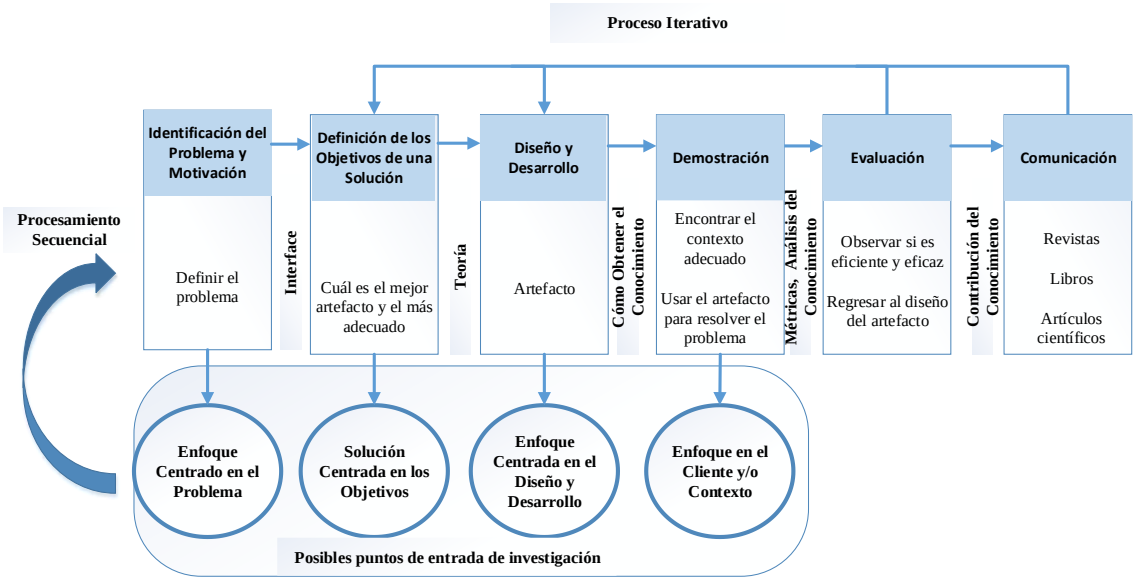


Figura 13. Modelo Design Science Research (DSR).
Fuente: Elaborado por las autoras con base en (Peffers, Tuunanen, Rothernberger, & Chatterjee, 2007)

a) **Identificación de problemas y motivación:** En esta fase se trata de definir la investigación de un problema específico y justificar el valor de dicha solución. Dado que la definición del problema será utilizada para desarrollar un artefacto que pueda

proporcionar una solución de manera efectiva y útil para atomizar el problema en forma conceptual para que la solución pueda capturar su complejidad. Al justificar el valor de una solución se logra dos cosas: motiva al investigador y al público de la investigación a buscar la solución. y para aceptar los resultados y la ayuda para entender el razonamiento asociado con la comprensión del investigador del problema. Los recursos requeridos para esta actividad será incluir el conocimiento del estado del problema y la importancia de la solución (Peffers & Tuunanen, 2008).

b) **Definir los objetivos para una solución:** La idea principal es inferir en los objetivos de una solución. Los objetivos pueden ser cuantitativos, por ejemplo, los términos en los que una solución deseable sería mejores que los actuales, o cualitativos, por ejemplo, una descripción de cómo es un nuevo artefacto Se espera que apoye soluciones a problemas que hasta ahora no se han abordado. Los objetivos deben inferirse racionalmente de la especificación del problema. Los recursos requeridos para esto incluyen el conocimiento del estado de los problemas y las soluciones actuales, si las hay, y su eficacia (Peffers & Tuunanen, 2008).

c) **Diseño y desarrollo:** El objetivo es crear el artefacto. Tales artefactos son potencialmente construcciones, modelos, métodos o ejemplificaciones (cada una definida ampliamente "nuevas propiedades de recursos técnicos, sociales y / o informativos". Conceptualmente, un artefacto de investigación de diseño puede ser cualquier objeto diseñado en el que la contribución de la investigación está incrustada en el diseño. Esta actividad o fase incluye la de determinar la funcionalidad deseada del artefacto y su arquitectura y luego crear el artefacto real. Los recursos necesarios para pasar de los objetivos al diseño y luego al

desarrollo será incluir el conocimiento de la teoría que se puede aplicar en una solución (Peffers & Tuunanen, 2008).

d) **Demostración:** Demostrar el uso del artefacto para resolver uno o más casos del problema. Esto podría implicar su uso en la experimentación, simulación, estudio de caso, prueba u otra actividad apropiada. Los recursos requeridos para la demostración incluyen un conocimiento efectivo de cómo usar el artefacto para resolver. el problema (Peffers & Tuunanen, 2008).

e) **Evaluación:** La idea es observar y medir qué tan bien soporta el artefacto en una solución al problema. Esta actividad consiste en comparar los objetivos de una solución a los resultados reales observados del uso del artefacto en la demostración. Eso requiere conocimiento de métricas relevantes y técnicas de análisis. Dependiendo de la naturaleza del lugar problemático y el artefacto, la evaluación podría tomar muchas formas. Eso podría incluir elementos como una comparación de la funcionalidad del artefacto con el objetivo de la solución de la actividad o fase dos, desempeño cuantitativo objetivo y medidas como los presupuestos o artículos producidos, los resultados de las encuestas de satisfacción, comentarios de clientes, o simulaciones. También podría incluir medidas cuantificables de sistema como es el rendimiento del tiempo de respuesta o disponibilidad. Conceptualmente, dicha evaluación podría incluir cualquier evidencia empírica apropiada o prueba lógica. Al final de esta actividad, los investigadores pueden decidir si repetir el paso tres para probar, para mejorar la efectividad del artefacto o para continuar con la comunicación (Peffers & Tuunanen, 2008).

f) **Comunicación:** La idea es comunicar el problema y su importancia, el artefacto, su utilidad y novedad, el rigor de su diseño y su eficacia para investigadores. En

publicaciones de investigación académica, los investigadores podrían usar la estructura de este proceso para estructurar el papel, al igual que la estructura nominal de un empírico proceso de investigación (definición del problema, revisión de la literatura, desarrollo de hipótesis, recopilación de datos, análisis, resultados, discusión y conclusión) es una estructura para trabajos de investigación empírica. La comunicación requiere conocimiento de la cultura disciplinaria (Peffers & Tuunanen, 2008).

2.4.5.2. Metodología de Minería de Datos

Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos (Miguel Moine , Haedo , & Gordillo).

2.4.5.2.1. Metodología Crisp DM

Cross-Industry Standard Process for Data Mining (CRISPDM)

Esta metodología involucra una guía constituida por **seis fases**, algunas de estas fases permitirán examinar parcial o totalmente las fases previas. (Goicochea A, 2011).

Las fases de esta metodología se muestran en la siguiente gráfica:

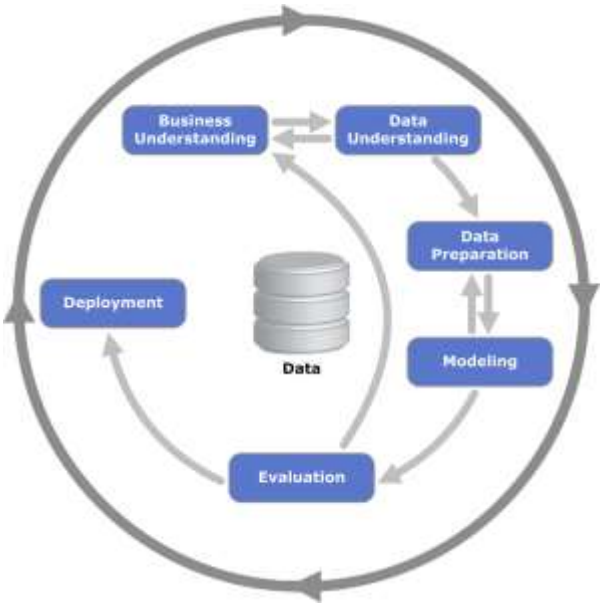


***Figura 14*:** Fases del Crisp DM
Fuente: Imagen obtenida de (Wikipedia, 2018)

a) **Fase 1. Comprensión del Negocio:** La metodología indica como fase primera a la comprensión del negocio, esta es probablemente la más importante y contiene en compendio las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional según sea el caso, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Para obtener el mejor provecho de Data Mining, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Data Mining y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. (Gallardo Arancibia,J A)
b) **Fase 2. Comprensión de los datos:** La segunda fase es Comprensión de los datos, ésta comprende la recolección inicial de datos con el objeto de familiarizarse con ellos,

identificar la calidad de los mismos y establecer las relaciones más visibles y evidentes y definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas. (Gallardo Arancibia,J A)**.**

c) **Fase 3 Preparación de los datos:** Una vez efectuada 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, técnicas de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. Esta fase incluye tareas de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza, generación de variables adicionales, integración de diferentes fuentes y cambios de formato. Esta fase interactúa de forma permanente con la fase de modelado, puesto que, en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas.

   Una descripción de las tareas involucradas en esta fase es la siguiente:

   - ✓ Selección de datos. En esta etapa, se escoge un grupo de los datos recolectados en la fase anterior, basándose en criterios establecidos con antelación en las primeras fases como son: calidad de los datos en cuanto a completitud y corrección y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas.

- ✓ Limpieza de los datos. Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: reducción del volumen de datos, normalización, discretización, tratamiento de valores ausentes, etc.
- ✓ Estructuración de los datos. Este paso involucra operaciones de preparación de datos como: generación de nuevos atributos, integración de nuevos registros, transformación de valores.
- ✓ La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.
- ✓ Formateo de los datos. Esta tarea consiste, en la transformación de los datos de acuerdo a la necesidad y sin llegar a alterar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo 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.). (Gallardo Arancibia, J A)

d) Fase 4. Modelado: En esta fase se requiere la elección de las técnicas de modelado más optimas y apropiadas de acuerdo al proyecto de minería a desarrollar. Las técnicas que se utilicen deben ser seleccionadas de acuerdo al cumplimiento de los siguientes supuestos:

- ✓ Apropiada para el problema.
- ✓ Tiempo adecuado para la obtención un modelo.
- ✓ Cumplimiento de los requisitos del problema.
- ✓ Poseer datos apropiados.

Los parámetros que se ocupen en la creación del modelo de minería, dependen directamente tanto de las características de los datos como de la precisión que se proponga alcanzar.

Una descripción de las principales tareas de esta fase es la siguiente:

- ✓ Selección de la técnica de modelado. Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos (CBR); si el problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.
- ✓ Generación del plan de prueba. En este paso es necesario generar un procedimiento que permita comprobar la validez y efectividad del modelo construido en las fases previas. Para lo cual generalmente, se dividen los datos en dos grupos, de los cuales uno es destinado para entrenamiento y el otro para prueba, Seguidamente se construye el modelo basándose en el grupo de entrenamiento y se realiza la medición de la calidad y efectividad utilizando el grupo de prueba.
- ✓ Construcción del Modelo. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de

modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

- ✓ Evaluación del modelo. En esta fase se interpretan los modelos de minería a razón del conocimiento y experticia sobre los criterios de éxito definidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc..). (Gallardo Arancibia, J A)

e) Fase 5. Evaluación: En esta fase se realiza la evaluación del modelo de minería de datos, enfocándose en el cumplimiento de los criterios de éxito preestablecidos.

Los pasos o tareas a realizarse en esta fase son:

- ✓ Evaluación de los resultados. En este paso se evalúa el modelo de minería de datos desarrollado previamente, la verificación es realizada en función de los objetivos del negocio, con el fin de determinar razones de negocio para las cuales el modelo sea insuficiente.
- ✓ Proceso de revisión. En este paso, se califica a todo el proceso de minería de datos, con la finalidad de encontrar elementos a los que se les pudiera proveer una mejora.
- ✓ Determinación de futuras fases. En el caso de determinarse que los pasos realizados previamente generaron resultados óptimos es posible avanzar a la siguiente fase, caso contrario se puede proceder a la selección de otra iteración en la fase de preparación de datos o al modelado con otros parámetros. (Gallardo Arancibia, J A)

f) **Fase 6. Implementación o Despliegue:** Posterior a que el modelo ha sido elaborado y verificado, se convierte el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como, por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc.

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

- ✓ Plan de implementación. Para implementar el resultado de DM en la organización, 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.
- ✓ Monitorización y Mantenimiento. Si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos.
- ✓ Informe Final. Es la conclusión del proyecto de minería de datos realizado. Dependiendo del plan de implementación, este informe puede constituirse por un resumen de los puntos importantes del proyecto juntamente con la experiencia adquirida, como también puede ser una presentación incluyendo y explicando los resultados alcanzados con el desarrollo del proyecto.
- ✓ Revisión del proyecto: En este punto se realiza un proceso de evaluación de lo correcto e incorrecto, lo qué es lo que se hizo bien y de lo qué es posible optimizar.

2.5. Categorización de las Variables de Investigación

El establecimiento de la fundamentación teórica tiene como finalidad la búsqueda de la congruencia teórica que fue determinada con la hipótesis, de ahí radica la importancia de crear una red de categorías desde lo general a lo específico la cual permitirá realizar un desarrollo teórico ordenado donde estarán incluidas las variables del problema. Con lo anteriormente indicado, el objetivo es llegar a la categoría que permita tener entendimiento y, que explique las variables dependientes e independientes del tema de estudio; por consiguiente, se plantea la siguiente red jerarquizada del estudio en la Figura 15.

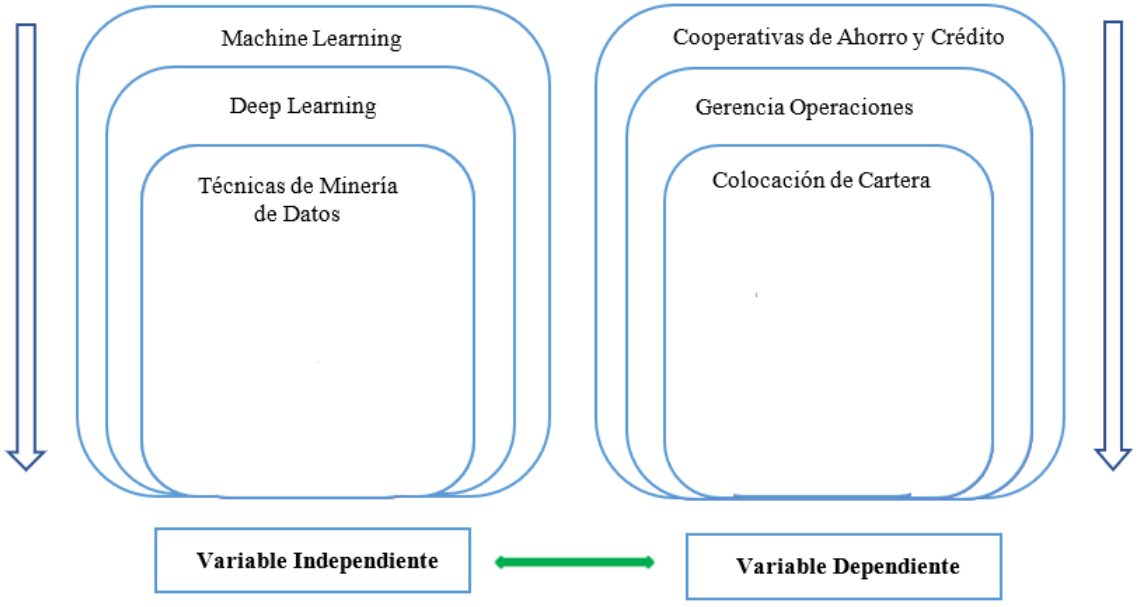


Figura 15. Determinación de Variables

2.5.1. Fundamentación de la Variable Independiente

2.5.1.1. Machine Learning

Conjunto de algoritmos que están diseñados especialmente para abordar problemas de reconocimiento de patrones en conjunto de datos extremadamente grandes. Estas técnicas incluyen una función de base real, clasificadores basados en árboles y máquinas de vectores de soporte, las cuales son ideales para el análisis de riesgo crediticio en base a los grandes tamaños de muestra y por la complejidad de las posibles relaciones entre las transacciones y las características de los clientes. (Khandani, Kim, & Lo, 2010)

2.5.1.2. Deep Learning

Aprende de grandes cantidades de datos no supervisados y, extrae representaciones y patrones significativos de grandes cantidades de datos a través de un proceso de aprendizaje jerárquico. Uno de los problemas de predicción en las finanzas se tiene en la gestión de riesgos por cuanto este tipo de dificultades incluyen grandes conjuntos de data con relaciones complejas entre datos y eventos. Los métodos de aprendizaje profundo pueden representar relaciones complejas entre datos con resultados más útiles que los métodos tradicionales en finanzas. (Hasan & Kalipsiz, 2017)

2.5.1.3. Técnicas de Minería de Datos

La minería de datos y el descubrimiento de conocimiento en bases de datos (KDD – Knowledge Discovery in Database) atraen una gran cantidad de investigación y de atención de los medios durante los últimos tiempos. La extracción de datos y el descubrimiento de conocimiento en la base de datos están relacionados entre sí y con campos relacionados como

aprendizaje automático, estadísticas y bases de datos. La minería de datos usa métodos robustos que ayudan a la reducción de los costos y riesgos para un negocio, de igual manera apoya el incremento de las rentas por cuanto extrae información crucial y estratégica a través de los datos disponibles. (Fayyad, Piatetsky, & Smyth, 1996)

2.5.2. Fundamentación de la Variable Dependiente

2.5.2.1. Cooperativas de Ahorro y Crédito

De acuerdo a los datos estadísticos de la Superintendencia de Economía Popular y Solidaria con corte a octubre 2015, Ecuador cuenta con 887 cooperativas de Ahorro y Crédito, lo cual lo constituyó en el segundo país a nivel latinoamericano, después de Brasil que sumaron 4.700.000 socios, llegando a alcanzar activos de 8.300 millones de dólares. El superintendente Hugo Jácome aseguró que el 66% del microcrédito otorgado en el país proviene del sistema cooperativo, manifestó también que en el país se ha venido elaborando una propuesta clara y oportuna para lograr el fortalecimiento y crecimiento del sector cooperativo basándose en objetivos base como son: el cumplimiento de solvencia, incremento de la gestión integral de riesgos, fortalecimiento patrimonial, gestión de riesgo de crédito. (Solidaria, s.f.)

2.5.2.2. Gerencia Operaciones

La finalidad de la administración de Operaciones; además de precalificar y evaluar el riesgo creditico para colocaciones de cartera para los socios, también, analiza y cuantifica los riesgos de sus operaciones. Para el tema de estudio, se enfocará parte de los análisis en los índices de morosidad que se originan de los socios que ya se les han otorgado créditos a través de la cooperativa.

2.5.2.3. Colocación de Cartera

Las empresas para mantenerse y crecer en el mercado, deben seguir ofertando sus productos a sus clientes que son aquellos que tienen un producto financiero vigente o aquellos que ya han adquirido anteriormente un producto. Dentro de este contexto, las bases de campañas deben tener como finalidad la fidelización de los antiguos clientes; por lo que, debe ofertar un producto financiero crediticio la cual debe ser ágil, oportuno y alineados a los tiempos de mercado, conjuntamente, debe tener un valor agregado en su oferta que le diferencie de sus competidores.

2.6. Trabajos Relacionados

2.6.1. Estado del arte

El presente estudio del arte responde a las actividades correspondientes a una Revisión Inicial de Literatura inspirada en las guías de revisiones sistemáticas de literatura propuestas por KITCHENHAM & OTROS, siendo el objetivo del presente estudio del arte resolver las preguntas de investigación, se realizó una búsqueda en la base digital IEEE acerca de la temática planteada, para ello se definieron dentro de los criterios de inclusión artículos científicos superiores al año 2010, publicados en el idioma inglés, referentes a técnicas de minería de datos aplicadas el análisis de riesgo crediticio en entidades financieras, dentro de los criterios de exclusión se señaló que los artículos no estén enfocados hacia tarjetas de crédito, no sean artículos empíricos y no incluyan las palabras “transmisión de datos”.

Se verifica que los estudios cumplan con los criterios de inclusión y exclusión, revisando títulos, resúmenes, conclusiones y palabras claves, con lo cual se obtienen el listado inicial de documentos académicos que conforman los estudios del grupo de control, mismos que se detallan en la tabla 2

Tabla 2.
Estudios de Candidatos

Estudios	Titulo	Palabras Clave
EC1	Credit Risk Assessment for Rural Credit Cooperatives based on Improved Neural network	Credit risk assessment Neural Network
EC2	Credit risk assessment based on Neural Network	Credit Risk, bank, neural network
EC3	A Machine Learning Approach for Predicting Bank Credit Worthiness	Machine Learning, Bank Credit
EC4	Prediction Analysis of Risky Credit Using Data Mining Classification Models	Data mining, credit, prediction analysis

2.6.2. Construcción de la cadena de búsqueda

En la construcción de la cadena de búsqueda se utilizaron las palabras que más se repitieron en cada contexto definido a partir de los estudios del grupo de control, se encontraron palabras comunes entre estudios y palabras propias direccionadas al objetivo de la presente investigación, para lo cual se formaron los siguientes contextos: financial, neural network, predictive. Los resultados del proceso de búsqueda mencionado pueden ser apreciados en la Tabla 3 siguiente.

Tabla 3.
Palabras recurrentes en grupos de control

Contexto	Palabra Clave	EC1	EC2	EC3	EC4	Cantidad Palabras Repetidas
FINANCIAL	Finance		x		x	2
	Credit risk	x	x	x	x	4
	Assesment	x	x	x	x	4
	Bank	x	x	x	x	4
	Promotional campaigns					0
MACHINE LEARNING	Data mining	x	x			2
	Neural network	x	x	x	x	4
	learning	x	x		x	3
PREDICTIVE	Forecasting		x			1
	Predictive analysis	x		x		2
	Time series	x				1

La cadena de búsqueda se formó con la combinación de las palabras que más se repiten en cada contexto, utilizando conectores como AND, OR, la misma que se aplicó en IEEE en búsquedas de título y contenido con el objetivo de encontrar un contexto adicional de métodos de solución del problema planteado, se realizaron varios intentos de cadena de búsqueda con diferente número de artículos resultantes tal como se puede observar en la tabla 4 siguiente:

Tabla 4.
Determinación de la Cadenas de Búsqueda

Cod	Cadena de búsqueda propuesta	Artículos encontrados	Artículos Candidatos encontrados
CB1	(((((Credit risk) AND bank loan) AND Bank) NOT credit card) AND data mining)	202	3
CB2	((((Credit risk) AND bank) NOT credit card) AND (data mining OR neural network))	53	3
CB3	((((Credit risk) AND bank) NOT credit card) AND data mining)	24	2
CB4	((((Credit risk) AND bank) NOT credit card) AND (data mining techniques OR neural network) NOT EMPIRICAL)	35	4

CONTINÚA

CB5	((((Credit risk) AND bank credit) NOT credit card) AND (data mining techniques OR neural network model) NOT EMPIRICAL)	13	4
CB6	((((Credit risk) AND bank credit) NOT credit card) AND (data mining techniques OR neural network model) NOT EMPIRICAL NOT STREAMING DATA)	8	3

Determinando que de las cadenas propuestas la que devolvió una cantidad de resultados manejable, como también la mayoría de los estudios candidatos es:

((((Credit risk) AND bank credit) NOT credit card) AND (data mining techniques OR neural network model) NOT EMPIRICAL NOT STREAMING DATA)

Una vez obtenidos los resultados se aplicó un filtro en donde se seleccionaron únicamente artículos publicados en el idioma inglés, y los artículos que fueron publicados a partir del año 2016 quedando una selección de 5 artículos que conformarían el grupo de control, mismos que se listan en la tabla 5.

Tabla 5.

Estudios del Grupo de Control

Estudios	Titulo	Año	Autor
GC1	Credit Risk Assessment for Rural Credit Cooperatives Based on Improved Neural Network	2017	Li Changjian ; Hu Peng
GC2	Prediction analysis of risky credit using Data mining classification models	2017	Archana Gahlaut ; Tushar ; Prince Kumar Singh
GC3	SR-based binary classification in credit scoring	2017	Pornwatthana Wongchinsri ; Werasak Kuratach
GC4	Application of BP neural network optimization algorithm based on genetic algorithm in credit risk early-warning of commercial bank	2017	Jie Su ; Ya-Ning Zhang

CONTINÚA

GC5	A machine learning approach for predicting bank credit worthiness	2016	ReginaEsi Turkson; Edward Yeallakuor Baagyere ; Gideon Evans Wenya

Se realizó la revisión de los documentos encontrados de lo cual se extrajo la información más relevante.

(LI & Hu, 2017) Credit Risk Assessment for Rural Credit Cooperatives based on Improved Neural Network

El presente artículo se centra en el análisis de riesgo en las cooperativa rurales en China, aplica el método de redes neuronales para solventar el análisis crediticio clasificando los índices en cinco categorías: rentabilidad, índice de flujo de caja, capacidad de operación, índice de liquidez , capacidad de pago; eligieron 12 indicadores del sistema de índice de evaluación para el análisis empírico, apareciendo 160 muestras de entrenamiento y 40 muestras de prueba, utilizaron Matlab para la visualización de los datos.

Al final realizaron experimentos para verificar el cumplimiento del modelo creado a partir de redes neuronales ya que el modelo puede proporcionar una referencia científica para la política de crédito en las cooperativas rurales

(Gahlaut,Tushar,Singh, 2017) Prediction analysis of risky credit using Data mining classification models

En este trabajo, se presenta un sistema de predicción de crédito con el objeto de ayudar a las organizaciones bancarias a tomar la decisión correcta con respecto a la aprobación o rechazo de la solicitud de préstamo de un cliente en función de sus antecedentes familiares, ocupación, situación financiera, estado civil y otros factores. Anteponen que entre todos los factores la edad, la duración y monto son los factores más importantes que pueden afectar a cualquier persona financieramente.

Los algoritmos como el árbol de decisión, Regresión Lineal, Random Forest y de redes neuronales se utilizan para construir modelos predictivos, para predecir y clasificar la solicitud de préstamo como bueno o malo, esto es debido a que los algoritmos analizan el comportamiento de los clientes y sus capacidades o el alcance para pagar el crédito. Concluyeron que el mejor algoritmo para la clasificación de crédito riesgoso es el algoritmo Random Forest, ya que aseguran tiene una alta precisión a pesar de ser más lento en tiempo de ejecución para grandes bases de datos dimensionales, no se degrada en el rendimiento y la precisión.

(Wongchinsri & Werasak ,2017) SR-based binary classification in credit scoring

Identifica claramente que existen variados métodos de evaluación de riesgo de crédito, denotando que un método puede ser beneficioso en ciertos casos y perjudicial en otros, describe el “Método crítico” el cual no requiere de cálculos matemáticos si no que utiliza el pensamiento crítico del asesor y entrevistas, explica que dicho método es ineficiente ya que esta sesgada a la experiencia del gestor. Analiza el método “Puntuación de crédito” indicando que es un procedimiento de modelo estadístico que permite la predicción del comportamiento futuro basado

en la evaluación de resultados anteriores, identifica que los procedimientos más populares para este proceso son: análisis discriminante, modelo logit y redes neuronales, El análisis desarrollado por los autores concluyo que el análisis discriminante y regresión logística implican más desventajas que las redes neuronales, ya que este último es un método con menos limitaciones y restricciones. Aunque las redes neuronales no siempre superan a los otros métodos.

(Su & Zhang, 2017) Application of BP neural network optimization algorithm based on genetic algorithm in credit risk early-warning of commercial bank

Identifica que para hacer frente a la competencia hay que asumir nuevos retos utilizando herramientas matemáticas y estadísticas cada vez más sofisticada para supervisar el rendimiento del cliente incluyendo la solvencia, riesgo de impago, riesgo de reembolso anticipado, probabilidad de deserción, segmentación de mercado entre otros. Analiza a las redes neuronales como un modelo complejo altamente eficaz, sin embargo, la complejidad indica que se desafía a la interpretación sencilla, siendo estos modelos mejores que los modelos lineales simples

(Turkson, Baagyere, Wenya, 2016) A machine learning approach for predicting bank credit worthiness

El riesgo crediticio es uno de los mayores desafíos que enfrenan las entidades financieras en China, la red neuronal BP tiene la capacidad de autoaprendizaje, adaptación y tolerancia a fallos, en el trabajo se adopta la red neuronal BP para evaluar el riesgo de crédito de empresas, el método de evaluación del riesgo de crédito que utiliza la red neural BP denota algunas carencias

y deficiencias, que muestran en los siguientes dos aspectos. Uno, BP modelo de red neuronal requieren una cierta cantidad de muestras de aprendizaje, la cantidad y la calidad de las muestras de aprendizaje tienen gran influencia en el modelo de red neuronal y los resultados finales de la evaluación.

2.6.3. Extracción de datos

En conclusión, posterior a realizar un estudio de literatura de los artículos relacionados al tema de investigación, se identifica que todos los autores coinciden en la importancia del análisis de riesgo crediticio como una de los factores prioritarios en toda institución financiera, en varios de los artículos hacen análisis sobre los métodos para llevar a cabo un control y análisis crediticio previo al otorgamiento de créditos, apareciendo métodos convencionales y métodos estadísticos , llegan a definir que las entidades financieras aplican diferentes métodos para realizar los análisis de clientes entre ellos redes neuronales, regresión logística, análisis discriminante, entre otros, siendo las más recomendadas las técnicas estadísticas ya que poseen mayor exactitud en sus resultados, para llegar a realizar el proceso de análisis predictivo se analizan varios indicadores como son rentabilidad, índice de flujo de caja, capacidad de operación, índice de liquidez , capacidad de pago.

CAPITULO III

DISEÑO DE LA INVESTIGACIÓN

3.1. Categoría de investigación

En el proyecto se realizará una **investigación no experimental**, por motivo de que las variables no son manipuladas, es decir, el objetivo es solo observar los fenómenos tal como se dan en el contexto natural para luego analizarlo. Dentro de la investigación no experimental seleccionaremos el **diseño transaccional o transversal** para recolectar datos de un tiempo único, cuyo propósito es describir las variables y analizar su incidencia e interrelación en un momento dado. Utilizaremos el **diseño transaccional descriptivo** que consiste en indagar la incidencia de las modalidades, categorías o niveles de una o más variables en una población, por ejemplo: Edad(años) y el estado civil (Soltero(a), Divorciado(a), Separado(a), viudo(a), unión libre) y el resultado es la descripción de cuantos solteros o casados conforman ciertas edades. También se elegirá el **diseño transaccional correlacional-causal** para relacionar dos o más categorías, conceptos o variables en un momento determinando (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014).

Al utilizar el diseño transaccional correlacional determinamos el grado de relación que existe entre las variables **independiente técnicas de minería de datos (X)** y la variable **dependiente colocación de cartera (Y).** El diseño se muestra gráficamente en la figura 16

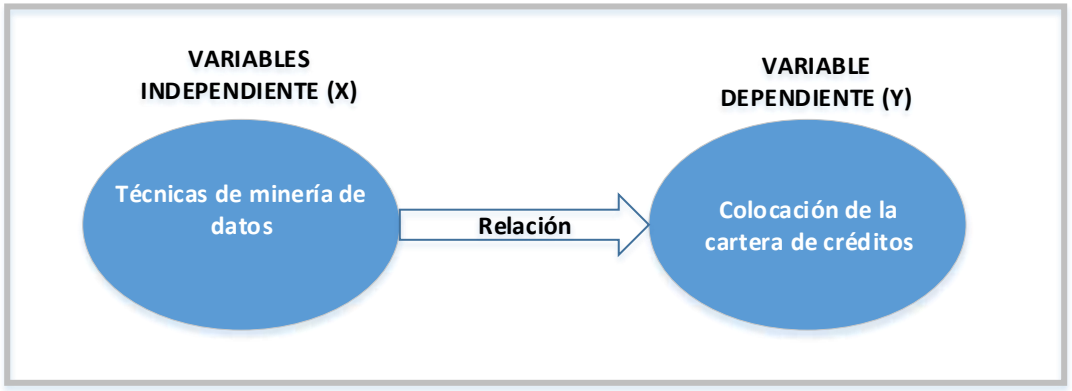


Figura 16. Relación entre la variable independiente y la variable dependiente

3.1.1. Enfoque de investigación

En la investigación se utilizará el enfoque mixto (cuantitativo y cualitativo).

Se eligió el enfoque cuantitativo por las siguientes características (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014):

- ✓ Predicciones iniciales(hipótesis) y de estudios previos(teoría)
- ✓ La recolección de los datos está fundamentada en la medición (medir las variables o conceptos contenido en la hipótesis)
- ✓ Los datos son productos de las mediciones, en la cual son representados en números(cantidades) y se deben analizar con métodos estadísticos.
- ✓ Generalizar los resultados encontrado en un segmento de un universo o población, que luego permita que los estudios puedan replicarse.
- ✓ Se pretende predecir los fenómenos investigados, buscando las relaciones causales entre los elementos, cuya meta es la formulación y las demostraciones de teorías

Se eligió el enfoque cualitativo por las siguientes características (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014):

- ✓ Se basa en una lógica y proceso inductivo (explorar y describir, y luego generar perspectiva teórica). Ir de lo particular a lo general, por ejemplo, una entrevista.
- ✓ La recolección de datos consiste en obtener perspectiva y punto de vista de los participantes (tendencias), es decir, comprender a las personas, procesos, eventos y sus contextos.

Enfoque de investigación aplicado al proyecto de investigación

El enfoque es cuantitativo porque se basa en el estudio de los datos y las variables que caracterizan los solicitantes de créditos, con el fin de obtener una metodología de clasificación como buen o mal cliente en función de la probabilidad de cumplimiento con la obligación crediticia contraída. Este estudio explora el rendimiento de los modelos de calificación de crédito utilizando técnicas de minería de datos: análisis de discriminantes, regresión logística, redes neuronales y árboles de clasificación y regresión.

El enfoque es cualitativo porque se basa en identificar el sector al cual pertenece la actividad de la empresa o negocio del cliente que solicita el crédito y se investigue sobre su historia, la situación actual, problemas y perspectivas del mismo. Además, es necesario tener información general de la empresa, productos o servicios ofrecidos, área de influencia, principales clientes, principales proveedores y posicionamiento en el mercado.

3.2. Metodología de investigación

La metodología de investigación en Ciencias del Diseño (DSR) fue elegida por su capacidad para estudiar la conexión entre la investigación y las prácticas profesionales mediante el diseño,

la implementación y la evaluación de artefactos (modelo de medición de riesgo crediticio) que abordan las necesidades específicas. Es una metodología que propone los artefactos para resolver problemas, evaluar lo que esta proyecta o lo que está funcionando y comunica los resultados obtenidos. En la figura 17 se muestra las fases del DSR.

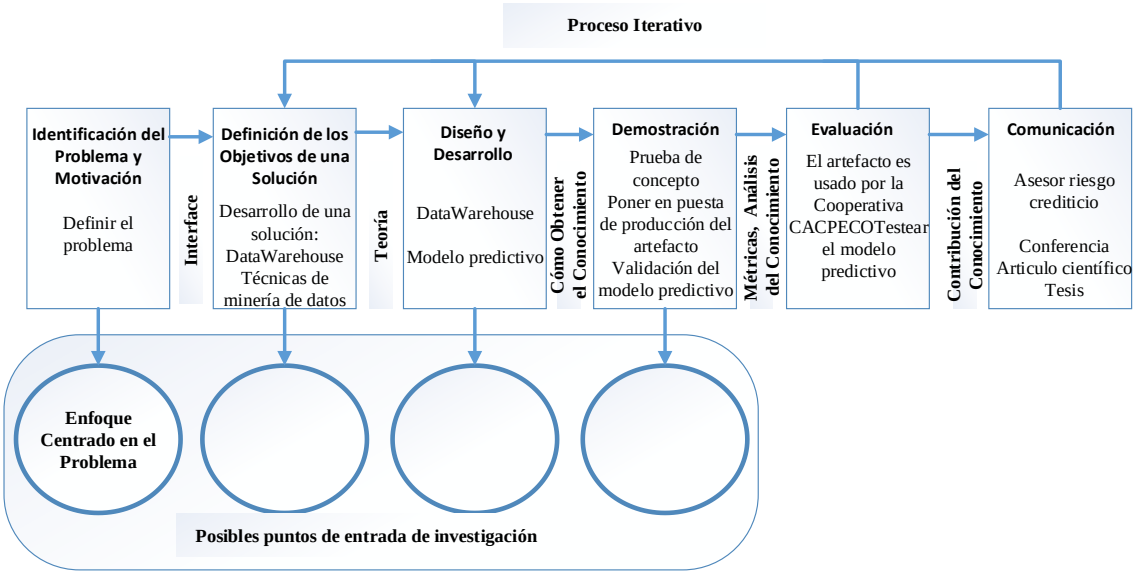


Figura 17. Modelo Design Research (DSR) aplicado a Minería de datos
Fuente: Elaborado por las autoras con base en (Peffers, Tuunanen, Rothernberger, & Chatterjee, 2007)

3.2.1. Pasos o fases

Las fases del DSR son los siguientes:

3.2.1.1. Fase 1: Identificación del problema y motivación

El objetivo en esta fase es responder las preguntas de investigación.

OE1 – RQ1.1: ¿Cuáles son las políticas, las normativas y las metodologías que utilizan las entidades financieras para el análisis de riesgos crediticio?

Se requiere de una investigación cualitativa en donde se realizará un estudio de las normativas y las políticas. En este caso, se elegirá como método de investigación la investigación documental.

El **método de investigación documental** está conformado por las siguientes etapas (Cruz García, 2014):

a) **Selección del tema**: En la elección del tema se delimitará en las normativas y políticas aplicadas al análisis financiero. En la delimitación del tema estará enfocada a las Cooperativas de Ahorro y Crédito.
b) **Recopilación de información:** Se buscará información en lo referente a las políticas y normativas sobre el análisis del riesgo crediticio.
c) **Análisis y sistematización de la información:** En esta etapa se realizará una compresión de los documentos recolectados.
d) **Integración, redacción y presentación del trabajo:** Se realiza una organización de todos los documentos recolectado para luego proceder a elaborar una redacción de los documentos.

En el método de investigación documental se aplicó **la técnica de la fuente de información primaria** y la **estrategia de investigación** utilizada son los siguientes:

a) ¿Dónde?

Se solicitará información al departamento legal de la Cooperativa de Ahorro y Crédito CACPECO LTDA.

b) ¿Cuándo?

La solicitud de información se realizará en el inicio del desarrollo de la tesis.

c) ¿Quién o qué?

El abogado de la Cooperativa de Ahorro y Crédito CACPECO LTDA. nos guiará durante la investigación documental.

d) ¿Cómo?

Sé obtendrá una fotocopia o impresión o medio digital de los documentos.

Se necesita de una investigación descriptiva para explorar la variedad de metodología, por lo que se utilizará el método de revisión bibliográfica.

El **método de revisión bibliográfica** constas de las siguientes etapas (Marta, 8):

a) Justificación de la revisión bibliográfica: Es necesario conocer las metodologías que está utilizando la Cooperativa de Ahorro y Crédito CACPECO LTDA. y las demás entidades financiera en el análisis de riesgo crediticio.

b) Recopilación de la bibliografía: Se buscará información en la Cooperativa de Ahorro y Crédito CACPECO LTDA. y la búsqueda de información online.

c) Evaluación y selección de la bibliografía: Se debe tener en cuenta los siguientes criterios:

- ✓ **Relevancia:** La publicación encontrada debe encajar con el tema y las preguntas de investigación, por lo cual se sugiere leer la introducción y la conclusión del mismo para juzgar si la publicación es relevante para tu revisión bibliográfica o no.
- ✓ **Calidad:** Se sugiere tomar en cuenta las publicaciones que estén afiliados a una institución académica, institución científica, institución financiera, gubernamental.

d) Elaboración de la revisión bibliográfica: Se desarrolla los resultados y las conclusiones de los diferentes planteamientos de los autores e intentar responder a la pregunta de investigación.

En la revisión bibliográfica se utilizó la **técnica exploratoria y analítica** para la recolección de información más relevante y actualizada. Mediante la **técnica comparativa** se sintetizó la información más relevante. La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

En la recolección de información online se tomó en cuenta la siguiente sugerencia:

IEEE: Es un recurso que proporciona información de contenido científico-tecnológico editado por IEEE y otros editores como la MIT y Wiley.

b) **¿Cuándo?**

La revisión literaria se realiza después de haber realizado una investigación documental sobre los factores que inciden en los riesgos crediticio.

c) ¿Quién o qué?

Se contará con la ayuda del analista de riesgo crediticio de la Cooperativas de Ahorro y Crédito CACPECO LTDA. para la revisión bibliográfica

d)) ¿Cómo?

Está claro que se debe considerar las metodologías de riesgo crediticio aplicado a las Cooperativas de Ahorro y Crédito. Sin embargo, también se puede examinar las metodologías aplicado en los Bancos para poder realizar una comparación entre ellas.

Se necesita de una investigación cualitativa y cuantitativa para definir las variables cuantitativas y cualitativas críticas a considerar en un proceso de gestión de créditos. En este caso, se elegirá como método de investigación el método de recolección de datos.

El **método de recolección de datos** permitirá explorar la problemática a profundidad. El método de recopilación de datos cuantitativo usa datos medibles para formular hechos y descubrir ciertos patrones. El método de recopilación de datos cualitativo examina las razones de la toma de decisiones.

En la **técnica de la entrevista semiestructurada** se sugerirá una propuesta de formato de documento para realizar la entrevista (Aguirre Baztán, 1995):

Preparación

- ✓ Identificación de la entrevista:
- ✓ Identificativo único.
- ✓ Preparada por: nombre(s) y cargo(s).
- ✓ Fecha de preparación.
- ✓ Fase en la que se encuadra.
- ✓ Documento(s) al que se hace referencia. (Si se hace referencia a alguno y modo en que hace referencia).
- ✓ Tiempo necesitado para la preparación.
- ✓ Identificación de los participantes previstos:
- ✓ Entrevistado(s): nombre(s) y cargo(s).
- ✓ Entrevistador(es): nombre(s) y cargo(s).
- ✓ Objetivos: Se identificarán mediante numeración, caracteres alfabéticos.

- ✓ Descripción de los puntos a tratar y/o preguntas de la entrevista: que también serán identificados.
- ✓ Previsiones respecto a la entrevista:
- ✓ Lugar.
- ✓ Fecha.
- ✓ Hora.
- ✓ Duración prevista.
- ✓ Recomendaciones a los entrevistadores:
- ✓ Información previa a recabar.
- ✓ Documentación a revisar.
- ✓ Informaciones pendientes de entrevistas anteriores.
- ✓ Consideraciones especiales sobre los participantes.
- ✓ Otras cuestiones.

Resultado

- ✓ Identificación de la entrevista:
- ✓ Identificativo de la preparación.
- ✓ Lugar.
- ✓ Fecha.
- ✓ Hora.
- ✓ Duración.
- ✓ Incidencias sobre los participantes: Modificaciones sobre las previsiones realizadas.
- ✓ Cuerpo de la entrevista: Anotaciones para cada punto y/o preguntas de la entrevista.
- ✓ Informe final sobre la entrevista:

- ✓ Información obtenida.
- ✓ Información pendiente:
- ✓ Documentos que los entrevistados deben entregar.
- ✓ Documentos que los entrevistadores deben entregar.
- ✓ Información olvidada en la entrevista.
- ✓ Grado de cobertura de los objetivos.
- ✓ Grado de participación y colaboración de los entrevistados.
- ✓ Notas y recomendaciones especiales.

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

La entrevista se realizará en el Cooperativa de Ahorro y Crédito CACPECO LTDA.

b)¿Cuándo?

Una vez terminada de revisar la revisión bibliográfica se realizará la formulación de preguntas para la entrevista

c) ¿Quién o qué?

Los participantes son la Gerencia de Operaciones y los ejecutivos de créditos

d)¿Cómo?

En el Anexo 1, se puede observar el formato de la entrevista y el temario, la cual consta de 10 preguntas.

Se requiere de una investigación cualitativa y cuantitativa para definir las variables cuantitativas y cualitativas críticas a considerar en un proceso de gestión de créditos. En este caso, se elegirá como método de investigación el método Delphi.

El **método Delphi** consta de los siguientes pasos (Comunicación, s.f.):

a) **Identificación de la problemática**: también es necesaria la identificación del objetivo perseguido con la aplicación de este método.
b) **Elaboración del cuestionario**: de acuerdo a los objetivos planteados, las preguntas deben ser claras concisas y cuantificables con el fin de facilitar su análisis.
c) **Definir el panel de expertos o participantes en la encuesta**: en base a una serie de características que los hayan determinado de importancia en el caso de estudio.
d) **Distribuir el cuestionario**: debe ser llenado de forma anónima de tal forma que no se vean afectados los resultados, es recomendable antes de aplicar el cuestionario informar a los participantes acerca de objetivos perseguidos con el cuestionario.
e) **Analizar resultados**: corresponde a la tabulación y análisis de los resultados obtenidos tras aplicar el cuestionario.
f) **Entrega de análisis a expertos y redistribución del cuestionario**: corresponde a la entrega del cuestionario conjuntamente con el análisis de las respuestas obtenidas anteriormente.
g) **Segundo análisis**: con la entrega de los nuevos cuestionarios, se realiza un segundo análisis para identificar tendencias, patrones, etc. eliminando los datos estadísticos más dispersos.

La **técnica del cuestionario** se basa en varias preguntas en donde el experto tratará de responder a las preguntas que más identifique el aspecto cualitativo del cliente en el momento de otorgar un crédito. La propuesta de un formato de cuestionario se muestra en la tabla 6.

Tabla 6.
Cuestionarios de preguntas (Ejemplo de propuesta)

No.	Preguntas	SI	NO
1	Antigüedad del cliente		
2	Más de un propietario		
3	Pronóstico de ventas		
4	Problemas legales		
5	Calidad de cartera de clientes		
6	Calidad de cartera de proveedores		
7	Referencias comerciales		
8	Referencia bancaria		
9	Personal contratado: Alto, Medio, Bajo		
10	Importancia del estado financiero		
11	Importancia del cliente		
12	Cumplimiento de pagos anteriores		
13	Ubicación del sector del negocio: Alto, Medio, Bajo		
14	Calificación en el central de riesgo: Alto, Medio, Bajo		
15	Tipo de sociedad: Corporación, compañía		

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

El cuestionario se realizará en el Cooperativa de Ahorro y Crédito CACPECO LTDA.

b) ¿Cuándo?

Se procede después de concluir la entrevista. Una vez revisado la entrevista se desarrollará el cuestionario.

c) ¿Quién o qué?

Los participantes son la Gerencia de Operaciones y los analistas de crédito

d) ¿Cómo?

El registro del cuestionario se realizará por medio físico. En el Anexo 2 se puede observar el formato del cuestionario y las 60 preguntas tanto en la Tabla 27 y la Tabla 28. Una vez finalizada el cuestionario se analizará las preguntas que mejor se identifiquen para el modelo de riesgo creditico.

3.2.1.2. Fase 2: Definición de los objetivos de una solución

El objetivo en esta fase es responder las preguntas de investigación.

OE1 – RQ1.2: ¿Cuáles son las técnicas de minería de datos aplicables al análisis de riesgo crediticio?

Se necesita de una investigación descriptiva para explorar la variedad de técnicas de minería de datos, por lo que se utiliza el método de revisión bibliográfica

El **método de revisión bibliográfica** constas de las siguientes etapas (Marta, 8):

a) Justificación de la revisión bibliográfica: Conocer las diferentes técnicas de minería de datos aplicado al análisis de riesgo crediticio.

b) Recopilación de la bibliografía: Se realizará la búsqueda de información online.

c) Evaluación y selección de la bibliografía: Se debe tener en cuenta los siguientes criterios:

- ✓ **Relevancia:** La publicación encontrada debe encajar con el tema y las preguntas de investigación, por lo cual es recomendable realizar una lectura tanto de la introducción como de la conclusión y de tal forma juzgar si la publicación es relevante para la revisión bibliográfica.
- ✓ **Calidad:** Se recomienda tomar en cuenta las publicaciones que estén afiliados a una institución académica, institución científica, institución financiera, gubernamental.

d) Elaboración de la revisión bibliográfica: Se desarrolla los resultados y las conclusiones de los diferentes planteamientos de los autores e intentar responder a la pregunta de investigación.

En la revisión bibliográfica se utilizó la **técnica exploratoria y analítica** para la recolección de información más relevante y actualizada. Mediante la **técnica comparativa** se sintetizó la información más relevante. La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

En la recolección de información online se tomó en cuenta la siguiente sugerencia:

IEEE: Es un recurso que proporciona información de contenido científico-tecnológico editado por IEEE y otros editores como la MIT y Wiley.

b) ¿Cuándo?

La revisión literaria se realiza después de haber terminado de evaluar los resultados del cuestionario.

c) ¿Quién o qué?

La investigación de la revisión literaria lo realizará las autoras bajo la orientación de la tutora académica

d) ¿Cómo?

Se planificará una o varias reuniones entre las autoras para realizar la revisión bibliográfica. Una vez elaborado la revisión bibliográfica se procede a enviar a la tutora académica para su revisión parcial y final.

Se requiere de una investigación cuantitativa e investigación analítica para valorar la técnica de minería más eficiente y adecuado para el análisis de riesgo crediticio, para ello se utilizará el método heurístico de análisis y selección.

El **método heurístico de análisis y selección** consta de tres etapas secuenciales (Silclir, Szyrko, Ruiz de Mendarozqueta, & Rubio):

a) **Definición de requerimientos y categorización**: El análisis de las técnicas de minería de datos se plantea sobre la comparación de un conjunto de requerimientos considerados como deseables. Cada uno de estos requerimientos es categorizado con el fin de establecer una jerarquía entre aquellos requerimientos de alto valor contra aquellos que representan aspectos deseables pero cuyo cumplimiento es ciertamente opcional.

b) **Análisis individual de cada técnica de minería de datos**: Una vez establecidos los requerimientos se analiza cada una de las técnicas de minería de datos bajo análisis valorando el cumplimiento de cada uno de esos requerimientos contra su implementación particular. Cada una de estas valoraciones tiene asociado un valor cuantitativo que será utilizado al aplicar el método cuantitativo de selección:

   - ✓ **Cumplimiento absoluto:** la técnica de minería de datos cumple con todas las especificaciones del requerimiento de forma directa. El valor cuantitativo asignado es 5.

- ✓ **Cumplimiento parcial:** la técnica de minería de datos cumple con las especificaciones del requerimiento ya sea en forma parcial, considerando sólo algunas condiciones o aspectos de dicho requerimiento. El valor cuantitativo asignado es 2.
- ✓ **No cumplimiento:** la técnica de minería de datos no proporciona ningún nivel de cumplimiento del requerimiento. El valor cuantitativo asignado es 0.

c) Análisis comparativo de las técnicas de minería de datos: En esta etapa se resumen los resultados obtenidos del análisis de cada una de las técnicas de minería de datos y se procede a aplicar las operaciones numéricas tendientes a seleccionar una de ellas, obteniendo una valoración general de cada una de las técnicas de minería de datos sobre la base de estas dos dimensiones:

- ✓ **Valor de cada requerimiento:** Si una técnica de minería de datos para un requerimiento particular categorizado como mandatorio ha sido valorado con un No Cumplimiento, queda automáticamente descartada.
- ✓ Cumplimiento del requerimiento en la implementación particular (técnica de minería de datos)

d) Selección de la técnica de minería de datos: Finalmente se comparan los valores cuantitativos de cada una de las técnicas de minería de datos y se selecciona la que mayor valor tenga.

Utiliza la técnica cuantitativa de recolección de datos que implica el utilizar datos numéricos para evaluar la información.

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

La investigación se centrará en las diferentes técnicas de minería de datos en el análisis de riesgo crediticio.

b) ¿Cuándo?

Después de evaluar la revisión bibliográfica de técnicas de minería de datos, se procederá a realizar un cuadro comparativo con el fin de seleccionar la mejor técnica de minería de datos en la gestión riesgo crediticio.

c) ¿Quién o qué?

El análisis comparativo y la selección de la técnica de minería de datos que permita analizar los riesgos crediticios lo realizará las autoras bajo la orientación de la tutora académica

d) ¿Cómo?

Para evaluar cada una de las técnicas de minería de datos se utilizará la siguiente formula (Silclir, Szyrko, Ruiz de Mendarozqueta, & Rubio):

CRi : Categoría asignada al requerimiento

i VRij: Valoración aplicada para el requerimiento i en la herramienta j

Vj: Valoración cuantitativa de la herramienta

$Vj = i=1...n \sum CRi * VRij$

3.2.1.3. Fase 3: Diseño y desarrollo

El objetivo en esta fase es responder las preguntas de investigación.

OE2 – RQ2.1: ¿Qué metodología es adecuada para realizar el pre procesamiento y el modelado de los datos?

Se requiere una investigación cuantitativa y cualitativa para el procesamiento de las variables, para ello se usará la metodología de minería de datos CRISP-DM (del inglés Cross Industry Standard Process for Data Mining)

La **metodología de minería de datos CRISP-DM** consta de seis fases, pero se enfatizará en 3 fases que son los siguientes (Galán Cortina, 2015):

a) Comprensión de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio)

- ✓ Recopilación inicial de datos
- ✓ Descripción de los datos
- ✓ Exploración de los datos
- ✓ Verificación de calidad de datos

b) Preparación de los datos (Obtener la vista minable o dataset)

- ✓ Selección de los datos
- ✓ Limpieza de datos
- ✓ Construcción de datos
- ✓ Integración de datos
- ✓ Formateo de datos

c) Modelado (Aplicar las técnicas de minería de datos a los dataset)

- ✓ Selección de la técnica de modelado
- ✓ Diseño de la evaluación
- ✓ Construcción del modelo
- ✓ Evaluación del modelo

La técnica para el diseño y desarrollo del modelo son las **técnicas de minería de datos**.

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

Se utilizará la base de datos transaccional de la Cooperativa de Ahorro y Crédito CACPECO LTDA., que posteriormente servirá para el diseño del modelo predictivo.

b) ¿Cuándo?

Una vez seleccionado las mejores técnicas de minería de datos se procederá a recolectar los datos y depurarlo para posteriormente modelar los datos utilizando técnicas de minería de datos.

c) ¿Quién o qué?

El pre procesamiento de los datos y el diseño del modelo lo realizarán las autoras bajo la orientación de la tutora académica

d) ¿Cómo?

Para el procesamiento de los datos se utilizará las herramientas ETL y el diseño de modelo se empleará las herramientas de minería de datos.

OE2 – RQ2.2: ¿Cuáles son las características, las variables y los métodos utilizado para construir el modelo predictivo?

Se requiere de una investigación cuantitativa en donde se cuantifica las relaciones entre la variable independiente o predictiva y la variable dependiente o resultado. Se utilizará el diseño no experimental para realizar estudios descriptivos y de correlación. Dentro del diseño no experimental se enfocará en el diseño transversal, ya que las variables serán consideradas en un punto en el tiempo y las relaciones entre las mismas.

Se utilizará el **método hipotético-deductivo** el cual está conformado por la complementación entre los métodos inductivo y deductivo. De tal forma el método inductivo permitirá descubrir hechos que sirvan para fundamentar los enunciados teóricos mismos que deberán derivar a hipótesis que serán verificadas mediante el método deductivo.

Este método científico consta de varias etapas que pueden ser desarrolladas tanto en paralelo como también en estricto orden cronológico (Wikiteka, 2011):

a) **Construir un modelo teórico** que se aproxime a la explicación del hecho, incluyendo la elaboración de proposiciones relativas a las variables relevantes.
b) **Formular la hipótesis** a razón del modelo teórico y de los datos obtenidos.
c) **Elaborar un plan** para recoger los datos y las observaciones para someterlo a prueba las predicciones.
d) **Recogida y análisis de los datos**. Una vez recogida los datos según la planificación del punto anterior, se procede a codificar y ordenar los datos para luego analizarlos mediante las técnicas estadísticas apropiadas.
e) **Toma de decisión acerca de la hipótesis.** Un adecuado análisis estadístico representa el pilar base para determinar la aceptación o el rechazo de la hipótesis, según el grado de probabilidad estas pueden ser confirmadas o rechazadas.
f) **Introducción de las conclusiones**. Como último se realizan evaluaciones del alcance y eficacia a los resultados obtenidos, En el caso que los resultados sean los esperados estos se deben concentrar a la teoría, en el caso opuesto se deberá plantear un nuevo problema que junto con las hipótesis serán una guía para estudios posteriores. El modelo teórico podrá ser modificado o corregidos, es decir, no es una teoría estática ni posee verdad absoluta, ya que requerirá que sea comprobada de forma sistemática.

Las **técnicas** para construir el modelo serán los siguientes:

- ✓ Análisis univariante
- ✓ Análisis multivariante
- ✓ Modelo de riesgo crediticio
- ✓ Técnicas de minería de datos

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

Se realizará la selección de variables de la base de datos transaccional de la Cooperativa de Ahorro y Crédito CACPECO LTDA.

b) ¿Cuándo?

Al finalizar el pre-procesamiento de datos y se procederá a construir el modelo predictivo

c) ¿Quién o qué?

El análisis de las variables, las características y el desarrollo del modelo predictivo lo realizará las autoras bajo la orientación de la tutora académica

d) ¿Cómo?

Con la ayuda de la minería de datos se realizará un estudio del análisis estadístico univariado y multivariado para seleccionar las variables que formará parte del modelo predictivo, con el fin de encontrar posibles patrones oculto para medir los riesgos de crédito.

3.2.1.4. Fase 4: Demostración

El objetivo en esta fase es responder las preguntas de investigación.

OE3 – RQ3.1: ¿Qué proceso se debe seguir para la validación del modelo de riesgo crediticio?

En esta fase se realizará una investigación cuantitativa de tipo no experimental basado en el diseño de correlacionales descriptivo predictivo para analizar las variables independientes (predictiva) y dependiente(resultado) sin que sean manipulada, es decir, que se dan en forma natural.

El **método de validación** de los modelos de minería de datos se tomará en cuenta las siguientes características (Microsoft, 2018):

- ✓ Utilizar varias medidas de validez estadística con la finalidad de poder determinar la existencia de problemas en los datos o en el modelo.
- ✓ Dividir los datos en grupos de entrenamiento y prueba con la finalidad de evaluar la precisión de las predicciones.
- ✓ Solicitar a los expertos en el área comercial la revisión de los resultados del modelo de minería de datos y de tal forma determinar si los patrones detectados tienen sentido en un escenario empresarial concreto.
- ✓ Todos los métodos estadísticos son útiles para la metodología de minería de datos, son utilizados iterativamente al generar, validar y depurar modelos con el fin de solucionar un problema determinado.

Durante el proceso de validación del modelo de minería de datos se tendrá en cuenta los siguientes criterios (Microsoft, 2018):

a) Precisión Esta medida muestra los puntos en los que el modelo pone en correspondencia un resultado con los atributos de los datos proporcionados.

Todas las medidas de precisión dependen de los datos utilizados, en la fase de análisis y generación del modelo, se podría aceptar una cierta cantidad de errores en los datos.

b) Confiabilidad Analiza el comportamiento de un modelo de minería en grupos diferentes de datos. Se considera confiable a un modelo de minería de datos cuando devuelve el mismo tipo de predicción o encuentra los mismos patrones independientemente de los datos de prueba proporcionados.

c) Utilidad Incluye métricas distintas con el fin de determinar la utilidad de la información que provee el modelo de minería de datos.

La **técnica de validación** se puede considerar lo siguiente (Vera Noguez):

- ✓ Validación simple
- ✓ Validación cruzada
- ✓ Validación cruzada con n pliegues
- ✓ Validación bootstrapping

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

Para la demostración del modelo se realizará en la Cooperativa de Ahorro y Crédito CACPECO LTDA.

b) ¿Cuándo?

Después de haber desarrollado un prototipo del modelo de medición de riesgo creditico.

c) ¿Quién o qué?

Participará la Gerencia de Operaciones, los analistas de gestión de riesgos y las autoras bajo la orientación de la tutora académica.

d) ¿Cómo?

Se utilizará las herramientas de minería de datos para la prueba y validación de modelos que permitirá estimar la probabilidad de error.

3.2.1.5. Fase 5: Evaluación

El objetivo en esta fase es responder las preguntas de investigación.

OE3 – RQ3.2: ¿Qué metodología se elige para evaluar un modelo de medición de riesgo crediticio?

Se realizará una investigación cuantitativa del tipo no experimental, cuyos estudios es realizar un test de correlación que permita examinara la relaciones propuesta por un modelo o teoría.

La metodología de minería de datos CRISP-DM consta de seis fases, pero se enfatizará en 1 fases que es el siguiente (Galán Cortina, 2015):

Evaluación de los resultados: Este paso evalúa el grado al que el modelo responde a los objetivos de negocio, y determina si hay alguna decisión de negocio que el modelo no cubra. El objetivo es resumir los resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final estableciendo si el proyecto ha alcanzado los objetivos iniciales de negocio.

- ✓ Comprender los resultados de la minería de datos.
- ✓ Interpretar los resultados en términos de su aplicación.
- ✓ Comprobar los efectos sobre los objetivos de minería de datos.
- ✓ Comprobar los resultados de minería de datos contra la base de un conocimiento determinado para ver si la información descubierta es nueva y útil.
- ✓ Evaluar y estimar los resultados en lo que respecta a criterios de éxito de negocio (esto es, el proyecto ha alcanzado los objetivos de negocio originales).
- ✓ Comparar los resultados de la evaluación y la interpretación.
- ✓ Clasificar los resultados en lo que respecta a criterios de éxito de negocio.
- ✓ Comprobar el efecto de los resultados sobre el objetivo de la aplicación inicial.

- ✓ Determinar si hay nuevos objetivos de negocio para abordar en una evolución del proyecto o en nuevos proyectos.
- ✓ Expresar recomendaciones para proyectos futuros de minería de datos.

Las **técnicas de evaluación** serán consideradas los siguientes:

- ✓ Test de Back Testing
- ✓ Test Kolmogorov-Smirnov
- ✓ Curva ROC
- ✓ Test de Hosmer-Lemeshow

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

Para la evaluación del modelo se realizará en la Cooperativa de Ahorro y Crédito CACPECO LTDA.

b) ¿Cuándo?

Después de realizar la validación de un prototipo del modelo de medición de riesgo creditico, se procederá a evaluar el modelo.

c) ¿Quién o qué?

Participará la Gerencia de Operaciones, los analistas de gestión de riesgos y las autoras bajo la orientación de la tutora académica

d) **¿Cómo?**

Se utilizará las herramientas de minería de datos para que realizar la prueba del modelo.

3.2.1.6. Fase 6: Comunicación

Para desarrollar el informe de investigación se utilizará la **metodología de diseño descendente**:

a) Escribir los objetivos del trabajo.

b) Realizar un esquema del texto, dividiendo cada apartado en secciones.

Escribir el objetivo de cada sección

- ✓ Dividir cada sección en bloques, y escribir el objetivo de cada uno de ellos
- ✓ Dividir cada bloque en párrafos, y escribir cada uno de ellos

c) En general, es más fácil empezar escribiendo el CUERPO del artículo (método, resultados y discusión).

- ✓ Comenzar a escribir como si el lector fuese un compañero de trabajo.

d) Después, escribir las conclusiones, el resumen (abstract) y la introducción.

Se debe tener en cuenta las siguientes **técnicas** al elaborar el informe de investigación (Ruíz Guerra & González R.):

a) Emplear oraciones cortas y concretas. Cada párrafo debe contener una idea principal en torno a la cual se organicen frases secundarias que contribuyan a la comprensión de la idea: argumentos, aclaratorias, ejemplos y contrastes.

b) Los párrafos y secciones del informe deben guardar la adecuada coherencia de tal manera que el informe constituya una unidad estructural, con hilación.

c) Evitar términos complicados, recordando que la finalidad del informe es comunicar los resultados de la investigación por lo que debe ser comprensible al lector. Exige sobre todo precisión, por lo que deben emplearse palabras que traduzcan fielmente lo que desee

expresar, debe evitarse el empleo de palabras de cuyo sentido no se está seguro por lo que conviene el uso de términos frecuentes. Se recomienda la consulta al diccionario.

d) Procurar uniformidad en el lenguaje en cuanto a tiempo y persona gramatical utilizada durante todo el escrito, es conveniente definir previamente quien habla en el informe; se sugiere la primera persona del plural o la forma impersonal (“se cree que...”).

e) Vigilar ortografía y puntuación. Las faltas no sólo desacreditan al escritor, sino también pueden despertar sospechas sobre los méritos profesionales; así como una puntuación defectuosa dificulta la lectura e incluso puede deformar el sentido de la oración.

f) Dar relevancia a las ideas, es decir, resaltar lo esencial sobre lo accesorio. Para ello pueden emplearse algunos procedimientos de uso común como subrayado, uso de subtítulos y divisiones en forma ordenada y en secuencia uniforme.

La **estrategia de investigación** utilizada son las siguientes:

a) ¿Dónde?

Se centra en elaborar el informe de investigación

b) ¿Cuándo?

Después de que se haya evaluado exitosamente el modelo de medición de riesgo crediticio de una institución financiera.

c) ¿Quién o qué?

El informe de investigación lo elaborará las autoras bajo la orientación de la tutora académica y luego el informe será entregado a la Gerencia de Operaciones y a los analistas de gestión de riesgos.

d) ¿Cómo?

Se utilizará la herramienta ofimática en la redacción del informe de investigación.

CAPITULO IV

DESARROLLO DEL MODELO

La Metodología a ser utilizada en el proceso de desarrollo del proyecto ha sido Design Science Research (DSR) ya que se acopla a nuestras necesidades.

4.1. Identificación del problema y motivación

En el capítulo I se ha realizado un análisis del problema de investigación. La cooperativa de ahorro y crédito CACPECO ha disminuido en porcentaje la cantidad de créditos otorgados mensualmente, una de las causas es la escasa promoción de créditos a clientes objetivo.

La selección de *candidatos* a los cuales promocionar productos u otorgar créditos pre-aprobados es realizado a través de la observación tornándose ineficiente e insuficiente, la motivación para el desarrollo del presente proyecto radica en la realización de análisis crediticios detallados a través de modelos de minería de datos que permitan identificar del total de socios, una mayor cantidad de *clientes objetivo* con altas posibilidades de retorno de capital, de esta forma se propone hacer campañas agresivas en promoción de créditos.

Cuando un socio solicita un crédito, la institución realiza el análisis del riesgo crediticio haciendo uso del buró de crédito el cual es ofrecido por la empresa *Equifax*, cuyos reportes de crédito reflejan el comportamiento de pago de una persona o empresa en distintas instituciones financieras y la situación crediticia del socio, para acceder a revisar dicha información es necesaria la aceptación del cliente mediante un documento firmado. También realizan visitas a los domicilios de los solicitantes de crédito con el objeto de evidenciar que el nivel de vida sostenido sea consecuente con los ingresos declarados a la institución, como también en un futuro poderlos ubicar y realizar gestión de mora en caso de requerirlo. Se constata también que el

monto solicitado no sea causal de sobreendeudamiento, y que la capacidad de pago sea suficiente para la cancelación de cuotas correspondientes a la operación de crédito.

La cooperativa CACPECO se rige a las normativas vigentes en el Ecuador. Las normativas que dictaminan las políticas de análisis de riesgos están definidas por la SEPS (organismo de control cooperativo en el ECUADOR) a través de la resolución No. *129-2015-F* la cual indica textualmente lo siguiente:

"Las entidades financieras públicas privadas, y las del segmento 1 del sector financiero popular y solidario no podrán realizar operaciones activas y contingentes con una misma persona natural o jurídica por una suma que exceda en conjunto el 10% del patrimonio técnico de la entidad. Este límite se elevará al 20% si lo excede del 10% corresponde a obligaciones caucionadas con garantía de bancos nacionales o extranjeros de reconocida solvencia o por garantías adecuadas, en los términos que dicte la Junta de Política y Regulación Monetaria y Financiera. Los límites de créditos establecidos se determinarán a la fecha de aprobación original de las operaciones o de cada reforma efectuada. En ningún caso la garantía podrá tener un valor inferior al valor total del exceso. "(SEPS,2015)

La resolución No. 129-2015-F denota los siguientes artículos:

ARTICULO 4. De la gestión del riesgo de crédito se dispone contemplar como mínimo lo siguiente:

- ✓ Límites de exposición al riesgo de crédito de la entidad, en los distintos tipos de crédito y de tolerancia de la cartera vencida por cada tipo de crédito, para las cooperativas de los segmentos 1 y 2.
- ✓ Criterios para la determinación de tasas para operaciones de crédito. Considerando entre otros: montos, plazos, garantías, tipo de productos, destino de financiamiento.

- ✓ Criterios para definir su mercado objetivo, es decir, el grupo de socios a los que se quiere otorgar créditos: zonas geográficas, sectores socio-económicos, para las cooperativas de los segmentos 1 y 2.
- ✓ Perfiles de riesgo: características de los socios con los cuales se va a operar, como edad, actividad económica, género entre otros, para los segmentos 1 y 2

ARTICULO 6. _ Responsabilidades del Comité de Administración de Riesgos: El comité de Administración Integral de Riesgos deberá:

- ✓ Aprobar y presentar al Consejo de Administración el informe de la unidad o administrador de riesgos, según corresponda, referido al cumplimiento de políticas y estado de la cartera vigente que incluya la situación de las operaciones refinanciadas, reestructuradas, castigadas y vinculadas.
- ✓ Aprobar y monitorear en las cooperativas de los segmentos 1 y 2 la implementación permanente de modelos y procedimientos de monitoreo de riesgos para la colocación y recuperación de cartera de crédito.
- ✓ Recomendar al Consejo de Administración la aprobación del Manual de Crédito propuesto por el área de Crédito: y,
- ✓ Evaluar los problemas derivados del incumplimiento de políticas, procesos y procedimientos para recomendar a los administradores de la entidad las medidas que correspondan.

ARTICULO7. _ Responsabilidades de la Unidad y del Administrador de Riesgos: La unidad o el administrador de riesgos según corresponda deberán:

- ✓ Revisar e informar al Comité de Administración integral de riesgos, las exposiciones de créditos reestructurados, refinanciados, operaciones castigadas, recuperaciones y las que se encuentren sometidas

- ✓ Informar al comité de Administración integral de riesgos la situación de las operaciones vinculadas, señalando las acciones realizadas para la recuperación de aquellas que se encuentren en estado vencido.

4.2. Definición de los objetivos de la solución

El objetivo de la presente solución es incrementar el porcentaje de adjudicación de créditos en el mercado, mediante la realización de un artefacto el cual deberá permitir identificar con exactitud *clientes objetivo* es decir clientes con altas probabilidades de pago a tiempo, a los cuales se les ofertarán créditos, mediante campañas publicitarias emitidas por el call center.

Será necesario también analizar la situación actual de la institución a través de la entrevista e investigación documental y evaluar las técnicas de minería de datos para medir el riesgo de crédito de una institución financiera con base en la revisión bibliográfica.

4.3. Diseño y desarrollo

En la actualidad los datos históricos son almacenados en una base de datos independiente de la base relacional del negocio, se registran en el esquema FBS_HISTORICOS, la información es registrada diariamente durante la noche mediante un procedimiento almacenado que realiza la copia de la información.

Para acceder a dicha información desde el modelo relacional se utilizan dblinks creados hacia la base de históricos.

Como alternativa de solución del presente proyecto se ha propuesto aplicar minería de datos directamente al modelo relacional de la institución, para evitar la ralentización del funcionamiento del sistema transaccional al aplicar la minería en tiempo real, se ha propuesto realizar una extracción de la información necesaria en archivo plano y proceder a analizarla.

La sistematización del proceso de minería de datos es un punto importante para la planificación y ejecución del presente proyecto, para el presente desarrollo se ha seleccionado la metodología CRISP-DM debido a que profundiza en mayor detalle las tareas y actividades a ejecutar en cada etapa del proceso de minería, se realiza un estudio más profundo en la sección 2.4.5.2.1

La metodología CRISP-DM propone las siguientes fases:

4.3.1. FASE 1: Comprensión del negocio

La metodología Crisp-DM están estipuladas en fases, en la figura 18 se observa en detalle la etapa de comprensión del negocio

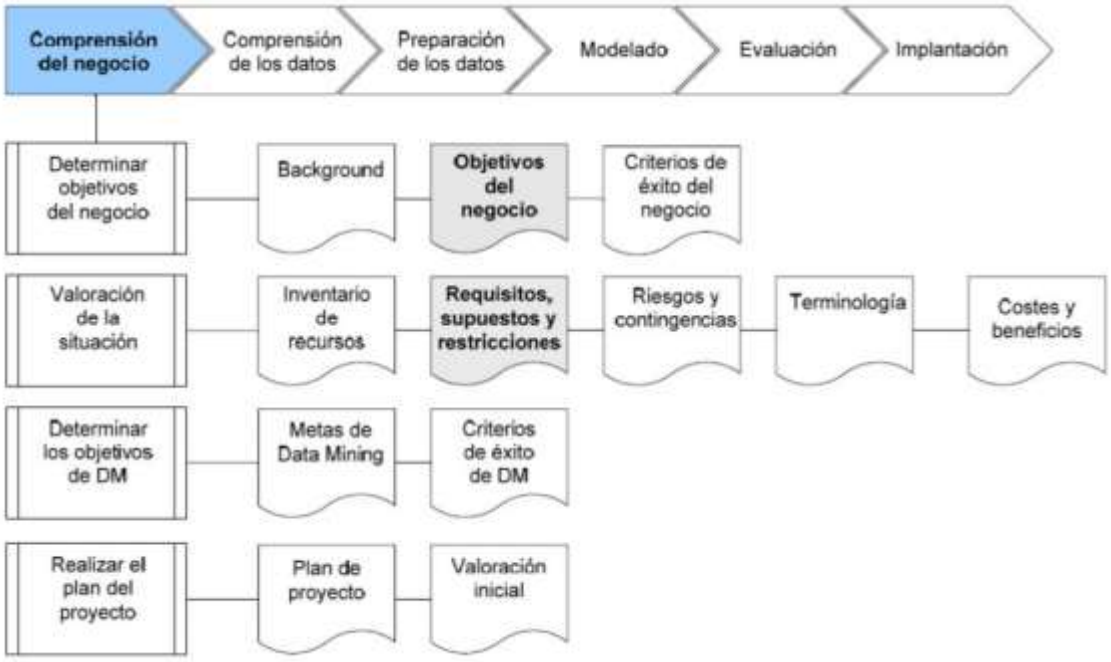


Figura 18. Fase de Comprensión del negocio (Metodología Crisp-DM)
Fuente: (Crisp-DM,2000)

a) Objetivos del Negocio

La cooperativa de ahorro y crédito de la pequeña empresa del Cotopaxi "CACPECO" para publicitarse como institución utiliza canales de información como radio, televisión, impresiones en vallas publicitarias, avisos en los periódicos, etc. Las campañas promocionales aplicadas están enfocadas principalmente al área de captaciones, siendo su debilidad la aplicación de campañas a el área de crédito, puesto que la selección de clientes a los cuales ofertar créditos está realizado a través de análisis subjetivos y manuales, de la información crediticia histórica de los clientes en la institución.

CACPECO se proyecta realizar estudios de datamining en sus bases de datos relacionales y de tal forma atacar el mercado mediante campañas publicitarias dirigidas, pudiendo incluirse las llamadas "preaprobaciones automáticas" ofertando montos de crédito pre-aprobados, lo cual podría ser publicitado a través del call center de la institución, de tal forma que los objetivos comerciales que la cooperativa persigue con la realización de minería de datos son:

- ✓ Optimizar campañas de marketing a través de oferta de créditos pre-aprobados
- ✓ Incrementar el número de créditos concedidos
- ✓ Emitir campañas de marketing dirigidas
- ✓ Incrementar la rentabilidad de la cooperativa

b) Valoración de la Situación Actual

La cooperativa de ahorro y crédito de la pequeña empresa del Cotopaxi tuvo sus inicios en el año 1988, su enfoque principal siempre ha sido generar crecimiento económico sectorial principalmente a través de microcréditos a pequeños empresarios que inician sus primeros emprendimientos, en la actualidad se encuentra clasificada por la entidad de control de

cooperativas SEPS en el segmento uno, lo cual indica que posee un patrimonio superior a 80 millones de dólares, cuenta con 19 agencias a nivel nacional distribuidas en las provincias de Cotopaxi, Tungurahua, Los Ríos, Pichincha, Chimborazo. Dispone de una calificación de riesgo AAA- lo que exterioriza que es muy poco el riesgo de llegar a un desequilibrio económico. La cooperativa trabaja con responsabilidad social lo que contribuye al crecimiento de sus áreas de influencia como clientes, proveedores y demás, está enfocada al crecimiento institucional desde la satisfacción de sus colaboradores, así es el caso que se encuentra calificada como uno de los mejores lugares para trabajar según apreciación de "Great place to work".

El departamento de mercadeo se encarga de publicitar por medios físicos y electrónicos a la cooperativa, dando a conocer los diferentes productos tanto de cartera como de captaciones que se ofrece, los productos de captaciones corresponden a ahorros a la vista, ahorro inversión, ahorro infantil, ahorro migrante, depósitos a plazo fijo, mientras que los productos de cartera se refieren a los variados tipos de crédito que se pueden otorgar, éstas campañas están dirigidas al público en general, sin distinción.

En el ámbito tecnológico el ERP que se utiliza en la institución es "Financial 2.0" el cual fue adquirido a través de su proveedor de servicios tecnológicos "Sifizsoft" desarrollado bajo la plataforma Visual Studio C#, las modificaciones necesarias para ir acorde a las necesidades del negocio son realizadas in-house por el área de desarrollo, el ERP contiene módulos de Contabilidad, Nomina, Tesorería, Inversiones, Captaciones Vista, Captaciones Plazo, Portafolio, Cartera, Crédito, Cobranzas, Proveeduría, Activos Fijos, Cajas.

Se disponen de varias bases de datos para el correcto funcionamiento tecnológico como son:

- ✓ Bdd del modelo relacional transaccional
- ✓ Bdd de históricos

- ✓ Bdd origen con corte a fin de mes para la obtención de reportería y estructuras a presentarse a las entidades de control.
- ✓ Bdd de desarrollo para la utilización de los analistas programadores
- ✓ Bdd de preproducción para las certificaciones de proyectos previos a salir a producción.

Se mantiene un servidor de aplicación adicional al servidor de producción para capacitación a personal nuevo y para pruebas por parte del área de calidad o auditoria.

La administración de la base de datos se la realiza a través del gestor Oracle 12C, se utilizan servidores Windows. Toda la reportería es administrada mediante Reporting Services con conexión directa a el modelo relacional o a la base de históricos dependiendo de la necesidad.

La data center de la institución está compuesto por Blades y se administran a través de máquinas virtuales con el software Virtual Machine

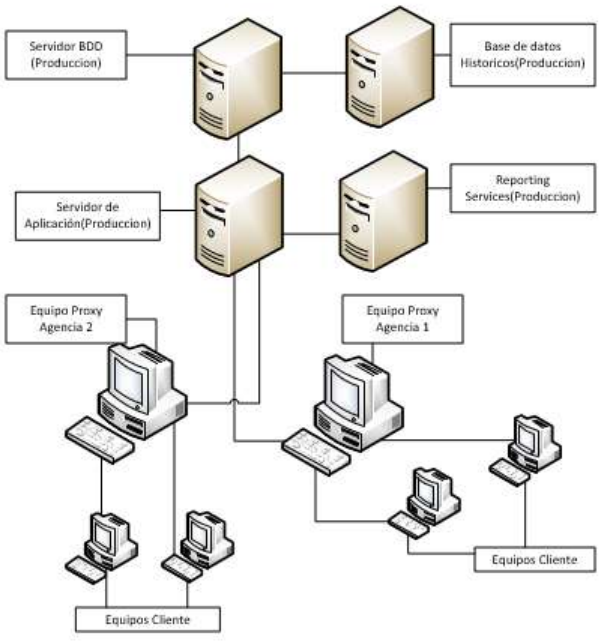


Figura 19. Arquitectura actual ambiente producción institución

c) Objetivos de la Minería de Datos

Los objetivos comerciales planteados por Cacpeco serán alcanzados tras el cumplimiento de los objetivos de minería de datos los cuales son:

- ✓ Clasificar los clientes en buenos y malos
- ✓ Descubrir patrones en clientes malos en términos de rentabilidad
- ✓ Descubrir patrones en morosidad de clientes
- ✓ Generar datos de entrenamiento

d) Plan del Proyecto de la Minería

En el presente plan de proyecto se engloba todas las etapas que se planea ejecutar con detalle de tiempos y recursos, tal como se muestra en la tabla siguiente:

Tabla 7.
Cronograma del plan de proyecto de minería de datos

Fase	Tiempo	Recursos
Comprensión del negocio	2 semana	Área negocios Cacpeco, Investigadoras
Comprensión de los datos	3 semanas	Área de tecnología Cacpeco, Investigadoras
Preparación de los datos	7 semanas	Investigadoras, Oracle, Knime
Modelado	2 semanas	Investigadoras

CONTINÚA

Evaluación	1 semanas	Área de negocios Cacpeco, Investigadoras
Implementación o Despliegue	1 semanas	Área de tecnología Cacpeco, Investigadoras

Para la solución propuesta, se requerirá de un servidor que cumpla los siguientes requerimientos:

Tabla 8.
Requerimientos de Hardware y Software

Hardware	Procesador Intel Core I7
	Velocidad del procesador 3.0 GHz o superior
	Disco Duro 2TB
	Memoria RAM 8GB
Software	Herramienta Knime

El recurso humano para la realización del proyecto se encuentra conformado por las Ingenieras Adriana Salinas y Jenny Chee.

4.3.2. FASE 2: Comprensión de los datos

El detalle de la fase de comprensión de los datos puede observarse a detalle en la figura 20:

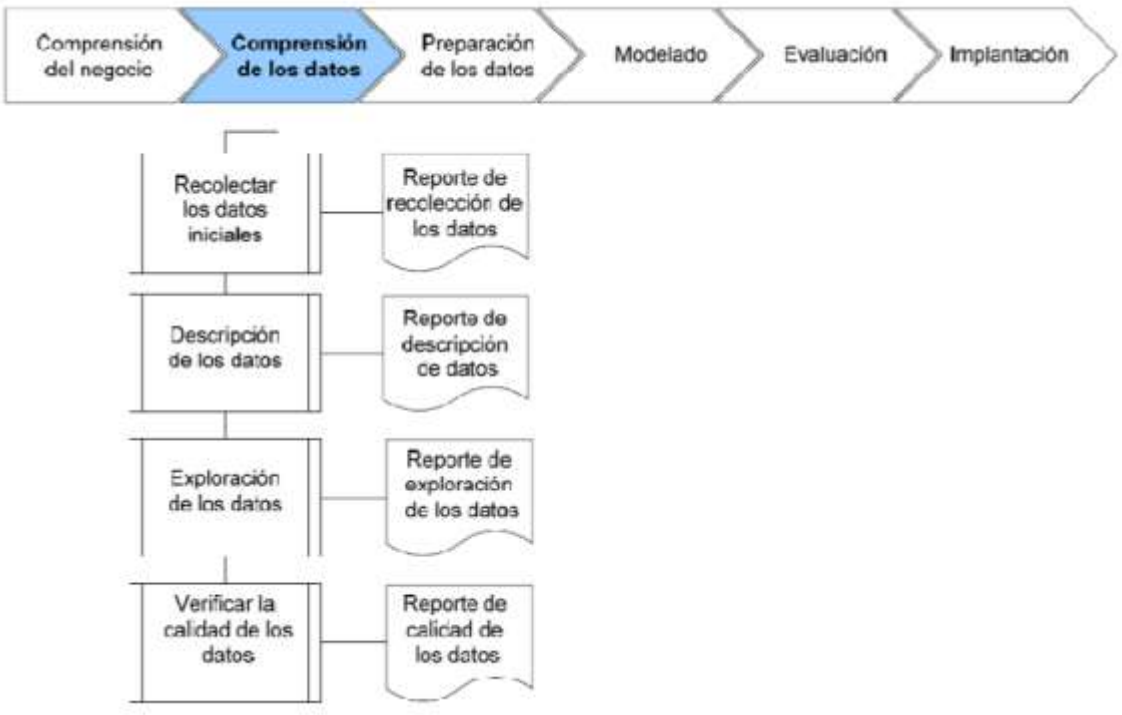


Figura 20. Fase de Comprensión de los datos (Metodología Crisp-DM)
Fuente: (Crisp-DM,2000)

a) Recolección de Datos Iniciales

Los datos necesarios para el desarrollo del presente proyecto fueron recolectados a partir de la base de datos de la cooperativa CACPECO, debido al sigilo bancario los datos han sido extraídos desde una de las bases cuyos datos se encuentran ofuscados.

Se utilizaron variables correspondientes al historial crediticio e información personal del socio como son:

Tabla 9.

Variables Demográficas e ingresos

VARIABLES DEMOGRÁFICAS E INGRESOS	
Sexo	Patrimonio
Estado Civil	Activos
Cantidad de cargas	Pasivos
Nivel educacional	Saldo vigente en créditos no hipotecarios
Actividad Económica	Saldo vigente en créditos hipotecarios
Profesión	
Edad	

Tabla 10.

Variables de vínculo con la institución

VARIABLES DE VINCULO CON LA INSTITUCIÓN
Antigüedad del cliente
¿Tiene créditos no hipotecarios?
Fecha más antigua de otorgamiento de crédito no hipotecario y vigente
Monto Original adeudado en créditos no hipotecarios y vigentes
Plazo máximo de créditos no hipotecarios vigentes
Fecha más antigua de otorgamiento de crédito hipotecario y vigente
Monto original adeudado en créditos hipotecarios y vigentes
Plazo máximo de créditos hipotecarios vigentes

Tabla 11.
Variables de ingresos generados

VARIABLES DE INGRESOS GENERADOS
Ingresos generados por intereses, comisiones y otros en créditos no hipotecarios
Ingresos generados por intereses, comisiones y otros ingresos en créditos hipotecarios

Tabla 12.
Variables de comportamiento

VARIABLES DE COMPORTAMIENTO
En que tramo de morosidad se encuentra el cliente (menos de treinta días de mora, entre 30 y 60 días, entre 60 y 90 días o más de 90 días)
Monto en mora
¿Presenta deuda vencida?
¿Presenta capital castigado?
Máximo de días de mora
¿Posee créditos reestructurados?
¿Posee créditos refinanciados?

b) Descripción de los datos

Los datos relevantes en los que se encuentra los datos que se analizaran se encuentran en los módulos de cartera personas, las tablas en las que se encuentran los datos a utilizar en el presente proyecto han sido representadas en el siguiente diagrama entidad relación.

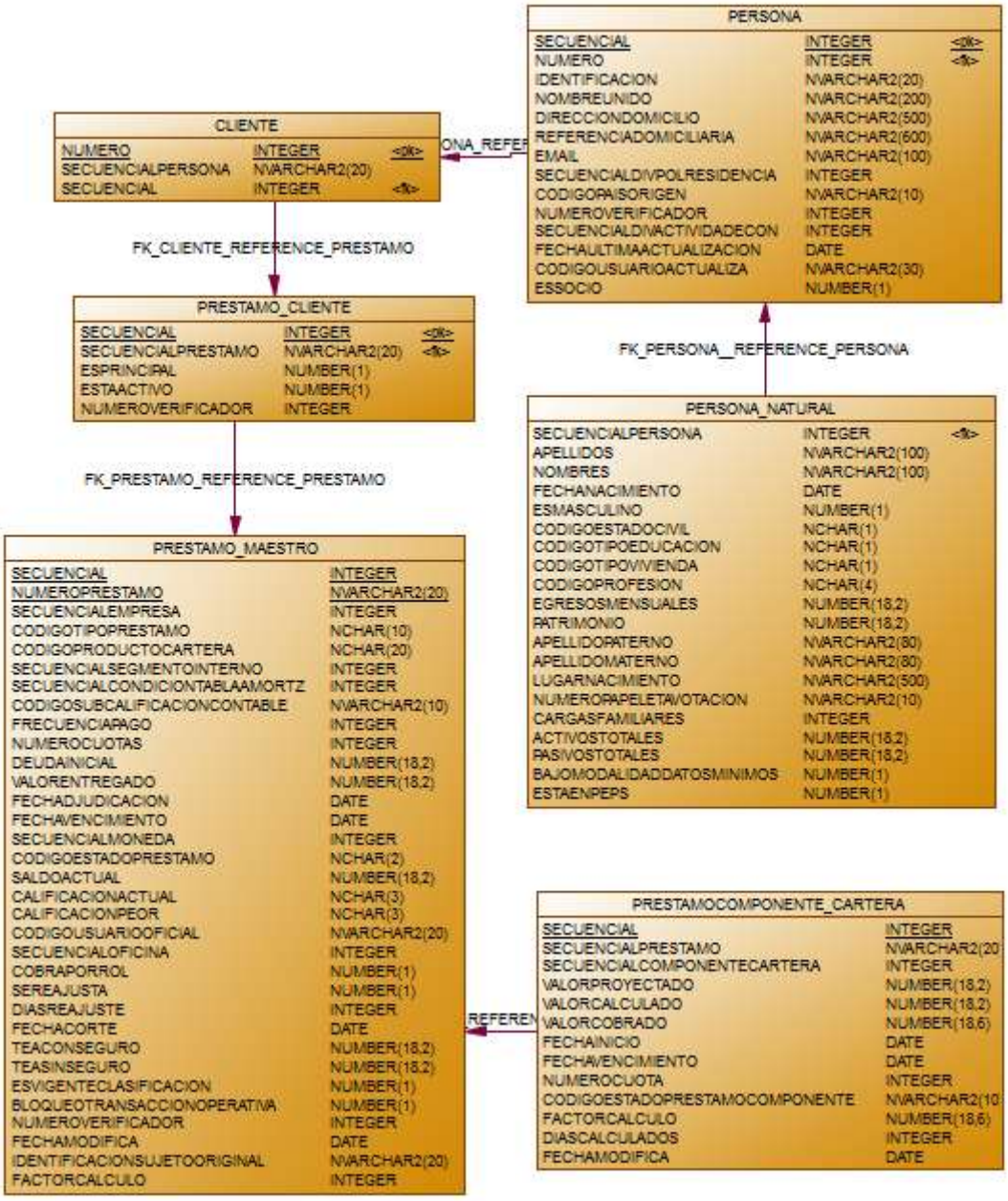


Figura 21. Esquema tablas requeridas

A continuación, se realiza una descripción de cada una de las tablas de la base de datos que se utilizaron para obtener la información para la elaboración de nuestro proyecto.

1. La tabla Préstamo Maestro se encuentra descrita a continuación a través del diccionario de datos.

Tabla 13.
Diccionario Préstamo Maestro

PRESTAMO MAESTRO		
NOMBRE	TIPO	DESCRIPCIÓN
SECUENCIAL	INTEGER	Clave Primaria
NUMEROPRESTAMO	NVARCHAR2(20)	Clave Única
SECUENCIALEMPRESA	INTEGER	Identificador de la empresa
CODIGOTIPOPRESTAMO	NCHAR (10)	Identificador del tipo de préstamo
CODIGOPRODUCTOCARTERA	NCHAR (20)	Identificador del producto
SECUENCIALSEGMENTOINTERNO	INTEGER	Clave Foránea
SECUENCIALCONDICIONTABLAAMORTZ	INTEGER	Clave Foránea
CODIGOSUBCALIFICACIONCONTABLE	NVARCHAR2(10)	Clave Foránea
FRECUENCIAPAGO	INTEGER	Frecuencia de pago pactada
NUMEROCUOTAS	INTEGER	Numero de cuotas pactadas

CONTINÚA

DEUDAINICIAL	NUMBER (18,2)	Deuda Inicial del socio
VALORENTREGADO	NUMBER (18,2)	Valor adjudicado en cuenta
FECHADJUDICACION	DATE	Fecha de concesión del crédito
FECHAVENCIMIENTO	DATE	Fecha que vence el crédito
SECUENCIALMONEDA	INTEGER	Clave Foránea
CODIGOESTADOPRESTAMO	NCHAR (2)	Clave Foránea
SALDOACTUAL	NUMBER (18,2)	Valor actual adeudado
CALIFICACIONACTUAL	NCHAR (3)	Calificación actual del socio
CALIFICACIONPEOR	NCHAR (3)	Calificación peor del socio
CODIGOUSUARIOOFICIAL	NVARCHAR2(20)	Asesor que gestiono el crédito
SECUENCIALOFICINA	INTEGER	Clave Foránea
COBRAPORROL	NUMBER (1)	Identifica si se cobra el crédito a través del rol
SEREAJUSTA	NUMBER (1)	Identifica si es reajustable la tasa

CONTINÚA

DIASREAJUSTE	INTEGER	Número de días de reajuste en caso de ser reajustable
FECHACORTE	DATE	Fecha que se procesa los cálculos correspondientes
TEACONSEGURO	NUMBER (18,2)	Identificador de tasas
TEASINSEGURO	NUMBER (18,2)	Identificador de tasa
ESVIGENTECLASIFICACION	NUMBER (1)	Identificador de clasificación
BLOQUEOTRANSACCIONOPERATIVA	NUMBER (1)	Indica si el crédito bloquea transacciones operativas
NUMEROVERIFICADOR	INTEGER	Valor de control
FECHAMODIFICA	DATE	Ultima fecha que se realiza modificaciones
IDENTIFICACIONSUJETOORIGINAL	NVARCHAR2(20)	Identificación del deudor
FACTORCALCULO	INTEGER	Tasa

2. La información de los campos de la tabla Prestamo_Cliente la cual es una tabla intermedia entre PrestamoMaestro y Cliente se muestran a continuación:

Tabla 14.
Diccionario Préstamo Cliente

PRESTAMO_CLIENTE		
NOMBRE	TIPO	DESCRIPCIÓN
SECUENCIAL	INTEGER	Clave Primaria
SECUENCIALPRESTAMO	INTEGER	Clave Foránea
SECUENCIALCLIENTE	INTEGER	Clave Foránea
ESPRINCIPAL	NUMBER (1)	Identificador de si es principal
ESTAACTIVO	NUMBER (1)	Indicador de esta activo
NUMEROVERIFICADOR	INTEGER	Numero interno de control

3. Los campos correspondientes a la tabla Persona, la misma que está relacionada directamente con la tabla clientes, y que se encuentra en el esquema FBS_Personas se muestra a continuación:

Tabla 15.
Diccionario Persona

PERSONA		
NOMBRE	TIPO	DESCRIPCIÓN
SECUENCIAL	INTEGER	Clave Primaria

CONTINÚA

IDENTIFICACION	NVARCHAR2(20)	Numero identificación de la persona
NOMBRE UNIDO	NVARCHAR2(200)	Nombre completo unido
DIRECCION DOMICILIO	NVARCHAR2(500)	Dirección domiciliaria
REFERENCIA DOMICILIARIA	NVARCHAR2(600)	Referencias Domiciliarias
EMAIL	NVARCHAR2(100)	Correo electrónico
SECUENCIALDIVPOLRESIDENCIA	INTEGER	Clave Foránea
CODIGOPAISORIGEN	NVARCHAR2(10)	Identificador del país de origen
NUMEROVERIFICADOR	INTEGER	Numero para control interno
SECUENCIALDIVACTIVIDADECON	INTEGER	Clave Foránea

CONTINÚA

FECHA ULTIMA ACTUALIZACION		DATE	Fecha que fue actualizada la información
CODIGO USUARIO ACTUALIZA		NVARCHAR2(30)	Asesor que actualiza la información
ES SOCIO		NUMBER (1)	Indicador de si es socio

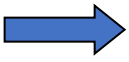
4. La tabla Persona_Natural se encuentra en el esquema FBS_Personas, se encuentra relacionada con la tabla Personas y almacena información correspondientes a personas naturales, sus campos son descritos a continuación:

Tabla 16.
Diccionario Persona Natural

PERSONA_NATURAL		
NOMBRE	TIPO	DESCRIPCIÓN
SECUENCIALPERSONA	INTEGER	Clave Foránea
APELLIDOS	NVARCHAR2(100)	Apellidos del cliente
NOMBRES	NVARCHAR2(100)	Nombres del cliente
FECHANACIMIENTO	DATE	Fecha de nacimiento

CONTINÚA

ESMASCULINO	NUMBER (1)	Indicador del genero
CODIGOESTADOCIVIL	NCHAR (1)	Estado civil del cliente
CODIGOTIPOEDUCACION	NCHAR (1)	Clave foránea
CODIGOTIPOVIVIENDA	NCHAR (1)	Clave foránea
CODIGOPROFESION	NCHAR (4)	Clave foránea
EGRESOSMENSUALES	NUMBER (18,2)	Monto de egresos mensuales
PATRIMONIO	NUMBER (18,2)	Patrimonio del cliente
APELLIDOPATERNO	NVARCHAR2(80)	Apellido paterno del cliente
APELLIDOMATERNO	NVARCHAR2(80)	Apellido materno del cliente
LUGARNACIMIENTO	NVARCHAR2(500)	Lugar de nacimiento del cliente

CONTINÚA 

NUMEROPAPELETAVOTACION	NVARCHAR2(10)	Numero de papeleta de votación del cliente
CARGASFAMILIARES	INTEGER	Numero de cargas familiares
ACTIVOSTOTALES	NUMBER (18,2)	Monto total de activos
PASIVOSTOTALES	NUMBER (18,2)	Monto total de pasivos
BAJOMODALIDADDATOSMINIMOS	NUMBER (1)	Identificador del número de campos que se almacena
ESTAENPEPS	NUMBER (1)	Identificador de si es peps

5. Para la obtención de información del comportamiento de pago de los clientes se hizo uso de la tabla PrestamoComponenteCartera la cual registra a través de cada uno de los componentes la forma prevista de cancelación de rubros, se encuentra en el esquema FBS_Cartera directamente relacionado al Préstamo y a la tabla de componentes. Su descripción se muestra a continuación:

Tabla 17.
Diccionario Préstamo Componente Cartera

PRESTAMOCOMPONENTE_CARTERA		
NOMBRE	TIPO	DESCRIPCIÓN
SECUENCIAL	INTEGER	Clave Primaria
SECUENCIALPRESTAMO	INTEGER	Clave Foránea
SECUENCIALCOMPONENTECARTERA	INTEGER	Clave Foránea
VALORPROYECTADO	NUMBER (18,6)	Valor que se proyecta a cobrarse
VALORCALCULADO	NUMBER (18,6)	Valor calculado para cobrar
VALORCOBRADO	NUMBER (18,6)	Valor que se cobro
FECHAINICIO	DATE	Fecha de inicio de la cuota
FECHAVENCIMIENTO	DATE	Fecha de vencimiento de la cuota
NUMEROCUOTA	INTEGER	Numero de cuota
CODIGOESTADOPRESTAMOCOMPONENTE	NVARCHAR2(10)	Clave Foránea
FACTORCALCULO	NUMBER (18,6)	Tasa
DIASCALCULADOS	INTEGER	Días calculados
FECHAMODIFICA	DATE	Ultima fecha de modificación

c) Exploración de los datos

Es posible identificar dentro de la base relacional de la institución datos que independientemente del grupo de clientes del que se trate son relevantes para construir un modelo predictivo de riesgo y rentabilidad, las variables requeridas para el desarrollo del presente proyecto y que fueron descritas en el apartado de Recolección de datos Iniciales.

En la tabla 18 se puede apreciar un enlace entre las variables requeridas para el análisis y los campos desde donde se puede obtener la información dentro de la base de datos relacional de la cooperativa.

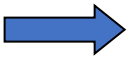
Tabla 18.

Exploración de datos

SQUEMA	TABLA	CAMPO	VARIABLE
FBS_GENERALES	DIVISION	NOMBRE	Región
FBS_GENERALES	DIVISION	NOMBRE	Provincia
FBS_PERSONAS	PERSONA_NATURAL	ESMASCULINO	Sexo
FBS_PERSONAS	PERSONA_NATURAL	CODIGOESTADOCIVIL	Estado Civil
FBS_PERSONAS	PERSONA_NATURAL	CARGASFAMILIARES	Cantidad de Cargas
FBS_PERSONAS	PERSONA_NATURAL	CODIGOTIPOEDUCACION	Nivel Educacional
FBS_PERSONAS	PERSONA	SECUENCIALDIV ACTIVIDADECON	Actividad Económica
FBS_PERSONAS	PERSONA_NATURAL	CODIGOPROFESION	Profesión
FBS_PERSONAS	PERSONA_NATURAL	FECHANACIMIENTO	Edad.
FBS_CLIENTES	CLIENTE	FECHASISTEMA	Antigüedad del cliente
FBS_CARTERA	PRESTAMOMAESTRO	CODIGOTIPOPRESTAMO	Tiene Créditos no Hipotecarias

CONTINÚA

FBS_CARTERA	PRESTAMOMAESTRO	FECHAADJUDICACION	Fecha más antigua de otorgamiento de crédito no hipotecario y vigente
FBS_CARTERA	PRESTAMOMAESTRO	SALDOACTUAL	Monto original adeudado en créditos no hipotecarios y vigentes
FBS_CARTERA	PRESTAMOMAESTRO	CODIGOTIPOPRESTAMO	¿Tiene créditos no hipotecarios?
FBS_CARTERA	PRESTAMOMAESTRO	FECHAADJUDICACION	Fecha más antigua de otorgamiento de créditos hipotecarios vigentes
FBS_CARTERA	PRESTAMOMAESTRO	VALORENTREGADO	Monto original adeudado en créditos hipotecarios y vigentes
FBS_CARTERA	PRESTAMOMAESTRO	NUMEROCUOTAS	Plazo máximo de créditos hipotecarios y vigentes
FBS_CARTERA	PRESTAMO COMPONENTE_CARTERA	VALORCOBRADO	Ingresos generados por concepto de intereses, comisiones, y otros ingresos en créditos no hipotecarios

CONTINÚA 

FBS_CARTERA	PRESTAMO COMPONENTE_CARTERA	VALORCOBRADO	Ingresos generados por concepto de intereses comisiones y otros ingresos en créditos hipotecarios
FBS_NEGOCIOS FINANCIEROS	MOVIMIENTO	FECHA	En que tramo de morosidad se encuentra el cliente
FBS_CARTERA	PRESTAMO COMPONENTE_CARTERA	VALORCALCULADO	Monto en mora
FBS_CARTERA	PRESTAMO COMPONENTE_CARTERA	CODIGOESTADO	Presenta deuda vencida
FBS_CARTERA	PRESTAMO COMPONENTE_CARTERA	CODIGOESTADO	Presenta capital castigado
FBS_NEGOCIOS FINANCIEROS	MOVIMIENTO	FECHA	Máximo de días de mora
FBS_CAPTACIONES VISTA	CUENTAMAESTRO	CODIGOESTADOCUENTA	Posee cuentas cerradas
FBS_CARTERA	PRESTAMOMAESTRO	CODIGOTIPOPRESTAMO	Posee créditos reestructurados
FBS_CARTERA	PRESTAMOMAESTRO	CODIGOTIPOPRESTAMO	Posee créditos refinanciados

La información a ser utilizada ha sido extraída a través de una consulta SQL aplicada directamente a la base de datos, el tiempo que se ha demorado el equipo en extraer la data ha sido de 3 horas debido a la cantidad de registros y al cruce de tablas necesario para obtener la información. Dicha información se obtuvo en un archivo de Excel de donde se registraron 144323 registros, correspondiente únicamente a comportamientos de pago del componente de capital de los distintos préstamos otorgados en un rango de tiempo.

Verificación de datos

En el proceso de verificación de datos, se analiza los datos obtenidos desde el modelo relacional a través de la consulta sql desarrollada en el punto de exploración de datos, dicha información debe ser consistente, desde donde se puede asegurar la completitud y su consistencia.

En el proceso de verificación se realizó las validaciones siguientes:

- ✓ Revisión de las llaves primarias, foráneas y los atributos en la base de datos.
- ✓ Análisis de los atributos que permiten elaborar conclusiones y que caen en conflicto con el sentido común.
- ✓ Chequeo de coincidencia entre los significados de los atributos con los valores que contienen.
- ✓ Validación del delimitador utilizado para la extracción de los datos desde la bdd hacia el archivo plano.
- ✓ Contraste de datos entre datos globales e individuales.

Se determina que el modelo relacional de base de datos manejada en la institución mantiene todas sus tablas estructuradas través de llaves primarias foráneas y claves únicas, cumple con reglas de normalización, y no existe valores duplicados.

En el proceso de ingreso de datos existen validaciones que evitan en su gran mayoría el ingreso de data inconsistente. Con lo que se puede constatar que la data es correcta.

4.3.3. FASE 3: Preparación de los datos

La figura 22 muestra las diferentes fases que deben ser llevadas a cabo en la etapa de preparación de los datos.

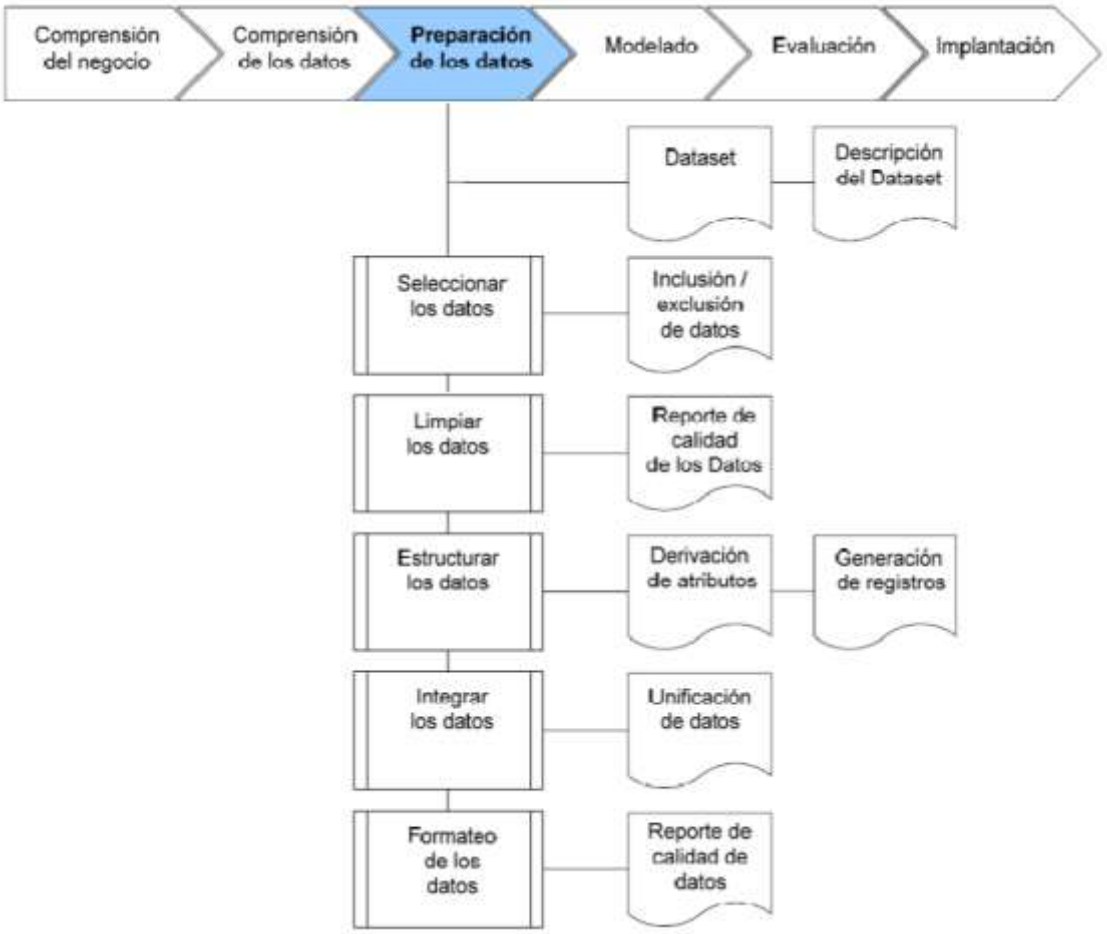


Figura 22. Fase de Preparación de los datos (Metodología Crisp-DM)
Fuente: (Crisp-DM,2000)

a) Selección de los datos

Los datos proporcionados por la institución para el desarrollo del presente proyecto están ofuscados ocultándose la información de nombres, identificaciones personales, entre otros, por concepto del sigilo bancario.

La data que se seleccionó para el desarrollo del presente proyecto fue proveniente de los últimos 4 años, es decir desde el año 2014 hasta el año 2018, para la obtención de dicha data se

consultó en los datos históricos de la institución, dicha información es almacenada en Blades en el centro de datos a través de un servidor virtualizado. En esta etapa se tomó en cuenta los datos relevantes y necesarios de análisis para el cumplimiento de los objetivos tanto comerciales como de minería de datos, se utilizó un total de 144323 registros para el proceso de análisis.

b) Limpieza de los datos

El sistema utilizado por la institución es desarrollado in-house, por lo que se acopla totalmente a las necesidades de la cooperativa, para el desarrollo del proyecto se utilizaron datos en archivo xls los cuales fueron obtenidos a través de una consulta SQL hacia la base de datos, a través de la herramienta Knime se realizaron los siguientes trabajos de limpieza en los datos obtenidos:

- ✓ En los casos de créditos que no se encontraban en mora aparecieron campos en blanco en columnas como *monto en mora* en dichos campos se rellenó con cero.
- ✓ En el campo *Fecha más antigua del otorgamiento de crédito vigente* aparecieron campos vacíos para los casos en que en que las personas no disponían créditos vigentes para esos casos hemos rellenado con el valor “01/01/1900”
- ✓ En el campo *Monto Original de crédito vigente* aparecieron campos vacíos para los casos en que las personas no disponían créditos vigentes para esos casos hemos rellenado con cero
- ✓ En el campo *Plazo máximo de crédito vigente* aparecieron campos vacíos para los casos en que en que las personas no disponían créditos vigentes para esos casos hemos rellenado con cero
- ✓ En el campo *monto en mora* han aparecido valores vacíos para los casos en que las personas no se encontraban en mora en esos casos hemos rellenado con cero

- ✓ En el campo edad existen valores inferiores a 18 años lo cual se considera proviene de un error de ingreso de información, dichos registros fueron suprimidos.

Se utilizaron componentes como Missing Value para realizar el llenado de los datos faltantes

Figura 23. Componente Missing Value
Fuente: Knime

La configuración que se realizó dentro del componente se muestra en la figura siguiente:

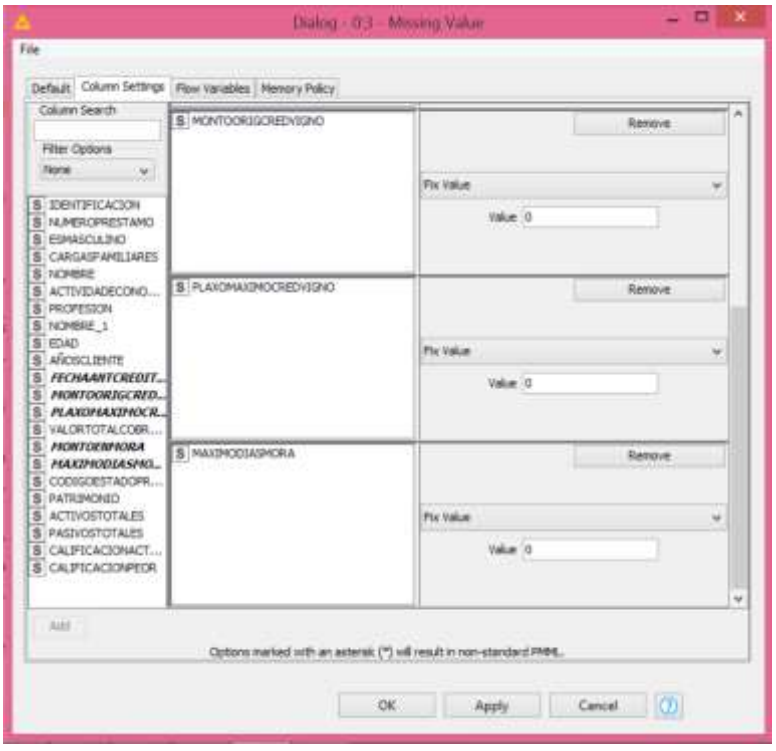


Figura 24. Configuración Missing Value
Fuente: Knime

Se utilizó el componente Row Filter que se muestra a continuación para realizar el filtrado de datos.

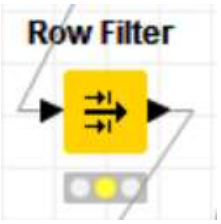


Figura 25. Componente Row Filter
Fuente: Knime

En el componente Row Filter, se filtró únicamente las personas que cumplen el rango de edad entre los 18 a 99 años ya que se detectó datos mal ingresados con edades que no se encontraban en ese rango, su configuración se muestra en la imagen siguiente.

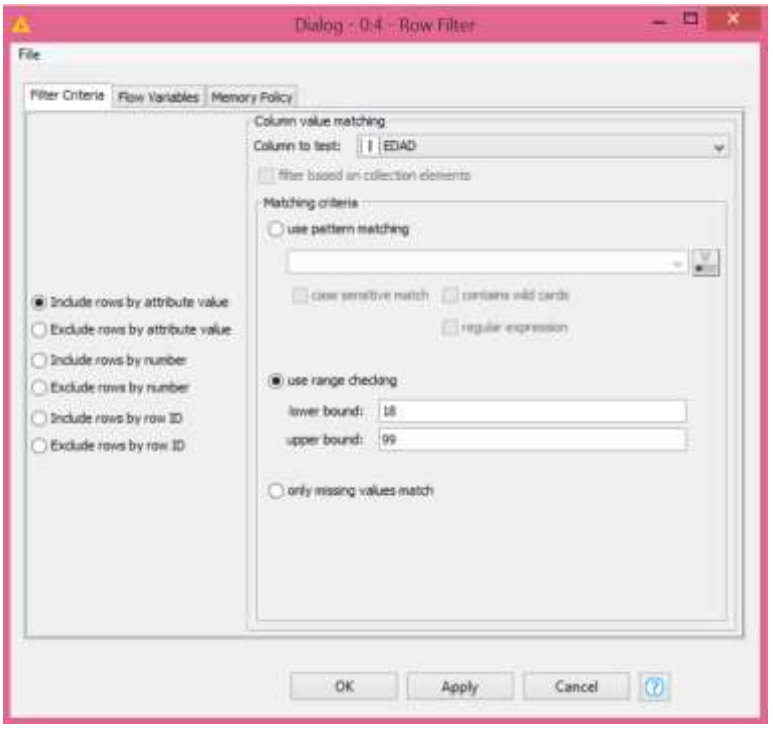


Figura 26. Configuración Row Filter
Fuente: Knime

La figura siguiente muestra el proceso de limpieza de datos que se realizó.

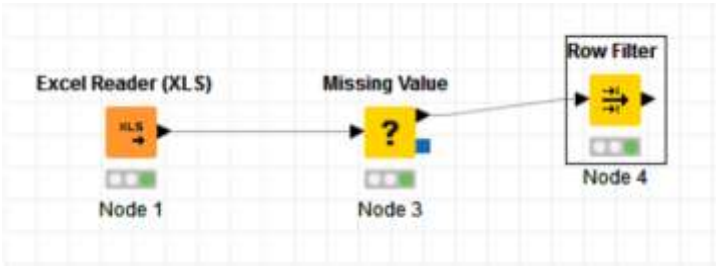


Figura 27. Flujo Limpieza de datos

c) Estructura de los datos

En la fase de estructuración de los datos se han generado atributos a partir de otros campos que nos permitiesen de mejor manera aplicar minería da datos, ya que los datos obtenidos fueron de créditos, se realizó una agrupación según identificación para poder determinar campos como:

- ✓ ¿Posee Créditos Hipotecarios?
- ✓ ¿Presenta capital castigado?
- ✓ Sumatoria del valor total cobrado
- ✓ Monto total en mora
- ✓ Posee deuda vencida
- ✓ Sumatoria de valor total adeudado

Para realizar el proceso de agregado de columnas de acuerdo a lo requerido se utilizó el componente String Manipulation, cuya imagen se muestra a continuación

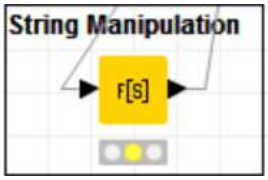


Figura 28. Componente StringManipulation
Fuente: Knime

La configuración del componente String Manipulation para realizar el agregado de las diferentes variables.

1. En la siguiente figura se muestra la configuración del componente String Manipulation a través del cual se crea la variable “tieneCreditosHipotecarios”

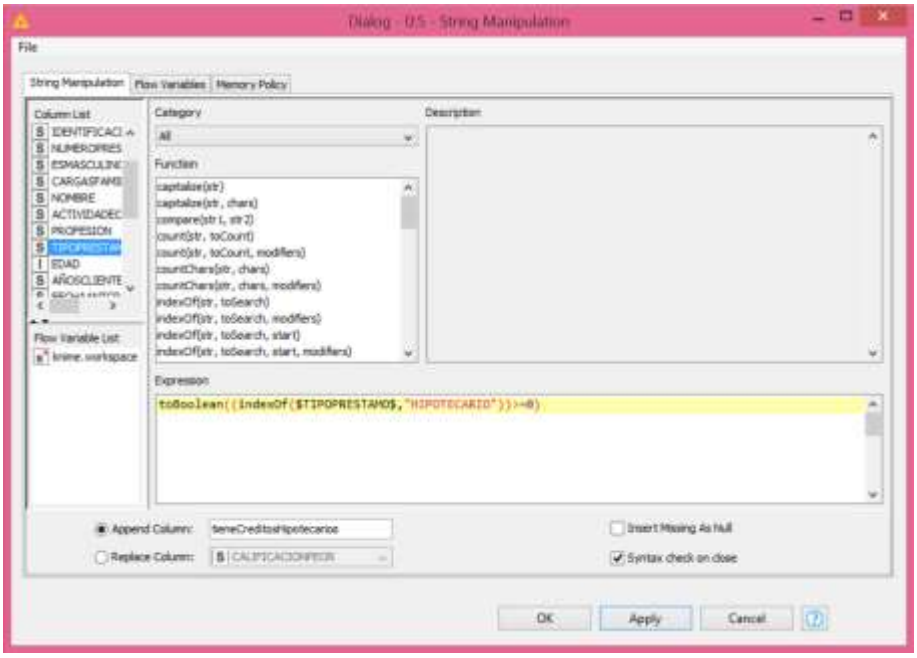


Figura 29. Creación Variable TieneCreditoHipotecario
Fuente**:** Knime

2. En la imagen siguiente se puede apreciar la configuración del componente string manipulation para la creación de la variable “PoseeDeudaVencida”

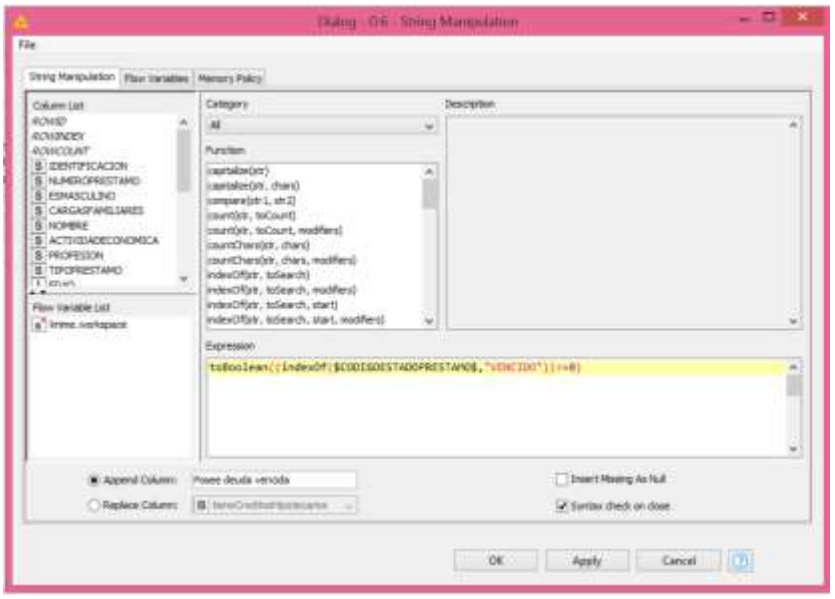


Figura 30. Creación Variable Posee deuda vencida
Fuente: Knime

3. En la figura siguiente se puede apreciar la configuración del componente string manipulation para la creación de la variable “presentaCapitalCastigado”

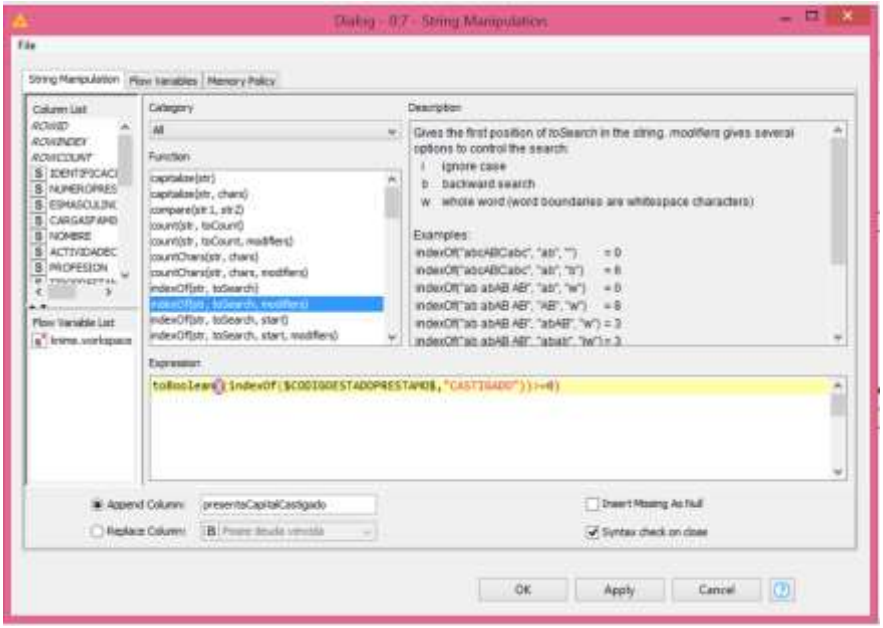


Figura 31. Creación de variable presentaCapitalCastigado
Fuente: Knime

Para obtener las sumatorias de montos totales adeudados y cancelados se realizó una agrupación por el identificador de cliente mediante el componente GROUP BY, se muestra el icono en la figura siguiente.

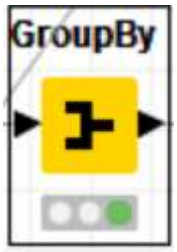


Figura 32. Componente Group by
Fuente: Knime

4. En la figura siguiente se puede observar la configuración necesaria del componente GroupBy para poder realizar la sumatoria del monto en mora, y del valor total cobrado agrupado por cedula.

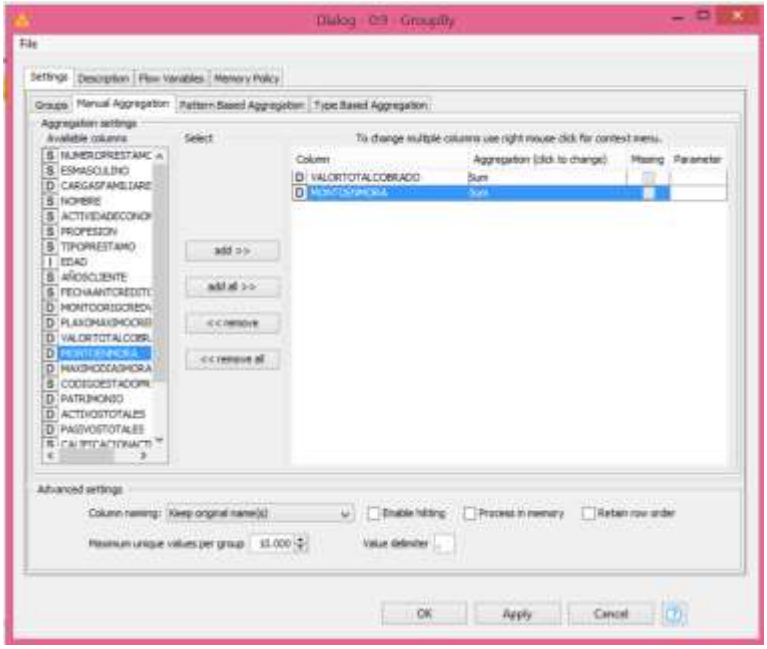


Figura 33. Configuración SumatoriaValorTotalCobrado
Fuente**:** Knime

Posterior a utilizar el componente Group By los datos se encuentran limpios y con agrupados por cliente para la siguiente fase.

En la figura siguiente se puede observar el flujo de datos completo que se aplicó.

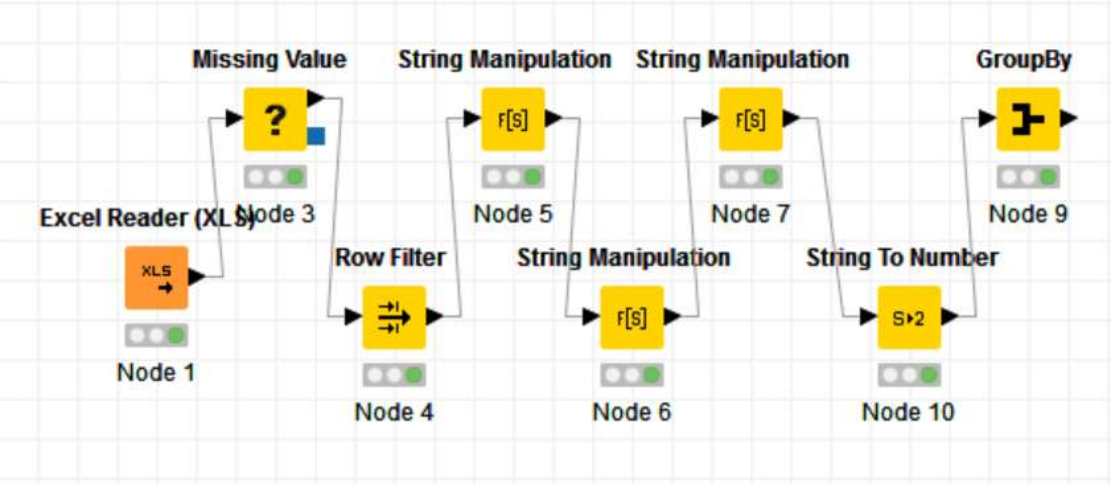


Figura 34. Flujo Estructura de datos

d) Integración de los datos

Se utilizó únicamente la información de la base de datos institucional proveniente de la fuente de datos Oracle por lo que no se realizó la integración con otras fuentes.

e) Formateo de los datos

En esta fase se realizó la transformación de datos numéricos a categorías es decir que se discretizaron los datos, se realizaron discretización de los campos:

- ✓ Edad
- ✓ Años como cliente

- ✓ Patrimonio
- ✓ Activos totales
- ✓ Máximo de días en mora

Para la realización de este proceso se utilizó la herramienta knime a través del componente Numeric Binner, cuya imagen se muestra a continuación

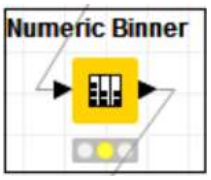


Figura 35. Componente Numeric Binner
Fuente: Knime

Se realizó una agrupación en bandas para el número máximo de días de mora en segmentos de 1-30, 30-60, 61-90 y mayor a 120 como se muestra en la imagen siguiente:

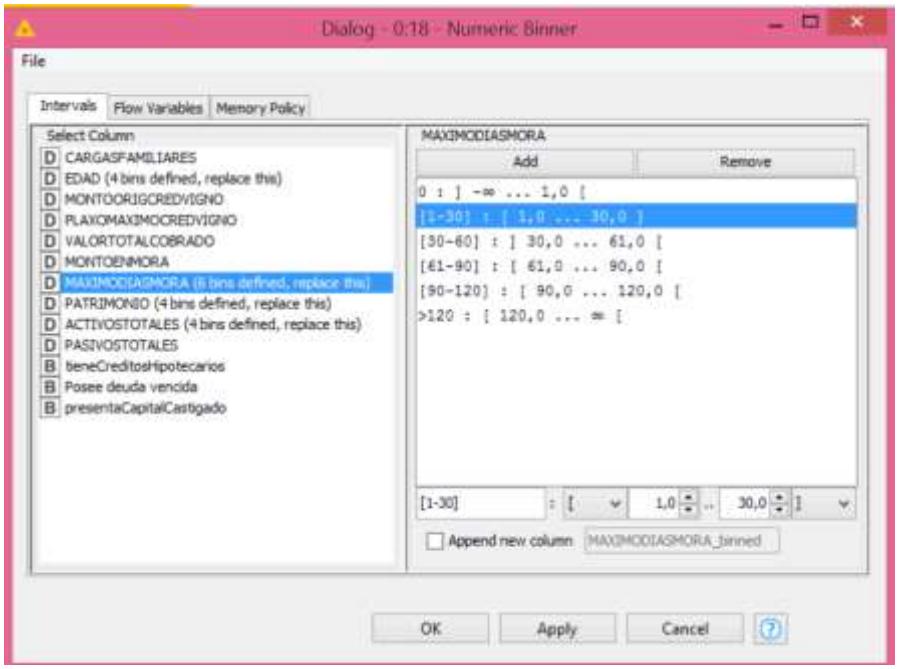


Figura 36. Discretizacion Variable DiasMora
Fuente: Knime

También se discretizó la variable edad en segmentos adulto joven, adulto, adulto mayor, considerando que un adulto joven se encuentra en el rango entre 18 a 35 años, un adulto en el rango 35 a 60 y adulto mayor quienes superan los 60 años, la configuración se muestra en la imagen siguiente:

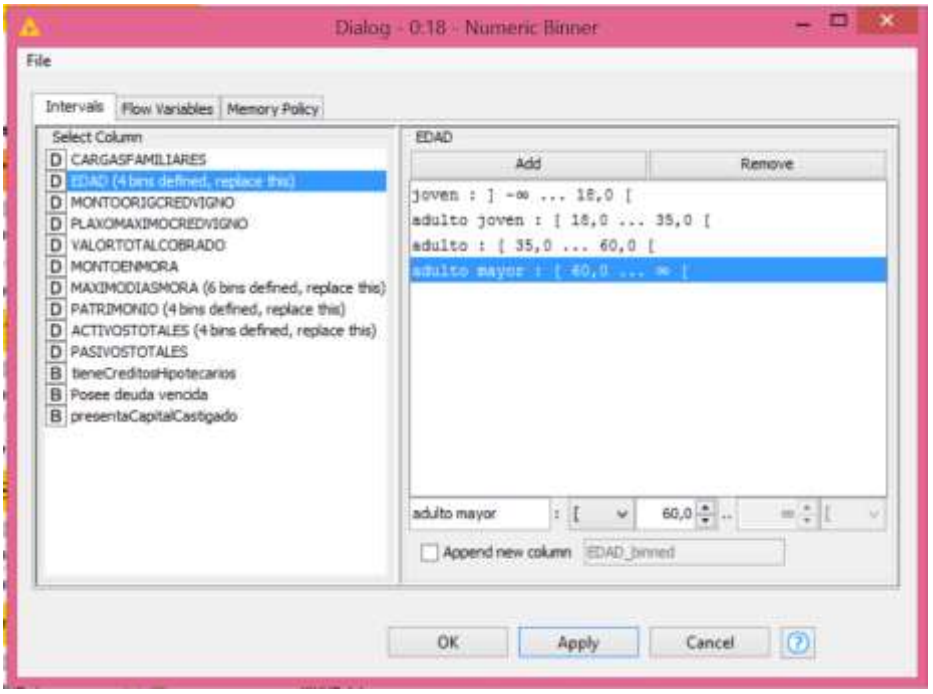


Figura 37. Discretización Variable Edad
Fuente: Knime

El género se mantenía en 0 y 1, se lo normalizo a masculino y femenino para una mayor visión, tal como se muestra en la imagen siguiente:

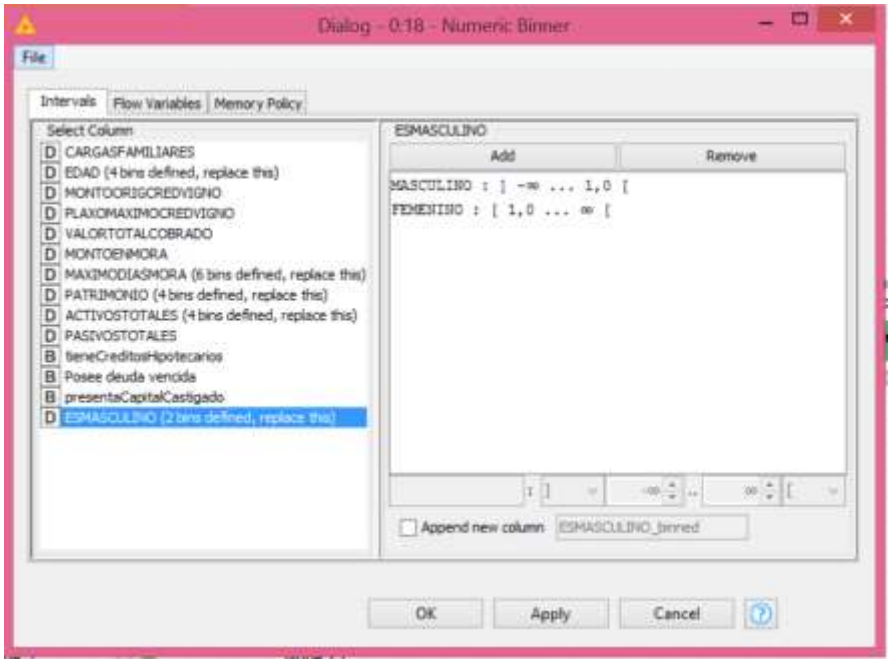


Figura 38. Discretización Variable Genero
Fuente: Knime

Se realizó una discretización según la cantidad de años que forma parte de la institución como cliente, clasificándolos en bajo, medio y alto, la figura siguiente muestra su configuración en el componente Numeric Binner

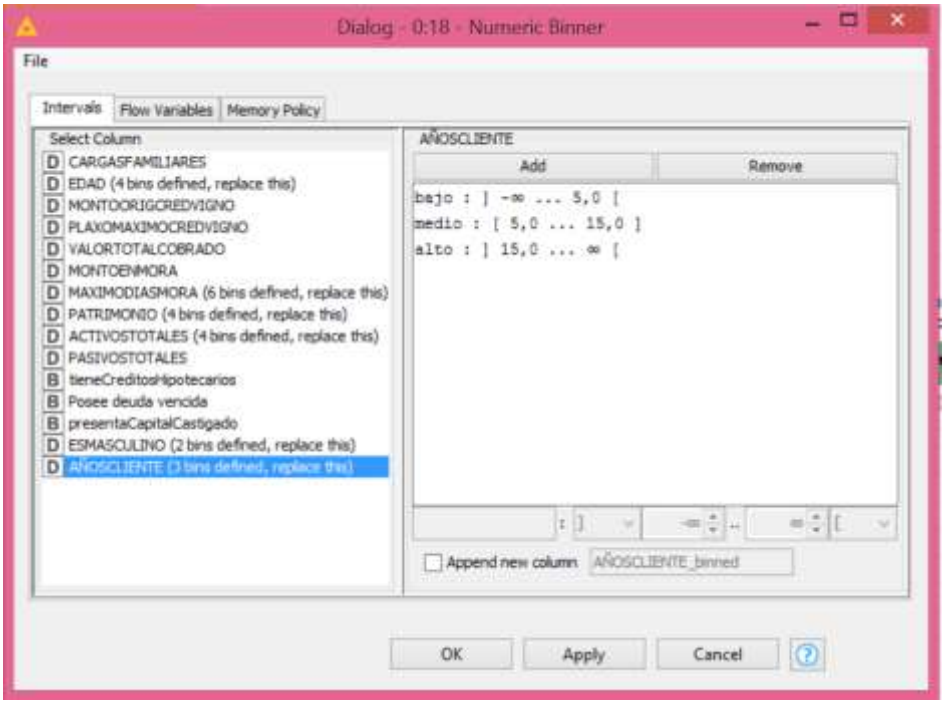


Figura 39. Discretización Variable Anos Cliente
Fuente: Knime

Las calificaciones de riesgo que se manejan en las instituciones financieras se detallan en la tabla siguiente:

Tabla 19.
Calificaciones Propias Riesgo

A1	**Créditos de riesgo normal categoría A-1**
A2	Créditos de riesgo normal categoría A-2
A3	Créditos de riesgo normal categoría A-3
B1	Créditos con riesgo potencial categoría B-1
B2	Créditos con riesgo potencial categoría B-2
C1	Créditos deficientes categoría C-1

CONTINÚA

C2	Créditos deficientes categoría C-2
D	Créditos de dudoso recaudo categoría D
E	Pérdidas categoría E

Fuente: SB –Estructuras del Sistema de Operaciones Activas y Contingentes

Para el desarrollo del presente proyecto se han determinado como clientes aptos aquellos que se encuentran con calificación de riesgo normal, los demás serán determinados como clientes no aptos. Para realizar dicho proceso de normalización se ha utilizado el componente Rule Engine, cuya figura representativa se muestra a continuación.

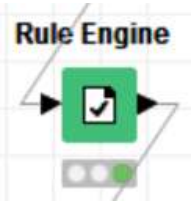


Figura 40. Componente Rule Engine
Fuente**:** Knime

Este nodo toma una lista de reglas definidas e intenta hacerlas coincidir con cada fila en la tabla de entrada. Si una regla coincide, su valor de resultado se agrega a una nueva columna.

En la figura siguiente se puede observar el flujo desde la carga de datos hasta el proceso de formateo de datos.

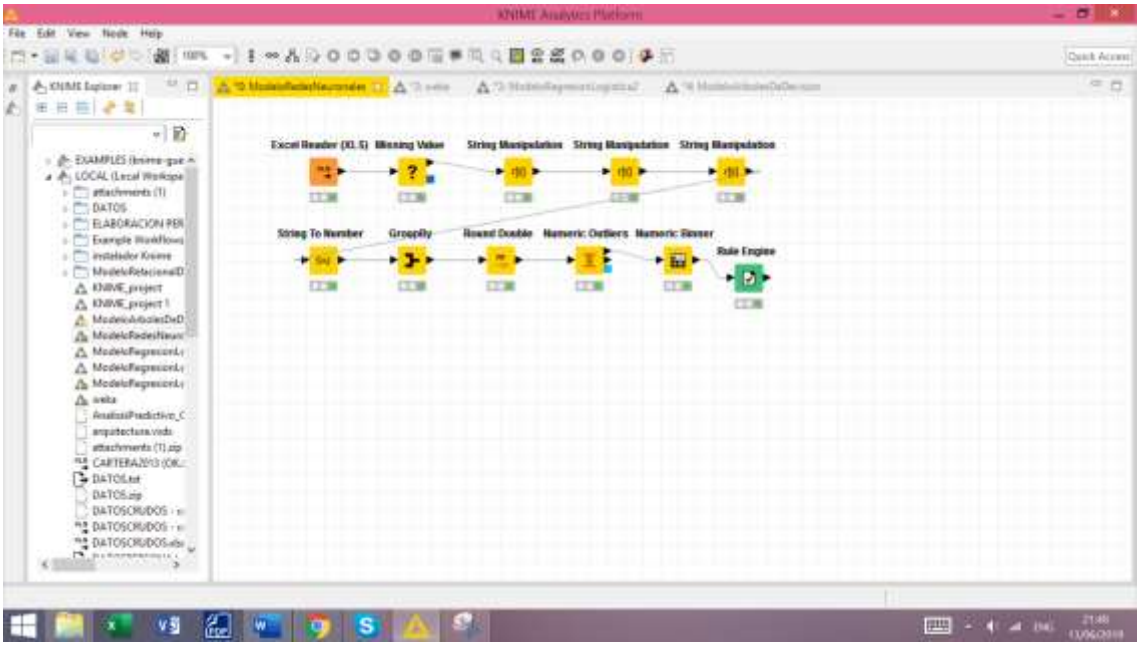


Figura 41. Formateo de los datos

4.3.4. FASE 4: Modelado

En la figura 42 denota cada uno de los pasos a seguir para cumplir con la etapa de modelado.

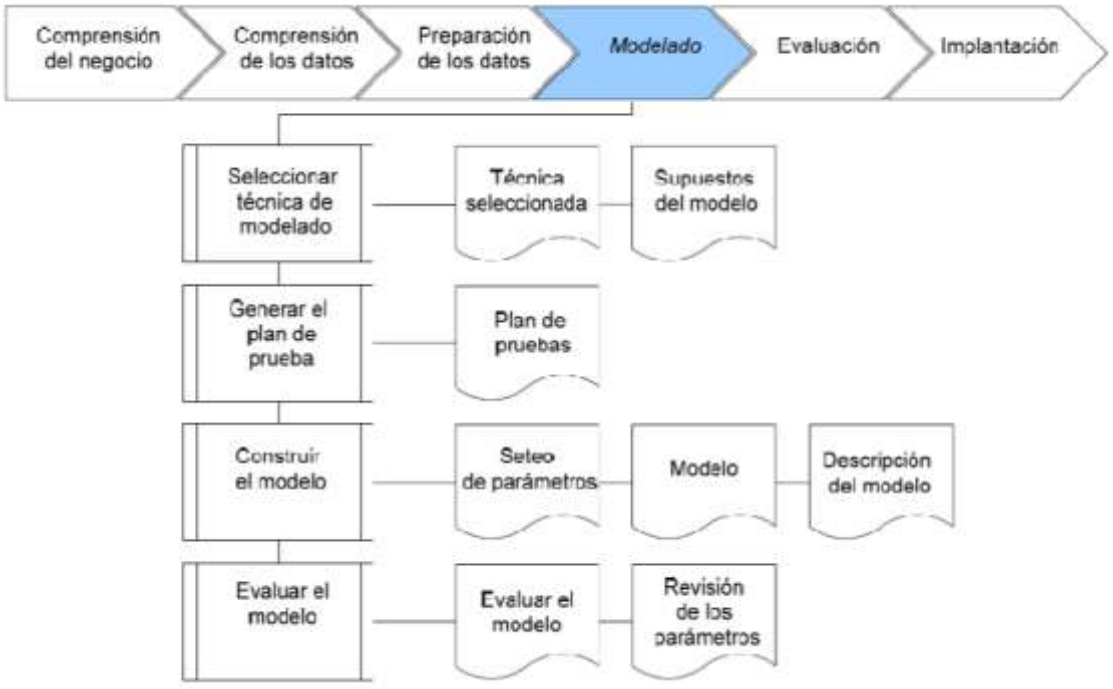


Figura 42. Fase de Modelado (Metodología Crisp-DM)
Fuente: (Crisp-DM,2000)

a) Selección de Técnicas

En esta fase es preciso seleccionar la técnica de minería de datos que más se ajuste a nuestras necesidades en este caso, se requiere una técnica de predicción de las cuales las que más se ajustan al problema de enfocar las campañas bancarias son:

- ✓ Regresión Logística ya que esta técnica es utilizada cuando la variable dependiente es binaria, como es el caso de la clasificación de los clientes en buenos y malos en términos crediticios o clientes con potencial y sin potencial en términos de rentabilidad.
- ✓ Redes neuronales ya que pueden representar cualquier tipo de función, incluyendo funciones probabilísticas y lógicas, y en particular pueden ser utilizadas para clasificar clientes y para predecir su comportamiento. Para asignar a las clientes notas de

rentabilidad será necesario dividir el output de la red neuronal en varios tramos, y asignar una nota particular si el output se encuentra dentro de un tramo

- ✓ Árboles de decisión ya que pueden ser traducidos en términos de conjuntos de reglas fácilmente comprensibles y a su vez pueden ser transformadas en políticas de crédito o manuales de procedimientos.
- ✓ Weka J48 ya que permite predecir la variable de destino de un nuevo registro de conjunto de datos, este componente nos permitirá encontrar patrones comunes en clientes no rentables.

Las tres son soportadas por la herramienta de minería de datos seleccionada la cual es Knime.

Las técnicas seleccionadas fueron descritas a detalle en el apartado 2.4.1.2.2 del presente documento.

b) Generación del plan de prueba

En esta sección se trata de valorar que el modelo desarrollado sea válido y efectivo, para evaluar el modelo se utilizará la técnica Cross Validation, en primera instancia en el software utilizado Knime aplicamos el componente **X-Partitioner,** cuya imagen se denota abajo.

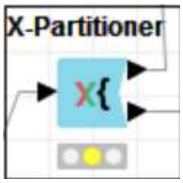


Figura 43. Elemento X-Partitioner
Fuente: Knime

Este componente es incorporado en la herramienta seleccionada el cual permite dividir en dos partes una de entrenamiento y otra de prueba al conjunto de datos seleccionado, de donde en el presente desarrollo el 80% de datos se determinará para el entrenamiento y el 20% para la prueba.

En la imagen siguiente se puede apreciar la configuración realizada en el componente X-Partitioner donde se señala 20 validaciones con un muestreo randómico

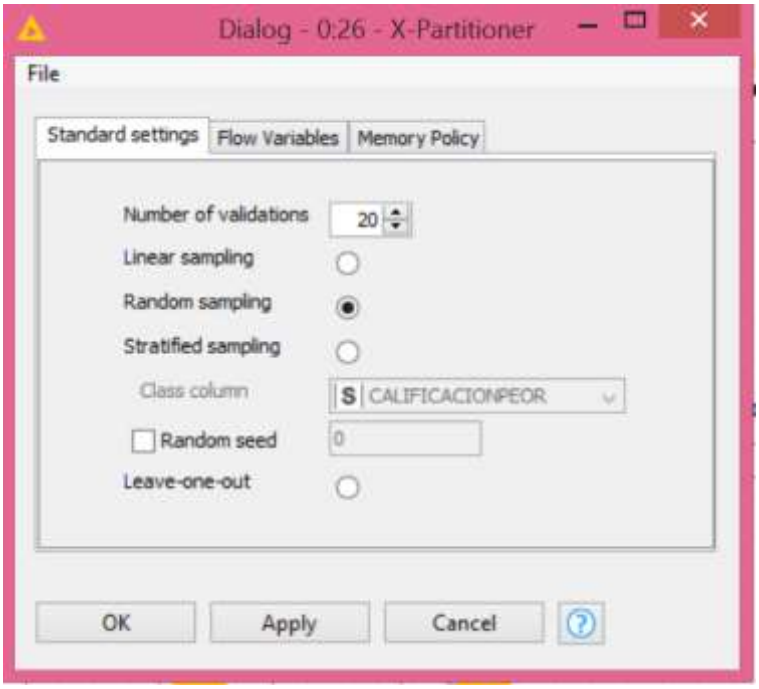


Figura 44. Configuración elemento X-Partitioner
Fuente: Knime

Se utiliza el componente X-Agregator el cual agrega el resultado de la validación cruzada, se muestra en la imagen siguiente

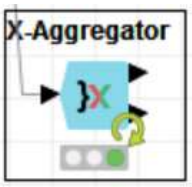


Figura 45. Elemento X-Aggregator
Fuente: Knime

En la siguiente grafica se puede visualizar la configuración ingresada en el componente X-Aggregator, señalando la variable objetivo que se CALIFICACIONPEOR y la variable predictora que es PREDICTION (CALIFICACION PEOR)

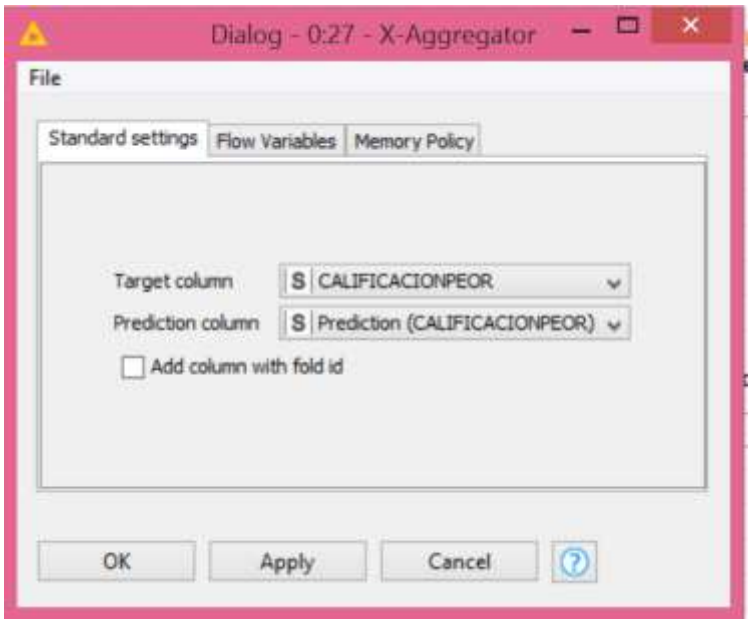


Figura 46. Configuración Elemento X-Agregator
Fuente: Knime

Finalmente, se utilizará también la matriz de confusión para medir la calidad del modelo, para lo cual se utilizar el componente ¨**Scorer**¨, se muestra en la imagen siguiente.

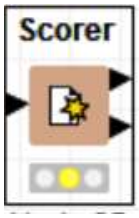


Figura 47. Componente Scorer
Fuente: Knime

c) Construcción del modelo

En esta sección se realiza la creación de modelos de minería de datos bajo la herramienta Knime, los cuales nos permitirán llegar al cumplimiento de nuestros objetivos. Se han seleccionado cuatro técnicas que serán aplicadas las cuales son: Arboles de decisión, Regresión Lineal, Weka J48 y Redes Neuronales

Árboles de Decisión

El presente modelo se ha desarrollado utilizando la técnica de minería de datos Arboles de Decisión, el cual permite cumplir los siguientes objetivos propuestos:

- ✓ Generación de datos de entrenamiento y prueba
- ✓ Detección de patrones de morosidad de clientes
- ✓ Determinación de si un cliente es bueno o malo.

Para el desarrollo del presente modelo se utilizaron los componentes siguientes:

X-Partitioner

El cual permite determinar el número de iteraciones de validación que se debe realizar, en este proceso la muestra original se divide en submuestras de igual tamaño, del total de las

submuestras, una sola se retiene como datos de validación para validar el modelo y las demás se retienen para entrenamiento, la imagen siguiente corresponde a la representación del componente X-Partitioner.

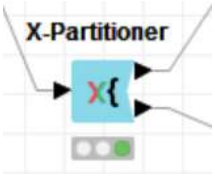


Figura 48. Elemento X-Partitioner
Fuente: Knime

Para el desarrollo del presente modelo se ha realizado la configuración siguiente:

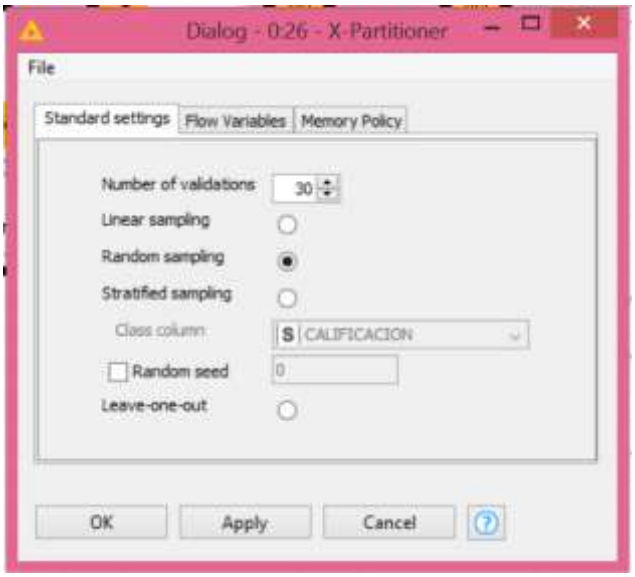


Figura 49. Configuración Elemento X-Partitioner
Fuente: Knime

Number of validations: Indica el número de iteraciones de validación cruzada que se deben realizar, en este caso hemos registrado el número 30 lo que implica que se utilizara el 30% de los datos para prueba y el 70% restante para el entrenamiento.

Decision Tree Learner el cual introduce un árbol de decisión en memoria, la variable objetivo debe ser nominal en nuestro caso la clasificación del cliente como bueno o malo, los demás atributos que se toman en cuenta son nominales y también numéricos, la imagen siguiente muestra el componente.

Figura 50. Componente Decision Tree Learner
Fuente: Knime

La configuración que se realiza en la herramienta fue la que se muestra en la siguiente figura:

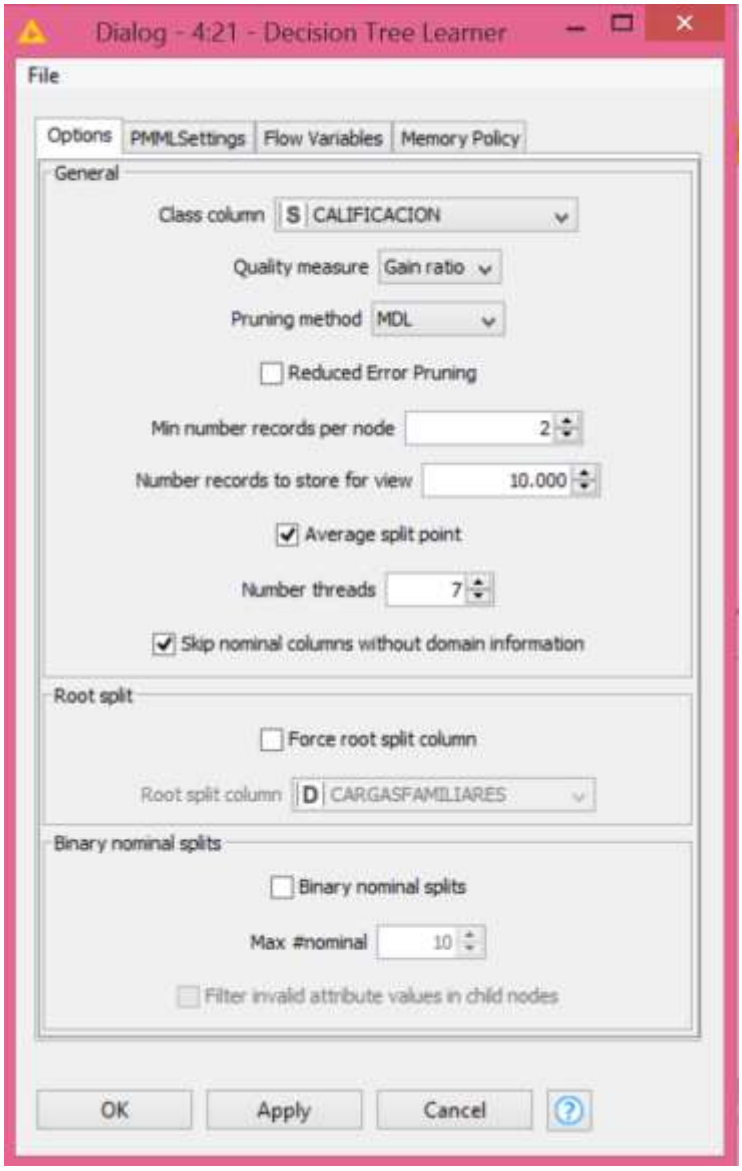


Figura 51. Configuración herramienta Decisión Tree Learner
Fuente: Knime

En la gráfica se observa las siguientes opciones:

class column la cual requiere se seleccione la columna a predecir en nuestro caso se ha seleccionado la columna CALIFICACION la cual nos indica si un cliente es bueno o malo.

Quality measure permite seleccionar el índice de calidad, se puede seleccionar gini index o gain ratio, para nuestro caso hemos seleccionado gain ratio.

Pruning method como método de poda se ha utilizado MDL, la poda reduce el tamaño del árbol y evita el ajuste excesivo, lo que aumenta el rendimiento de la generalización y, por lo tanto, la calidad de la predicción.

Min number record per node el número mínimo de registro por nodo se ha ingresado el 2

Number records to store for view el nivel de profundidad del árbol se han registrado 1000 niveles

Number of threads se registra 7 como número de hilos

Decision Tree Predictor el cual usa un árbol de decisión existente para predecir el valor de la clase para los nuevos patrones, la imagen siguiente muestra el componente en la herramienta.

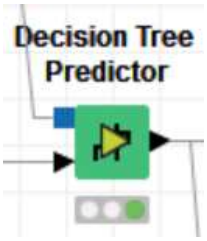


Figura 52. Elemento Decision Tree Predictor
Fuente: Knime

Decision Tree to Ruleset en español arboles de decisiones para el conjunto de reglas, este componente convierte (un único) modelo de árbol de decisión en el modelo de regla de PMML (variedad de técnicas de estadísticas) y también en una tabla que contiene las reglas en una forma textual.

Las reglas resultantes son independientes entre sí, el orden de las reglas no se especifica, se puede cambiar, la imagen siguiente representa al componente en mención.

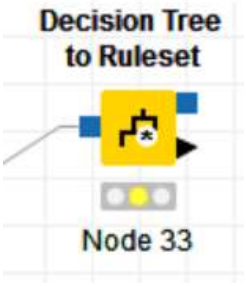


Figura 53. Elemento Decision Tree to Ruleset
Fuente: Knime

El modelo de árbol de decisión realizado se detalla a continuación en la figura 54.

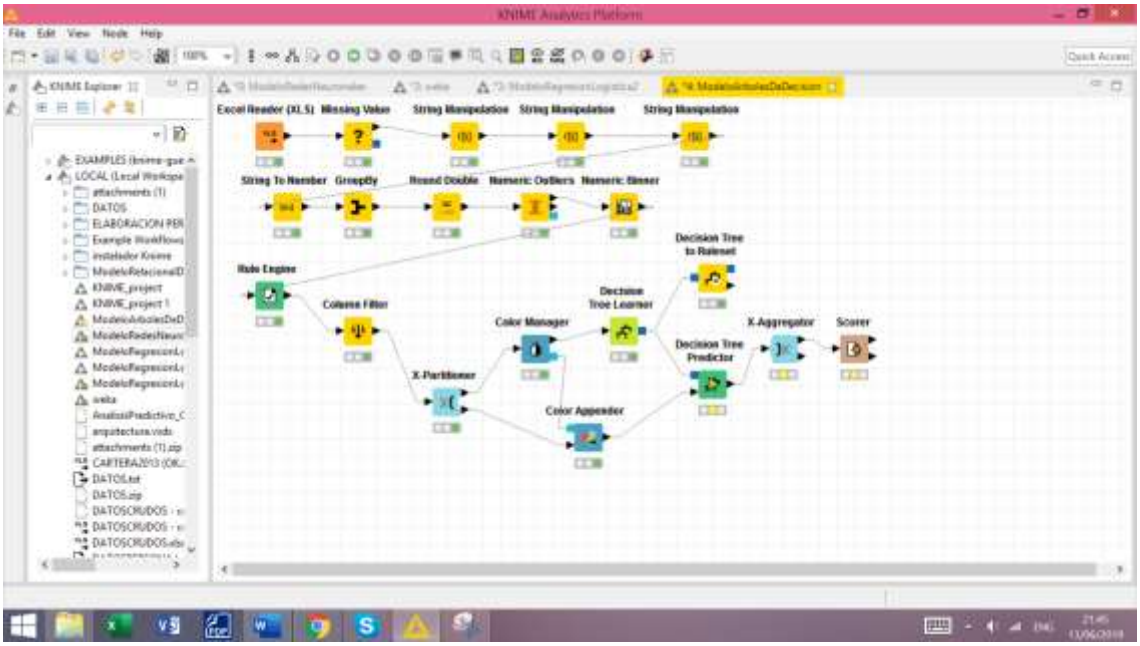


Figura 54. Modelo Arboles de decisión
Fuente: Knime

En la gráfica siguiente como resultado de la aplicación del algoritmo se observa el árbol de clasificación de acuerdo al atributo de mayor peso que son los clientes con potencial, de la clasificación general de los socios con cartera dentro de la institución el 80.8% son clientes con potencial y el 19.2% son clientes sin potencial.

En la gráfica siguiente se visualiza el árbol de decisión devuelto por la herramienta, allí se observa que una variable dominante en la determinación es el capital adeudado por el socio, de tal forma que si presenta capital castigado inmediatamente se le reconoce como cliente no rentable, de la clasificación realizada el 100% de las personas que poseen capital castigado son malos clientes.

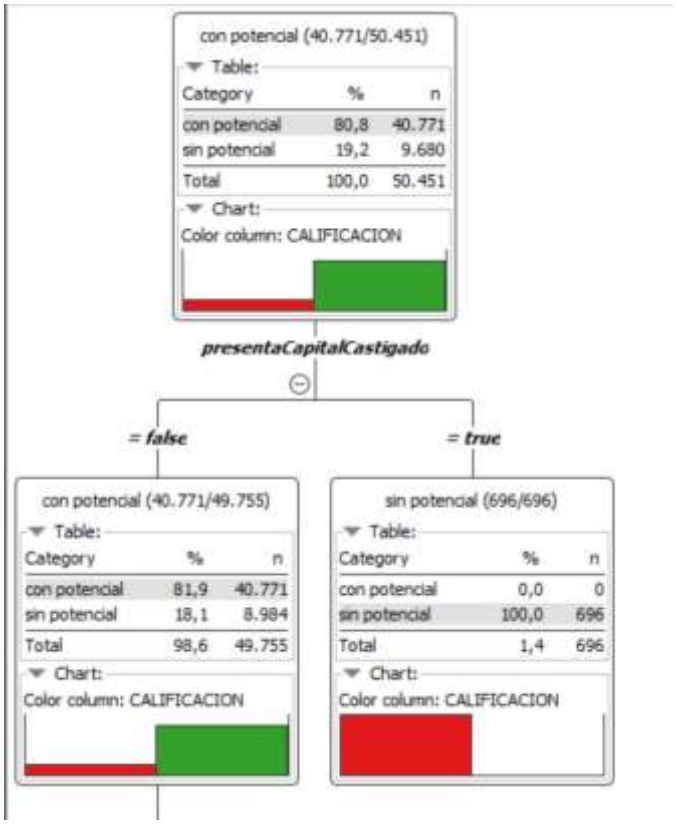


Figura 55. Árbol de Decisión Parte Superior
Fuente: Knime

La siguiente variable importante en el análisis es MaximoDiasMora esta variable recoge la cantidad máxima de días en mora del socio en las diferentes cuotas de la obligación de crédito, donde se visualiza que mientras menor fue la mora mayor cantidad de clientes se clasificaron como potenciales, lo cual puede ser visualizado en la figura 56:

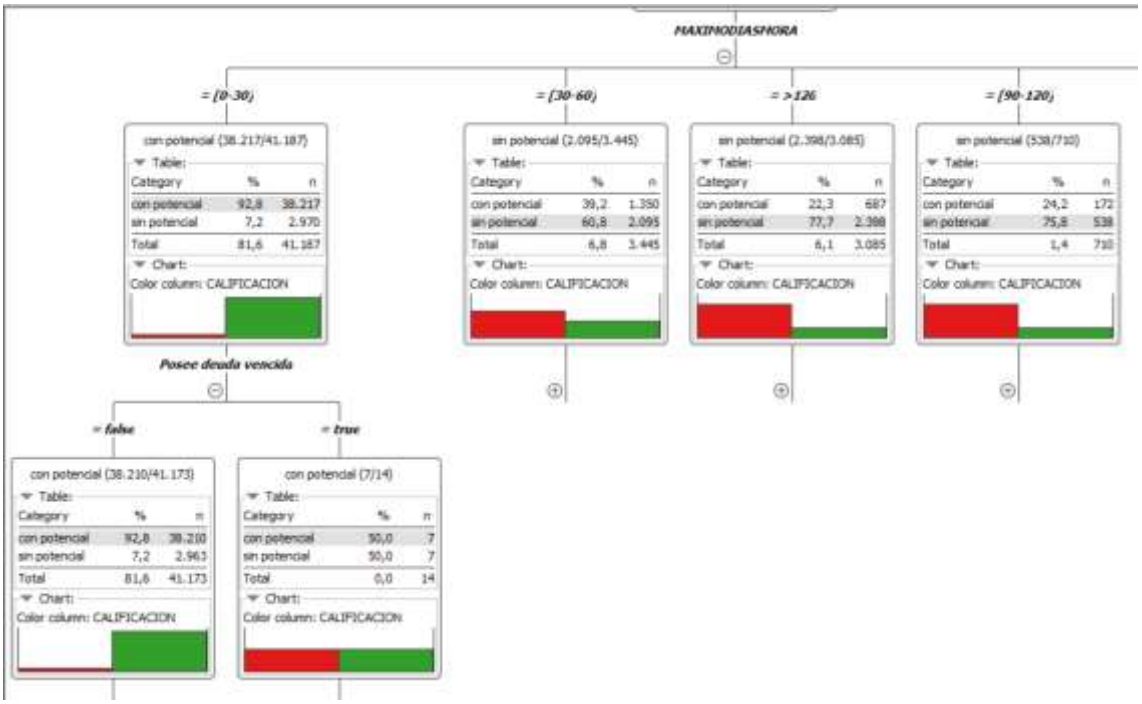


Figura 56. Árbol Decisión Parte Media
Fuente: Knime

Posterior en el árbol aparece la variable MONTOENMORA la cual almacena el monto actual que posee el socio en mora, se visualiza que mientras mayor es el monto en mora menos clientes son clasificados como potenciales, se pude visualizar en la figura 57:

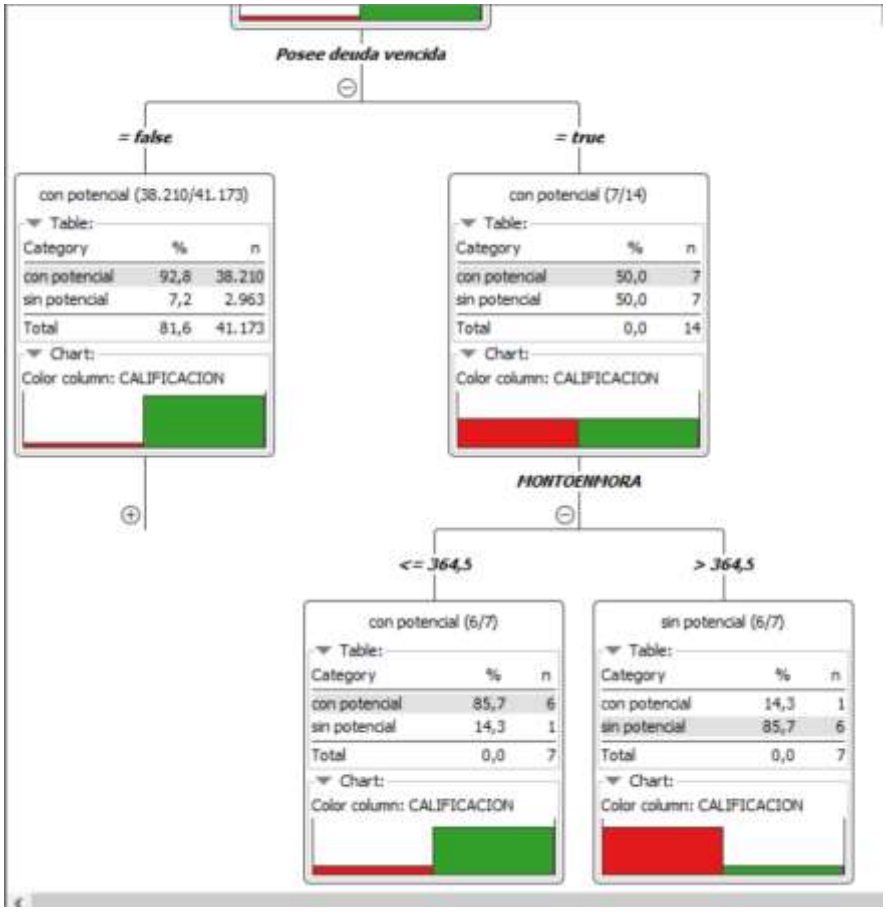


Figura 57. Árbol de Decisión Parte Inferior
Fuente: Knime

Como se observa en las gráficas anteriores la distribución de datos está determinado por la cantidad máxima de días que haya caído en mora el socio, los años que pertenece a la institución, los pasivos, la edad.

Otro de los objetivos de la minería de datos es determinar cuáles son los patrones para que un cliente sea no rentable, para el cumplimiento de este objetivo se utiliza el componente **Decision Tree to rule set**, el cual devolvió como resultado un conjunto de reglas que se muestra en la tabla 20.

Tabla 20.
Reglas para determinar clientes

Se generaron un total de 39 reglas de las cuales la que posee mayor cantidad de aciertos son las siguientes:

$MONTOENMORA$ <= 2816.0 AND $Posee deuda vencida$ = 0.0 AND $MAXIMODIASMORA$ = "[0-30]" AND $presentaCapitalCastigado$ = 0.0 => "con potencial"

$VALORTOTALCOBRADO$<= 14147.0 AND $MONTOORIGCREDVIGNO$ <= 54205.0 AND $VALORTOTALCOBRADO$ <= 40953.5 AND $MAXIMODIASMORA$ = "[30-60]" AND $presentaCapitalCastigado$ = 0.0 => "sin potencial"

Regresión Logística

Este método estadístico es utilizado con frecuencia cuando la variable dependiente es binaria, como es el caso de la clasificación de los clientes en buenos y malos en términos crediticios

Para realizar el modelo de regresión logística se utilizó el componente Logistic Regresion Learner, el cual realiza una regresión logística multinomial, además que permite seleccionar qué

solucionador se debe usar para el problema, las dos opciones que se pueden aplicar son (Stochastic average gradient (Gradiente Promedio Estocastico) o Iteratively reweighted least squares (Cuadrados minimos ponderados iterativamente)).

Cuadrados mínimos ponderados iterativamente

Este solucionador utiliza un enfoque de optimización iterativo que a veces también se denomina puntuación de Fisher, para calcular el modelo. Funciona bien para tablas pequeñas con solo ver columnas, pero falla en tablas más grandes. Es el solucionador más propenso a errores porque no puede calcular un modelo si los datos son linealmente separables, este solucionador tampoco es capaz de manejar tablas donde hay más columnas que filas porque no admite la regularización.

Gradiente promedio estocástico (SAG)

Este solucionador implementa una variante del descenso del gradiente estocástico que tiende a converger considerablemente más rápido que el descenso del gradiente estocástico de vainilla. Funciona bien para tablas grandes y también para tablas con más columnas que filas.

La figura que la representa se muestra a continuación

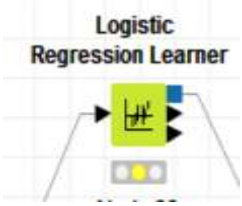


Figura 58. Elemento Logistic Regresion Learner
Fuente: Knime

La configuración que se aplicó al componente fue utilizando el método de solución Stochastic average gradient, se seleccionaron todas las variables para la predicción, su configuración se muestra a continuación.

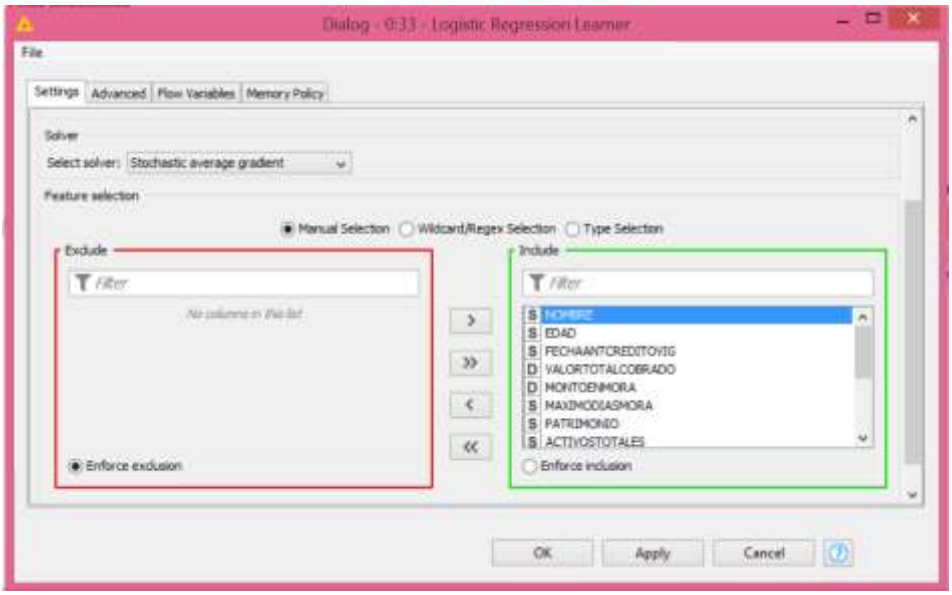


Figura 59. Configuración Elemento Logistic Regression Learner
Fuente: Knime

Se configuraron 100 como número máximo de iteraciones, una épsilon de 1.0E-5, una regularización de Gass con una varianza del 0.1, en la figura 60 se puede observar cómo se parametrizo el componente.

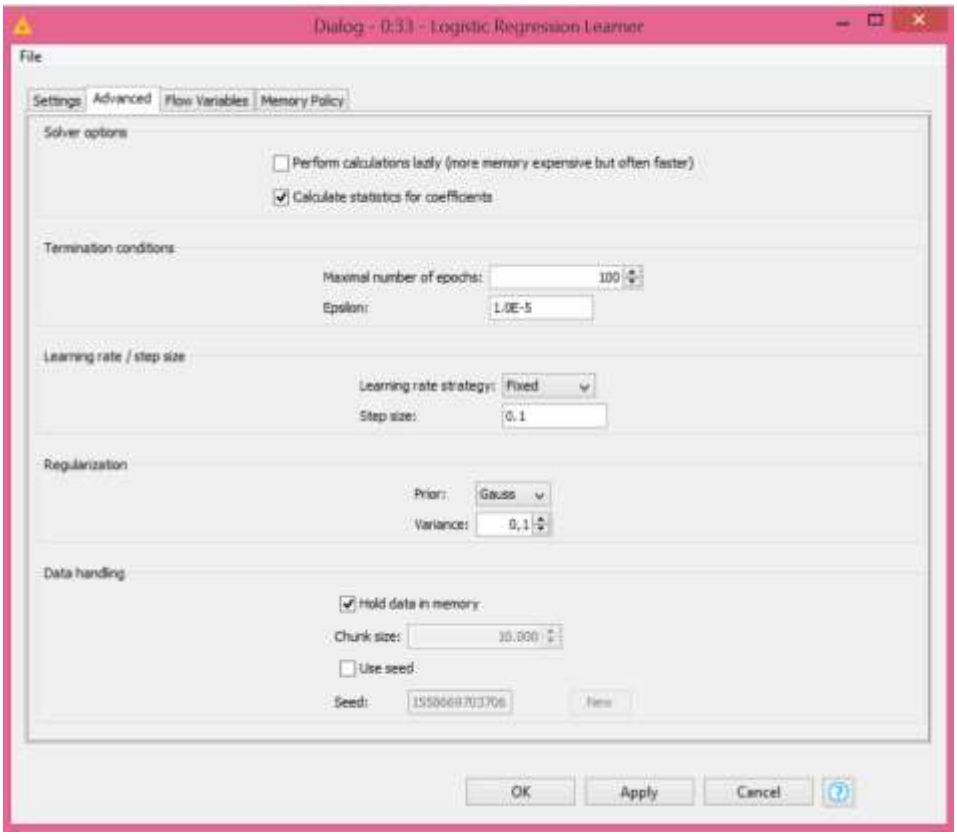


Figura 60. Configuración Avanzada Logistic Regression Learner
Fuente: Knime

El nodo de aprendizaje utilizado en el presente modelo es Logistic Regresion Predictor el cual predice la respuesta utilizando un modelo de regresión logística. El nodo está conectado a un modelo de nodo de regresión logística y algunos datos de prueba. Solo es ejecutable si los datos de prueba contienen las columnas que utiliza el modelo de aprendiz. Este nodo agrega una nueva columna a la tabla de entrada que contiene la predicción para cada fila.

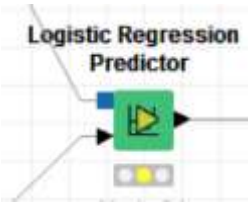


Figura 61. Elemento Logistic Regression Predictor
Fuente**:** Knime

Dentro de este componente se realizó el registro de la variable a predecir en este caso la calificación de riesgo

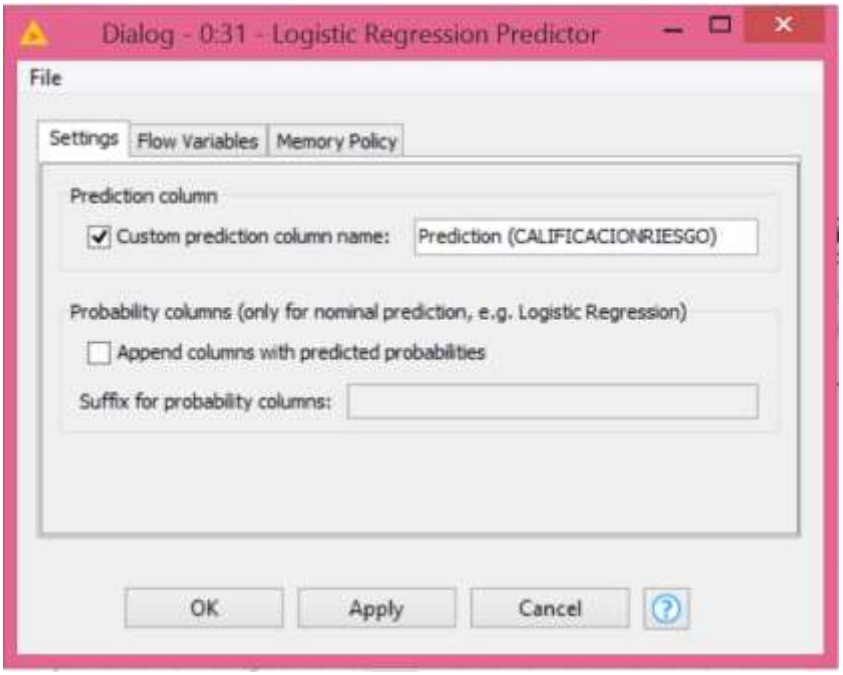


Figura 62. Configuración elemento Logistic Regression Predictor
Fuente: Knime

Se utilizó el nodo X-Partitioner el cual permite definir el porcentaje de entrenamiento y el porcentaje de prueba para el presente modelo se ha definido el 10% de los valores de la muestra

como datos de prueba y el 90% de datos restantes se utilizan como datos de entrenamiento. En la figura 63 se puede observar la configuración ingresada en el componente

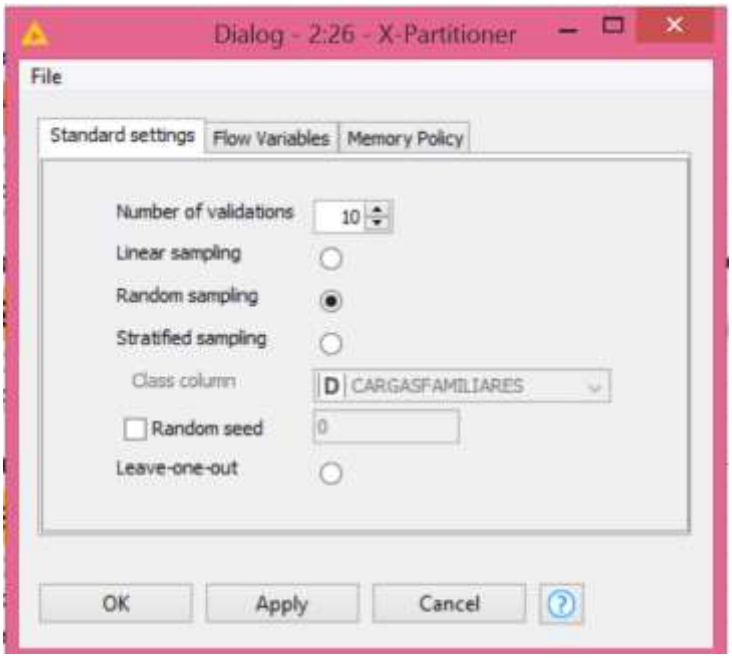


Figura 63. Configuración elemento X-Partitioner
Fuente: Knime

Recopila el resultado de un nodo predictor, compara la clase predicha y la clase real y genera las predicciones para todas las filas y las estadísticas de iteración, a continuación, se puede observar la configuración interna realizada dentro del componente X-Aggregator

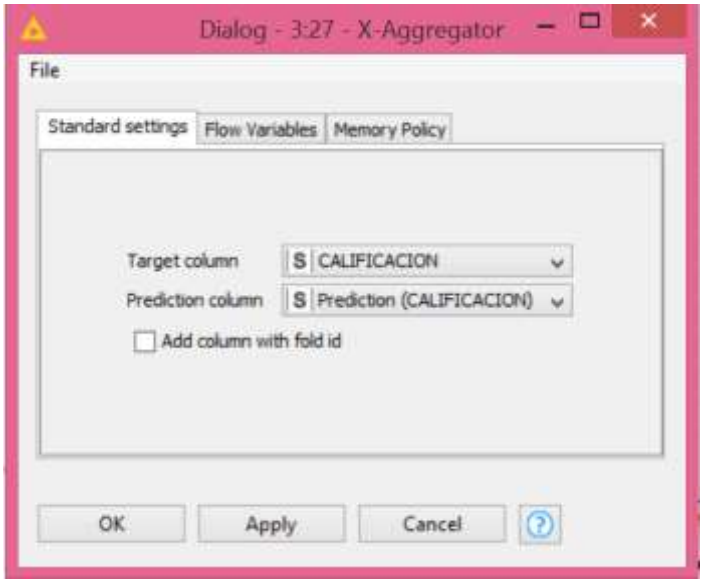


Figura 64. Configuración nodo X-Aggregator
Fuente: Knime

El modelo de regresión logística final se muestra en la figura 65.

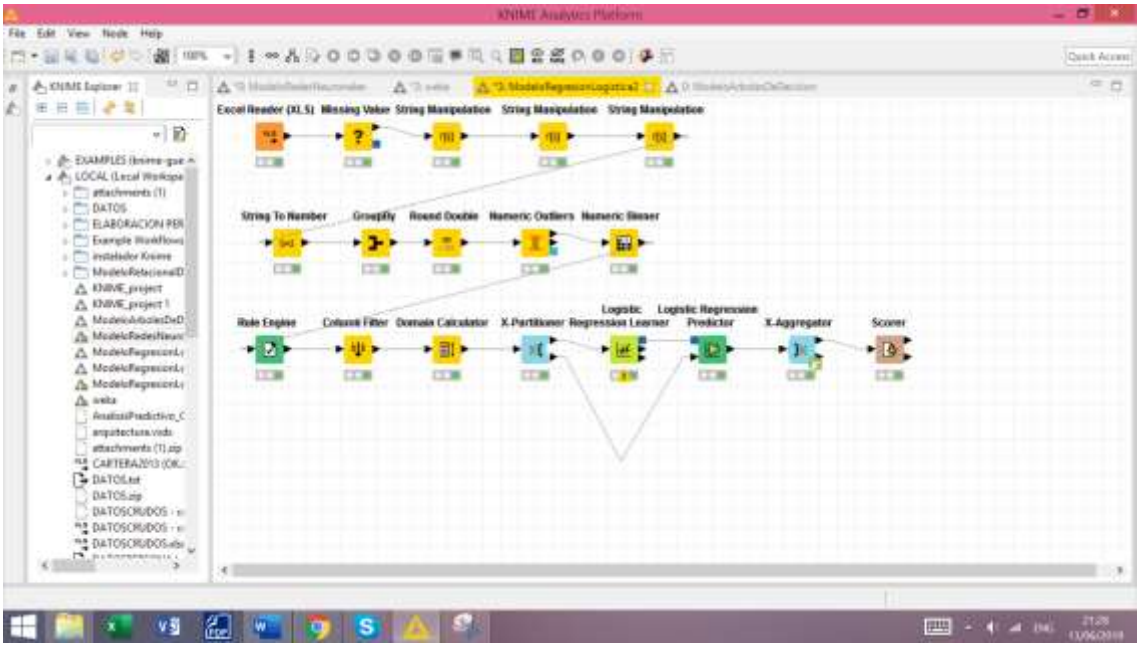


Figura 65. Modelo de regresión logística final
Fuente: Knime

Con esta configuración se obtuvo un porcentaje de acierto del 63.28% y un error del 36.7%

Redes Neuronales

En la figura 66 se muestra el modelo aplicado a las redes neuronales:

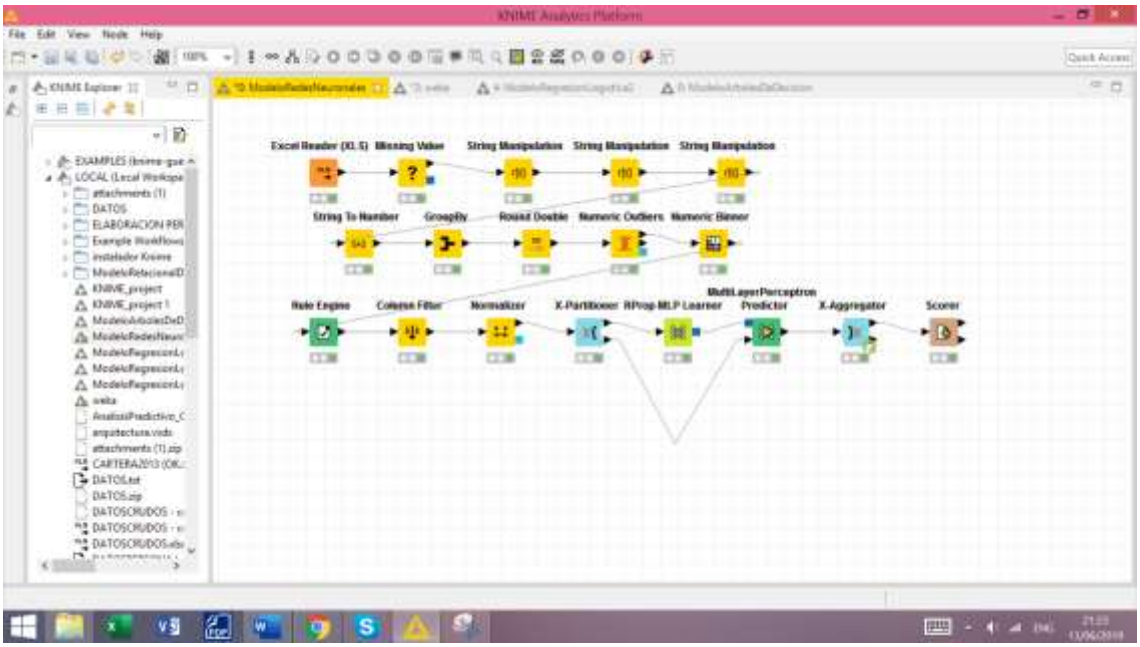


Figura 66. Modelo Redes Neuronales
Fuente: Knime

Para lo cual se utilizó el componente RProp MLP Learner el cual realiza una adaptación local de las actualizaciones de peso de acuerdo con el comportamiento de la función de error, se lo puede observar en la figura 67.

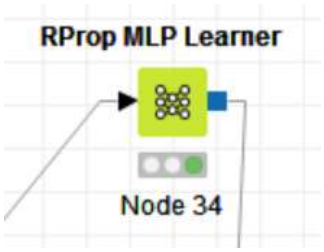


Figura 67. Elemento Multilayer Perceptron Predictor
Fuente: knime

La configuración interna que se realizó dentro de la herramienta es la que se muestra en la figura 68.

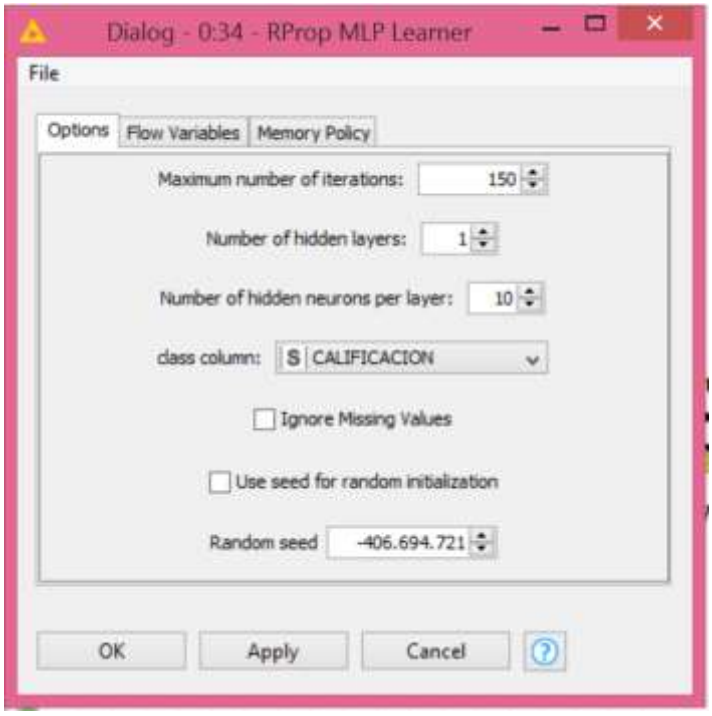


Figura 68. Configuración componente RProp MLP Learner
Fuente: Knime

Se utilizó también el componente MultiLayerPerceptronPredictor el cual se muestra en la figura 69.

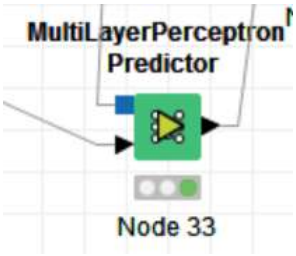


Figura 69. Nodo Multilayer Perceptron Predictor
Fuente**:** Knime

Sobre la base de un modelo MultiLayerPerceptron capacitado dado en el informe de modelo de este nodo, se calculan los valores de salida esperados. Si la variable de salida es nominal, se producen la salida de cada neurona y la clase de la neurona ganadora. De lo contrario, se calcula el valor de regresión.

Se utilizo el componente X-Partitioner a travez del cual se definió el 10% para datos de prueba y el 90% para entrenamiento

Con este modelo se obtuvo un 83.181% de acierto y 16.81% de error

WEKA J48

El algoritmo J48 de weka es un algoritmo de clasificación que nos permitirá determinar los patrones para que un cliente sea no rentable para la institución, en la elaboración del modelo se utilizaron los siguientes componentes:

Partitioning

El componente Partitioning permite dividir la tabla de entrada en dos particiones, en este será utilizado para generar datos de entrenamiento y datos de prueba, la imagen del componente se muestra en la imagen siguiente:

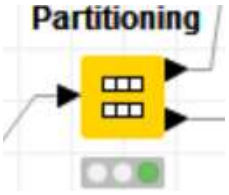


Figura 70. Nodo Partitioning
Fuente: Knime

La configuración que se realizó en el componente se muestra en la figura 71.

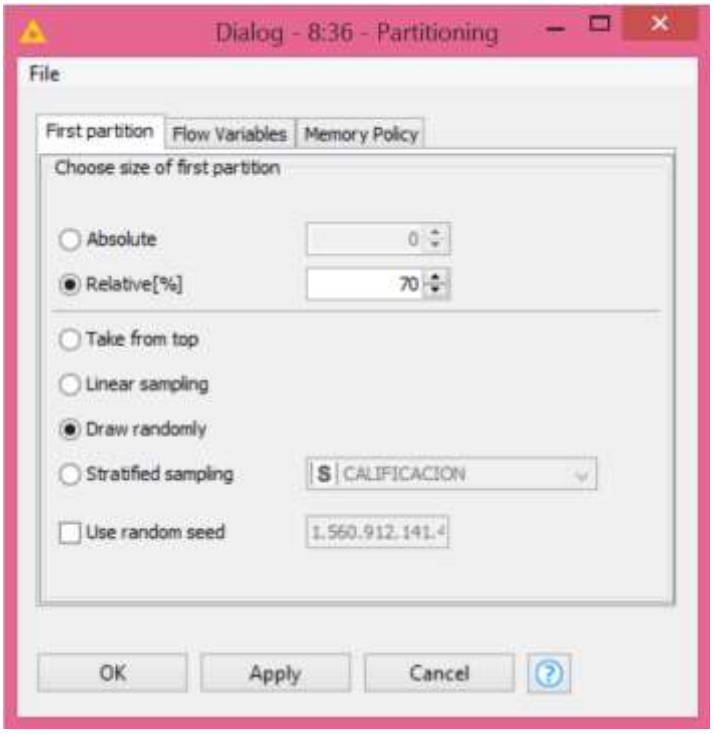


Figura 71. Configuración Nodo Partitioning
Fuente: Knime

Se ha configurado el 70% para la primera partición la cual es la partición de entrenamiento quedando el 30% restante para las validaciones.

J48(3.7)

El componente J48 permite generar un árbol de decisiones C4.5 eliminado o no ajustado, la imagen del componente se muestra en la figura 72.

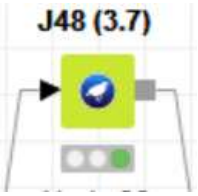


Figura 72. Nodo J48(3.7)
Fuente: Knime

Weka Predictor

El componente Weka Predictor toma el modelo generado con el componente anterior J48 y clasifica los datos de prueba que son proporcionado por el componente partitioning, la imagen del componente se muestra en la imagen siguiente:

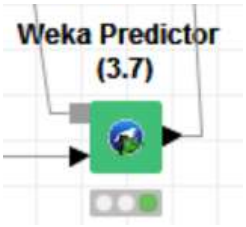


Figura 73. Nodo Weka Predictor
Fuente: Knime

El modelo aplicado completo se muestra en la figura 74:

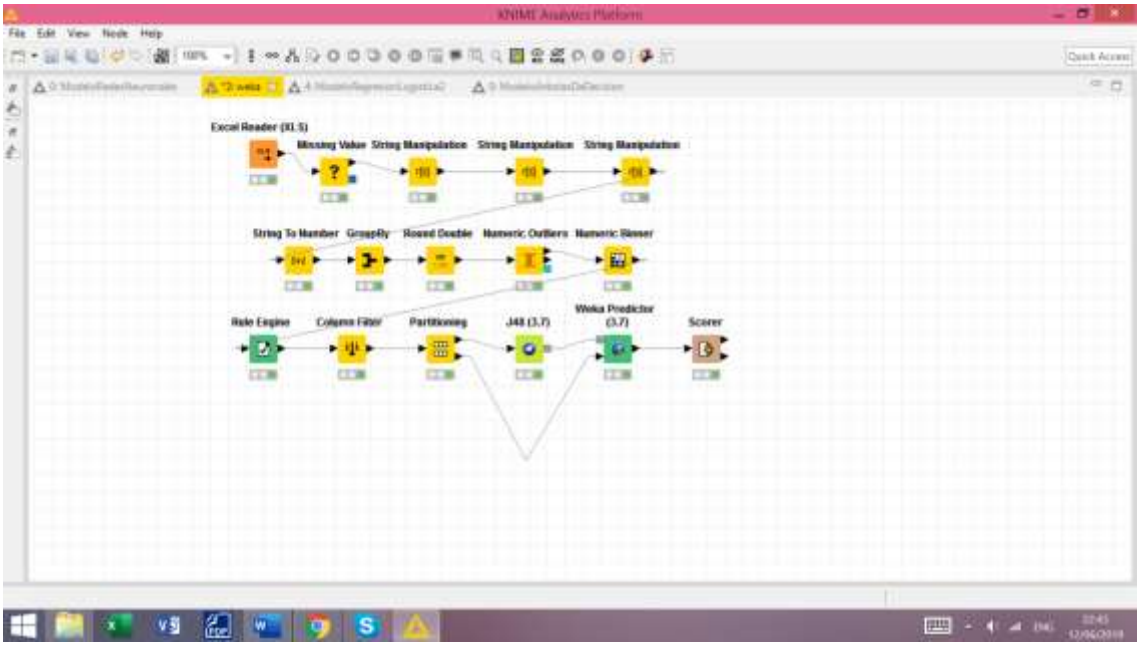


Figura 74. ModeloWekaJ48
Fuente: Knime

A través de este modelo obtuvimos reglas de clasificación que nos indican que características puede tener un cliente para ser considerado como sin potencial, en la figura 75 podemos visualizar el resultado de la aplicación del modelo.

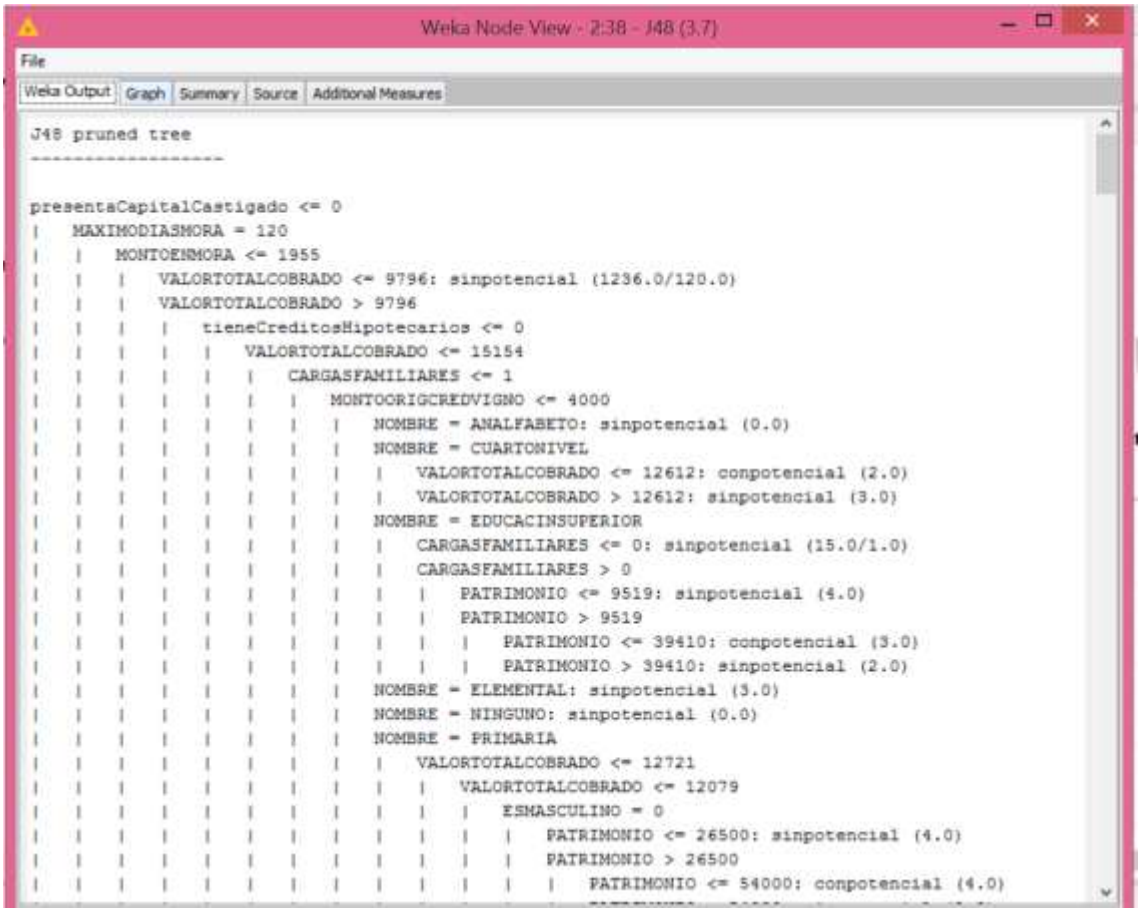


Figura 75. Patrones obtenidos Weka J48
Fuente: Knime

Con este modelo se obtuvo un porcentaje de error de 9.893% y un porcentaje de acierto de 90.107% de acuerdo al componente scorer

4.3.5. FASE 5: Evaluación

La figura 76 indica cada uno de los pasos a seguir en la etapa de evaluación.

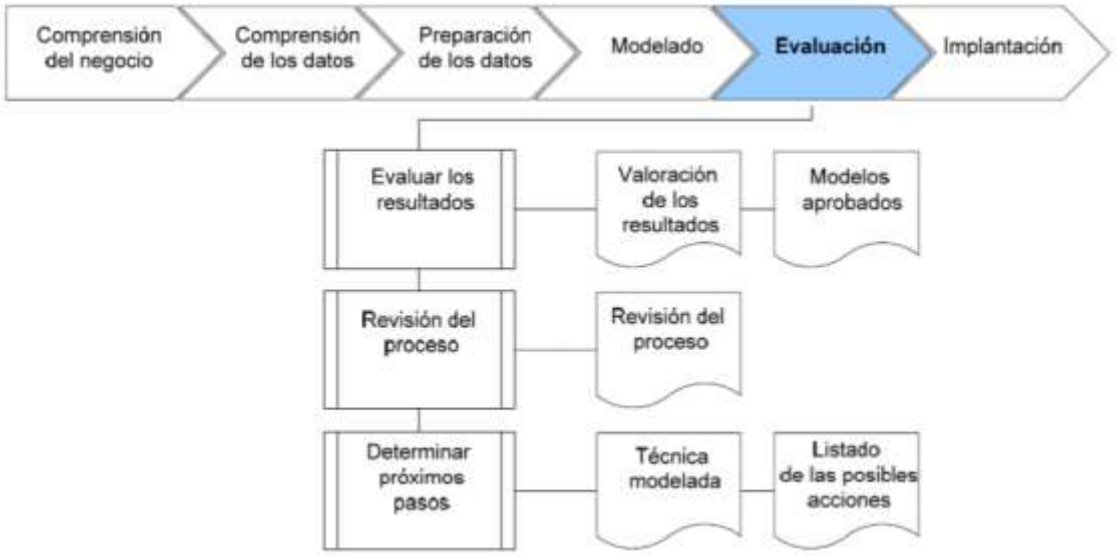


Figura 76. Fase de Evaluación (Metodología Crisp-DM)
Fuente**:** (Crisp-DM,2000)

En el desarrollo del presente proyecto se aplicaron tres modelos que fueron: árboles de decisión, regresión logística y redes neuronales, el porcentaje de acierto fue diferente en cada una de ellas, es así que a través del modelo de árboles de decisión se obtuvo un 90.59% de acierto y un 9.8% de error, tal como se puede apreciar en la figura 77.

Figura 77. Matriz de Confusión-Arboles de Decisión
Fuente: Knime

Al aplicar el modelo de regresión logística se obtuvo un porcentaje de acierto del 63.28% y un porcentaje de error del 36.71% tal como se puede observar en a la figura 78.

Figura 78. Matriz de Confusión - Regresión Logística
Fuente: Knime

Con el modelo de redes neuronales se obtuvo un porcentaje de acierto del 83.181% y un porcentaje de error del 16.81% tal como se puede apreciar en la figura 79.

Figura 79. Matriz de Confusión - Redes Neuronales
Fuente: Knime

Con el modelo J48 de Weka se obtuvo un porcentaje de acierto de 90.107% y un porcentaje de error de 9.893%, tal como se puede apreciar en la figura 80.

Figura 80. Matriz de Confusión - Weka J48
Fuente: Knime

En la tabla siguiente se puede apreciar una comparativa entre las cuatro técnicas de minera de datos aplicadas, en relación a los porcentajes de acierto y error que se obtuvieron, se pudo evidenciar que a través de la técnica de Arboles de Decisión se obtiene un porcentaje de acierto superior a las otras técnicas

Tabla 21.
Comparativo Resultados Aplicación Técnicas

Técnica de Minería	Porcentaje Acierto	Porcentaje Error
Arboles de decisión	90.59%	9.8%
Regresión Logística	63.28%	36.71%
Redes Neuronales	83.181%	16.819%
Weka J48	90.107%	9.893%

De acuerdo a los objetivos de minería de datos planteada se puede observar en la siguiente tabla la comparativa correspondiente.

Tabla 22.
Cumplimiento de objetivos

Objetivos	Arboles de Decisión	Regresión Logística	Redes Neuronales	J48 Weka
Clasificar los clientes en buenos y malos en términos crediticios	SI	SI	SI	SI
Descubrir patrones en morosidad de clientes	SI	NO	NO	SI
Generar datos de entrenamiento y de prueba	SI	SI	SI	SI

Tabla 23.
Resultados Técnicas Aplicadas

Técnica	Clasificados Correctamente	Exactitud	Coeficiente Kappa	Clasificados Incorrectamente	Error
Arboles de Decisión	47054	90.159%	0.661	5.136	9.8%
Regresión logística	33.027	63.282%	0.049	19.163	36.7%
Redes Neuronales	43412	83.181%	0.199	8778	16.8%
J48 Weka	14108	90.10%	0.648	1549	9.89%

4.3.6. FASE 6: Implementación o Despliegue

La figura 81 explica cada uno de los pasos a seguir en la etapa de implantación

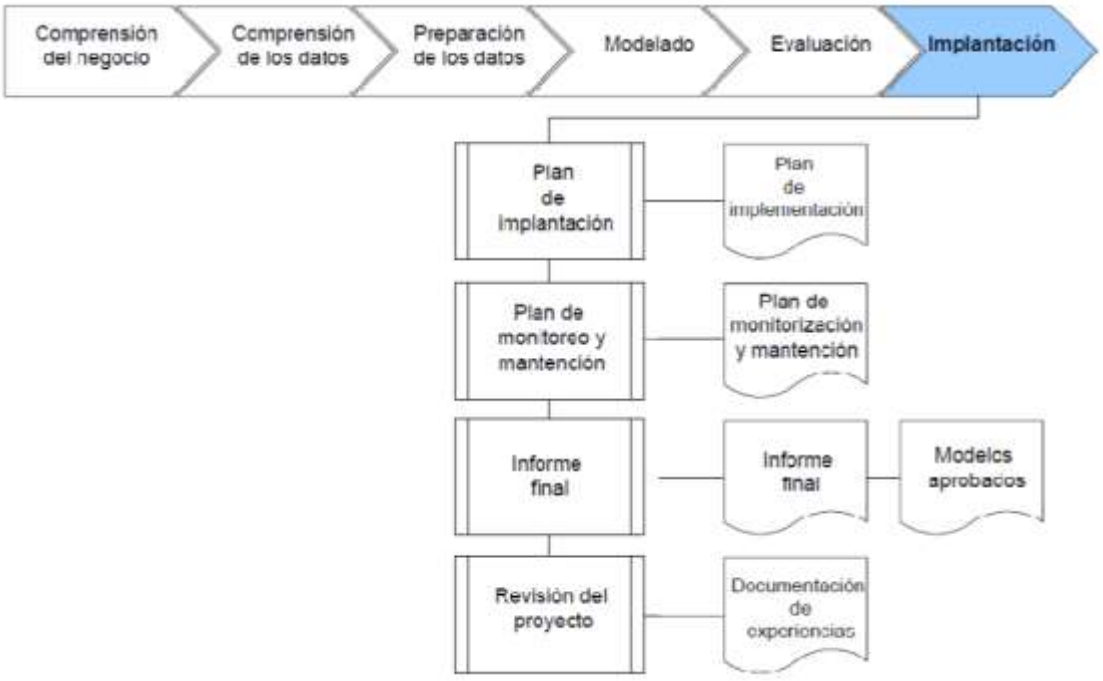


Figura 81. Fase de Implantación (Metodología Crisp-DM)
Fuente**:** (Crisp-DM,2000)

Para poder realizar el despliegue de los modelos es necesario los siguientes requisitos:

- ✓ Información extraída de la base de datos
- ✓ Instalación de la herramienta de minería knime
- ✓ Computador de alta gama

Posterior a que se disponga de los requisitos mencionados se podrá proceder a cargar la información en los modelos generados, verificar los algoritmos, registrar la información y llevar una bitácora para control

4.4. Demostración

En esta fase se realiza una explicación a las personas encargadas de tomar decisiones en la institución como utilizar la información resultante del proyecto, presentando la información de forma clara y fácil de entender.

4.5. Evaluación

La evaluación de los tres modelos desarrollados fue realizada en el apartado 4.3.5 haciendo uso del plan de pruebas diseñado en la sección 4.3.4.2

4.6. Comunicación

En esta fase ser realiza la comunicación de los resultados obtenidos a el área de mercadeo quienes retienen la información relevante de los clientes con gran potencial.

CAPITULO V

DISCUSIÓN DE RESULTADOS

5.1. Introducción

En este capítulo se realiza una explicación de los resultados obtenidos a través de la ejecución de los modelos, se realizó la aplicación de tres modelos cada uno con diferente técnica de minería de datos, se aplicaron arboles de decisión, redes neuronales y regresión logística

5.2. Evaluación de los resultados obtenidos

En el presente apartado se realiza un análisis de los resultados obtenidos a través de los tres modelos aplicados.

El objetivo principal del negocio es clasificar a los clientes en potenciales y no potenciales en términos de riesgo crediticio, en los tres modelos se obtuvo un porcentaje de error bajo, lo cual puede estar determinado a que las variables que se utilizaron fueron seleccionadas de acuerdo a el conocimiento de expertos del negocio.

5.2.1. Árboles de Decisión

El primer modelo aplicado fue árboles de decisión, con un 30% de datos de prueba y 70% de datos de entrenamiento, mediante este modelo se obtuvo un porcentaje de error del 9.8%, para este modelo se clasificaron correctamente 47054 mientras que se clasificaron erróneamente 5136 registros.

En donde:

- ✓ 40464 clientes han sido clasificados de forma correcta como potenciales

- ✓ 6590 cliente han sido clasificados de forma correcta como no potenciales
- ✓ 1689 clientes fueron clasificados erróneamente como sin potenciales
- ✓ 3447 clientes fueron clasificados erróneamente como con potencial

Como se puede observar en la matriz de confusión proporcionado por el componente Scorer cuya imagen se muestra en la figura 77 en el apartado 4.3.5, el coeficiente Cohen's kappa(k) es de 0.661 lo que indica que la mayoría de las predicciones son correctas

La tabla 24 se muestra una estadística de precisión en donde la columna Sensivity mantiene 0.65 indicando que el 65% de los valores analizados se clasificaron erróneamente, la columna Specifity mantiene 0.96 lo cual indica que el 96% de los valores analizados fueron clasificados de manera correcta.

Tabla 24.

Estadísticas - Arboles de Decisión

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy	D Cohen'...
con potencial	40464	3447	6590	1689	0.96	0.922	0.96	0.657	0.94	?	?
sin potencial	6590	1689	40464	3447	0.657	0.796	0.657	0.96	0.72	?	?
Overall	?	?	?	?	?	?	?	?	?	0.902	0.661

A través de este modelo se han podido llegar al cumplimiento de los objetivos de minería de datos propuestos como:

- ✓ Clasificar los clientes en buenos y malos
- ✓ Descubrir patrones en clientes malos
- ✓ Descubrir patrones en morosidad de clientes
- ✓ Generar datos de entrenamiento

5.2.2. Regresión Logística

El segundo modelo aplicado fue el de regresión logística con un 10% de datos de prueba y 90% de datos de entrenamiento, de acuerdo a la matriz de confusión determinada por el componente scorer que se puede observar en el apartado 4.3.5 en la figura 78, se realizó una clasificación correcta de 33027 clientes y una clasificación incorrecta de 19163 clientes de tal forma:

- ✓ 3661 clientes se clasificaron correctamente como sin potencial
- ✓ 29366 clientes se clasificaron correctamente como con potencial
- ✓ 6376 clientes se clasificaron erróneamente como con potencial
- ✓ 12787 clientes se clasificación erróneamente como sin potencial

El porcentaje de acierto adquirido fue del 63.282% mientras que el porcentaje de error fue del 36.718% y un coeficiente kappa(k) de Cohen igual a 0.049, este coeficiente se encarga de medir la concordancia entre dos examinadores en sus correspondientes clasificaciones.

En la tabla 25 se puede observar la tabla de estadísticas de precisión de los resultados correspondientes al modelo de regresión logística

Tabla 25.
Estadísticas de Precisión - Regresión Logística

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
sin potencial	1380	121	42032	8657	0.137	0.919	0.137	0.997	0.239	?	?
con potencial	42032	8657	1380	121	0.997	0.829	0.997	0.137	0.905	?	?
Overall	?	?	?	?	?	?	?	?	?	0.832	0.199

En la columna Sensitivity se encuentra 0.137 indicando que el 0.13% de los valores analizados fueron clasificados de manera errónea, mientras que en la columna Specifity se encuentra el valor de 0.997 lo cual muestra que el 99% se clasifico de manera correcta.

A través de este modelo se han podido llegar al cumplimiento de los objetivos de minería de datos propuestos como:

- ✓ Clasificar los clientes en buenos y malos
- ✓ Generar datos de entrenamiento

5.2.3. Redes Neuronales

El tercer modelo aplicado fue redes neuronales con un 10% de datos de la muestra para datos de prueba y 90% para datos de entrenamiento, de acuerdo a la matriz de confusión determinada por el componente scorer mismo se puede observar en el apartado 4.3.5 en la figura 79, se realizó una clasificación correcta de 43412 socios y una clasificación incorrecta de 8778 socios, de tal forma:

- ✓ 1380 socios fueron clasificados correctamente como sin potencial
- ✓ 42032 socios fueron clasificados correctamente como con potencial
- ✓ 121 socios fueron clasificados erróneamente como sin potencial
- ✓ 8657 socios fueron clasificados erróneamente como con potencial

En la tabla 26 se muestra las estadísticas de precisión obtenidas para el modelo de redes neuronales a través de la herramienta scorer

Tabla 26.
Estadísticas de precisión - Redes Neuronales

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
sin potencial	1380	121	42032	8657	0.137	0.919	0.137	0.997	0.239	?	?
con potencial	42032	8657	1380	121	0.997	0.829	0.997	0.137	0.905	?	?
Overall	?	?	?	?	?	?	?	?	?	0.832	0.199

Se observa que existen un total de 121 falsos positivos en lo correspondiente a la clasificación como clientes sin potencial, 8657 falsos positivos en la clasificación de clientes con potencial. Adicional aparecen 8657 falsos negativos en la clasificación de clientes sin potencial y 121 falsos negativos en la clasificación de clientes con potencial.

Hay una precisión del 91% para determinar clientes sin potencial y del 82% para determinar clientes con potencial

A través de este modelo se han podido llegar al cumplimiento de los objetivos de minería de datos propuestos como:

- ✓ Clasificar los clientes en buenos y malos
- ✓ Generar datos de entrenamiento

5.2.4. Componente J48

El cuarto modelo aplicado fue una extensión de árboles de decisión de weka con el componente J 48, en el cual se utilizaron 30% de los datos de la muestra para prueba y 70% para entrenamiento, de acuerdo a la matriz de confusión obtenida (el cual se puede observar en la figura 80 del apartado 4.3.5), Se observa que se utilizó 36533 datos entrenamiento y 15657 datos

de prueba, el modelo realizó una clasificación correcta de 14108 socios y una clasificación incorrecta de 1549 socios, de tal forma:

- ✓ 1858 socios fueron clasificados correctamente como sin potencial
- ✓ 12250 socios fueron clasificados correctamente como con potencial
- ✓ 424 socios fueron clasificados erróneamente como sin potencial
- ✓ 1125 socios fueron clasificados erróneamente como con potencial

En la tabla 27 se muestra las estadísticas de precisión obtenidas para el modelo de weka J48 a través de la herramienta scorer.

Tabla 27.
Estadísticas de precisión - WekaJ48


File Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
con potencial	12250	1125	1858	424	0.967	0.916	0.967	0.623	0.941	?	?
sin potencial	1858	424	12250	1125	0.623	0.814	0.623	0.967	0.706	?	?
Overall	?	?	?	?	?	?	?	?	?	0.901	0.648

En la gráfica anterior se observa que hay una precisión del 96% para determinar clientes con potencial y del 81% para determinar clientes sin potencial.

El presente modelo permitió observar un total de 225 hojas y un tamaño de árbol de 392, y un total de 225 reglas que señalan que comportamientos previos tiene un cliente antes de ser considerado como un mal cliente o un buen cliente

A través de este modelo se han podido llegar al cumplimiento de los objetivos de minería de datos propuestos como:

- ✓ Clasificar los clientes en buenos y malos

- ✓ Descubrir patrones en clientes malos
- ✓ Descubrir patrones en morosidad de clientes
- ✓ Generar datos de entrenamiento

CAPÍTULO VI

CONCLUSIONES Y RECOMENDACIONES

6.1. CONCLUSIONES

- ✓ En el análisis de la situación actual de la Cooperativa CACPECO LTDA. con respecto a la calificación de crédito de los clientes se han encontrado cierta deficiencia en el manejo de la información para definir si es un buen cliente o mal cliente.
- ✓ La calificación de crédito de los clientes entregada por una empresa de Buró de crédito ha sido motivo para denegar a varios posibles clientes en el otorgamiento de un crédito.
- ✓ El uso de herramienta ofimática para analizar todos los parámetros del cliente en la toma de decisiones para otorgar un crédito es de motivo de preocupación, por la razón de que el conocimiento o la experticia radica en el personal que trabaja en el área de crédito. La aprobación o la negación de un crédito está establecido en la herramienta ofimática manejada por el área de crédito, es decir, no hay una metodología o técnicas que indique qué parámetro seleccionó para definir si es un buen o mal cliente.
- ✓ La experticia personal en el área de crédito sin el uso de la Inteligencia de negocio o Business Intelligence (BI) conlleva a la incertidumbre de las condiciones del préstamo de un cliente en un futuro inmediato.
- ✓ La deficiencia de la promoción de los productos crediticio es debido a la falta de análisis de la información interna de sus actuales clientes y de sus posibles clientes. Al no manejar e interpretar la información, corre el riesgo de no conocer los deseos y el comportamiento

del cliente, lo que provoca que los clientes no estén interesados en el producto ofrecido o el producto no es el adecuado para ello.

- ✓ La incertidumbre de que un cliente es bueno o malo y el no poseer una metodología o técnicas que sustenten las decisiones del área de crédito, provoca la inseguridad en asumir o medir los riesgos de créditos, y eso a su vez trae como consecuencia el bajo porcentaje de colocación de crédito.
- ✓ El uso de la técnica de la entrevista al personal de área de crédito con 10 preguntas elaborada por las investigadoras ha permitido conocer a breve rasgo el proceso de análisis de los clientes en el otorgamiento de un crédito y las incidencias que suelen ocurrir durante el proceso de análisis. También se compartió por parte de las investigadoras al personal del crédito sobre el desarrollo de un modelo de medición de riesgo basado en la minería de datos.
- ✓ En el proceso de investigación para la elección de las técnicas de minería de datos se utilizó la revisión bibliográfica y documental, por lo cual se estableció 4 algoritmos (árbol de decisión, regresión logística, redes neuronales y algoritmo J48) que son utilizado para medir el riesgo crediticio.
- ✓ La elección de una herramienta se basó en el cumplimiento de 2 requisito que son que posea la capacidad de realizar proceso de ETL (Extract, Transform and Load) y minería de datos. Al principio se eligió la herramienta RapidMiner para el desarrollo del modelo, pero en su versión gratuita hay un limitante de máximo 10000 registro, por lo cual se descartó y se buscó otra alternativa. La alternativa elegida es la herramienta Knime que cumple con los requisitos señalados con anterioridad.

- ✓ Se desarrolló 4 modelos, en la que cada modelo corresponde a cada técnica de minería de datos que son árboles de decisión, regresión logística, redes neuronales y weka j48.
- ✓ Para la validación de las técnicas de minería de datos de cada modelo se utilizó el método de la matriz de confusión en donde valida si el valor asignado de cliente bueno y malo corresponde a la predicción realizada por la técnica de minería de datos.
- ✓ De acuerdo a las evaluaciones realizadas la técnica que mejor se adapta a las necesidades del negocio fue árboles de decisión, ya que adicional a que se obtuvo el porcentaje más alto de precisión en la predicción de los clientes como buenos o malos, también permitió identificar patrones o reglas que indican que comportamientos tienen en común los socios que se consideran buenos o malos.
- ✓ La minería de datos será un apoyo tecnológico al área de crédito que dará mejoras en el análisis de los datos del cliente. Reduciendo el tiempo de procesamiento de información, lo cual conlleva a agilizar la toma de decisiones en la aprobación o negación de un crédito.
- ✓ La utilización de la minería de datos por parte de la Cooperativa de ahorro y crédito CACPECO permitirá prevenir los malos clientes que generan costos legales y tiempo del personal de crédito.
- ✓ Al invertir la Cooperativa de ahorro y crédito CACPECO minería de datos mejorará la relación de la institución con los clientes al entender la información y las necesidades del cliente.

- ✓ La minería de datos ayudará a la Cooperativa de ahorro y crédito CACPECO a recudir los riesgos de perder los clientes actuales y generará mayor posibilidad de atraer más cliente a la instrucción lo que traería beneficio de rentabilidad a la Cooperativa CACPECO.

6.2. Recomendaciones

- ✓ El trabajo de investigación propuesta por las investigadoras será una guía para los directivos de la Cooperativa de ahorro y crédito CACPECO, y al personal de área de crédito, con la finalidad de que puedan mejorar sus procesos en el análisis de los clientes basado en minería de datos.
- ✓ En el desarrollo de los modelos de medición de riesgo creditico no se incluyó variables crítico, en referencia a la calificación otorgada por los buros de crédito o la calificación interna que realiza el área de crédito a cada cliente, esto es debido a sigilo y reserva bancario. Se recomienda utilizar dichas variables críticas ya que influiría en un modelo más real en la predicción de un cliente.
- ✓ Implementar una solución de Data Warehouse, que junto con la herramienta de minería de datos mejoraran el proceso de análisis para la toma de decisiones.
- ✓ Realizar capacitación al personal de los beneficios de la aplicación de minería de dato
- ✓ Diseñar una infraestructura tecnológica de soluciones de inteligencia de negocios que involucre todas las áreas de negocio. Esto ayudará a planificar y organizar los proyectos de minería de datos.
- ✓ Invertir en equipos de cómputo de última generación, que posea alta capacidad de procesamiento y almacenamiento destinado a la minería de datos.

- ✓ Apoyar el marketing digital con inteligencia de negocio para incrementar la promoción de los productos crediticio.

- ✓ Analizar los resultados del modelo crediticio aplicado en los clientes, y evaluarlo en un periodo sea bimestral, semestral o anual, con el fin de realizar una retroalimentación en el mejoramiento del modelo predictivo.
- ✓ El proyecto de crear un modelo basado en minería de datos se puede aplicar en predecir que clientes van a solicitar un nuevo préstamo o qué clientes abandonará la institución.

6.3. Futuras líneas de investigación

- ✓ En el futuro serian interesante realizan una investigación y análisis de los patrones que inciden las captaciones de clientes como ahorristas e inversionistas.
- ✓ Investigación sobre los patrones de comportamiento de los clientes en el uso de la tarjeta de crédito y su capacidad de pago.
- ✓ Desarrollar un modelo que mida qué cliente dará más beneficio a la institución al usar la billetera electrónica, es decir, cuáles son los clientes que generará mayor cantidad de transacciones con la billetera electrónica para el otorgamiento de un crédito.

REFERENCIA BIBLIOGRÁFICA

García Herrero, J. (n.d.). *Universidad Carlos III de Madrid.* Retrieved from Predicción Numérica: http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/regresion.pdf

Schreiner, M. (2002). *Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas.* Washington, Estados Unidos: Washington University in St. Louis .

Acosta Henríquez, G. (2015). Aplicaciones empresariales de minería de datos usando software libre. *Anuario de Investigación 2016, Universidad Católica de El Salvador* , 290.

Aguirre Baztán, Á. (1995). *metodología cualitativa en la investigación sociocultural.* España: Marcombo.

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Questiió*, 479-498.

Asamblea Nacional, R. d. (2011). Ley Orgánica de la Economía Popular y Solidaria y del Sector Financiero Popular y Solidario.

Asamblea Nacional, R. d. (2012). Reglamento a ley organica economia popular y solidaria.

Barrezueta, H. E. (2014, Junio 19). *Registro Oficial 444 10-may-2011.* Retrieved from Ley Orgánica de Economía Popular y Solidaria del Sistema Financiero.

Barrientos, F., & Ríos, S. (2013). Aplicación de minería de datos para predecir fuga de clientes en la industria de las telecomunicaciones. *Revista ingenieria de sistemas*, 88.

Bunge, M. (1980). *Epistemología.* Buenos Aires: siglo xxi editores.

Carranza Rueda, D. R. (2008). *Aplicación de la lógica difusa para la ubicación de especies faunísticas y florísticas, y su comparación con otros métodos geoestadísticos.* Quito.

Carrasquilla-Batista, A., Chacon Rodriguez, A., Núñez Montero, K., & Gómez Espinoza, O. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Tecnología en Marcha*, 36.

Carstensen, A.-K., & Bernhard, J. (2015). Design Science Research – an engineering research approach to improve methods for engineering education research. *REES organisers and educational*, 3.

Cendejas Valdez, J., Acuña Lopez, M., & Cortes Morales, G. (2017). El uso de modelos y metodologías de minería de datos para la inteligencia de negocios. *Revista de Sistemas Computacionales y TIC's*, 56.

Chaglla Rodríguez, L. (2015). *Arquitectura de ejecución de experimentos de minería de datos.* Castilla-La Mancha, España: Universidad Castilla-La Mancha.

Comunicación, C. (n.d.). *¿Qué es el método Delphi? Aplicación y usos*. Retrieved from Cícero Comunicación: https://www.cicerocomunicacion.es/que-es-el-metodo-delphi/

Cortés Cortés, M., & Iglesias León, M. (2004). *Generalidades sobre Metodología de la Investigación.* Carmen, Campeche, México: Universidad Autónoma del Carmen.

Cortez Cortez, G. (2011). Análisis de la gestión del riesgo de la banca múltiple en el Perú: 2000–2010. *Pensamiento Crítico*, 7-19.

Cruz García, L. (2014). *Metodología de Investigación.* Colima, Mexico: Universidad Multecnica Profesional.

Durán, A. (2017, 12 11). *¿Qué es Pentaho Data Integraton (PDI)?* Retrieved from OpenWebinars: https://openwebinars.net/blog/que-es-pentaho-data-integraton-pdi/?cat=big-data

Fayyad, U., Piatetsky, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Al Magazine Volumen 17*, Number 3.

Fernández Castaño, H., & Pérez Ramírez, F. (2005). El modelo logístico: Una herramienta estadística para evaluar el riesgo de crédito. *Revista Ingenierías Universidad de Medellín*, 58.

Fernandez, D., & Soares Netto, R. (n.d.). *Valor en riesgo de las carteras de préstamos bancarios.* Retrieved from Universidad de la Republica: http://biblioteca.fcea.edu.uy/QUANTUM/Vol3/No1/Fernandez.pdf

Fisher, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 179-188.

Galán Cortina, V. (2015). *Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario.* Madrid, España: Universidad Carlos III de Madrid.

Galicia, M. (2003). *Los enfoques del riesgo de crédito.* Instituto del Riesgo Financiero.

Gambin, D., & Pallotta, E. (2009). *Minería de datos aplicada a cultivos de maíz.* Buenos Aires, Argentina: Universidad de Buenos Aires.

García Bermúdez, J., & Acevedo Ramirez, Á. (2010). *Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies.* Universidad Tecnológica de Pereira.

García González, F. (2013). *Aplicación de técnicas de Minería de Datos datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA).* Granada, España: Universidad de Granada.

García Morate, D. (2011). *Manual de Weka.* Madrid, España. Retrieved from Programa de Doctorado Formación en la Sociedad del Conocimiento, Universidad de Salamanca: https://knowledgesociety.usal.es/sites/default/files/MANUAL%20WEKA.pdf

Garre, M., Cuadrado, J. J., & Sicilia, M. Á. (n.d.). *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software*. Retrieved from ETS Ingeniería Informática: http://www.sc.ehu.es/jiwdocoj/remis/docs/GarreAdis05.pdf

Ghobadi, F., & Rohani, M. (2016). Cost Sensitive Modeling of Credit Card Fraud. *IEEE*.

Gonzalez Bernal, J. A. (2011, Febrero 17). *Instituto Nacional de Astrofísica, Optica y Electrónica*. Retrieved from https://ccc.inaoep.mx/~jagonzalez/ML/principal/node69.html

Güereca Tijerina, M. J. (2013, Febrero 18). *Prezi.* Retrieved from Prezi: https://prezi.com/ucgya1hfb9fu/algoritmos-de-md-para-descubrir-conocimiento-iii/

Hand, D., & Jacka, S. (1998). Consumer credit and statistics. *Statistics in Finance*, 69-81.

Hasan, A., & Kalipsiz, O. (2017). Predicting Financial Market in Big Data: Deep Learning. *Internacional Conference on Computer Science and Engineering*.

Hernández Cáceres, J. (2016). Clustering basado en el algoritmo K-means para la identificación de grupos de pacientes quirúrgicos. *Congreso Académico UDI - 2016.* Bucaramanga, Santander, Colombia.

Hernández Orallo, J., Ramírez Quintana, M., & Ramírez Ferri, C. (2004). Introducción a la Minería de Datos. Prentice Hall.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). *Metodología de la investigación.* Mexico: McGraw-Hill.

Informáticos, D. d. (n.d.). *Universidad de Granada*. Retrieved from Apéndice técnicas de entrevista: http://lsi.ugr.es/~mvega/docis/entre1.html

jortilles. (2017, mayo 31). *Knime. Herramienta de Análisis y todo lo que se te ocurra*. Retrieved from Jortilles: http://www.jortilles.com/knime-herramienta-de-analisis/

Khandani, A., Kim, A., & Lo, A. (2010). Consumer Credit Risk Models via Machine-Learning Algorithms.

Kishinani, K. (2016, octubre 01). *Prueba de Kolmogorov–Smirnov*. Retrieved from Simulacion de sistemas: https://simulacionutp2016.wordpress.com/2016/10/01/prueba-kolmogorov-smirnov/

La Junta de politica y regulacion monetaria. (2015, 4 1). *Normas que regulan la segmentación de la cartera de crédito de las entidades del sistema financiero nacional.* Retrieved from Banco Central del Ecuador: https://contenido.bce.fin.ec/documentos/Estadisticas/SectorMonFin/TasasInteres/RegTasas043.pdf

La Rotta Mendoza -, J., & Celis Torres , E. (n.d.). *Investigación evaluativa*. Retrieved from Escuela de Formación- Infantería de Marina - Colombia: https://sites.google.com/site/ciefim/investigaci%C3%B3nevaluativa

Lewis, E. (1992). An Introduction to Credit Scoring. *Fair, Isaac and Co., Inc., San Rafael, CA*, 86.

Llorente Lopez , M. A. (2012). *Programación genética en mercados financieros.* Barcelona, España.

Marin Castro, H. M. (2014). Minería de Datos. *Universidad Politécnica de Victoria*, 2.

Marta. (8, 3 2017). *¿Cómo hacer una revisión bibliográfica?* Retrieved from Scribbr: https://www.scribbr.es/category/revision-bibliografica/

Martínez Abad, F. (n.d.). *Minería de datos con software Weka.* Salamanca, España.

Mays, E. (1998). Credit Risk Modeling, Design and Application. *Fitzroy Dearborn Publishers*.

Medina Patrón, J., Ortiz Servin, J., Castillo, A., Montes Tadeo, J., & Perusquía, R. (2015). Minería de Datos en el estudio de Celdas de Combustible Nuclear. *XIV Congreso Nacional de la Sociedad Mexicana de Seguridad Radiológica*, 4-5.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 449-470.

Metodología de la investigación documental. (n.d.). Retrieved from EcuRed: https://www.ecured.cu/Metodolog%C3%ADa_de_la_investigaci%C3%B3n_documental

Microsoft. (2018, 05 07). *Prueba y validación (minería de datos)*. Retrieved from Microsoft Docs: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/testing-and-validation-data-mining?view=sql-server-2017

Miguel Moine , J., Haedo , A., & Gordillo, S. (n.d.). *Estudio comparativo de metodologías para minería de datos.* La Plata, Buenos Aires: Grupo de investigación en Minería de Datos, Universidad Nacional de Buenos Aires .

Miranda Moles, A. (2003). *El método de remuestreo y su aplicación en la investigación biométrica.* Habana, Cuba: Escuela Nacional de Salud pública "Carlos J. Finlay".

Molina López, J. M., & García Herrero, J. (2006). *Técnicas de análisis de datos.* Madrid.

Mora Maqueda, M., & Luque Calvo, L. (2017). *Introducción a la Inteligencia de Negocios con ayuda de R.* Sevilla, España: Universidad de Sevilla.

Ochoa P., J., Galeano M., W., & Agudelo V., L. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de coyuntura económica*, 207.

Peffers, K., & Tuunanen, T. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* , 45-77.

Peffers, Tuunanen, Rothernberger, & Chatterjee. (2007). A DSR Methodology for IS Research.

Ponsot, E., Sinha, Surendra, & Goitía, A. (2009). Sobre la agrupación de niveles del factor explicativo en el modelo logit binario. *Revista Colombiana de Estadística*, 157-187.

PowerData. (2013, Julio 12). *Los acuerdos de Basilea y la gestión de los datos*. Retrieved from PowerData: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/307125/qu-son-los-acuerdos-de-basilea-basilea-i-basilea-ii-y-basilea-iii

Provider, S. (n.d.). *XLMiner*. Retrieved from Indiamart: https://www.indiamart.com/proddetail/xlminer-6082745097.html

Rapidminer. (2019). *Rapidminer*. Retrieved from Rapidminer: https://rapidminer.com/

Rayo Cantón, S., Lara Rubio, J., & Camino Blasco, D. (2010). *Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea Ii.* Granada, España: Universidad de Granada.

Rodriguez Jaume, M., & Morar Catala, R. (2001). *ESTADISTICA INFORMATICA: CASOS Y EJEMPLOS CON EL SPSS.* Alicante, España: UNIVERSIDAD DE ALICANTE. SERVICIO DE PUBLICACIONES.

Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: A survey. *Operations Research*, 589-613.

Ruíz Guerra, A., & González R., O. (n.d.). Informe de investigación. *Universidad de Carabobo*, 5-6.

Ruiz Limón, R. (n.d.). *Historia y evolución del pensamiento científico*. Retrieved from eumed.net: http://www.eumed.net/libros-gratis/2007a/257/7.1.htm

Ruza, C., & Curbera, P. (2013). *El riesgo de crédito en perspectiva.* Madrid, España: Universidad Nacional de Educación a distancia.

Saavedra García, M., & Saavedra García, M. (2010). *Modelos para Medir el riesgo de crédito de la banca.* México: Universidad La Salle, Dirección de Posgrado e Investigación.

Seijas Giménez, M., Vivel Búa, M., Lado Sestayo, R., & Fernández López, S. (2017). La evaluación del riesgo de crédito en las instituciones de microfinanzas: estado del arte. *Compendium*, 36-52.

Self Bank. (2017, Mayo 26). *Los acuerdos de Basilea ¿Cómo afectan a ahorradores e inversores?* Retrieved from El Blog de Self Bank: https://blog.selfbank.es/los-acuerdos-de-basilea-como-afectan-ahorradores-e-inversores/

Semerena, Y. (n.d.). *¿Qué es la Investigación Exploratoria?* Retrieved from QuestionPro: https://www.questionpro.com/blog/es/investigacion-exploratoria/

Sepúveda Rivillas, C., Reina Guitiérrez, W., & Guitierrez Bentacur, J. (2012). *Estimación del riesgo de crédito en empresas del sector real en Colombia.* Bogotá, Colombia: Universidad de Antioquia.

Silclir, M., Szyrko, P., Ruiz de Mendarozqueta, Á., & Rubio, D. (n.d.). *Un Método Heurístico para el Análisis y Selección de Herramientas de Modelado de Procesos de Desarrollo de Software.* Cordoba, Argentina: Universidad Tecnológica Nacional.

Solarte Martínez, G. R., & Ocampos, C. A. (2009). Técnicas de clasificación y análisis de representación de diagnóstico . *Scientia et Technica Año XV, No 42*, 181.

Solidaria, S. d. (n.d.). *Superintendencia de Economía Popular y Solidaria*. Retrieved from Superintendencia de Economía Popular y Solidaria: http://www.seps.gob.ec

Superintendencia de bancos. (n.d.). *Glosario de Términos*. Retrieved from Superintendencia de bancos: https://www.superbancos.gob.ec/bancos/glosario-de-terminos/

Superintendencia de bancos y seguros. (2003, 12). *Libro I.- Normas generales para las instituciones del sistema financiero, Titulo X.- de la gestión y administración de riesgos, Capitulo II.- de la administración del riesgo de crédito.* Retrieved from Superintendencia de bancos y seguros: https://www.superbancos.gob.ec/bancos/wp-content/uploads/downloads/2017/06/L1_X_cap_II.pdf

Tableau. (2019). *Tableau*. Retrieved from Tableau: https://www.tableau.com/

Téllez Cabrera, M. (2010). *Medición del riesgo en crédito: Implementación y cálculo del Var y el CVaR en tres modelos de incumplimiento.* Ciudad de Mexico, Mexico: Universidad Autónoma Metropolitana.

Thomas, L., Crook, J., & Edelman, D. (1992). Credit Scoring and Credit Control. *Oxford University Press*.

Universidade de Santiago de Compostela. (2012). Fundamentos de biología aplicada I: Estadistica. In U. d. Compostela, *Práctica 6: Regresión Logística I.* Santiago de Compostela, Galicia, España: Universidade de Santiago de Compostela. Retrieved from Universidade de Santiago de Compostela: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140142_practicaRegLogI_1112.pdf

Vaishnavi, V., & Kuechler, B. (2013). Design Science rersearch in Information System. *Association for Information Systems*, 1.

Vaishnavi, V., Kuechler, B., & Petter, S. (2017). Design Science Research in Information System. *Association for information systems*, 3.

Vaishnavi, V., Kuechler, B., & Petter, S. (2017). Design Science Research in Information System. *Association for Information System*, 8.

Valcárcel Asencios, V. (2004). Data Mining y el Descubrimiento del Conocimiento. *Revista de la Facultad de Ingeniería Industrial*, 83-86.

Valle Carrascal, J. (2015). *Modelos de medición del riesgo de crédito.* Madrid: Universidad Complutense de Madrid.

Velandia Ortega, R. A., & Hernández Suárez, F. L. (2010). *Evaluación de algoritmos de extracción de reglas de decisión para el diagnóstico de huecos de tensión.* Bucaramanga.

Vera Noguez, S. (n.d.). *Minería de datos.* Retrieved from Universidad Aútonoma del Estado de México: http://ri.uaemex.mx/bitstream/handle/20.500.11799/64109/secme-12408.pdf?sequence=1

Wikipedia. (n.d.). Retrieved from https://es.wikipedia.org/wiki/Algoritmo_apriori

Wikipedia. (2018). *Cross Industry Standard Process for Data Mining*. Retrieved from Wikipedia: https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Wikipedia. (2018, Noviembre 16). *Modelo probit*. Retrieved from Wikipedia: https://es.wikipedia.org/wiki/Modelo_probit

Wikipedia. (2019, 12 9). *Knime*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/KNIME

Wikipedia. (2019, 10 17). *Orange (Software)*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Orange_(software)

Wikipedia. (2019, 11 19). *Pentaho*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Pentaho

Wikipedia. (2019, 11 19). *Pentaho*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Pentaho

Wikipedia. (2019, 12 6). *R (programing language)*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/R_(programming_language)

Wikipedia. (2019, 11 21). *Tanagra (machine learning)*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Tanagra_(machine_learning)

Wikiteka. (2011, Enero 25). *Variables dependientes e independientes*. Retrieved from Wikiteka: https://www.wikiteka.com/apuntes/variables-dependientes-e-independientes/