Desarrollo de un modelo predictivo para la evaluación del riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne

Chiluiza Molina, Oscar Wladimir

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnológica

Centro de Posgrados

Maestría en Ingeniería en Software

Trabajo de titulación, previo a la obtención de título de Magister en Ingeniería en Software

Msc. Guevara Vega, Cathy Pamela

25 de junio del 2021

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Desarrollo de un Modelo Predictivo para la Evaluación del Riesgo Crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne**" fue realizado por el señor **Chiluiza Molina Oscar Wladimir** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Latacunga, 25 junio de 2021

CATHY PAMELA GUEVARA VEGA

Firmado digitalmente por CATHY PAMELA GUEVARA VEGA
Fecha: 2021.09.24 13:27:04 -05'00'

Msc. Guevara Vega, Cathy Pamela

Directora

C.C.: 1002334835

REPORTE URKUND

Curiginal

Document Information

Analyzed document	Tesis-Oscar Chiluiza-Final-V03.pdf (D109971098)
Submitted	6/30/2021 8:42:00 PM
Submitted by	
Submitter email	owchiluiza@espe.edu.ec
Similarity	3%
Analysis address	cguevara.utn@analysis.urkund.com

Sources included in the report

SA	**TESIS_Salinas_VersionFinal.doc** Document TESIS_Salinas_VersionFinal.doc (D54630529)	3
W	URL: http://201.159.222.180/bitstream/3317/11838/1/T-UCSG-POS-MAE-332.pdf Fetched: 6/24/2021 5:42:50 AM	1
SA	**ANDRADE SORIANO RONALD LEONEL.pdf** Document ANDRADE SORIANO RONALD LEONEL.pdf (D63151072)	2
SA	**PARA URKUND Tesis Cooperativa Maquita revisada V4.docx** Document PARA URKUND Tesis Cooperativa Maquita revisada V4.docx (D25714222)	2
W	URL: https://repositorio.ucci.edu.co/bitstream/handle/001/773/Diet%C3%A9tica%20de%20un%20Modelo%20Predictivo%20de%20Riesgo%20de%20Clientes.pdf?sequence=1&isAllowed=y Fetched: 6/29/2021 8:42:00 PM	2
SA	**M2.878_20201_PEC4 - Redacci◆n de la memoria_13884760.txt** Document M2.878_20201_PEC4 - Redacci◆n de la memoria_13884760.txt (D90841465)	3
W	URL: http://repositorio.utc.edu.ec/bitstream/27000/2136/1/T-UTC-3172.pdf Fetched: 1/5/2021 6:13:54 PM	1
SA	**20190719 Oscar Montalvo .pdf** Document 20190719 Oscar Montalvo .pdf (D54402505)	1
SA	**TESIS_versionfinal01082019.doc** Document TESIS_versionfinal01082019.doc (D54660967)	1
W	URL: https://www.researchgate.net/publication/337480778_Propuesta_de_modelo_de_machine_learning_para_la_evaluacion_de_riesgo_de_credito_utilizando_algoritmos_de_prediccion_para_la_Cooperativa_de_Ahorro_y_Credito_La Fetched: 3/30/2020 1:21:29 AM	2
SA	**Tesis_LilianaGonzalezLapo_SIG1.docx** Document Tesis_LilianaGonzalezLapo_SIG1.docx (D54799820)	1
W	URL: https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/IEMensual/in2067not4_monetaria.pdf	1

1/35

CATHY PAMELA GUEVARA VEGA

Firmado digitalmente por CATHY PAMELA GUEVARA VEGA
Fecha: 2021.09.24 12:27:31 -05'00'

Msc. Guevara Vega Cathy Pamela

Directora

C.C.: 1002334835

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo, **Chiluiza Molina, Oscar Wladimir,** con cédula de ciudadanía Nº 0503264459, declaro que el contenido, ideas y criterios del trabajo de titulación**: "Desarrollo de un modelo predictivo para la evaluación del riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne"** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Latacunga, junio de 2021

.....

Chiluiza Molina Oscar Wladimir

C.C.: 0503264459

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Chiluiza Molina Oscar Wladimir,** con cedula de ciudadanía 0503264459, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Desarrollo de un modelo predictivo para la evaluación del riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Latacunga, junio de 2021

.....
Chiluiza Molina Oscar Wladimir

C.C.: 0503264459

DEDICATORIA

Este trabajo está dedicado a mis padres Luis y María quienes día a día me han demostrado que con esfuerzo y trabajo duro se consigue las metas.

A mis angelitos Emilio, Martin, Zoé y Stefano, quienes son y serán mis motores que impulsan mi vida.

A mis hermanos y todas las personas que me apoyaron y me dieron la fortaleza para culminar este escrito.

AGRADECIMIENTO

A DIOS por permitir seguir con vida y dejarme disfrutar cada instante junto a mis seres queridos.

A mis padres por apoyarme infinitamente en cada etapa de mi vida.

Al Centro de Posgrados de la Universidad de las Fuerzas Armadas sede Latacunga, por permitirme continuar con mis estudios y especializarme.

A mi tutora MSc. Guevara Vega, Cathy Pamela por guiarme en el desarrollo de este trabajo de titulación y cumplir con los objetivos planteados.

Tabla de contenido

Índice de tablas

Índice de figuras

Resumen

Sin duda en la actualidad existe falencia en el análisis de información para otorgar un crédito o un préstamo, provocando pérdidas a la institución financiera que involucra gastos de cobranza, notificaciones, pago a abogados entre otros. Gracias a la transformación digital y el avance tecnológico hoy en día se puede utilizar la Inteligencia Artificial y en especial la rama de Machine Learning como estudio para el análisis de datos de los clientes y predecir el incumplimiento en el pago de sus obligaciones con la institución. El objetivo de este trabajo de investigación fue realizar un análisis de los modelos de Machine Learning de aprendizaje supervisado como Random Forest (XGBoots), Regresión Logística y Redes Neuronales y aplicar la metodología CRISP-DM para implementar un modelo predictivo que permita la evaluación del riesgo crediticio. Con este resultado podemos concluir que la utilización de herramientas de Machine Learning ayudan a optimizar la evaluación del riesgo de crédito en la entidad financiera. Con base en esta experiencia, se marca el camino para que, en futuros trabajos se implemente estos modelos en otras áreas como: detención de fraudes, segmentación de clientes o un motor de recomendaciones que pueda sugerir productos y servicios financieros basados en las necesidades y comportamientos de los clientes.

Palabras Claves:

- **REDES NEURONALES**
- **REGRESIÓN LOGÍSTICA**
- **CRÉDITOS BANCARIOS**
- **RIESGO CREDITICIO**

Abstract

Undoubtedly, at the present time exists deficiency in the analysis of information to grant a credit or a loan by causing economic losses to the financial institution involving collection costs, payment notices, payment for covering lawyers' fees, and so on. Thanks to digital transformation and technological advance today, artificial intelligence can be used and especially the branch of Machine Learning that supports the study for customers' data analysis to predict non-compliance with the payment of their obligations with the institution. The objective of this research work was to perform an analysis of Machine Learning Models for supervised learning as Random Forest (XGBoots), logistics regression and neuronal networks and apply the CRISP-DM methodology to implement a predictive model that allows an assessment of credit risk. As a result, it can be concluded that the use of Machine Learning tools help to optimize the assessment of credit risk in the financial establishment. Based on this experience, the way is marked, for future work to implement these models in other areas such as: fraud prevention, customer segmentation or a referral engine that can suggest financial products and services based on customer needs and behaviour.

Key Words:

- **NEURAL NETWORKS**
- **LOGISTIC REGRESSION**
- **BANK CREDITS**
- **CREDIT RISK**

Capítulo I

1 Introducción

1.1 Antecedentes

Los riesgos son la posibilidad de ocurrencia de eventos adversos que puedan generar pérdidas en la entidad o entorpecer el normal desarrollo de las diferentes funciones establecidas, generando el retraso en el cumplimiento de los objetivos trazados. En específico al hablar de riesgo crediticio, según (Elizondo & Altman, 2004, p. 21), indican que: “En su configuración ideal permiten cuantificar la probabilidad de incumplimiento de los deudores con sus obligaciones y la severidad de las perdidas en caso de incumplimiento” (Elizondo & Altman, 2004, p. 21).

La SEPS (Superintendencia de Economía Popular y Solidaria) como entidad de control y regularización para las cooperativas de ahorro y crédito (COAC) del Ecuador, no ha establecido alguna normativa que ayude como lineamiento al control del riesgo de crédito esto ha provocado que algunas cooperativas cierren siendo esta una de las causas a la hora de mitigar los riesgos de crédito, además como lo afirma Ltda et al. (2019) las COAC reguladas por la SEPS, no han desarrollado alguna norma o política que permita el uso de modelos para la gestión crediticia, razón por la cual muchas COAC han sido cerradas a causa del aumento de morosidad en sus carteras de crédito y la falta de liquidez.

Esta afirmación es la principal motivación para dar inicio al presente trabajo de investigación, la limitación de herramientas tecnológicas que ayuden en forma general a controlar el riesgo, ha generado que las entidades financieras con más recursos y estables busquen sus propias alternativas para mitigar el riesgo de crédito. Para esta investigación se ha considerado como entidad financiera a la Cooperativa de Ahorro y Crédito Virgen del Cisne de Ecuador, cuyo proceso de gestión de riesgo de crédito realiza a través de la calificación de crédito. Este proceso no es suficiente para valorar

si un socio es sujeto o no a crédito. Para lograr una mayor evaluación es necesario utilizar herramientas tecnológicas y en específico modelos de predicción del riesgo crediticio.

La propuesta de este modelo predictivo para la evaluación del riesgo crediticio, es el estudio de los modelos y técnicas de Machine Learning enfocados en el análisis de datos predictivos, con el propósito de identificar los aspectos más importantes que caracterizan a cada uno de ellos, y con una evaluación a la técnica elegir el modelo más idóneo que ayude a dar respuestas de forma precisa y con un porcentaje de error mínimo.

1.2 Planteamiento del problema

Los analistas de créditos en la Cooperativa de Ahorro y Crédito Virgen del Cisne presentan un déficit en el análisis de datos, al no contar con herramientas especializadas como minería de datos, inteligencia de negocios o alguna otra herramienta tecnológica como apoyo al análisis. Al presentar esta carencia la entidad financiera cae en índices altos de morosidad o aumento en la cartera vencida esto implica generar gastos de provisión.

Además, por el desconocimiento de los manuales, reglamentos y políticas internas con que cuentan la entidad financiera, no todos los socios tienen el mismo trato en el análisis de crédito lo cual no justifica si esta persona es sujeto o no de crédito, lo que también ocasiona un déficit al realizar los cobros generando como gastos extrajudiciales o contratan empresas especializadas en recuperación de cartera generando más gastos para la entidad.

Además, todo esto ha generado que el tiempo de análisis y respuesta al pedido de un crédito en la entidad financiera sea muy alta. Por lo cual, se plantea el siguiente problema: ¿Cómo el modelo de predicción evaluará el riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne?

1.3 Objetivos de la investigación

1.3.1 Objetivo General

Desarrollar un modelo predictivo para la evaluación del riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne.

1.3.2 Objetivos Específicos

OE1: Construir el Marco Teórico que fundamente el modelo predictivo y el proceso del riesgo crediticio.

OE2: Desarrollar el modelo predictivo.

OE3: Implementar el modelo predictivo para la evaluación del riesgo crediticio.

OE4: Validar el modelo predictivo.

1.4 Justificación e Importancia

Actualmente, la Cooperativa de Ahorro y Crédito Virgen del Cisne cuenta con un proceso manual para el análisis del otorgamiento de un crédito, este proceso comienza con el ingreso de una solicitud de crédito (información como monto, tasa, tiempo, etc.) para luego ingresar a un comité de crédito (personas que analizan y evalúan la solicitud del crédito con información de garantías, fuentes de pago, solvencia patrimonial, etc.) y dando como resultado final el estado de la solicitud de crédito (negada o aprobada). Sin embargo, mediante los informes que presenta el administrador de riesgos a la gerencia muestra indicadores de morosidad muy altos, siendo este indicador el calificador que le dan a las cooperativas la entidad de control.

Ante esta perspectiva se decide realizar el presente proyecto enfocado en la investigación, diseño e implementación de un modelo predictivo para la evaluación del riesgo crediticio; esto contribuirá al proceso de toma de decisiones sobre si los nuevos y/o actuales clientes pueden presentar un riesgo de no pago o de fraude, en forma ágil, oportuna y segura.

El presente proyecto contribuirá de primera mano a un análisis de tipo predictivo y pronóstico del riesgo crediticio que actualmente no se los evalúa con ninguna herramienta. Esto será realizado mediante la aplicación de técnicas de Machine Learning y predicción siendo la parte novedosa de esta investigación, además el uso de estas técnicas dentro de las áreas de Ciencias de la Computación e Ingeniería del software.

Los principales beneficiarios del proyecto serán la Cooperativa y los socios que al obtener esta herramienta tecnológica obtendrán un resultado rápido de su crédito.

1.5 Hipótesis

H0: Si se desarrolla un modelo predictivo entonces se evaluará el riesgo crediticio optimizando el otorgamiento de créditos en la Cooperativa de Ahorro y Crédito Virgen del Cisne.

1.6 Variables de la investigación

Variable Independiente: Desarrollar un modelo predictivo

Variable Dependiente: Evaluar el riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne.

1.7 Metodología de investigación

1.7.1 Tipo de investigación

El presente trabajo de investigación está enmarcado dentro del tipo de investigación aplicada porque utiliza conocimientos previos y los aplica a una realidad concreta.

1.7.2 Niveles de investigación

Investigación Exploratoria: La investigación es de nivel exploratorio porque se analiza e identifica los problemas de la entidad financiera con el objetivo de obtener la información inicial para el desarrollo del trabajo.

Investigación Descriptiva: Esta investigación tiene nivel descriptivo porque se realiza un análisis para determinar los datos relevantes del proceso manual de un otorgamiento de crédito, con el objetivo de encontrar las variables que ayuden con el diseño y construcción del modelo predictivo.

1.7.3 Métodos y técnicas de investigación

Método Histórico Lógico: Para determinar los antecedentes históricos de las técnicas para la optimización del riesgo crediticio.

Método Analítico Sintético: Para desarrollar la situación problemática, marco teórico, y para procesar la información obtenida en la investigación de campo.

Método Hipotético Deductivo: Presente en la formulación de la hipótesis y en todo el proceso de investigación

Método de Modelación: Determina el enfoque, para poder establecer que va a solucionar el modelo predictivo, determinando el alcance de este.

Capítulo II

2 Marco Teórico

2.1 Antecedentes Históricos

2.1.1 Acuerdos de Basilea

Los llamados Acuerdos de Basilea consisten en leyes bancarias y recomendaciones regulatorias emitidas por el Consejo de Supervisión Bancaria de Basilea, que incluye a los directores de los principales bancos del mundo (Wolters Kluwer, 2019).

El propósito del consejo de Supervisión bancaria de Basilea, es dar a conocer los lineamientos y recomendación que ayuden a mejorar el control en el sector financiero del estado o los estados que se acojan al mismo, de esta forma supervisarán el control y flujo de los bancos o entidad financiera.

2.1.1.1 Basilea I

El acuerdo es el resultado del consenso de los bancos centrales de las naciones del G10 (grupo de países que accedieron participar en el Acuerdo General de Préstamos) para ejercer estándares de capital mínimos en común en las diversas industrias bancarias de todo el mundo. Se logró publicar en 1988 y sitúa al capital como el primordial pilar en la regulación bancaria. En este acuerdo se definió el requerimiento mínimo de 8% entre el capital y los activos ponderados por peligro. Para septiembre de 1993, todas las naciones del G10 ya habían alcanzado esta meta. Para principios del año 1996, ya se habían realizado 2 enmiendas al acuerdo, en la primera, se hacía una aclaración más descriptiva sobre las provisiones en general (noviembre 1991) y en la segunda, se ponía en prueba el riesgo de las compensaciones bilaterales de las exposiciones crediticias de los bancos (abril 1995). En el primer mes del año de 1996 la junta emitió una enmienda al acuerdo de capital para poder ingresar al riesgo de mercado (Asobanca, 2019).

2.1.1.2 Basilea II

En junio de 1999, la junta de Basilea emitió una iniciativa para llevar a cabo un nuevo marco de requerimientos de capital y suplir el acuerdo del año 1988. Esto conllevó a la ejecución de un marco de capital inspeccionado en junio de 2004 el cual principalmente se sabe cómo Basilea II. Este nuevo acuerdo cuenta con 3 pilares: a) requerimientos mínimos de capital, b) examen del supervisor y c) disciplina de mercado. El acuerdo de Basilea II no se centra sólo en bancos que operan internacionalmente sino en todo un conjunto financiero diversificado; o sea, holdings bancarios en los cuales se integran entidades de seguros, bancos en el sitio de residencia y las que son internacionalmente activos. Si es comparable este nuevo acuerdo con la versión previa, destacan los próximos puntos de vista: el mantenimiento de la definición de interacción mínima de capital entre activos ponderados por riesgo (no menor al 8%) y que el Tier 2 (capital de grado 2) está reducido al 100% del Tier 1 (capital ordinario de Grado 1). Además, se introducen novedosas metodologías para medir el riesgo de crédito y el operacional; mientras tanto que, con interacción al riesgo de mercado, el acuerdo especifica que los bancos tienen que medir, una y otra vez, los costos de las herramientas por medio de organizaciones especializadas o mediante sus propios modelos. Al final, los Pilares 2 y 3 de este acuerdo promueven el mantenimiento de un diálogo activo entre el supervisor y las entidades reguladas; y el valor de tener una política formal de divulgación de información aprobada por el directorio de las mismas, respectivamente (Asobanca, 2019).

2.1.1.3 Basilea III

El marco de Basilea III nace como una contestación del Comité de Supervisión Bancaria a la crisis financiera mundial del año 2008 (aunque anteriormente a este acontecimiento ya se debatía la necesidad de llevar a cabo nuevas y novedosas reformas). Entre los aspectos más relevantes de cambio destacan: el mejoramiento de

la calidad e incremento del grado de los requerimientos de capital para que los bancos sean más resilientes para encarar pérdidas, la unión de colchones de capital que se alimenten a lo largo de las etapas de coyuntura positiva para determinar prociclicidad, el establecimiento de un sistema para las exposiciones al riesgo para mitigar riesgos sistémicos derivados de las interacciones entre entidades financieras, la integración de un colchón de capital para encarar a las externalidades que proceden de bancos sistémicamente relevantes, y la adhesión de recursos regulatorios y estándares para el control y manejo de la liquidez (Asobanca, 2019).

2.2 Antecedentes Referenciales

Un primer tema corresponde a Salinas Pérez & Chee Tse (2020), quien realizó el “Modelo de medición de riesgo crediticio en entidades financieras basado en minería de datos. Caso práctico: CACPECO LTDA.”, en este trabajo se manejaron teorías sobre el proceso de otorgamiento de crédito, la problemática a la hora de seleccionar un potencial sujeto de crédito ya que este proceso de análisis se realizaba de forma manual y con herramientas ofimáticas convencionales, en la parte del desarrollo se aborda conceptos de minería de datos, técnicas de predicción, herramientas ETL y metodologías para el proceso de minería de datos.

La investigación se enmarcó dentro de un proyecto factible, La muestra de estudio fue de 144323 registros, la metodología utilizada fue CRISP-DM la cual abarca 6 faces: desde la compresión del negocio hasta la implementación o despliegue del modelo evaluado, validado y aceptado. El estudio confirmó que la técnica que mejor se adapta a las necesidades del negocio fue árboles de decisión, ya que obtuvo el porcentaje más alto de precisión en la predicción de los clientes buenos o malos.

Este trabajo se relaciona con la investigación en curso, ya que proporciona información para llevar a cabo un proyecto de minería de datos y las técnicas que se

deben tomar en cuenta a la hora de seleccionar el modelo más idóneo que ayude a solucionar el problema de esta investigación.

Un segundo tema correspondiente a Toscano Palomo (2019), en su trabajo denominado “Modelo predictivo del comportamiento de la cartera crediticia para cooperativas de ahorro y crédito”, habla cerca de las cooperativas de ahorro y crédito conceptos y actividades, crédito, riesgo de crédito y todos los procesos que se lleva a cabo a la hora de entregar un crédito.

Este estudio demostró que se puede crear una metodología propia que aborda el diseño, la implementación y la validación del modelo, la muestra de estudio está dada desde el año 2015 hasta el 2020. Ratificando así que el modelo árbol de decisión es el que más acierta en la predicción del comportamiento del socio. Este trabajo se relaciona con la investigación planteada ya que muestra desde otro punto de vista, el desarrollo del modelo pero que a la vez llega a un mismo objetivo que es la predicción.

Un tercer tema corresponde a Fernández Hidalgo (2013) , que lleva por título “Diseño de un modelo de scoring de crédito para la cooperativa de ahorro y crédito Pujili ltda. Ubicada en el cantón Pujilí, provincia de Cotopaxi”, se trata de un proyecto donde menciona conceptos relevantes sobre el riesgo de crédito y las técnicas de evaluación del riesgo crediticio, pero ya en el ámbito Financiero. Este estudio genera puntos de vista Financieros que se deben tomar en cuenta a la hora de generar el modelo predictivo tales como:

- Incrementos de índices de morosidad.
- Limitada liquidez y rentabilidad.
- Incumplimiento de los objetivos Institucionales de Cartera y Crédito.
- Limitada evaluación de capacidad de pago.
- Incremento de provisiones incobrables.

- Limitada capitalización de patrimonio.

Este trabajo es pertinente con la investigación planteada, ya que aborda temas y conceptos básicos que ayudan a solucionar los objetivos del negocio y minería de datos.

2.3 Antecedentes Conceptuales

2.3.1 Crédito

Un crédito es una operación financiera donde una persona acreedor (entidad financiera), presta una determinada cifra monetaria a otro, denominado deudor, quien asegura al acreedor que retornará esta cifra monetaria en el plazo previamente estipulado más un valor adicional de interés (Montes de Oca, 2015).

2.3.1.1 Tipos de Crédito

Crédito de Consumo. Son aquellos créditos otorgados a personas naturales, destinado a la compra de bienes, servicios o gastos no relacionados con una actividad productiva, comercial y otras compras y gastos, incluidos los créditos prendarios de joyas y la adquisición de vehículos livianos que no sean de uso para una actividad productiva y comercial (Junta de Política y Regulación Monetaria y Financiera, 2015).

En la figura 1, se muestra un cuadro resumen de los diferentes tipos de créditos de la línea de consumo en la entidad financiera.

Figura 1

Tipos de Créditos de Consumo

Tipos de créditos de Consumo	Montos	Plazos Máximos	Tasas de Interés	Garantías	Periodicidad de Pago	Aporte para el fondo irrepartible de reserva legal	Encaje	Capacidad de Pago
Consumo Estándar	Desde 300 dólares hasta 35000 dólares	84 meses	Políticas de Tasas de Interés	Auto - liquidable; Garantia DPF, personales o solidarias	Mensual	2.50%	0	Solo socio con rol mecanizado 60%; socio y cónuge con rol mecanizado 70%
Credi Nómina	Desde 300 dólares hasta 35000 dólares	84 meses	Políticas de Tasas de Interés	Auto - liquidable; Garantia DPF, personales o solidarias	Mensual	2.50%	0	Solo socio con rol mecanizado 60%; socio y cónuge con rol mecanizado 70%
Consumo Garantia DPF	120% del valor del DPF	60 meses	Políticas de Tasas de Interés	Garantias DPF	Mensual	2.50%	0	Solo socio con rol mecanizado 60%; socio y cónuge con rol mecanizado 70%
Autoconsumo (Auto liquidable)	Equivalente al 90% del DPF	Al vencimiento del DPF	16%	Auto - liquidable (DPF)	Al Vencimiento	0%	0	N/A
Consumo emergente	Desde 300 dólares hasta 10000 dólares	Tiempo máximo de vencimiento crédito vigente	Políticas de Tasas de Interés	Auto - liquidable; personales o solidarias	Mensual	2.50%	0	Solo socio con rol mecanizado 60%; socio y cónuge con rol mecanizado 70%

Crédito Inmobiliario. Son créditos otorgados con garantía hipotecaria a personas naturales para la construcción, reparación, remodelación y mejora de inmuebles propios; para la adquisición de terrenos destinados a la construcción de vivienda propia; y, para la adquisición de vivienda terminada para uso del deudor y su familia, no categorizada en el segmento de crédito vivienda de interés social y público (Junta de Política y Regulación Monetaria y Financiera, 2015).

En la figura 2, se muestra un cuadro resumen de los diferentes tipos de segmentos de la línea de crédito inmobiliario en la entidad financiera.

Figura 2

Segmentación de Créditos Inmobiliarios

Segmentación	Monto Máximo	Plazo Máximo	Tasa de Interés	Garantía	Periodicidad de Pago	Aporte Futuras Capitalizaciones	Porcentaje de Financiamiento del avalúo	Capacidad de Pago
Adquisición de vivienda nueva	35000	84 meses	Tasas de Interés propias	Hipotecaria	Mensual	3%	80%	Solo socio con rol de pagos 60%; socio y cónyuge con rol de pagos 70%; sin relación de dependencia 50%.
Construcción en terreno propio	35000						Indistinto	
Adquisición de Vivienda Usada	35000						80%	
Mejoramiento, remodelación y /o ampliación	20000	66 meses					Indistinto	

Microcrédito. Son créditos otorgados a una persona natural o jurídica con un nivel de ventas anuales inferior o igual a USD 100,000.00 o a un grupo de prestatarios con garantía solidaria, destinado a financiar actividades de producción y/o comercialización en pequeña escala, cuya fuente principal de pago la constituye el producto de las ventas o ingresos generados por dichas actividades verificados adecuadamente por la entidad (Junta de Política y Regulación Monetaria y Financiera, 2015).

En la figura 3, se muestra un cuadro resumen de los diferentes tipos de créditos de la línea de microcréditos en la entidad financiera.

Figura 3

Tipos de MicroCréditos

Tipos de microcrédito	Montos	Encaje	Plazos Máximos	Tasa de Interés	Garantías	Periodicidad de Pago	Aporte Patrimonial	Capacidad de Pago
Micro estándar	Desde 300 Hasta 60000 dólares	Tasas de Interés propias	96 meses	Cuadro de Tasas de Interés	Auto - liquidable; garantía DPF, personales, solidarias, hipotecarias	Mensual, bimensual, trimestral Y cuatrimestral	Minorista y Acumulación Simple al 2.10 % Acumulación Ampliada al 2%	50%
Microproductivo	Desde 300 Hasta 10000 dólares		48 meses		Auto - liquidable; personales o solidarias	Mensual, bimensual, trimestral y cuatrimestral	3%	
Microinclusivo	Desde 300 Hasta 8000 dólares		36 meses	16%	Personales o solidarias	Mensual, bimensual	2.10%	
Microemergente	Desde 300 Hasta 10000 dólares		Tiempo máximo de vencimiento crédito vigente	Tasa de interés de acuerdo al segmento sumado las dos operaciones de crédito, a excepción de los créditos otorgados a los cónyuges	Auto - liquidable; personales o solidarias	Mensual	2.10%	
Microgrupal	Desde 300 Hasta 5000 dólares por integrante.		18 meses	Tasas de interés microcrédito	Personales o solidarias	Mensual	Minorista y Acumulación Simple al 2.10 % Acumulación Ampliada al 2%	
MicroCrediauto	Desde 10001 hasta 30000 dólares		72 meses	Tasas de interés microcrédito	Prenda Industrial	Mensual	2%	
Micro FOGEPS	desde 300 hasta 15000 dólares		60 meses	Tasas de interés microcrédito	certificado de CONAFIPS	Mensual	Minorista y Acumulación Simple al 2.10 % Acumulación Ampliada al 2%	
MicroAgro	Desde 300 hasta 20000 dólares		48 meses	Cuadro de Tasas de Interés	Personales, solidarias, hipotecaria, autoliquidable	Mensual Trimestral Cuatrimestral Semestral	Minorista y Acumulación Simple al 2.10 % Acumulación Ampliada al 2%	

Microcrédito Agrícola y Ganadero (MicroAgro). Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del Sistema

Financiero Nacional, sea menor o igual a USD 100,000.00, incluyendo el monto de la operación solicitada para financiar actividades agrícolas y ganaderas (Junta de Política y Regulación Monetaria y Financiera, 2015).

Para el Microcrédito se establecen los siguientes sub-segmentos de crédito:

a) Microcrédito Minorista. Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del sistema financiero, sea menor o igual a USD 1.000, incluyendo el monto de la operación solicitada (Junta de Política y Regulación Monetaria y Financiera, 2015).
b) Microcrédito de Acumulación Simple. Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del sistema financiero sea superior a USD 1.000 y hasta USD 10.000, incluyendo el monto de la operación solicitada (Junta de Política y Regulación Monetaria y Financiera, 2015).
c) Microcrédito de Acumulación Ampliada. Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del sistema financiero sea superior a USD 10.000 incluyendo el monto de la operación solicitada (Junta de Política y Regulación Monetaria y Financiera, 2015).

2.3.2 Riesgo Crediticio

El Riesgo de Crédito se define como el riesgo derivado de la incertidumbre de la capacidad del deudor frente a sus obligaciones contractuales. Este surge cuando los flujos de caja comprometidos por préstamos y valores pueden no ser pagados oportuna o totalmente según lo estipulado en el contrato, resultando así una pérdida financiera para la Cooperativa.

Según la SBS (2003) por medio de su normativa (No JB-2003-602), define al riesgo crediticio como la probabilidad de pérdida a causa del incumplimiento del prestatario o la contraparte en operaciones directas, indirectas o de derivados que conlleven el no pago, el pago parcial o la carencia de posibilidad en el pago de las obligaciones pactadas.

2.3.2.1 Elementos del Riesgo de Crédito

2.3.2.1.1 Riesgo de Incumplimiento

Es la probabilidad de que el valor principal o los intereses, o los dos a la vez no sean pagados total o parcialmente por el acreditado, por lo que el riesgo lo asume quien lo proporciona, declarándose como incumplimiento de pago una vez que un pago programado no se ha llevado a cabo en un tiempo definido, o se efectúa con posterioridad a la fecha en que estaba programado el pago (García Sánchez & Sánchez Barradas, 2005).

2.3.2.2 Riesgo de Exposición

Según García Sánchez & Sánchez Barradas (2005) menciona que este riesgo se produce por la incertidumbre en relación a los montos futuros en riesgo, una vez que los créditos tienen la posibilidad de pagarse total o parcialmente de forma anticipada, en particular una vez que no existe penalización, se muestra el peligro de exposición, debido a que no se sabe con precisión el plazo de liquidación y por esto se dificulta la estimación de los montos en riesgo.

2.3.2.2.1 Riesgo de Recuperación

Este riesgo está sujeto a las recuperaciones del crédito mediante la activación de las garantías, dichas garantías son las que aseguran el cumplimiento de una obligación directa o indirectamente, según García Sánchez & Sánchez Barradas (2005) menciona que la recuperación no se puede presagiar debido a que es dependiente del tipo de incumplimiento y del número de componentes involucrados con las garantías

que se hayan recibido. Una garantía disminuye el riesgo de crédito si esta puede desarrollarse simple e inmediatamente, a un valor conveniente acorde al valor adeudado, incluyendo los costos (moratorios, judiciales, etc.).

2.3.2.3 Modelos de medición del riesgo de crédito

2.3.2.3.1 Modelos Tradicionales

Estos modelos se basan en la experiencia del analista de crédito, utilizando factores de aprendizaje denominado las C del crédito, según García & García (2010) “Este tipo de modelos involucra el criterio subjetivo de cada analista que se hace basándose en valoraciones de acuerdo con la experiencia adquirida en la asignación de créditos”(p. 300).

2.3.2.3.2 Sistemas Expertos

Los sistemas expertos intentan captar la experticia de los analista de crédito de esta forma tener una base de conocimiento para evaluar el estado de una solicitud de crédito, a pesar de ello quedan limitados solamente a la fase de calificación, debido a que no tienen la posibilidad de entablar un vínculo que puedan identificar la posibilidad de impago y la gravedad de la pérdida esperada (García & García, 2010).

El modelo más importarte de los sistemas expertos son los cinco C del crédito, son los elementos que se debe tener en cuenta para obtener un valor de clasificación y decidir si se otorga el crédito o no, se detalla cada uno:

- Capacidad: Es el elemento más relevante en la elección de la entidad financiera. Se basa en evaluar la capacidad y vivencia en los negocios que tenga el individuo u organización, su gestión y resultados prácticos. Para esta evaluación se toma presente la antigüedad, el incremento de la organización, sus canales de repartición, ocupaciones, giro del negocio, región de predominación, número de empleados, sucursales, etc., debido a que es necesario saber cómo pagará el préstamo y para eso es necesario establecer el flujo de efectivo del comercio;

inclusive requieren el historial del crédito del propietario y sus deudas pasadas y presentes (García & García, 2010).

- Capital: Tiene relación con los valores invertidos en el comercio del deudor, así como obligaciones, es un análisis de las finanzas. Para la evaluación es preciso el estudio de su situación financiera. El estudio financiero descriptivo posibilita conocer por completo las maneras de pago, el flujo de ingresos y egresos, así como la capacidad de endeudamiento (García & García, 2010).
- Colateral: Son todos esos recursos que dispone el deudor para asegurar el cumplimiento del pago en el crédito (las garantías). Se evalúa por medio del costo y la calidad de sus bienes, debido a que en el estudio del crédito se establece que no tendrá que otorgarse un crédito sin tener prevista una segunda fuente de pago (García & García, 2010).
- Carácter: Son las condiciones (credibilidad, confiabilidad) del deudor con que cuenta para cumplir con el crédito. Se busca datos acerca de sus hábitos de pago y comportamiento en operaciones crediticias pasadas y presentes, relacionadas con sus pagos. La evaluación del carácter o solvencia moral de un socio debería hacerse desde recursos contundentes, cuantificables y verificables (García & García, 2010).
- Condiciones: Son los componentes externos que tienen la posibilidad de dañar la marcha del comercio del deudor, como las condiciones económicas, política y económica del territorio. Aun cuando estos componentes no permanecen bajo el control del deudor, se piensan en el análisis de créditos para prever sus probables efectos (García & García, 2010).

2.3.2.3.3 Sistemas de Calificación

Según García & García (2010) manifiesta que el sistema de calificación más antiguo de créditos es el que desarrolló la Oficina de Control de Moneda (OCC, por su

sigla en inglés) estadounidense de Norteamérica, el cual ha usado reguladores y banqueros en muchas naciones con la intención de evaluar la adecuación de sus reservas para pérdidas crediticias. En la actualidad muchas entidades financieras han ampliado este método de calificación incrementando más elementos, a la forma de cálculo en el otorgamiento de un crédito.

2.3.2.4 Modelos Modernos

Estos modelos usan técnicas más sofisticadas para evaluar el riesgo de crédito, se detalla los más importantes.

2.3.2.4.1 Modelo KVM (Kealhofer, McQuown y Vasicek) de monitoreo de crédito

Según García & García (2010) menciona que, mediante este modelo se estima la probabilidad de incumplimiento utilizando técnicas de simulaciones tanto del capital, como de la volatilidad del rendimiento deseado de los activos y su costo presente. Es decir que la posibilidad de impago de la deuda depende del incremento de sus activos a futuro mas no de su flujo de capital, este análisis se lo puede detallar en las siguientes etapas:

Primera etapa: estimación del valor del activo y la volatilidad del rendimiento: Los modelos financieros principalmente piensan un costo de mercado de los activos y no el costo en libros que solamente representan los precios históricos de los activos tangibles, netos de depreciación. El cálculo del costo de mercado de los activos de la empresa y su volatilidad podría ser bastante sencillo si cada una de sus obligaciones se evaluaran a costo del mercado cada día (García & García, 2010).

Segunda etapa: cálculo del riesgo de los activos, en el que se incluyen el riesgo del negocio y del sector en el que trabaja la empresa: Este riesgo se mide por la volatilidad de los activos, no obstante, esta volatilidad está relacionada con los valores de una compañía, sin embargo, no es exactamente la misma, debido a que el adeudo de una compañía impacta la volatilidad de los activos de la organización. De modo que,

las compañías con baja volatilidad de los activos, tienden a estar muchísimo más endeudadas que las que poseen altas (García & García, 2010).

Tercera etapa: derivación de la probabilidad de incumplimiento: Este modelo pretende descubrir la interacción entre la distancia al incumplimiento y la posibilidad de que se genere, por lo que se prepara una tabla que relaciona la posibilidad de incumplimiento con los niveles de distancias de default (García & García, 2010).

2.3.2.4.2 Modelo desarrollado para mercados emergentes: CyRCE

Este modelo se divide en dos parámetros de evolución que son la probabilidad de incumplimiento del crédito de forma individual y la concentración de cartera es decir a mayor número de créditos hay una concentración de riego de impago individual alta, así como lo menciona García & García (2010) El modelo llega a una expresión para el tamaño de riesgo que posibilita entablar la interacción directa entre el peligro de crédito y los límites más relevantes: capital solicitado para encarar riesgos y parámetros personales de cada segmento de la cartera para fines de diversificación.

2.3.3 Machine Learning

Machine Learning (ML) o Aprendizaje Automático es una rama de la IA (Inteligencia Artificial), que hace referencia a la capacidad de análisis de datos que enseña a una máquina o software para aprender mediante la adaptación de ciertos algoritmos, algo que es natural de los humanos.

Además como menciona Pinto Galindo (2020) El Machine Learning o Aprendizaje Automático construye algoritmos o modelos que predicen una acción o situación, en base al aprendizaje de datos históricos, mediante la utilización de la informática y las estadísticas.

Según Rouhiainen (2018) El aprendizaje automático usa algoritmos para aprender de los patrones de datos. Por ejemplo, los filtros de spam de correo electrónico utilizan este tipo de aprendizaje con el fin de detectar qué mensajes son

correo basura y separarlos de aquellos que no lo son. Éste es un sencillo ejemplo de cómo los algoritmos pueden usarse para aprender patrones y utilizar el conocimiento adquirido para tomar decisiones (p. 20).

2.3.3.1 Tipos de Aprendizaje Automático

2.3.3.1.1 Aprendizaje Supervisado

Según Rouhiainen (2018) menciona que, en el aprendizaje supervisado, los algoritmos utilizan datos previamente etiquetados u organizados mediante la intervención del humano, para indicar como esta categorización mostrará una nueva información. El desarrollo de este modelo parte del ingreso de un conjunto de datos categorizados, de este modo el proceso de entrenamiento se realiza con más efectividad y el resultado de salida sea más óptimo.

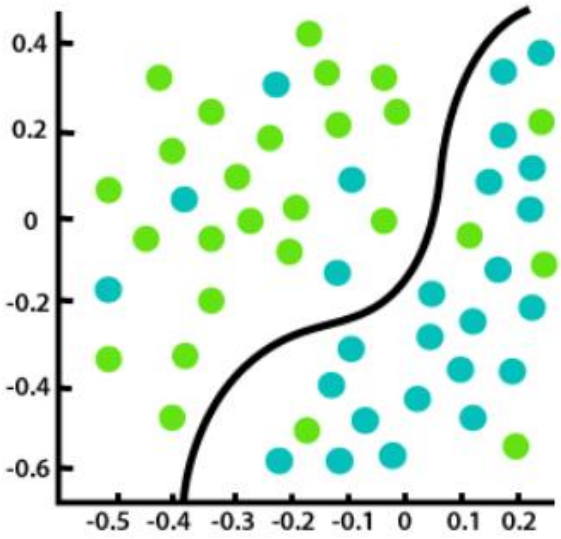
El desarrollo de estos modelos está divido en dos categorías:

Clasificación: Estos algoritmos clasificación toman datos previamente etiquetados y aprenden de estos patrones, de los datos que tienen la posibilidad de usarse para predecir una variable categórica de salida, siendo esta una variable de agrupación de dos o más conjuntos de salida (Pinto Galindo, 2020).

Como se muestra en la figura 4, el valor resultante esta agrupado o clasificado en un conjunto de datos.

Figura 4

Muestra de Clasificación

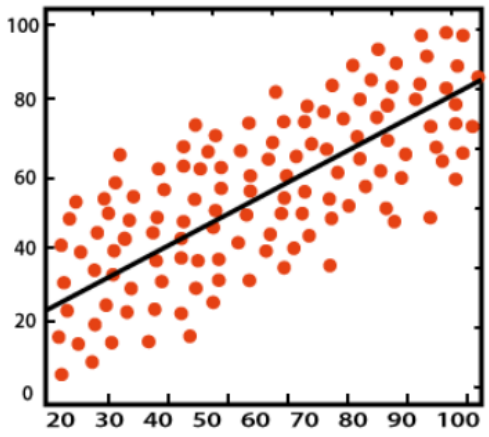


Regresión: Este algoritmo es eficaz para pronosticar productos consecutivos, la predicción se muestra como una variable de salida (Pinto Galindo, 2020). El costo predicho puede usarse para detectar la interacción lineal entre atributos.

Como se muestra en la figura 5, aquí se representa un ejemplo de una regresión.

Figura 5

Muestra de Regresión



2.3.3.1.2 Aprendizaje no Supervisado

En este tipo de aprendizaje, los algoritmos no utilizan datos previamente etiquetados u organizados para indicar cómo tendría que ser categorizada la nueva información, al contrario estos algoritmos deben descubrir la forma de clasificarlos, en otras palabras, este procedimiento no necesita la intervención humana (Rouhiainen, 2018). La finalidad del aprendizaje no supervisado es reestructurar los datos de ingreso en novedosas propiedades o un conjunto de objetos con patrones semejantes.

Los modelos de aprendizaje no supervisado se pueden categorizar en:

Reducción Dimensional: Estos algoritmos de reducción de dimensiones toman datos que no están etiquetados de altos volúmenes, datos con muchas variables y aprenden una forma de representar en un menor número de dimensiones, estas técnicas tienen la posibilidad de utilizar la exploración de datos, o como un paso de pre procesamiento en la técnica de aprendizaje automático (Pinto Galindo, 2020).

Agrupación: Según Pinto Galindo (2020) menciona que los algoritmos de agrupamiento toman datos sin estar previamente etiquetados y aprenden patrones de agrupamiento en los datos.

2.3.3.1.3 Aprendizaje de Refuerzo

Según Rouhiainen (2018) menciona que “En el aprendizaje por refuerzo, los algoritmos aprenden de la experiencia. En otras palabras, tenemos que darles «un refuerzo positivo» cada vez que aciertan” (p. 21). El objetivo en el aprendizaje por refuerzo es aprender a mapear situaciones de acciones para maximizar una cierta función de recompensa.

2.3.3.2 Modelos de Machine Learning de Aprendizaje Supervisado

2.3.3.2.1 Regresión Logística

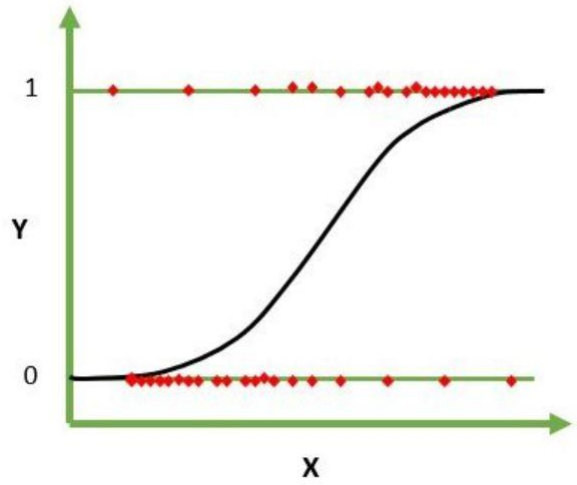
Los modelos de regresión logística permiten medir la posibilidad que tiene un individuo de pertenecer a un grupo de pagadores o no. La clasificación se hace según el

comportamiento de las variables independientes de cada individuo (Ticona Carpio, 2016).

Es un algoritmo de clasificación de aprendizaje supervisado, se utiliza para estimar valores discretos (valores binarios como 0/1, sí / no, verdadero / falso) basados en un conjunto dado de variables independientes. Es uno de los algoritmos más simples y más utilizados para la clasificación de dos clases, es de fácil implementación para resolver problemas de clasificación binaria. La figura 6 representa una regresión logística, donde los conjuntos (puntos rojos) están clasificados en el rango de 1 o 0.

Figura 6

Regresión Logística



2.3.3.2.2 Support Vector Machine (SVM)

Es un algoritmo de clasificación y regresión desarrollado en la década de los 90, dentro del campo de la ciencia computacional. Según Ticona & Cesar (2016) menciona “Las máquinas de soporte vectorial (Support Vector Machines) surgieron como un método de clasificación basado en la teoría de minimización del riesgo estructural de Vapnik.” (p. 29)

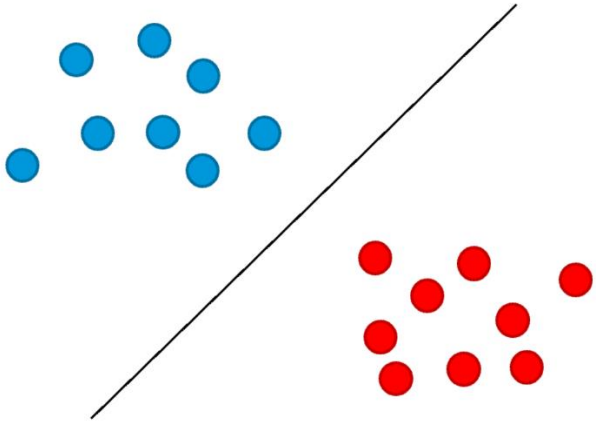
Aunque inicialmente se desarrolló como un método de clasificación binaria, su aplicación se ha extendido a problemas de clasificación múltiple y regresión. SVMs ha

resultado ser uno de los mejores clasificadores para un amplio abanico de situaciones, por lo que se considera uno de los referentes dentro del ámbito de aprendizaje estadístico y machine learning.

Como se puede ver en la Figura 7, el SVM encuentra el hiperplano que maximiza el margen de separación entre clases (puntos azules y rojos).

Figura 7

Ejemplo básico de un SMV



2.3.3.2.3 Artificial Neural Network (ANN)

La Red Neuronal Artificial (Artificial Neural Network) es un algoritmo que está basado en el funcionamiento de las redes de neuronas biológicas. Las neuronas que todos tenemos en nuestro cerebro están compuestas de dendritas, el soma y el axón: Las dendritas se encargan de captar los impulsos nerviosos que emiten otras neuronas. Estos impulsos, se procesan en el soma y se transmiten a través del axón que emite un impulso nervioso hacia las neuronas contiguas.

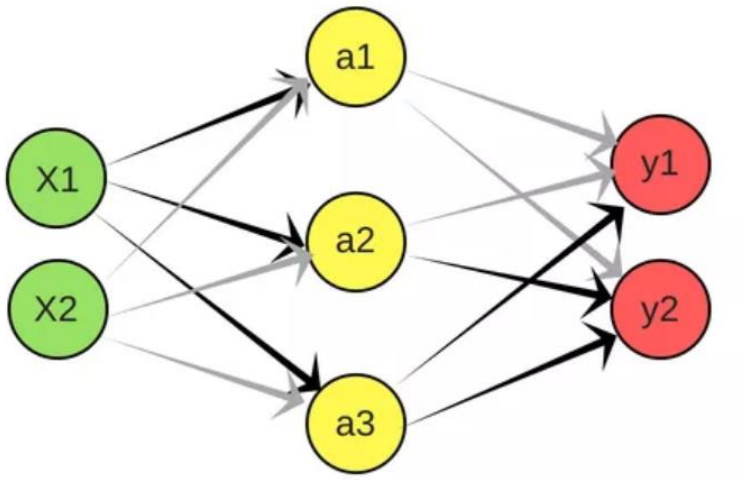
Tienen la posibilidad de aplicarse a problemas de predicción, categorización o control en un extenso espectro de campos como las finanzas, la psicología cognitiva/neurociencia, la medicina, la ingeniería y la física. Las redes neuronales se usan una vez que no se sabe la naturaleza precisa de la interacción entre los valores de

ingreso y de salida. Una característica clave de las redes neuronales es que aprenden la interacción entre los valores de acceso y salida por medio del entrenamiento (Timón, 2017).

Como se aprecia en la Figura 8, **x** son las entradas, **a** las capas de la red, **y** las salidas.

Figura 8

Ejemplo básico de una Red Neuronal



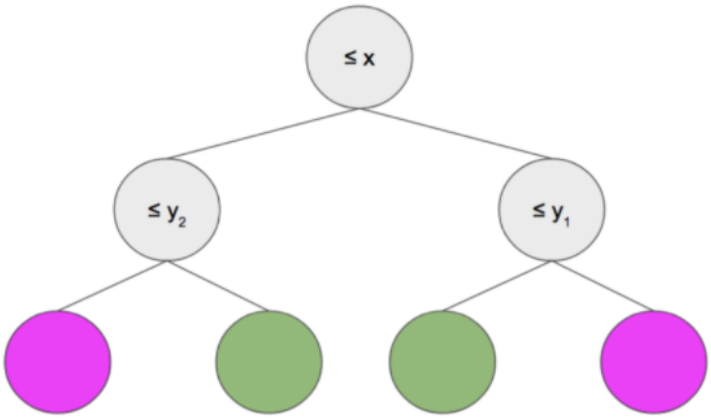
2.3.3.2.4 Decision Tree

Es un algoritmo de aprendizaje supervisado (que tiene una variable objetivo predefinida) que se utiliza principalmente en problemas de clasificación, donde se pueden utilizar variables de entrada, salida, categóricas y continuas. Según Ticona & Cesar (2016) menciona que un Árbol de Decisión (Decision Tree) tiene entradas de un objeto o de un grupo de atributos y desde esto regresa una respuesta de decisión. En esta técnica, dividimos la población o muestra en dos o más conjuntos homogéneos (o subpoblaciones) basados en el divisor / diferenciador más significativo en las variables de entrada.

Como se aprecia en la Figura 9, cada circulo representa una condición, generando así los nodos del árbol.

Figura 9

Ejemplo básico de un Árbol de Decisión



2.3.3.2.5 K-Nearest Neighbors (kNN)

Es un algoritmo de aprendizaje supervisado, se utiliza para problemas de clasificación y regresión. Sin embargo, se usa más ampliamente en problemas de clasificación en la industria.

El algoritmo vecino más próximo k-NN (Nearest Neighbor) forma parte de la clase de procedimientos estadísticos de reconocimiento de patrones. El procedimiento no obliga a priori ni una suposición sobre el reparto de la que se extrae la muestra de modelado (Timón, 2017).

K vecinos más cercanos es un algoritmo simple que almacena todos los casos disponibles y clasifica los casos nuevos por mayoría de votos de sus k vecinos, este proceso viene influenciado por tres factores importantes:

- La medida de distancia de cada vecino.
- La regla de decisión para la clasificación.
- El número de vecinos para clasificar.

2.3.3.2.6 Random Forest (RF)

Está formado por un conjunto (ensemble) de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante bootstrapping). Esto implica que cada árbol se entrena con unos datos ligeramente distintos. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

XGBoost: (Extra Gradient boosting) basado en arboles de desicion que se implementa en minería de datos para clasificar o profetizar sobre una variable objetivo (y), por medio, del aprendizaje automático que se hace sobre un set de datos (Espinosa Zúñiga, 2020). Sus principales ventajas son:

- Se puede utilizar como clasificador o para predecir.
- Es un modelo simple de implementar.
- Las predicciones son muy acertadas pese al volumen de datos sin excluir ninguna.

2.3.3.3 Técnicas para evaluar el desempeño de los Modelos de Machine Learning

Para la validación de los modelos verificando su porcentaje de predicción y su grado de certeza se han creado varias técnicas de evaluación, se detalla las más importantes.

2.3.3.3.1 Validación cruzada (Cross-Validations)

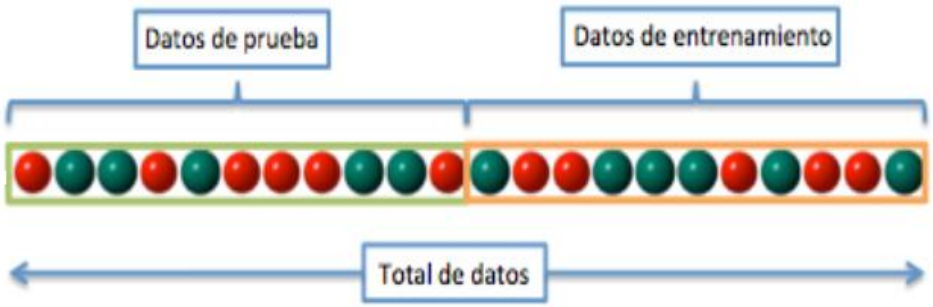
La validación cruzada es una técnica para evaluar modelos de ML mediante la división de datos de entrenamiento y datos de testeo.

“La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.” (Ticona Carpio, 2016, p. 31)

Esta técnica de evaluación consiste en dividir el set de datos total en dos subconjuntos, el primero denominado datos de entrenamiento o training y el segundo denominado datos de prueba o testing este último valida el análisis del primero, de forma que la función de aproximación sólo se ajusta con el conjunto de datos de entrenamiento y a partir de aquí calcula los valores de salida para el conjunto de datos de prueba o testing, en la figura 10 se observa el funcionamiento de esta técnica, donde se divide el set de datos en los subconjunto de datos antes mencionado.

Figura 10

Ejemplo Validación Cruzada



2.3.3.3.2 Matriz de Confusión o Matriz de error

La matriz de confusión es un instrumento bastante eficaz para verificar que tan bueno es un modelo de aprendizaje supervisado, mediante el desempeño de este se puede verificar unas clases de otras, lo que posibilita laborar de manera separada con diversos tipos de error.

Así como menciona (Ticona Carpio, 2016, p. 34) “La matriz de clasificación es una herramienta importante para evaluar los resultados de la predicción, ya que hace que resulte fácil entender y explicar los efectos de las predicciones erróneas.” (p. 34)

Como se observa en la figura 11, una matriz de confusión y sus elementos por columna y fila.

Figura 11

Matriz de Confusión

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

Cada columna de la matriz representará el número de predicciones para cada clase realizadas por el modelo, y cada fila los valores reales por cada clase. Con lo cual los conteos quedan divididos en 4 clases, VP, FN, FP y VN, que significan lo siguiente:

- VP es el número de clases positivas que fueron clasificados correctamente como positivos por el modelo.
- VN es el número de clases negativas que fueron clasificados correctamente como negativos por el modelo.
- FN es el número de clases positivas que fueron clasificados incorrectamente como negativos.
- FP es el número de clases negativas que fueron clasificados incorrectamente como positivos.

Por medio de estas cuatro categorías se puede calcular métricas más elaboradas, como las siguientes:

Exactitud (Acurracy): Porcentaje total de los aciertos de nuestro modelo, su fórmula es la siguiente.

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

Tasa de Error: Porcentaje de errores del modelo, su fórmula es la siguiente.

$$Tasa\ de\ error = \frac{FP + FN}{VP + FN + FP + VN}$$

Sensibilidad: También se la llama recall o tasa de verdaderos positivos. Da la probabilidad de que, dada una observación realmente positiva, el modelo la clasifique así, su fórmula es la siguiente.

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

Especificidad: También llamado tasa de verdaderos negativos. Da la probabilidad de que, dada una observación realmente negativa, el modelo la clasifique así, su fórmula es la siguiente.

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

Precisión: También llamado valor de predicción positiva. Da la probabilidad de que, dada una predicción positiva, la realidad sea positiva también, su fórmula es la siguiente.

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

Valor de predicción Negativa: Da la probabilidad de que, dada una predicción negativa, la realidad sea también negativa, su fórmula es la siguiente.

$$Prediccion\ negativa = \frac{VN}{VN + FN}$$

2.3.3.3.3 Curva ROC (Receiver Operating Characteristic)

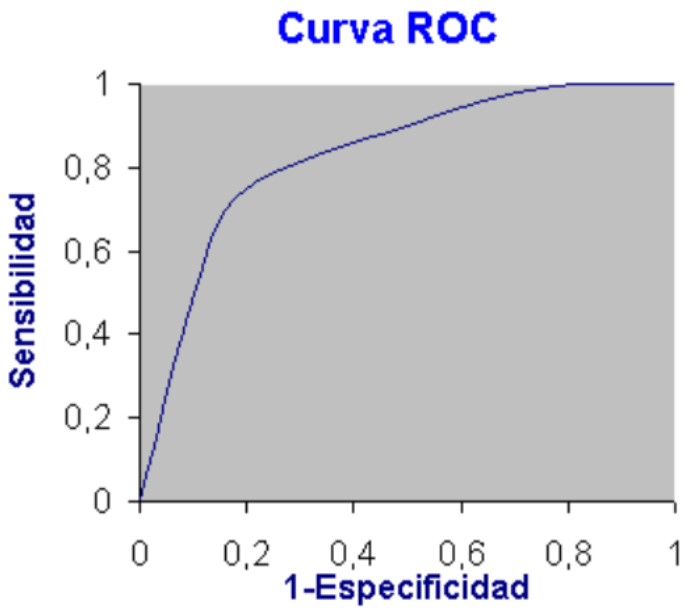
La curva ROC (Receiver operating characteristic o característica operativa del receptor) es una manera eficaz de evaluar la exactitud de las predicciones de modelo al dibujar la sensibilidad ante la especificidad de una prueba de categorización. El sector completa bajo una curva ROC definida, o AUC, formula un estadístico fundamental que representa la posibilidad de que la predicción se encuentre en el orden adecuado una vez que se observa una variable de prueba este puede ser un individuo seleccionado

aleatoriamente del conjunto de casos, y el otro seleccionado aleatoriamente del conjunto de control.

Como se observa en la figura 12, hay una representación de sensibilidad y la especificidad bajo la curva ROC.

Figura 12

Análisis mediante Curva ROC



2.3.4 Metodología CRISP-DM (Cross Industry Standard Process for Data Mining)

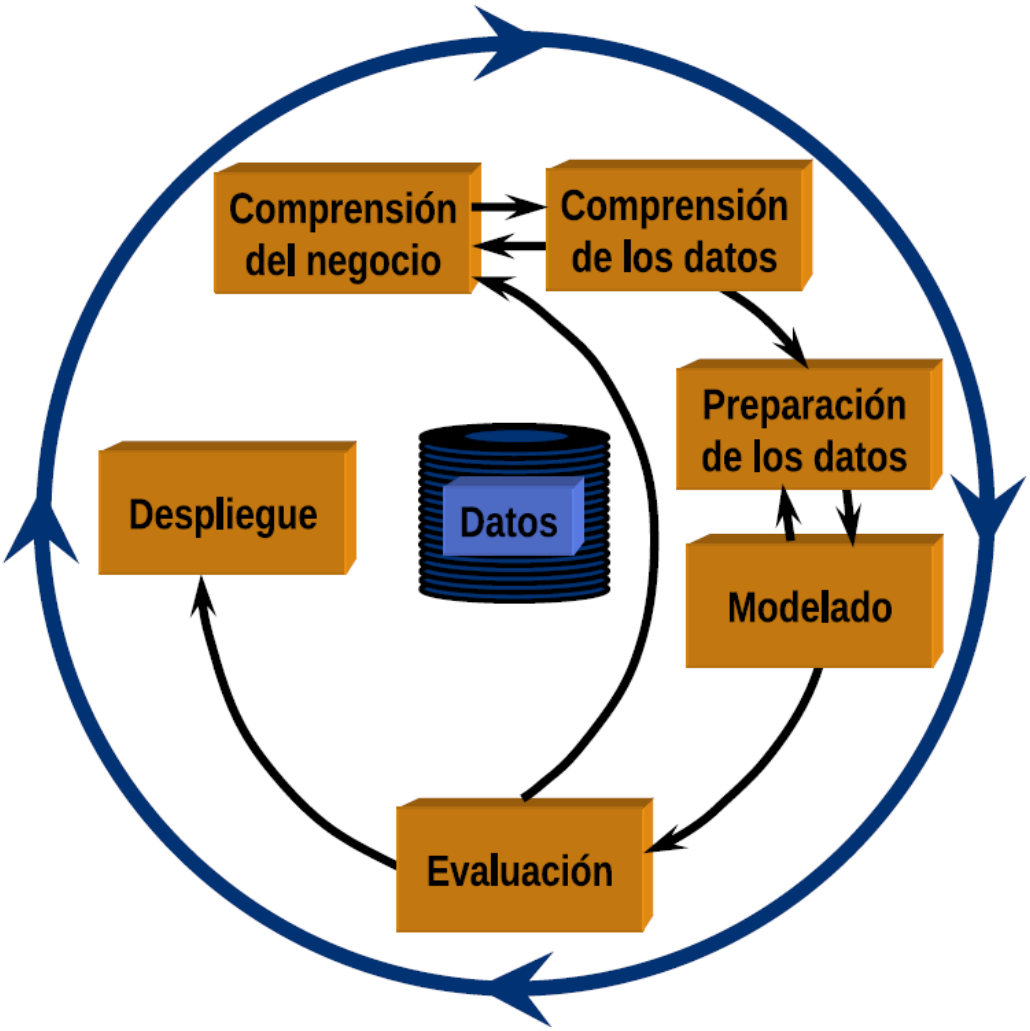
CRISP-DM, es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema, proporciona una descripción regulada del ciclo de vida de un proyecto estándar de análisis de datos.

El modelo CRISP-DM cubre las fases de un proyecto de minería de datos, desde el inicio de las actividades hasta las tareas específicas en cada iteración (Pinto Galindo, 2020).

En la Figura 13, se aprecia el ciclo de vida y las diferentes etapas de la metodología CRISP-DM.

Figura 13

Metodología CRIPS-DM



Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 5), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

2.3.4.1 Fases de la Metodología CRISP-DM

2.3.4.1.1 Fase I. Compresión del negocio

Determinar los objetivos empresariales: Esta es la primera actividad a desarrollar y tiene como metas establecer cuál es el problema que se quiere solucionar para la

empresa, por qué la necesidad de usar la minería de datos y conceptualizar los criterios de triunfo.

Evaluar la situación inicial: En esta actividad se debe evaluar la situación antes del inicio del proceso de minería de datos, tomando en cuenta puntos como, por ejemplo: ¿se cuenta con la proporción de datos solicitada para solucionar el problema?, ¿cuál es la interacción coste beneficio de la aplicación de minería de datos?, etc. Además, se realiza una revisión de los inventarios de recursos, los costos, si tiene algún riesgo y su contingencia y una lista de términos que aclaren su significado.

Establecer los objetivos de minería de datos: Esta actividad tiene como meta identificar los objetivos del negocio en términos y metas como un proyecto de minería de datos como por ejemplo patrones para el aumento de liquidez.

Redactar el plan del proyecto: La función de esta última actividad es realizar un plan para el proyecto donde se detalle los pasos a seguir y las técnicas que se van a implementar.

2.3.4.1.2 Fase II. Comprensión de los datos

Recoger datos iniciales: Esta actividad tiene como meta presentar un informe con todos los datos adquiridos, técnicas que se utilizaron para la recolección y los posibles errores a la hora de realizar esta actividad.

Describir datos: Esta actividad detalla un informe completo de todos los datos con una descripción de los mismos con el objetivo de comprender los datos iniciales.

Explorar datos: Esta actividad implica un informe del análisis de los datos con el objetivo de encontrar una estructura u organización general para los datos.

Verificar la calidad de los datos: Esta actividad muestra un informe donde efectúan verificaciones sobre los datos para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos.

2.3.4.1.3 Fase III. Preparación de los datos

Seleccionar datos: En esta actividad se selecciona el conjunto de datos con los cuales se va a trabajar.

Limpiar datos: Esta actividad tiene como meta presentar un informe detallado del proceso que se usó para la limpieza de los datos, donde se muestre que técnicas se utilizó para optimizar la calidad de los mismos.

Construir datos: Esta actividad tiene como meta generar y preparar los datos a partir de la actividad anterior de limpieza de datos, en la cual se crea la nueva data a partir de la anterior, con fin de utilizarlos en el desarrollo de los modelos.

Integrar datos: Esta actividad implica la creación de nuevas estructuras a partir de los datos seleccionados.

Formatear datos: Esta actividad consiste principalmente en la realización de transformaciones sintácticas de los datos de tal forma que se permita y se facilite utilizar alguna técnica de minería de datos.

2.3.4.1.4 Fase IV. Modelado

Seleccionar técnicas de modelado: Esta actividad consisten en la selección de la técnica de minería de datos más apropiada al tipo de problema que se quiere resolver, tomando en cuenta el objetivo principal del proyecto y la utilización de las técnicas de minería de datos existentes.

Diseñar las pruebas del modelo: Esta actividad consiste en seleccionar las técnicas para evaluar el modelo una vez que esté construido, por ejemplo, la validación cruzada.

Construir los modelos: En esta actividad tiene como meta desarrollar o ejecutar las técnicas previamente seleccionadas sobre los datos preparados para generar uno o más modelos, en cada uno de los modelos se definen los parámetros más aceptados para llegar con el objetivo del proyecto.

Evaluar los modelos: En esta actividad se pone a prueba el modelo de acuerdo a las técnicas anteriormente seleccionadas.

2.3.4.1.5 Fase V. Evaluación

Evaluar los resultados: En esta actividad los resultados de los modelos se evalúan de acuerdo a los objetivos del negocio planteados anteriormente y como resultado a ello se aprueba el o los modelos con el puntaje de evaluación más alto.

Revisar el proceso: En esta actividad se muestra un informe donde se detalla toda la revisión del proceso para generar el modelo o los modelos.

Determinar los siguientes pasos: En esta actividad, si el modelo ha generado resultados satisfactorios se podría continuar con la siguiente fase, de no ser así se podría decidirse por hacer otra iteración desde la fase de preparación de los datos o de modelado con distintos parámetros.

2.3.4.1.6 Fase VI. Despliegue

Planificar despliegue: En esta actividad se detalla las estrategias en forma de plan para poder implementar el proyecto de minería de datos en la organización.

Planificar la monitorización y mantenimiento: En esta actividad se detalla un plan para actualizar la información de los datos.

Redactar el informe final: En esta actividad se presenta un informe final con todos los procesos y los obstáculos que hubo al momento de generar el o los modelos.

Hacer una revisión final de todo el proyecto: En esta actividad se evalúa que cosas se hicieron correctamente y cuales fueron incorrectas, así como aquellos puntos que se podrían mejorar en el proyecto.

2.4 Antecedentes Contextuales

En el año 1996, en el Barrio Chantán de la Parroquia Eloy Alfaro perteneciente al Cantón Latacunga, se forma el banco Comunal “Salud y Progreso”, como parte de un proyecto iniciado por la Diócesis de Latacunga, con 20 mujeres emprendedoras, con

deseos de superación en buscar mejores días para sus familias, al frente una directiva entusiasta y todo el equipo colaborador, mismas que son beneficiarias del crédito masivo, que entrega por primera vez esta institución a través del Banco Comunal en forma igualitaria entre todas las socias la cantidad de 114.000 sucres a cada una, la mayoría de ellas invirtieron en negocios pequeños, y en la producción de bloques. Además de prestar el servicio de crédito, el Banco Comunal inicia una campaña de concientización sobre el valor que tiene el ahorro, dirigido a todos los socios en forma programada, es decir el socio debía cancelar la cuota de su crédito y ahorrar cada mes, estos ahorros se fueron capitalizando y en el año 2002, se decide iniciar a trabajar con fondos propios; En el año 2002, el grupo del Banco Comunal contaba con un valor considerable y deciden trabajar con estos fondos propios sin adquirir créditos de la Diócesis de Latacunga, pero siempre en base a estatutos y políticas que manejaba la misma. A medida que fue desarrollándose, también se fue expandiéndose a sus alrededores, donde se presenta la demanda de créditos, por este motivo se empieza a crear estrategias y métodos para obtener más ahorros, y para la obtención de los mismos no se contaban con las respectivas garantías para sus depositantes. Ante esta necesidad a mediados del año 2005 se decide crear la Cooperativa de Ahorro y Crédito “Virgen del Cisne” para responder a aquellas necesidades del sector, trabajando con lealtad, compromiso, responsabilidad y noble ejemplo de una vida dedicada a sembrar la semilla de la Cooperación, otorgando servicios de ahorro y crédito en la comunidad.

En febrero del 2006 obtuvo la personería jurídica con su Acuerdo Ministerial 00708 de febrero 10 del 2006 inscrita en el Registro General de Cooperativas con Nº de Orden 6857 del 04 de mayo del 2006; La Cooperativa de Ahorro y Crédito “Virgen del Cisne”, en el Ecuador, es una sociedad de hecho dedicada a los microcréditos productivos. En la actualidad cuenta con más de 40000 socios, en su mayoría: comerciantes, agricultores, artesanos, empleados públicos y privados, transportistas; Su

capital está formado con las aportaciones de todos los socios y ahorros de sus depositantes, quienes han ido aportando en base a la confianza generada durante todos estos años de trabajo; En la actualidad la Cooperativa de Ahorro y Crédito “Virgen del Cisne” cuenta con 12 Agencias y su oficina Matriz en la ciudad de Latacunga.

La institución ante los avances tecnológicos ha decidido actualizar sus métodos o proceso para la entrega de créditos, así mediante una herramienta que genere un modelo predictivo que optimice el riesgo crediticio podrán controlar los indicadores de morosidad que es la consecuencia más peligrosa que tiene una entidad financiera para la liquidación de sus activos.

Capítulo III

3 Desarrollo del modelo predictivo

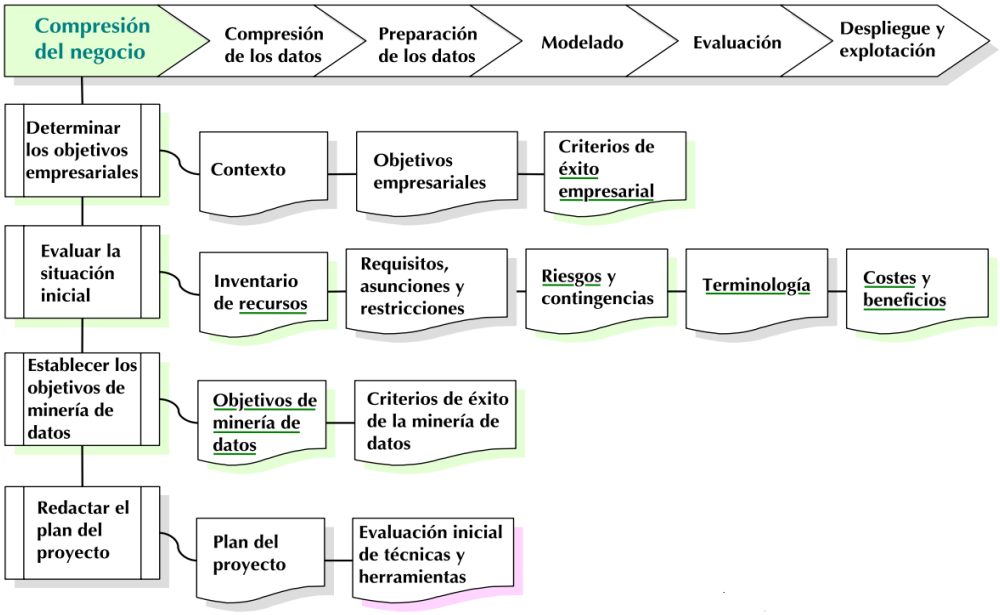
Una metodología es una herramienta que facilita la entrada, la práctica y la gestión en el campo que desea lograr y es posiblemente uno de los recursos más importantes para comunicar y aclarar el progreso de un proyecto. Se utilizará CRIPS-DM.v3 (Cross-Industry Standard Process for Data Mining)(García Osorio, 2019), para desarrollar este proyecto de investigación, ya que es una metodología completa que va desde el análisis empresarial hasta la implementación y presentación de resultados.

3.1 Fase 1: Compresión del negocio

En la Figura 14, se aprecia las tareas y documentos que se desarrollará en la fase de compresión del negocio.

Figura 14

Metodología CRIPS-DM, Fase de Compresión del Negocio



Nota: Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 7), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

3.1.1 Determinar los objetivos empresariales

Contexto

De acuerdo a la situación empresarial de la entidad financiera, se puede decir que desde el inicio de este trabajo de investigación se cuenta con una base de datos de créditos actuales e incluso históricos. Sin embargo, no existen estudios del comportamiento de los socios que puedan proporcionar conclusiones y patrones que hagan predicciones sobre futuros socios que sean sujetos o no de crédito.

Objetivos empresariales

La Cooperativa de Ahorro y Crédito Virgen del Cisne es una entidad financiera del segmento 2, que busca integrarse a esta nueva tecnología de Data Mining, la cual proyecta un análisis de datos históricos y actuales que tiene almacenado la institución, con este proceso la institución busca mejorar los siguientes objetivos:

- ✓ Mejorar la evaluación de un socio para saber si es apto o no para un crédito.
- ✓ Agilizar el proceso de calificación y entrega de un crédito.
- ✓ Minimizar la probabilidad de incumplimiento en los pagos del crédito.

Criterios de éxito empresarial

Desde una perspectiva empresarial, predecir que un nuevo socio es confiable puede reducir la tasa de impagos o morosos, esto se considera un criterio de éxito, otra medida de éxito es el aumento del porcentaje de agilidad a la hora de conceder un préstamo.

3.1.2 Evaluar la situación inicial

La Cooperativa de Ahorro y Crédito “Virgen del Cisne”, en el Ecuador, es una sociedad de hecho dedicada a los microcréditos productivos. En la actualidad cuenta con más de 40000 socios, en su mayoría: comerciantes, agricultores, artesanos, empleados públicos y privados, transportistas; su capital está formado con las

aportaciones de todos los socios y ahorros de sus depositantes, quienes han ido aportando en base a la confianza generada durante todos estos años de trabajo, la Cooperativa de Ahorro y Crédito “Virgen del Cisne” cuenta con 12 Agencias y su oficina Matriz en la ciudad de Latacunga.

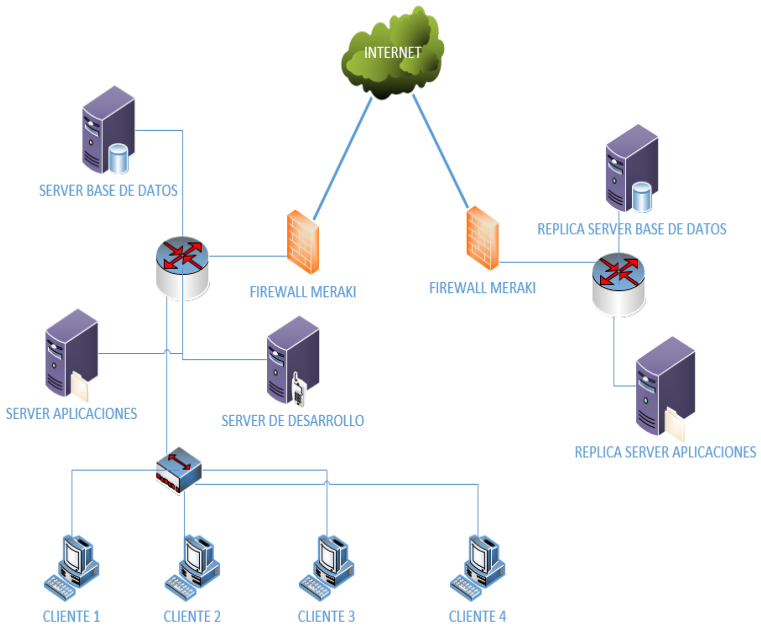
Inventario de recursos

En la parte tecnológica la institución cuenta con un core Financiero llamado eConx, adquirido al proveedor Avmei, está construido bajo lenguajes de desarrollo de última generación y en especial la utilización de productos denominados de CODIGO ABIERTO, utiliza una gama de programas para el ambiente WEB como son: PHP, HTML, JAVA, JAVA-SCRIPT, AJAX, CSS, DOOM; para el almacenamiento de información se utiliza el motor de base de datos INFORMIX 11.50, además maneja un proceso de espejo para respaldo de información en tiempo real para contingencias, en infraestructura tecnológica cuentan con un data center principal en la agencia matriz y un alterno en la ciudad de Quito, el cual consta de servidores Hp Proliant de generación 6, 8 y 10.

En la figura 15, se muestra un esquema que resume la infraestructura física que dispone la entidad financiera.

Figura 15

Esquema de la Infraestructura Tecnológica de la Institución.



Requisitos, asunciones y restricciones

Para el desarrollo de este proyecto se requiere las bases de datos que almacenan información actual e histórica de todos los créditos otorgados en los estados de cancelados, activos, reclasificados y vencidos. Por sigilo bancario se restringe la presentación de datos como identificación personal, nombres y apellidos.

Riesgos y contingencias

Los riesgos existentes que se debe tomar en cuenta para este trabajo son los siguientes:

El tiempo de desarrollo del proyecto, una vez aceptada la propuesta por la Gerencia General, se procedió a descargar toda la información de créditos de la entidad financiera y se guardó en otro motor de base de datos (SQL Server 2016).

Costos adicionales o sobresalientes, con los datos iniciales se trabajará, por lo que no se necesita costos adicionales por parte de la entidad o del investigador.

La calidad de datos, el proceso de otorgamiento de crédito inicia desde la actualización de la cuenta, es decir el socio actualiza los datos tales como información personal, vivienda, cónyuge, trabajo y estado financiero, con esto podemos decir que la calidad de datos es buena.

Terminología

Ver Anexo 1: Glosario de términos.

Costes y beneficios

Los datos de este proyecto son propios de la entidad financiera desde el momento de que el socio ingresa hasta cuando realiza la solicitud de crédito, por lo que esto significa ningún costo a la entidad financiera.

Sobre los beneficios de este proyecto se asegura un mejor análisis de crédito de un socio y la agilidad para otorgar el crédito, de esta manera mejorar los servicios en la entidad financiera.

3.1.3 Establecer los objetivos de minería de datos

Objetivos de minería de datos

Los objetivos planteados para el desarrollo del proyecto son los siguientes:

- ✓ Generar patrones que ayuden a la evaluación de un cliente para verificar si es apto o no para un crédito.
- ✓ Generar patrones que ayuden con la problemática del incremento de mora y baja de rentabilidad.

Criterios de éxito de la minería de datos

Como criterios de minería de datos será el de predecir con las variables o datos ya entrenados, si un socio es sujeto o no a crédito con un porcentaje del 90% de acierto y efectividad

3.1.4 Redactar el plan del proyecto

Plan del proyecto

Para el desarrollo del presente proyecto se contempla el desarrollo de todas las fases, en la tabla 1, se detalla la planificación con el tiempo que se tardará en realizar cada fase.

Tabla 1

Planificación de las fases de la metodología

Fase	Tiempo	Equipo
Comprensión del negocio	2 semanas	Jefe Operativo, Investigador
Comprensión de los datos	2 semanas	Departamento de TI, Investigador
Preparación de los datos	3 semanas	Investigador
Modelado	3 semanas	Investigador
Evaluación	2 semanas	Jefe Operativo, Investigador
Implementación o despliegue	1 semana	Departamento de TI, Investigador

Evaluación inicial de técnicas y herramientas

Para la generación de esta propuesta se tendrá en cuenta las siguientes herramientas:

Hardware:

- ✓ Procesador Core I7 4 núcleos o mas
- ✓ Velocidad de procesador 2.0 Ghz o superior
- ✓ Disco Duro 500 GB estado solido
- ✓ Memoria RAM 16

Software:

- ✓ Suite de Anaconda
- ✓ Jupyter Notebooks

En la parte del recurso humano para la realización de este proyecto se contará con el Ing. Oscar Chiluiza.

En cuanto a técnicas para la extracción de datos se utilizará el lenguaje de programación SQL para generar los scripts que permitan consultar la información y extráela, también se utilizará la tarea de clasificación para la generación de modelos predictivos, y las siguientes técnicas de modelado:

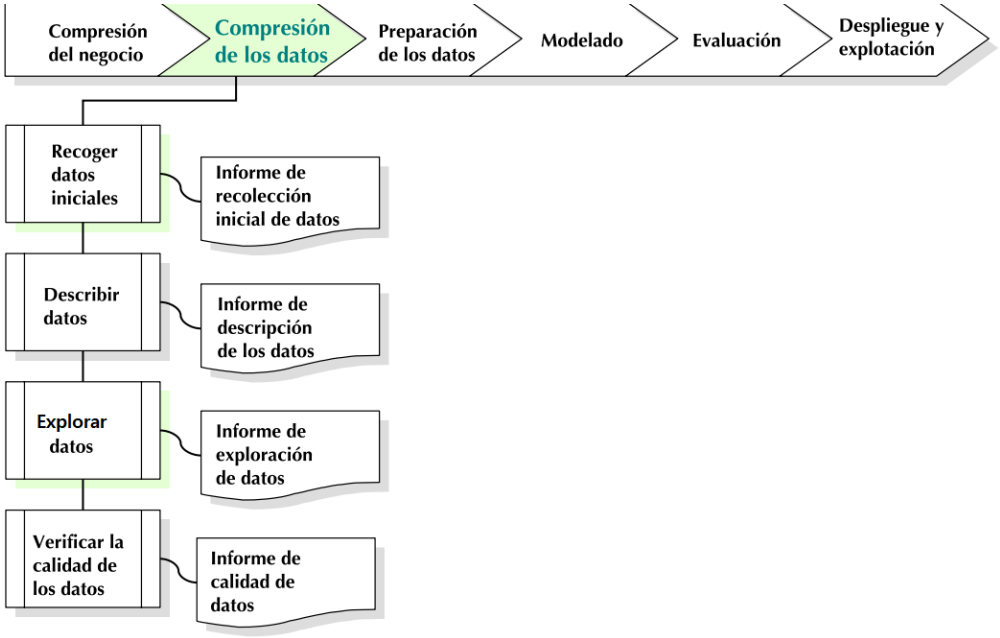
- ✓ Ramdon forest.
- ✓ Regresión logística.
- ✓ Redes neuronales.

3.2 Fase 2: Compresión de los datos

En la Figura 16, se aprecia las tareas y documentos que se desarrollará en la fase de compresión de los datos.

Figura 16

Metodología CRIPS-DM, Fase de Compresión de los Datos



Nota: Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 8), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

3.2.1 Recoger datos iniciales

Informe de recolección inicial de datos

Para el desarrollo de este proyecto se recolectaron datos sobre información del socio, tipo de crédito, monto, frecuencia de pago, actividad económica, etc., estos datos están almacenados en tablas relacionadas con el proceso de otorgamiento de créditos, además por motivos legales y sigilo bancario se omitieron los datos de identificación, nombres y apellidos. Como el objetivo del proyecto es realizar predicciones, se clasificó al socio en bueno y malo de acuerdo a la última calificación de crédito que obtuvieron al momento de cancelar el mismo.

Por la gran cantidad de registros que son necesarios para el desarrollo del proyecto, se optó por desarrollar sentencias SQL que ayuden a generar toda la información en base a los datos investigados en conjunto con la entidad financiera, estas sentencias se pueden consultar en el Anexo 2.

A continuación, se detalla la información de tablas utilizadas para la extracción de datos:

- Cliente: Tabla donde se registra los datos personales, vivienda, trabajo, referencias y estados financieros.
- Crédito: Tabla donde se registra los datos del monto, tipo de crédito, plazo, tasas de interés, frecuencia de pago, valor de cuota y la calificación de crédito.
- Periodicidad: Tabla donde se detalla la frecuencia de pago de un crédito.
- Línea Crédito: Tabla donde se detalla los tipos de créditos que existe en la entidad financiera.
- Destino financiero crédito: Tabla donde se detalla los diferentes destinos que el socio ha solicitado en el crédito.

- Garantía: Tabla donde se detalla las diferentes garantías que maneja la entidad financiera.
- Estado civil: Tabla donde se detalla el estado civil del socio.
- Instrucción: Tabla donde se detalla el nivel académico del socio.

Los datos recolectados fueron revisados conjuntamente con el Jefe de Negocios de la institución encargado del Área Operativa, de acuerdo a un análisis y a la experticia del mismo se logró recabar las siguientes variables para el desarrollo del modelo; en la tabla 2, se aprecia la lista de variables que fueron seleccionadas para el trabajo de investigación.

Tabla 2

Lista de variables

No	Variable
1	Cargas familiares
2	Destino del Crédito
3	Edad
4	Estado Civil
5	Genero
6	Garantía
7	Monto
8	Nivel Instrucción
9	Préstamos Instituciones
10	Número de créditos
11	Tipo de crédito
12	Tipo de Vivienda
13	Ingresos
14	Egresos
15	Activos

No	Variable
16	Pasivos
17	Tasa de Interés
18	Antigüedad del socio en la institución
19	Antigüedad laboral
20	Zona geográfica
21	Número de Cuotas
22	Frecuencia de Pago
23	Actividad Económica
24	Valor cuota
25	Capacidad de pago
26	Tipo de cliente
27	Calificación del cliente

3.2.2 Describir datos

Informe de descripción de los datos

De acuerdo al análisis tanto del investigado como del representante de la entidad financiera se da a conocer las variables, con las que se va a trabajar para el desarrollo del modelo. A continuación, se detalla cada una de ellas determinadas en variables de entrada y variables de salida:

Variables de entrada:

- Cargas familiares: es el número de personas que dependen económicamente del deudor.
- Destino del Crédito: indica en que se utilizará el capital desembolsado.
- Edad: es el número de años que tiene el deudor.
- Estado Civil: estado civil de la persona.
- Género: sexo definido del socio.

- Garantía: consiste en el respaldo ofrecido por el deudor a la entidad financiera.
- Monto: consiste en el valor inicial de dinero otorgado al socio en calidad de préstamo.
- Nivel Instrucción: nivel de instrucción del cliente que pueden ser sin estudios, primaria, secundaria o superior.
- Préstamos Instituciones: si el cliente tiene préstamos en otras instituciones.
- Número de créditos: corresponde al número de créditos que tiene o ha tenido el socio con la entidad.
- Tipo de crédito: tipo de crédito que está solicitando consumo, inmobiliario o microcrédito.
- Tipo de Vivienda: es el tipo de vivienda en la que vive el deudor.
- Ingresos: ingresos mensuales por sus actividades realizadas.
- Egresos: reducción mensual de los ingresos del cliente.
- Activos: la suma de todos los bienes que tiene el cliente.
- Pasivos: monto total de deudas al momento.
- Tasa de Interés: el porcentaje de uso del dinero prestado.
- Antigüedad del socio en la institución: corresponde al número de créditos que tiene o ha tenido el cliente con la entidad.
- Antigüedad laboral: consiste en el tiempo que lleva el deudor vinculado como empleado de una empresa.
- Zona geográfica: lugar domicilio del socio donde se divide en dos categorías urbano y rural.
- Número de Cuotas: el plazo crédito.
- Frecuencia de Pago: frecuencia con que se pagan las cuotas.

- Actividad Económica: actividad a que se dedica la persona.
- Valor cuota: valor de la cuota de acuerdo a la frecuencia de pago.
- Capacidad de pago: evalúa su capacidad de pago mensual.
- Tipo de cliente: tipo de cliente en la cual puede ser nuevo o recurrente.
- Calificación del cliente: se refiere a la última calificación de riesgo según su comportamiento de pago Crediticio.

Variable de salida:

- La variable de salida es el resultado de la predicción.

3.2.3 Explorar datos

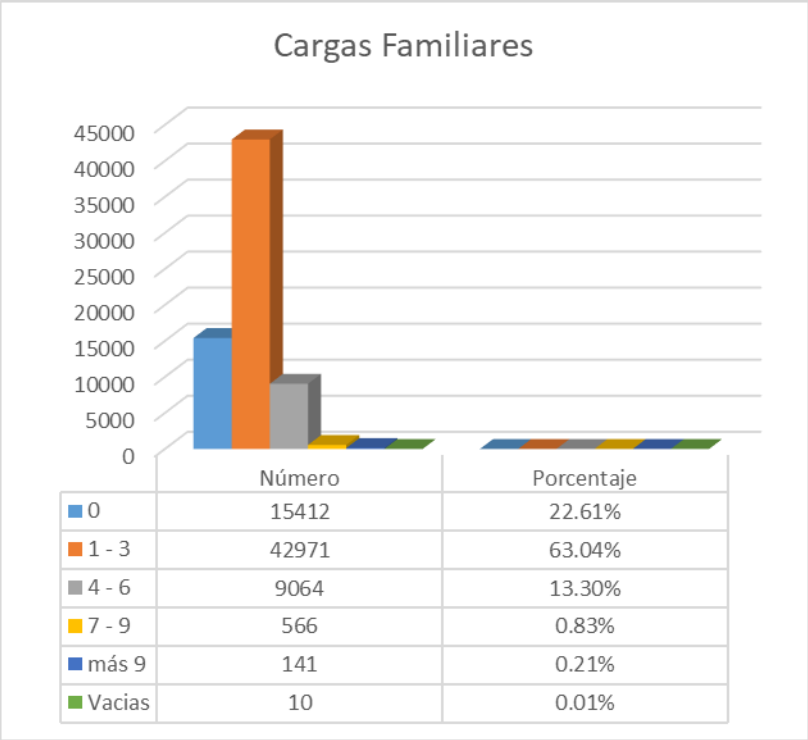
Informe de exploración de datos

Con las variables obtenidas se realizó un sondeo a la base de datos, con el fin de identificar toda la información que se puede obtener y será necesario para la realización de este proyecto de investigación, a continuación, se muestra un estadístico de cada variable a ser procesada:

Cargas Familiares. –Se identifica que el 63% de cargas familiares están dentro del rango de 1 a 3 personas por socio. (Ver Figura 17)

Figura 17

Exploración de Cargas familiares

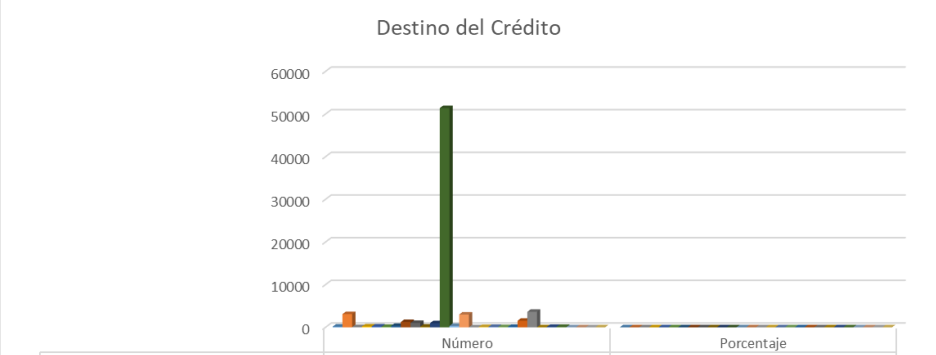


	Número	Porcentaje
0	15412	22.61%
1 - 3	42971	63.04%
4 - 6	9064	13.30%
7 - 9	566	0.83%
más 9	141	0.21%
Vacias	10	0.01%

Destino del Crédito. Se identifica que el 75% de los socios utiliza el dinero de un crédito para Capital de trabajo. (Ver figura 18)

Figura 18

Exploración de Destino de crédito

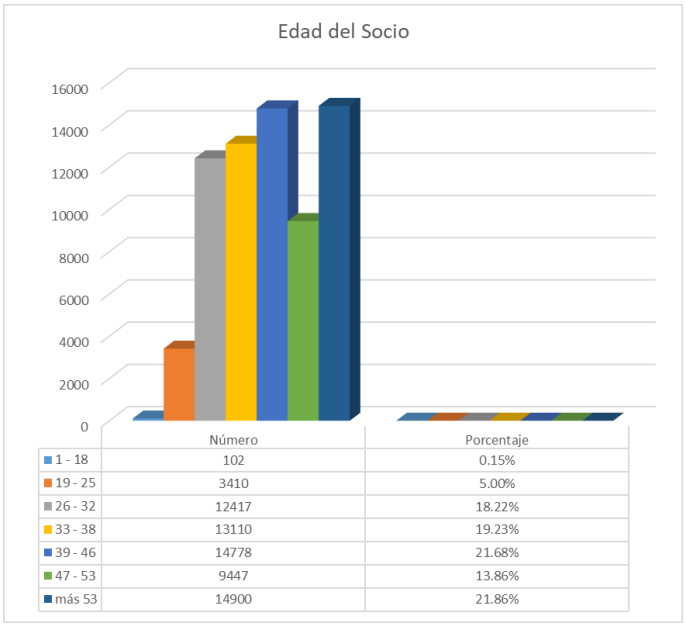


	Número	Porcentaje
Activos Fijos Intangibles	180	0.264%
Activos Fijos Tangibles	3107	4.558%
Adq. De vivienda para fines productivos	57	0.084%
Adq. De equipos para fines productivos y	203	0.298%
Adq. De maquinaria para fines productivo	193	0.283%
Adq. De otros bienes de capital para fin	94	0.138%
Adq. De servicios	350	0.513%
Adq. De terreno para fines productivos y	1256	1.843%
Adq. De vehículos livianos para fines p	1029	1.510%
Adq. De vehículos pesados para fines pro	167	0.245%
Alimentación	977	1.433%
Capital de trabajo	51402	75.409%
Compra de Vehículo	348	0.511%
Construcción de infraestructura para fin	3004	4.407%
Construcción de vivienda	3	0.004%
Educativo	82	0.120%
Hogar	101	0.148%
Microcredito de Consumo	65	0.095%
Microcredito de Vivienda	110	0.161%
Otros	1549	2.272%
Otros (para operaciones de consumo y viv	3642	5.343%
Reestructuracion de Pasivos y Pago de Ob	11	0.016%
Reestructuración de Pasivos y Pago de Ob	97	0.142%
Salud	107	0.157%
Tecnología	17	0.025%
Vestimenta y accesorios de vestir	4	0.006%
Viajes	6	0.009%
Vivienda terminada nueva	3	0.004%

Edad. Se identifica que el 21% de los socios que solicitan un crédito tienen una edad entre 39 y 46 años. (Ver figura 19)

Figura 19

Exploración de Edad

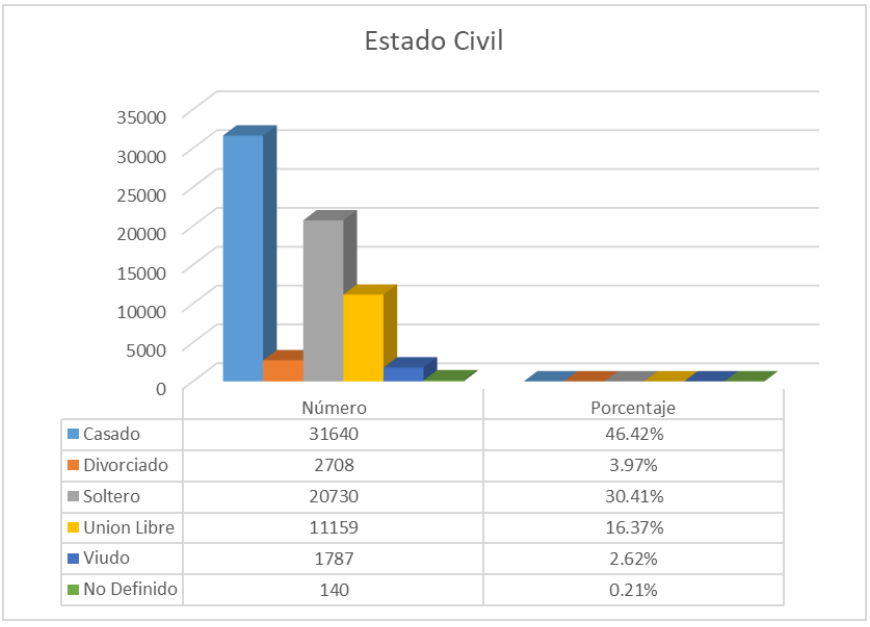


	Número	Porcentaje
1 - 18	102	0.15%
19 - 25	3410	5.00%
26 - 32	12417	18.22%
33 - 38	13110	19.23%
39 - 46	14778	21.68%
47 - 53	9447	13.86%
más 53	14900	21.86%

Estado Civil. Se identifica que el 46% de los socios que solicitan créditos están casados. (Ver figura 20)

Figura 20

Exploración de Estado civil

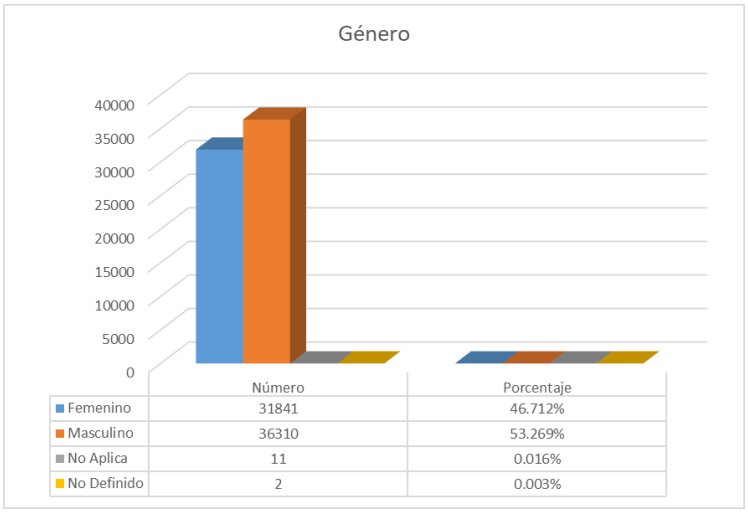


	Número	Porcentaje
Casado	31640	46.42%
Divorciado	2708	3.97%
Soltero	20730	30.41%
Union Libre	11159	16.37%
Viudo	1787	2.62%
No Definido	140	0.21%

Género. Se identifica que el 53% de socios que realizan créditos son Hombres. (Ver figura 21)

Figura 21

Exploración de Género

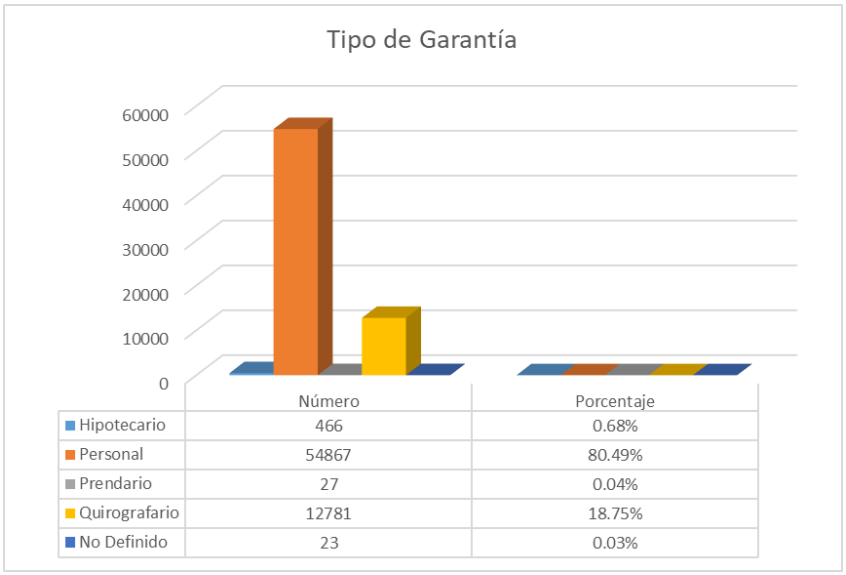


	Número	Porcentaje
Femenino	31841	46.712%
Masculino	36310	53.269%
No Aplica	11	0.016%
No Definido	2	0.003%

Garantía. Se identifica que el 80% de créditos tiene una garantía personal. (Ver figura 22)

Figura 22

Exploración de Garantía

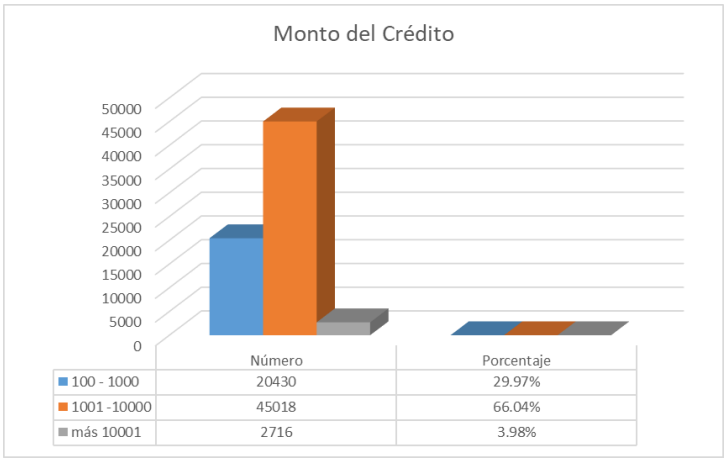


	Número	Porcentaje
Hipotecario	466	0.68%
Personal	54867	80.49%
Prendario	27	0.04%
Quirografario	12781	18.75%
No Definido	23	0.03%

Monto. Se identifica que el 66% de socios que solicitan créditos piden montos de entre 1000 a 10000 dólares. (Ver figura 23)

Figura 23

Exploración de Monto

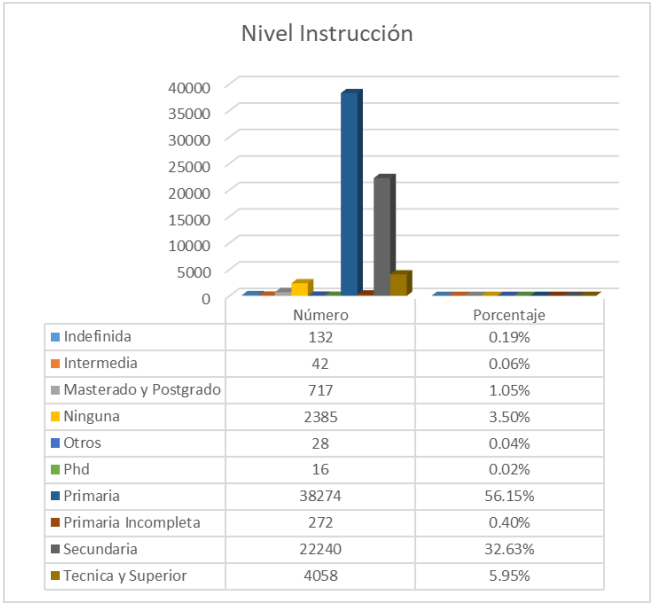


	Número	Porcentaje
100 - 1000	20430	29.97%
1001 -10000	45018	66.04%
más 10001	2716	3.98%

Nivel Instrucción. Se identifica que el 56% de socios que solicitaron créditos tiene un nivel de estudio primario. (Ver figura 24)

Figura 24

Exploración de Nivel instrucción

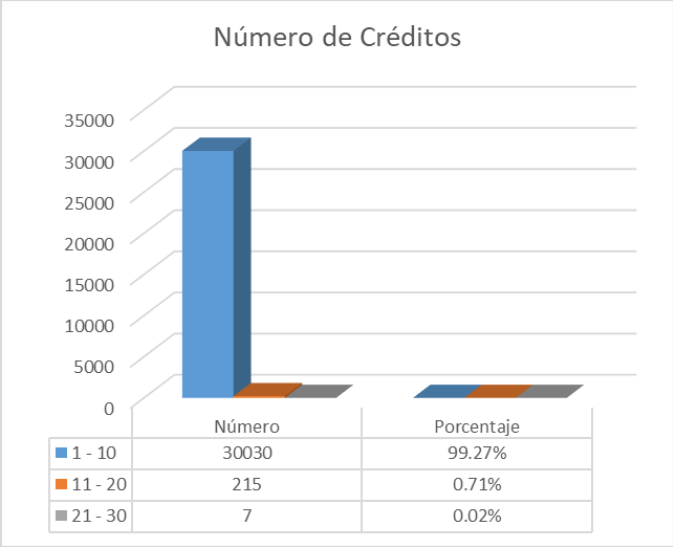


	Número	Porcentaje
Indefinida	132	0.19%
Intermedia	42	0.06%
Masterado y Postgrado	717	1.05%
Ninguna	2385	3.50%
Otros	28	0.04%
Phd	16	0.02%
Primaria	38274	56.15%
Primaria Incompleta	272	0.40%
Secundaria	22240	32.63%
Tecnica y Superior	4058	5.95%

Número de Créditos. Se identifica que el 99% de socios ya han solicitados créditos dentro del rango de 1 a 10, donde se puede etiquetar al socio como recurrente. (Ver figura 25)

Figura 25

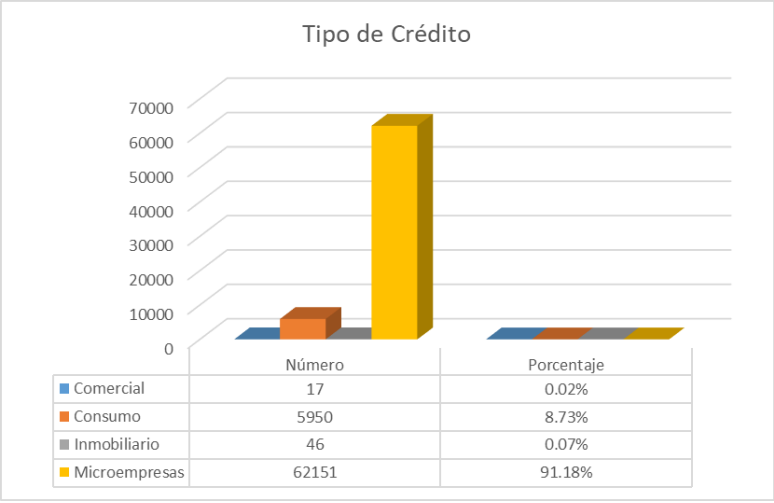
Exploración de Número de créditos



Tipo de Crédito. Se identifica que el 91% de créditos solicitados fueron otorgados dentro de la línea de crédito Microempresas. (Ver figura 26)

Figura 26

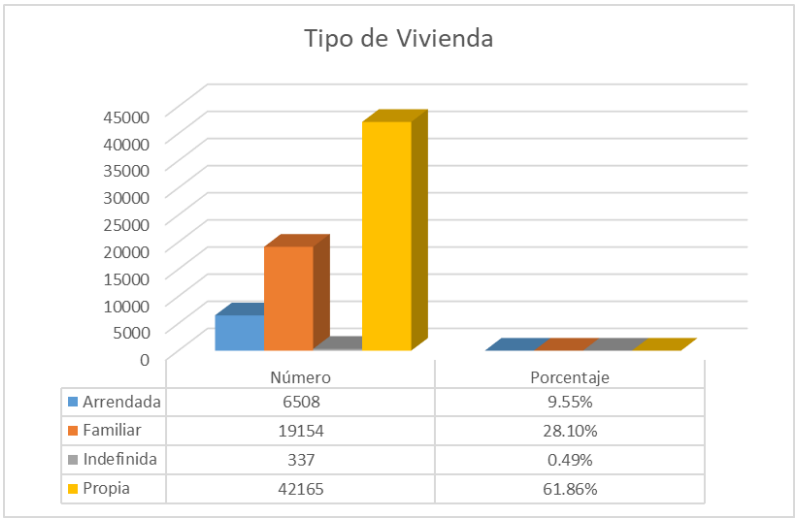
Exploración de Tipo de crédito



Tipo de Vivienda. Se identifica que el 61% de socios que solicitan créditos tienen su vivienda propia. (Ver figura 27)

Figura 27

Exploración de Tipo de vivienda

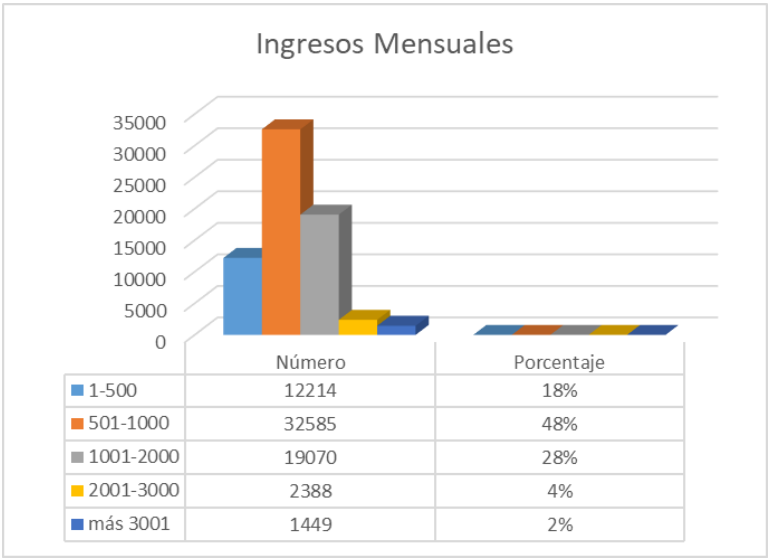


	Número	Porcentaje
Arrendada	6508	9.55%
Familiar	19154	28.10%
Indefinida	337	0.49%
Propia	42165	61.86%

Ingresos. Se identifica que el 48% de los socios cuenta con ingresos de entre 500 a 1000 dólares. (Ver figura 28)

Figura 28

Exploración de Ingresos

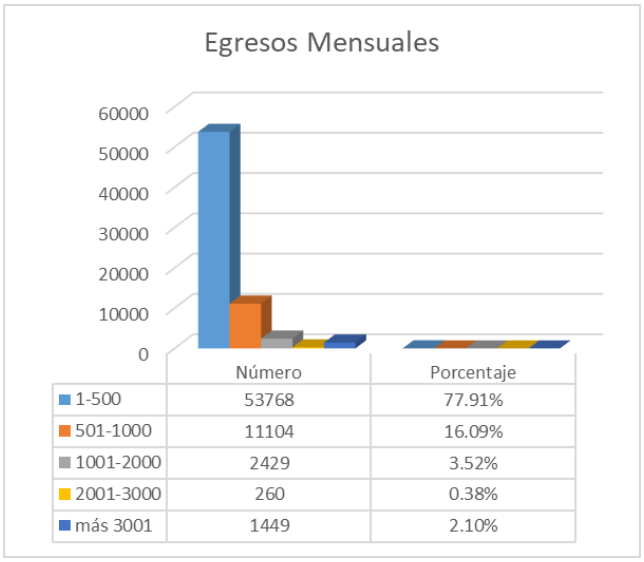


	Número	Porcentaje
1-500	12214	18%
501-1000	32585	48%
1001-2000	19070	28%
2001-3000	2388	4%
más 3001	1449	2%

Egresos. Se identifica que el 77% de los socios cuenta con egresos menores a 500 dólares. (Ver figura 29)

Figura 29

Exploración de Egresos

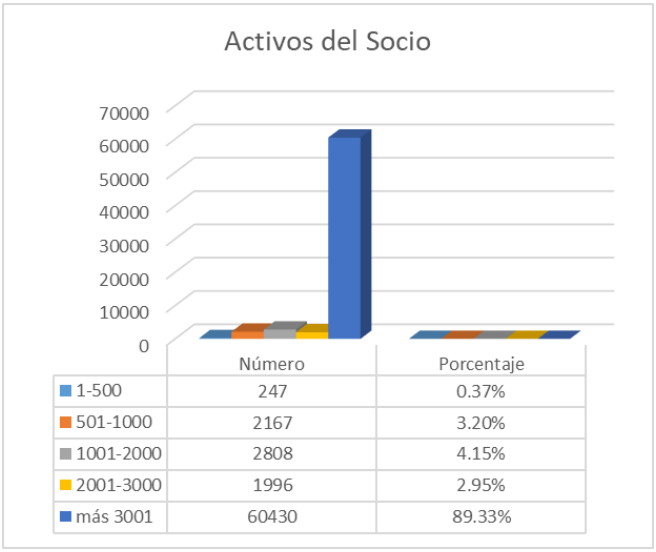


	Número	Porcentaje
1-500	53768	77.91%
501-1000	11104	16.09%
1001-2000	2429	3.52%
2001-3000	260	0.38%
más 3001	1449	2.10%

Activos. Se identifica que el 89% de socios tienen activos de más de 3000 dólares. (Ver figura 30)

Figura 30

Exploración de Activos

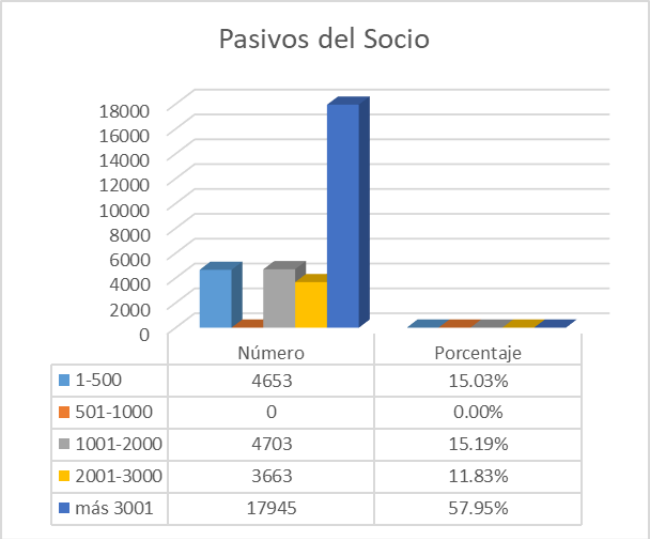


	Número	Porcentaje
1-500	247	0.37%
501-1000	2167	3.20%
1001-2000	2808	4.15%
2001-3000	1996	2.95%
más 3001	60430	89.33%

Pasivos. Se identifica que el 57% de socios tienen pasivos de más de 3000 dólares. (Ver figura 31)

Figura 31

Exploración de Pasivos

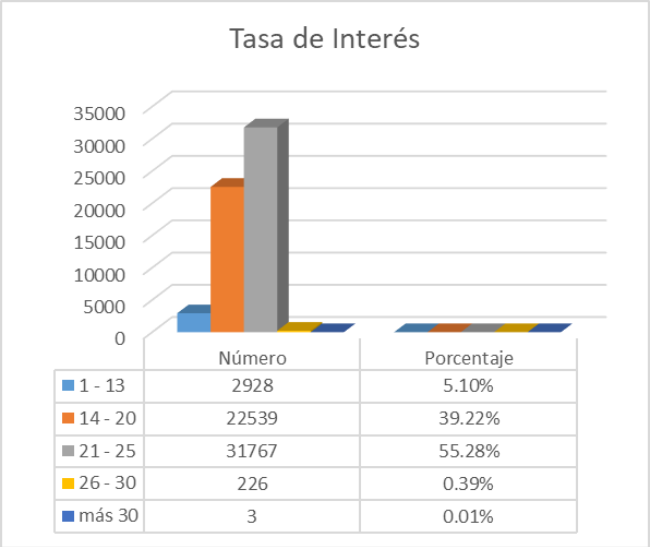


	Número	Porcentaje
1-500	4653	15.03%
501-1000	0	0.00%
1001-2000	4703	15.19%
2001-3000	3663	11.83%
más 3001	17945	57.95%

Tasa de Interés. Se identifica que el 55% de los créditos entregados comprenden una tasa dentro del rango de 21 a 25. (Ver figura 32)

Figura 32

Exploración de Tasas de interés

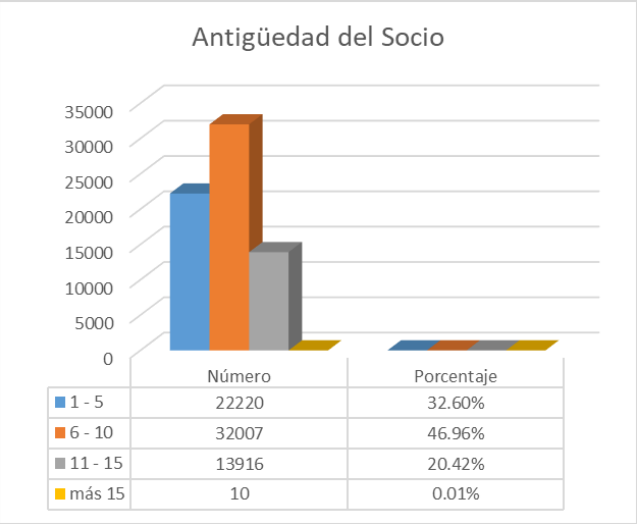


	Número	Porcentaje
1 - 13	2928	5.10%
14 - 20	22539	39.22%
21 - 25	31767	55.28%
26 - 30	226	0.39%
más 30	3	0.01%

Antigüedad del Socio en la Institución. Se identifica que el 46% de los socios mantienen una relación con la cooperativa de 6 a 10 años. (Ver figura 33)

Figura 33

Exploración de Antigüedad del socio

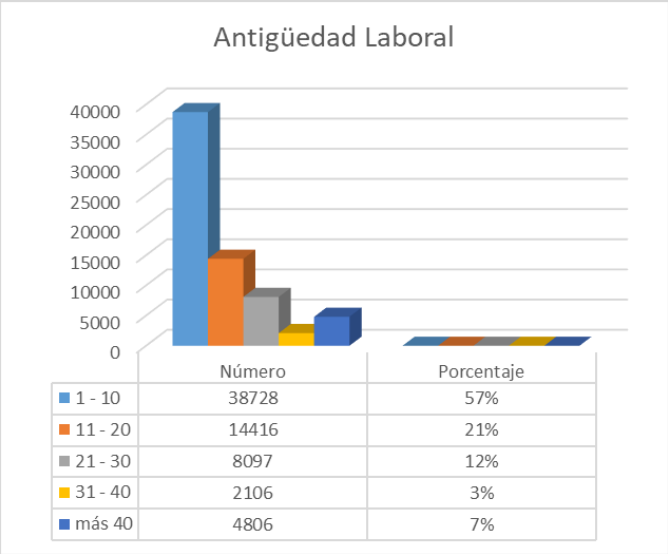


	Número	Porcentaje
1 - 5	22220	32.60%
6 - 10	32007	46.96%
11 - 15	13916	20.42%
más 15	10	0.01%

Antigüedad laboral. Se identifica que el 57% de los socios mantiene un tiempo de trabajo de 1 a 10 años. (Ver figura 34)

Figura 34

Exploración de Antigüedad laboral

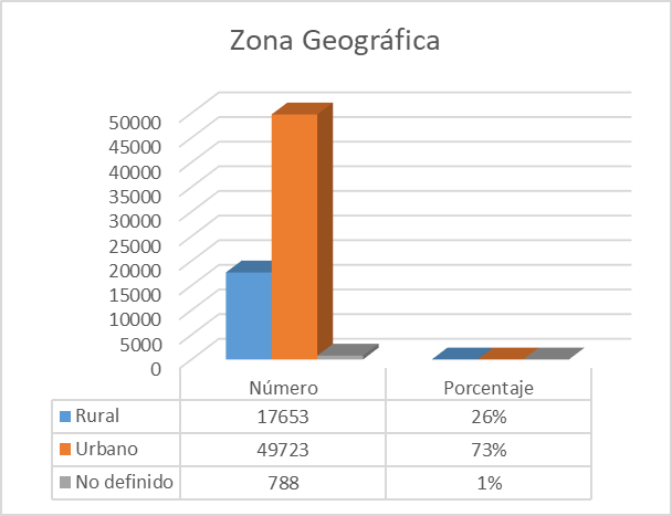


	Número	Porcentaje
1 - 10	38728	57%
11 - 20	14416	21%
21 - 30	8097	12%
31 - 40	2106	3%
más 40	4806	7%

Zona Geográfica. Se identifica que el 73% de socios viven dentro del sector urbano. (Ver figura 35)

Figura 35

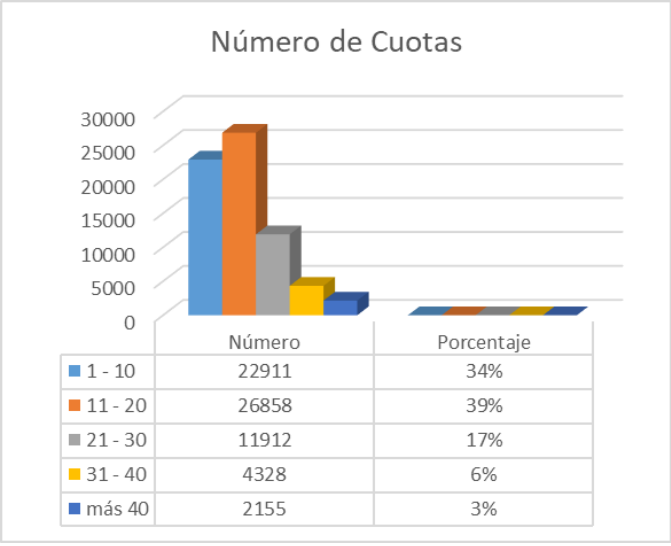
Exploración de Zona geográfica



Número de Cuotas. Se identifica que el 39% de socios que solicitan un crédito, piden pagar cuotas de 11 a 20. (Ver figura 36)

Figura 36

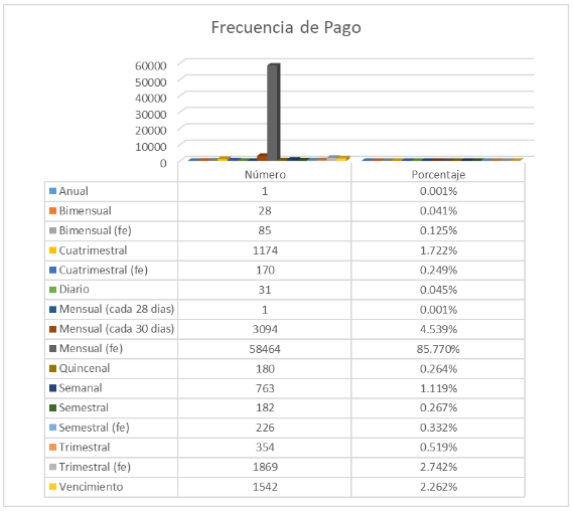
Exploración de Número de cuotas



Frecuencia de Pago. Se identifica que el 85% de socios que solicitan un crédito piden pagar mensualmente y a la fecha de corte. (Ver figura 37)

Figura 37

Exploración de Frecuencia de pago

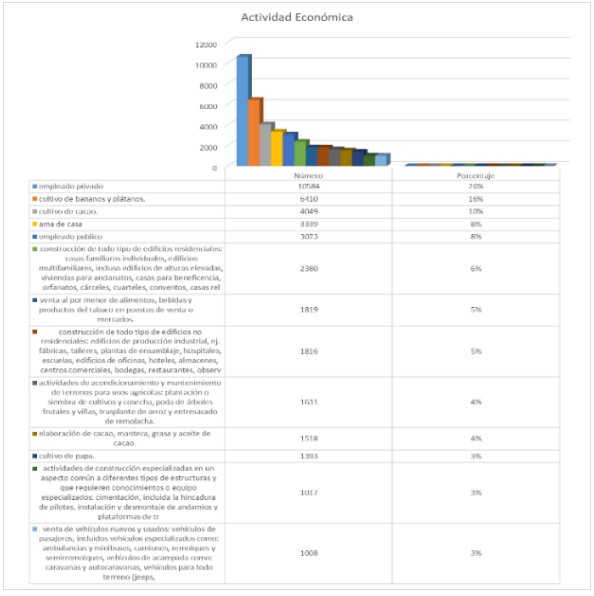


	Número	Porcentaje
Anual	1	0.001%
Bimensual	28	0.041%
Bimensual (fe)	85	0.125%
Cuatrimestral	1174	1.722%
Cuatrimestral (fe)	170	0.249%
Diario	31	0.045%
Mensual (cada 28 dias)	1	0.001%
Mensual (cada 30 dias)	3094	4.539%
Mensual (fe)	58464	85.770%
Quincenal	180	0.264%
Semanal	763	1.119%
Semestral	182	0.267%
Semestral (fe)	226	0.332%
Trimestral	354	0.519%
Trimestral (fe)	1869	2.742%
Vencimiento	1542	2.262%

Actividad Económica. Se identifica que el 26% de socios trabajan como empleados privados. (Ver figura 38)

Figura 38

Exploración de Actividad económica

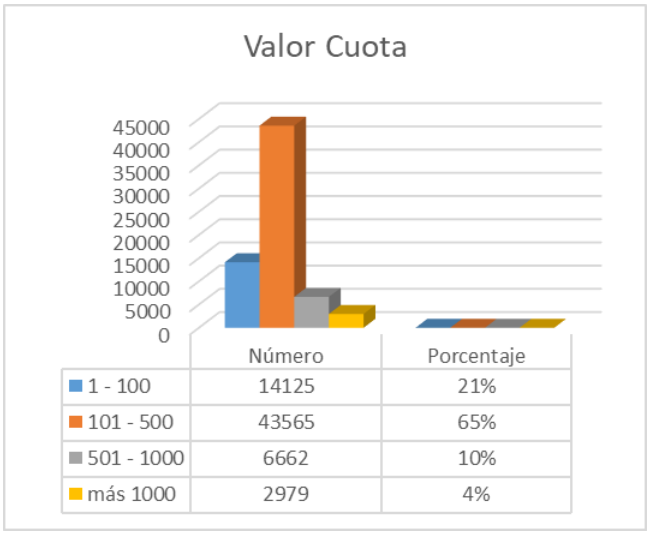


	Número	Porcentaje
empleado privado	10584	26%
cultivo de bananos y plátanos.	6410	16%
cultivo de cacao.	4049	10%
ama de casa	3339	8%
empleado publico	3073	8%
construcción de todo tipo de edificios residenciales: casas familiares individuales, edificios multifamiliares, incluso edificios de alturas elevadas, viviendas para ancianatos, casas para beneficencia, orfanatos, cárceles, cuarteles, conventos, casas rel	2380	6%
venta al por menor de alimentos, bebidas y productos del tabaco en puestos de venta o mercados.	1819	5%
construcción de todo tipo de edificios no residenciales: edificios de producción industrial, ej. fábricas, talleres, plantas de ensamblaje, hospitales, escuelas, edificios de oficinas, hoteles, almacenes, centros comerciales, bodegas, restaurantes, observ	1816	5%
actividades de acondicionamiento y mantenimiento de terrenos para usos agrícolas: plantación o siembra de cultivos y cosecha, poda de árboles frutales y viñas, trasplante de arroz y entresacado de remolacha.	1631	4%
elaboración de cacao, manteca, grasa y aceite de cacao.	1518	4%
cultivo de papa.	1393	3%
actividades de construcción especializadas en un aspecto común a diferentes tipos de estructuras y que requieren conocimientos o equipo especializados: cimentación, incluida la hincadura de pilotes, instalación y desmontaje de andamios y plataformas de tr	1017	3%
venta de vehículos nuevos y usados: vehículos de pasajeros, incluidos vehículos especializados como: ambulancias y minibuses, camiones, remolques y semirremolques, vehículos de acampada como: caravanas y autocaravanas, vehículos para todo terreno (jeep),	1008	3%

Valor Cuota. Se identifica que el 65% de socios han pagado o pagan cuotas de entre 100 a 500 dólares. (Ver figura 39)

Figura 39

Exploración de Valor Cuota

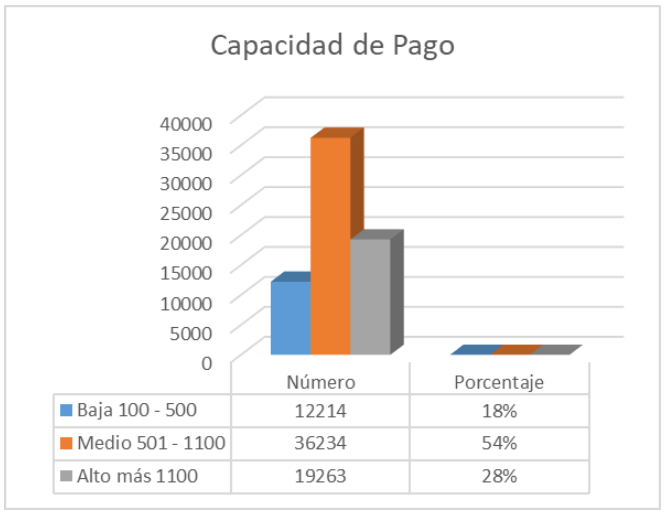


	Número	Porcentaje
1 - 100	14125	21%
101 - 500	43565	65%
501 - 1000	6662	10%
más 1000	2979	4%

Capacidad de Pago. Se identifica que el 54% de los socios tienen una capacidad de pago media. (Ver figura 40)

Figura 40

Exploración de Capacidad de pago

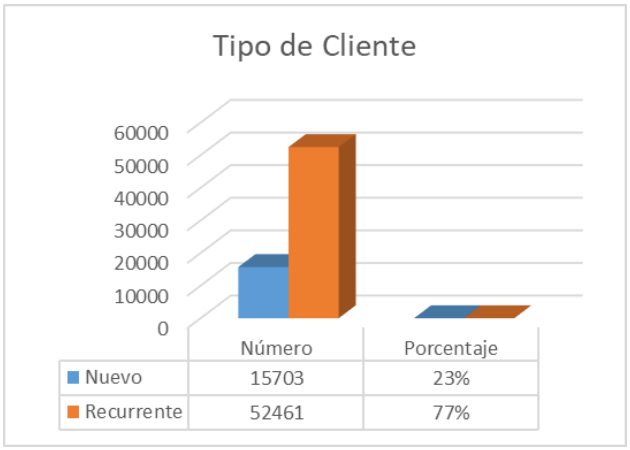


	Número	Porcentaje
Baja 100 - 500	12214	18%
Medio 501 - 1100	36234	54%
Alto más 1100	19263	28%

Tipo Socio. Se identifica que el 77% de los socios que han solicitado crédito son recurrentes. (Ver figura 41)

Figura 41

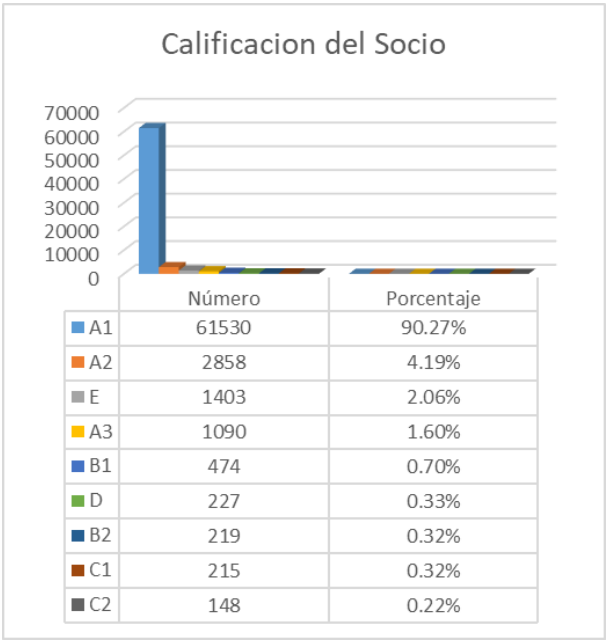
Exploración de Tipo socio



Calificación de Riesgo del Socio. Se identifica que el 90% de socios que han cancelado el crédito su última calificación de riesgo es potencialmente buena A1. (Ver figura 42)

Figura 42

Exploración de Riesgo del Socio



3.2.4 Verificar la calidad de los datos

Informe de calidad de datos

Para la verificación de la calidad de datos se utilizará la norma ISO/IEC 25012 – “Data Quality Model” que especifica un modelo general de calidad para aquellos datos que se encuentran definidos en un formato estructurado dentro de un sistema informático, mediante una matriz se valorará los datos buenos y malos.

Características que se tomará para la matriz de verificación de datos:

Accesibilidad (AC). Donde se especifica el grado en que los datos pueden ser accedidos en un contexto específico.

Conformidad (CO). Donde se verifica que los datos correspondientes cumplen con estándares, convenciones o normativas vigentes.

Confidencialidad (CF) (asociada a la seguridad de la información). Donde se asegura que los datos solo son accedidos e interpretados por usuarios específicos autorizados.

Eficiencia (EF). Donde se analiza el grado en el que los datos pueden ser procesados y proporcionados con los niveles de rendimiento esperados.

Precisión (PR). Donde los datos requieren de valores exactos o con discernimiento en un contexto específico.

Trazabilidad (TZ). Donde se analiza si los datos proporcionan un registro de los acontecimientos que los modifican.

Comprensibilidad (CP). Donde los datos son expresados utilizando lenguajes, símbolos y unidades apropiados y pueden ser leídos e interpretados por cualquier tipo de usuario.

En la tabla 3, se muestra la ponderación de cada variable de acuerdo a la característica de validez de datos.

Tabla 3

Matriz de verificación de datos

VARIABLES/CARACTERÍSTICAS	AC	CO	CF	EF	PR	TZ	CP	TOTAL
Cargas familiares	1	1	1	0	0	1	1	5
Destino del Crédito	1	1	1	1	1	1	1	7
Edad	1	1	1	1	0	1	1	6
Estado Civil	1	1	1	1	1	1	1	7
Genero	1	1	1	1	1	1	1	7
Garantía	1	1	1	0	0	1	1	5
Monto	1	1	1	1	1	1	1	7
Nivel Instrucción	1	1	1	0	0	1	1	5
Préstamos Instituciones	1	1	1	0	0	1	1	5
Número de créditos	1	1	1	1	1	1	1	7
Tipo de crédito	1	1	1	1	1	1	1	7
Tipo de Vivienda	1	1	1	1	1	1	1	7
Ingresos	1	1	1	1	0	1	1	6
Egresos	1	1	1	1	0	1	1	6
Activos	1	1	1	1	0	1	1	6
Pasivos	1	1	1	1	0	1	1	6
Tasa de Interés	1	1	1	1	1	1	1	7
Antigüedad del socio en la institución	1	1	1	1	1	1	1	7
Antigüedad laboral	1	1	1	0	0	1	1	5
Zona geográfica	1	1	1	1	1	1	1	7
Número de Cuotas	1	1	1	1	1	1	1	7
Frecuencia de Pago	1	1	1	1	1	1	1	7
Actividad Económica	1	1	1	1	1	1	1	7
Valor cuota	1	1	1	1	0	1	1	6
Capacidad de pago	1	1	1	0	0	1	1	5

VARIABLES/CARACTERÍSTICAS	AC	CO	CF	EF	PR	TZ	CP	TOTAL
Tipo de cliente	1	1	1	1	1	1	1	7
Calificación del cliente	1	1	1	1	1	1	1	7

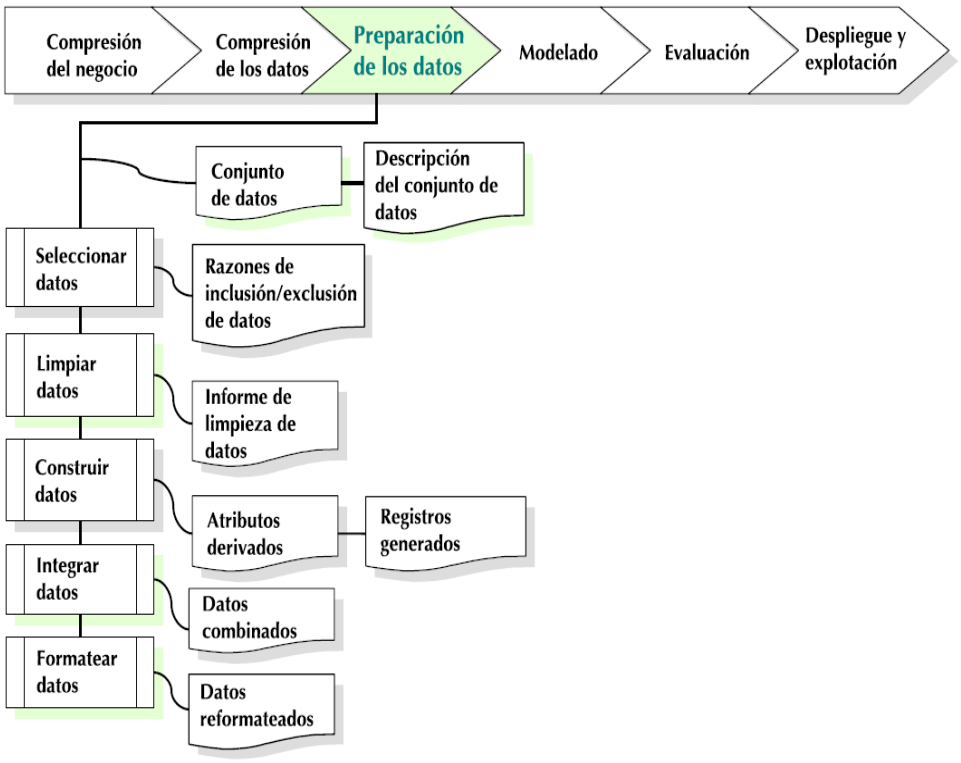
Mediante la ponderación se puede identificar que se necesita mejorar la calidad de datos en Eficiencia y Precisión.

3.3 Fase 3: Preparación de los datos

En la Figura 43, se aprecia las tareas y documentos que se desarrollará en la fase de preparación de los datos.

Figura 43

Metodología CRIPS-DM, Fase de Preparación de los Datos



Nota: Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 9), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

Conjunto de datos

El conjunto de datos a utilizar se enlista en la fase de compresión de datos.

Descripción del conjunto de datos

El conjunto de datos se enlista en la fase de compresión de datos.

3.3.1 Seleccionar datos

Razones de inclusión/exclusión de datos

La obtención de la data es la suma de todos los clientes con historial crediticio dentro de la institución, da un total de 68164 registros almacenados en su base de datos de producción e histórica y compuesta por 27 variables.

Además, para protección de los datos por concepto de sigilo bancario se ocultaron los nombres e identificaciones personales.

3.3.2 Limpiar datos

Informe de limpieza de datos

Uno de los puntos más importantes es el proceso de limpieza y eliminación de la información innecesaria, inconsistente, redundante o errónea en la extracción de los datos de las variables, se observa que existen información con valores nulos, que puede ser debido a fallas en el momento del llenado de la ficha del cliente, además existen incoherencias u omisión en el llenado de la información.

A continuación, se realiza la limpieza de las siguientes variables:

Cargas Familiares: los valores null se dio el valor de 0, además existen valores superiores a 13 (rango de identificación) esto se debe a la cartera migrada en absorciones que ha tenido la institución, estos valores fueron suprimidos.

Edad: existen valores menores a 18 años, esto se debe a errores de ingreso de información en la ficha de clientes, estos valores fueron suprimidos.

Estado Civil: los valores No definidos se les dio el valor de Soltero.

Género: los valores No Aplica y No Definido se les dio el valor de Masculino.

Garantía: los valores No definidos se les dio el valor de Personal.

Nivel de Instrucción: los valores Indefinida y Otros se les dio el valor de Ninguna.

Tipo de Vivienda: los valores Indefinida se les dio el valor de Propia.

Ingresos: existen valores iguales a 0 o null, esto se debe a cartera migrada de absorciones que ha tenido la institución, estos valores fueron suprimidos.

Egresos: existen valores iguales a 0 o null, esto se debe a cartera migrada de absorciones que ha tenido la institución, estos valores fueron suprimidos.

Activos: los valores null se le dio el valor de 0.

Pasivos: los valores null se le dio el valor de 0.

Tasa de Interés: existen valores iguales a 0, esto se debe a cartera migrada de absorciones que ha tenido la institución, estos valores fueron suprimidos.

Antigüedad Laboral: existen valores mayores a 56, esto se debe a cartera migrada de absorciones que ha tenido la institución, estos valores fueron suprimidos.

Zona Geográfica: los valores No definido se les dio el valor de Rural.

Valor de la Cuota: existen valores iguales a 0, esto se debe a cartera migrada de absorciones que ha tenido la institución, estos valores fueron suprimidos.

3.3.3 Construir datos

Atributos derivados

Se realizó un conteo de rangos donde se especifica los mínimos y máximos, que ayudarán a transformar los datos categóricos a números facilitando el entrenamiento del algoritmo y mejorar la interpretación de datos.

En la tabla 4, se aprecia el valor numérico que tendrá cada variable, esto ayudará a la generación del modelo.

Tabla 4

Ponderaciones numéricas

Variables	Rango	Detalle
Cargas familiares	0 a 13	Numero de cargas familiar ejemplo, cónyuge, hijos.
Destino del Crédito	1 a 31	A donde va dirigido el crédito Ver figura 18
Edad	18 a 100	Representa la edad del socio a la hora de solicitar el crédito
Estado Civil	1 a 5	1 Soltero 2 Casado 3 Union Libre/Union De Hecho 4 Divorciado 5 Viudo
Genero	1 - 2	1 Masculino 2 Femenino
Garantía	0 a 3	0 Personal 1 Quirografario 2 Prendario 3 Hipotecario
Monto	100 a 50000	El valor del crédito que puede ir desde 100 hasta 50000 dólares
Nivel Instrucción	1 a 6	1 Primaria 2 Secundaria 3 Tecnica Y Superior 4 Masterado Y Postgrado 5 Phd 6 Ninguna
Préstamos Instituciones	0 - 1	0 no 1 si
Número de créditos	1 a 30	Representa el número histórico de créditos
Tipo de crédito	1 a 4	1 Comercial 2 Consumo 3 Inmobiliario 4 Microempresas
Tipo de Vivienda	1 a 3	1 Propia 2 Arrendada 3 Familiar
Ingresos	1 a 50000	Representa el valor de los ingresos
Egresos	1 a 50000	Representa el valor de los egresos
Activos	1 a 80000	Representa el valor de los activos

Variables	Rango	Detalle
Pasivos	1 a 50000	Representa el valor de los pasivos
Tasa de Interés	9 a 36	Es el número o porcentaje de interés de la entidad financiera
Antigüedad del socio en la institución	0 a 21	Es el número de años del socio desde la apertura de cuenta hasta la solicitud de crédito
Antigüedad laboral	0 a 56	Es el número de años que el socio está laborando
Zona geográfica	1 - 2	1 Rural 2 Urbano
Número de Cuotas	1 a 180	Es el plazo a pagar del crédito
Frecuencia de Pago	1 a 10	1 Mensual (Cada 30 días) 2 Mensual (Fe) 3 Semanal 4 Quincenal 5 Mensual (Cada 28 días) 6 Bimensual 7 Trimestral 8 Semestral 9 Anual 10 Diario
Actividad Económica	1 a 21	Actividad del socio al momento de solicitar el crédito Ver figura 38
Valor cuota	3 a 35000	Es el monto a pagar por cuota
Capacidad de pago	0 a 2	0 Baja 1 Media 2 Alta
Tipo de cliente	0 - 1	0 Nuevo 1 Recurrente
Calificación del cliente	1 a 9	1 Riesgo Normal (A1) 2 Riesgo Normal (A2) 3 Riesgo Normal (A3) 4 Riesgo Potencial (B1) 5 Riesgo Potencial (B2) 6 Deficiente (C1) 7 Deficiente (C2) 8 Dudoso Recaudo (D) 9 Perdida (E)

Registros generados

A parte de las variables anteriores transformadas en numéricas, para un mejor desempeño del modelo no hubo nuevos atributos o cambios, toda la información está completa para el desarrollo de las actividades de la siguiente fase.

3.3.4 Integrar datos

Datos combinados

Se utilizó únicamente los datos almacenados en el Core proveniente del motor de base de datos Informix por lo que no fue necesario la integración con otras fuentes.

3.3.5 Formatear datos

Datos reformateados

Realizada la categorización, como resultado se obtiene una data procesada con valores numéricos que refleja el valor de cada registro, como se muestra en la figura 44, este tipo de procesamiento es parte de una normalización de datos.

Figura 44

Muestra de las variables normalizados

num_cargas	dest_credito	edad	estd_civil	genero	garantia	monto	instruccion	pre_institucionales	num_creditos	tip_crredito	tip_vivienda	ingresos	egresos	activos
4	1	35	2	2	0	3550	1	1	2	4	1	840	350	2000
4	1	35	2	2	0	4000	1	1	2	4	1	840	350	2000
1	1	70	2	1	0	1000	1	0	1	4	1	400	130	11600
4	1	31	3	1	0	1500	1	0	1	4	1	450	250	2250
1	1	68	2	1	0	1000	1	0	1	4	1	250	66	2480
3	1	38	3	1	0	2000	1	1	2	4	1	700	291	4050
3	1	38	3	1	0	800	1	1	2	4	1	700	291	4050
0	1	24	1	1	1	5000	2	0	2	4	3	500	140	150
0	1	24	1	1	1	1500	2	0	2	4	3	500	140	150
5	9	45	1	2	0	5000	1	1	1	4	1	1300	415	5390
1	1	68	2	1	0	3000	3	0	1	4	2	700	330	7600
4	8	43	2	1	1	15000	3	1	1	2	1	1934	810	12780
3	1	60	3	1	0	1200	1	1	1	4	1	800	460	6700
0	1	23	1	1	1	700	2	0	2	4	3	400	110	100
0	1	23	1	1	1	400	2	0	2	4	3	400	110	100
0	1	45	1	2	0	2000	1	0	1	4	1	300	120	1200
2	1	46	1	1	0	1600	1	1	1	4	1	1800	830	20500
1	1	27	2	2	0	3000	2	0	1	4	3	800	120	4300
3	1	54	5	2	0	5000	2	1	1	4	1	1050	235	12700

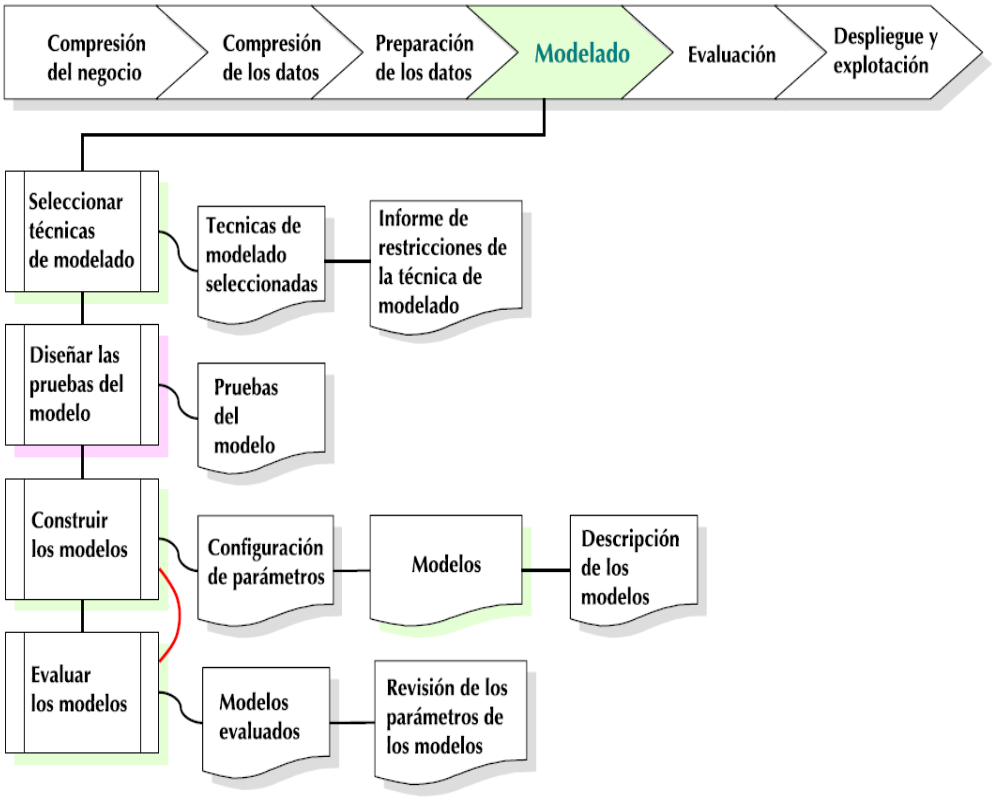
El orden de las variables no afecta en el desarrollo del proyecto así que no es necesario cambiar, la normalización de variables es una gran ayuda a la hora de generar un modelo.

3.4 Fase 4: Modelado

En la Figura 45, se aprecia las tareas y documentos que se desarrollará en la fase de modelado.

Figura 45

Metodología CRIPS-DM, Fase de Modelado



Nota: Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 10), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

3.4.1 Seleccionar técnicas de modelado

Técnicas de modelado seleccionadas

Para la selección de los modelos se enfocaron en modelos predictivos orientados a resolver problemas de ámbitos bancarios tales como predicciones de impago, predicciones de mora, las técnicas seleccionadas son las siguientes:

- Arboles de Decisión (Random Forest)
- Redes Neuronales
- Regresión Logística

Las tres técnicas están soportadas por Python como herramienta que se utilizará para la generación del modelo y la evaluación del mismo con el fin de encontrar el modelo más preciso.

Informe de restricciones de la técnica de modelado

De acuerdo al análisis en apartados anteriores y específicos en las técnicas de aprendizaje supervisado muestra los argumentos necesarios que se debe tomar en cuenta a la hora de desarrollar el modelo, en base a estos análisis en los modelos seleccionados no se encontraron restricciones para su uso, se puede decir que no hay ninguna restricción.

3.4.2 Diseñar las pruebas del modelo

Pruebas del modelo

Antes de la construcción de los modelos para el trabajo de investigación se debe considerar un plan de pruebas, es decir que procedimientos se va a tomar para validar la calidad y la exactitud de cada modelo, para el diseño del plan de pruebas se contempla dos etapas:

La primera etapa es dividir la data en datos de entrenamiento que abarca el 70% y datos para test que son el 30%.

Y la segunda etapa es la validación de los modelos para lo cual se utilizará la Técnica de Evaluación de Matriz de Confusión detallada en el Apartado 2 de este trabajo de investigación, se enlista las métricas para evaluar los modelos:

- Exactitud
- Tasa de error

- Sensibilidad
- Especificidad
- Precisión
- Valor de predicción negativa

3.4.3 Construir los modelos

Configuración de parámetros

Random Forest

Para la construcción del modelo se utilizará la herramienta XGBoost, (Extreme Gradient Boosting), donde se definirá los siguientes parámetros; Se utilizará una clasificación binaria como objetivo principal del modelo, una profundidad máxima de 150, un peso mínimo de observación de 25, una sub-muestra de 0.85, la muestra de columna por árbol de 0.8, una reducción de perdida mínima necesaria de 5, numero de trabajo 16, una tasa de aprendizaje de 0.025 y una velocidad de 1305.

Regresión Logística

Para la construcción del modelo se utilizará los parámetros que vienen definidos por defecto.

Red Neuronal

Para la construcción de este modelo se definirá los siguientes parámetros; el número de neuronas ocultas de (10,10,10), un número máximo de iteraciones de 500, un parámetro de regularización de 0.0001, un optimizador de solución basado en Adam para grandes volúmenes de datos, un número aleatorio para las ponderaciones de 21 y una tolerancia de 0.000000001.

Modelos

Random Forest

Se importa las librerías a utilizar y se conecta a la base de datos donde está almacenado la data.

```
import pyodbc as sqls
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
# Variables para la conexion a la base de datos
driver = '{SQL Server}'
server = 'DESKTOP-50ITGIH'
database = 'bddEconxDW'
trusted = 'yes'
```

```
# Denifir conexion usndo las variables
conexion = sqls.connect(
    Driver = driver,
    Server = server,
    Database = database,
    Trusted_Connection = trusted
)
```

```
sql = '''select * from DataSetY'''
```

```
df = pd.read_sql(sql, con=conexion)
```

```
df.head()
```

Se elimina el campo que no se utiliza en el entrenamiento.

```
X = df.loc[:,df.columns!='calf_socio']
```

```
y = df.loc[:,'calf_socio']
```

```
X.shape, y.shape, df.shape
```

```
((64001, 26), (64001,), (64001, 27))
```

```
X.head()
```

Se divide la data para el entrenamiento y testeo para lo cual el 70% para train y 30% para test.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 1)
```

```
X_train.shape
```

```
(44800, 26)
```

```
X_test.shape
```

```
(19201, 26)
```

Se importa XGBoost para el modelado del algoritmo, para ello se utilizará algunos parámetros que se consideró para definir el algoritmo.

```
import xgboost as xgb
from sklearn.metrics import roc_auc_score
from sklearn.metrics import auc, accuracy_score, confusion_matrix, mean_squared_error

xgtrain = xgb.DMatrix(X_train, label=y_train)

clf = xgb.XGBClassifier(objective = 'binary:logistic',
                        max_depth = 150,
                        min_child_weight = 25,
                        subsample = 0.85,
                        colsample_bytree = 0.8,
                        gamma = 5,
                        n_jobs = 16,
                        learning_rate = 0.025,
                        seed = 1305)

xgb_param = clf.get_xgb_params()
```

```
print('Start cross validation')
cvresult = xgb.cv(xgb_param, xgtrain, num_boost_round=5000, nfold=6, metrics=['auc'],early_stopping_rounds=25, stratified=True,
print('Best number of trees = {}'.format(cvresult.shape[0]))
```

Seguidamente se entrena el modelo utilizando la instrucción fit y los datos tanto de “X” como de “y” de entrenamiento.

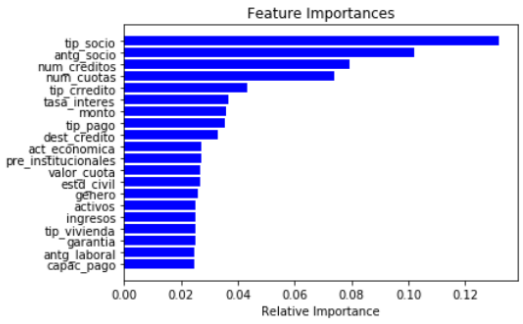
```
clf.set_params(n_estimators=cvresult.shape[0])
print('Fit on the trainingsdata')
xgbs = clf.fit(X_train, y_train, eval_metric='auc')
```

Y finalmente se realiza una predicción, utilizando la instrucción predict y los datos de test.

```
pred = clf.predict_proba(X_test, ntree_limit=cvresult.shape[0])[:,1]
```

Una vez realizada la predicción se obtiene las variables más importantes para este modelo.

```
features = X_train.columns
importances = clf.feature_importances_
indices = np.argsort(importances)[-20:]
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices], color='b', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```



Regresión Logística

Se importa las librerías a utilizar y se conecta a la base de datos donde está almacenado la data.

```
import pyodbc as sqls
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
# Variables para la conexion a la base de datos
driver = '{SQL Server}'
server = 'DESKTOP-50ITGIH'
database = 'bddEconxDW'
trusted = 'yes'
```

```
# Denifir conexion usndo las variables
conexion = sqls.connect(
    Driver = driver,
    Server = server,
    Database = database,
    Trusted_Connection = trusted
)
```

```
sql = '''select * from DataSetY'''
```

```
df = pd.read_sql(sql, con=conexion)
```

```
df.head()
```

Se elimina la columna a predecir, esta columna servirá como referencia para validar el modelo propuesto.

```
X = df.loc[:,df.columns!='calf_socio']
```

```
y = df.loc[:,'calf_socio']
```

```
X.shape, y.shape, df.shape
```

```
((64001, 26), (64001,), (64001, 27))
```

```
X.head()
```

Se separa la data para entrenamiento y testeo.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 1)
```

```
X_train.shape
```

```
(44800, 26)
```

```
X_test.shape
```

```
(19201, 26)
```

Se define el modelo y se entrena.

```
#Defino el algoritmo a utilizar
from sklearn.linear_model import LogisticRegression
algoritmo = LogisticRegression()
```

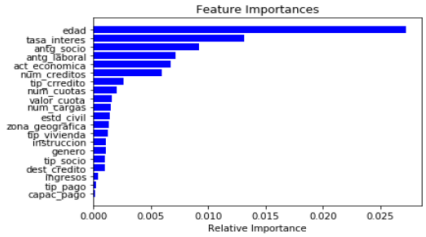
```
#Entreno el modelo
algoritmo.fit(X_train, y_train)
```

Se realiza la predicción utilizando los datos para el test.

```
#Realizo una predicción
y_pred = algoritmo.predict(X_test)
```

Se saca las variables importantes para este modelo.

```
features = X_train.columns
importances = np.abs(algoritmo.coef_[0])
indices = np.argsort(importances)[-20:]
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices], color='b', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```



Red Neuronal

Se importa las librerías y se conecta a la base de datos.

```
import pyodbc as sqls
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
# Variables para la conexion a la base de datos
driver = '{SQL Server}'
server = 'DESKTOP-50ITGIH'
database = 'bddEconxDW'
trusted = 'yes'
```

```
# Denifir conexion usndo las variables
conexion = sqls.connect(
    Driver = driver,
    Server = server,
    Database = database,
    Trusted_Connection = trusted
)
```

```
sql = '''select * from DataSetY'''
```

```
df = pd.read_sql(sql, con=conexion)
```

```
df.head()
```

Se elimina la columna no utilizada para el entrenamiento.

```
X = df.loc[:,df.columns!='calf_socio']
```

```
y = df.loc[:,'calf_socio']
```

```
X.shape, y.shape, df.shape
```

```
((64001, 26), (64001,), (64001, 27))
```

```
X.head()
```

Se separa la data para entrenamiento y testeo.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 1)
```

```
X_train.shape
```

```
(44800, 26)
```

```
X_test.shape
```

```
(19201, 26)
```

Se genera el modelo y se entrena

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

```
from sklearn.neural_network import MLPClassifier
mlp=MLPClassifier(hidden_layer_sizes=(10,10,10), max_iter=500, alpha=0.0001,
                    solver='adam', random_state=21,tol=0.000000001)
```

```
mlp.fit(X_train, y_train)
```

Se realiza una predicción utilizando el modelo entrenado previamente y los datos de test.

```
y_pred = mlp.predict(X_test)
```

Descripción de los modelos

Random Forest

Para este modelo su construcción fue de acuerdo a la definición de los parámetros iniciales, comenzando desde la importación de las librerías, la conexión a la base de datos, la división de la data en entrenamiento y testeo hasta la generación de las variables más importantes, para este modelo fue tipo de socio es decir si un socio es nuevo o ya ha tenido créditos en la entidad financiera con anterioridad (recurrente).

Regresión Logística

La construcción de este modelo está basada en los parámetros por defecto ya que los mismos cumplen con el objetivo del proyecto que es verificar si un socio es sujeto o no de crédito, para este modelo la variable más importante es la edad.

Red Neuronal

Los parámetros definidos para este modelo fueron los más acertados, con esto el modelo fue más preciso y acertado al objetivo.

3.4.4 Evaluar los modelos

Modelos Evaluados

Modelo de Arboles de Decisión (Random Forest)

Para la verificación de la Matriz de Confusión se importa confusion_matrix de la librería sklearn, y se utiliza junto a los datos de reales y los que se ha predicho con anterioridad.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 100)

import xgboost as xgb
from sklearn.metrics import roc_auc_score
from sklearn.metrics import auc, accuracy_score, confusion_matrix, mean_squared_error

xgtrain = xgb.DMatrix(X_train, label=y_train)

clf = xgb.XGBClassifier(objective = 'binary:logistic',
                        max_depth = 150,
                        min_child_weight = 25,
                        subsample = 0.85,
                        colsample_bytree = 0.8,
                        gamma = 5,
                        n_jobs = 16,
                        learning_rate = 0.025,
                        seed = 1305)

xgb_param = clf.get_xgb_params()

#Realizo una predicción
y_pred = clf.predict(X_test)

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred ))

#Verifico la matriz de Confusión
from sklearn.metrics import confusion_matrix

matriz = confusion_matrix(y_test, y_pred)
print('Matriz de Confusión:')
print(matriz)
```

```
Matriz de Confusión:
[[17226    80]
 [ 1785   110]]
```

Exactitud:

$$AC = \frac{17226 + 110}{17226 + 80 + 1785 + 110} = 0.90$$

Tasa de Error:

$$T\_ERROR = \frac{1785 + 80}{17226 + 80 + 1785 + 110} = 0.10$$

Sensibilidad:

$$TP = \frac{17226}{17226 + 80} = 0.996$$

Especificidad:

$$TN = \frac{110}{110 + 1785} = 0.06$$

Precisión:

$$P = \frac{17226}{17226 + 1785} = 0.91$$

Predicción Negativa:

$$VPN = \frac{110}{110 + 80} = 0.58$$

La matriz de error indica de manera general que el grado de clasificación es bastante bueno con un 90% de exactitud y una tasa de error o de clasificación incorrecta del 10%.

También el modelo indica que clasifica los casos positivos con una probabilidad del 99.6% y los casos negativos con una probabilidad del 6%.

Además, si el clasificador dice que un socio es bueno, es que lo es con un 91% de probabilidad. Y si dice que no lo es, entonces el socio es malo con una probabilidad del 58%.

Modelo de Regresión Logística

Para la validación del modelo se verifica la métrica de Matriz de confusión, donde se representa los valores positivos, valores negativos y valores falsos positivos

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 1)

#Defino el algoritmo a utilizar
from sklearn.linear_model import LogisticRegression
algoritmo = LogisticRegression()

#Entreno el modelo
algoritmo.fit(X_train, y_train)

#Realizo una predicción
y_pred = algoritmo.predict(X_test)

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred ))

#Verifico la matriz de Confusión
from sklearn.metrics import confusion_matrix

matriz = confusion_matrix(y_test, y_pred)
print('Matriz de Confusión:')
print(matriz)
```

```
Matriz de Confusión:
[[17334    10]
 [ 1852     5]]
```

Exactitud:

$$AC = \frac{17334 + 5}{17334 + 5 + 1852 + 10} = 0.90$$

Tasa de Error:

$$T\_ERROR = \frac{1852 + 10}{17334 + 5 + 1852 + 10} = 0.10$$

Sensibilidad:

$$TP = \frac{17334}{17334 + 10} = 0.99$$

Especificidad:

$$TN = \frac{5}{5 + 1852} = 0.003$$

Precisión:

$$P = \frac{17334}{17334 + 1852} = 0.90$$

Predicción Negativa:

$$VPN = \frac{5}{5 + 10} = 0.33$$

La matriz de error indica de manera general que el grado de clasificación es bastante bueno con un 90% de exactitud y una tasa de error o de clasificación incorrecta del 10%.

También el modelo indica que clasifica los casos positivos con una probabilidad del 99% y los casos negativos con una probabilidad del 0.3%.

Además, si el clasificador dice que un socio es bueno, es que lo es con un 90% de probabilidad. Y si dice que no lo es, entonces el socio es malo con una probabilidad del 33%.

Modelo Red Neuronal

Del modelo estimado se presenta los siguientes resultados:

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 1)
```

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

```
from sklearn.neural_network import MLPClassifier
mlp=MLPClassifier(hidden_layer_sizes=(10,10,10), max_iter=500, alpha=0.0001,
                  solver='adam', random_state=21,tol=0.000000001)
```

```
mlp.fit(X_train, y_train)
y_pred = mlp.predict(X_test)

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred ))
```

```
#Verifico la matriz de Confusión
from sklearn.metrics import confusion_matrix

matriz = confusion_matrix(y_test, y_pred)
print('Matriz de Confusión:')
print(matriz)
```

```
Matriz de Confusión:
[[17248    96]
 [ 1747   110]]
```

Exactitud:

$$AC = \frac{17248 + 110}{17248 + 110 + 1747 + 96} = 0.90$$

Tasa de Error:

$$T\_ERROR = \frac{1747 + 96}{17248 + 110 + 1747 + 96} = 0.10$$

Sensibilidad:

$$TP = \frac{17248}{17248 + 96} = 0.99$$

Especificidad:

$$TN = \frac{110}{110 + 1747} = 0.06$$

Precisión:

$$P = \frac{17248}{17248 + 1747} = 0.91$$

Predicción Negativa:

$$VPN = \frac{110}{110 + 96} = 0.53$$

La matriz de error indica de manera general que el grado de clasificación es bastante bueno con un 90% de exactitud y una tasa de error o de clasificación incorrecta del 10%.

También el modelo indica que clasifica los casos positivos con una probabilidad del 99% y los casos negativos con una probabilidad del 6%.

Además, si el clasificador dice que un socio es bueno, es que lo es con un 91% de probabilidad. Y si dice que no lo es, entonces el socio es malo con una probabilidad del 53%.

Revisión de los parámetros de los modelos

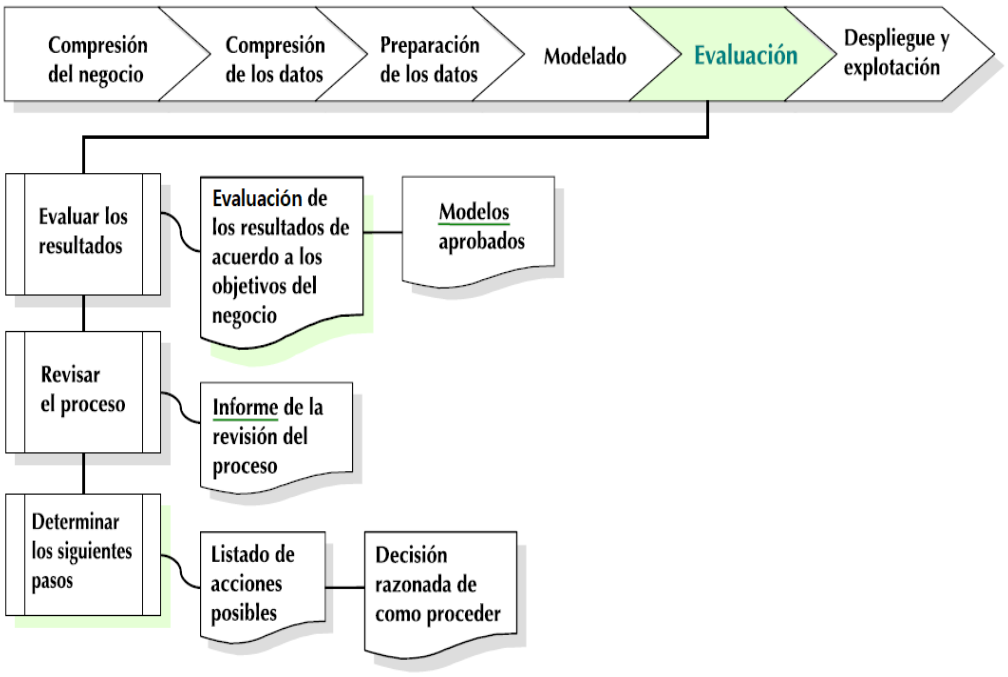
Para el desarrollo de cada uno de estos modelos se realizó una investigación previa de cómo debe funcionar, que debe tener y que no debe tener al momento de definir los parámetros y cuáles serían los resultados a obtener, por esta razón los aciertos al inicio del desarrollo de los modelos fueron los correctos, para justificar este argumento en las siguientes fases de detallan las evaluaciones que se dieron a cada una de los modelos.

3.5 Fase 5: Evaluación

En la Figura 46, se aprecia las tareas y documentos que se desarrollará en la fase de evaluación.

Figura 46

Metodología CRIPS-DM, Fase de Evaluación



Nota: Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 11), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

3.5.1 Evaluar los resultados

Evaluación de los resultados de acuerdo a los objetivos del negocio

Para el desarrollo del presente trabajo de investigación se aplicaron tres modelos de clasificación de Machine Learning que fueron: Random Forest, Regresión Logística y Redes Neuronales.

La siguiente tabla representa una comparativa de los datos con las métricas de nuestros modelos, dando como resultado general modelo con un porcentaje de exactitud muy bueno del 90%, como muestra en la Tabla 5.

Tabla 5

Matriz comparativa de los modelos

Modelo/Métrica	Exactitud	Tasa de Error	Sensibilidad	Especificidad	Precisión	Predicción Negativa
Random Forest (ML1)	0.90	0.10	0.99	0.06	0.91	0.58
Regresión Logística (ML2)	0.90	0.10	0.99	0.003	0.90	0.33
Red Neuronal (ML3)	0.90	0.10	0.99	0.06	0.91	0.53

Ahora se analiza los datos de los modelos aplicados a los objetivos del negocio y los objetivos de Data Mining.

Objetivos del Negocio:

a) Mejorar la evaluación de un socio para saber si es apto o no para un crédito: si se observa la tabla de comparación el modelo ML1 y ML3 predice con una precisión del 91% si un socio es bueno, pero solo el modelo ML1 predice que un socio es malo con el 58%.
b) Agilizar el proceso de calificación y entrega de un crédito: los tres modelos cumplen con este objetivo ya que al ingresar los datos de las variables y ejecutar el modelo hay un tiempo muy corto en el cual resuelve el modelo y da un resultado bueno o malo.

c) Minimizar la probabilidad de incumplimiento en los pagos del crédito: tanto el modelo ML1 como el ML3 dan esa probabilidad, pero el ML1 tiene el 58% más alta de los dos modelos.

Objetivos de Data Mining:

a) Generar patrones que ayuden a la evaluación de un cliente para verificar si es apto o no para un crédito: los modelos ML1 y ML2 muestran un resumen de las Variables Importantes (Features Importances, ver apartado construir los modelos) de cada modelo indicando que patrón es el más sobresaliente a la hora de generar un modelo y con esto verificar si un socio es apto o no.
b) Generar patrones que ayuden con la problemática del incremento de mora y baja de rentabilidad: los modelos ML1 y ML2 muestran un resumen de las Variables Importantes (Features Importances, ver apartado construir los modelos) de cada modelo indicando que patrón es el más sobresaliente a la hora de generar un modelo y con esto verificar la mora y la baja rentabilidad.

Modelos aprobados

Al revisar los objetivos tanto del negocio como de Data Mining se puede indicar que los tres factores más importantes para elegir el mejor modelo es la Exactitud, la Precisión y la Predicción Negativa, resumiendo un poco estas métricas las dos primeras representan el porcentaje de acierto y calidad del modelo al momento de realizar las predicciones verdaderas, mientras que la tercera métrica nos indica el porcentaje de predicción negativa es decir el porcentaje de incumplimiento en el pago del crédito; Se concluye que el modelo de Random Forest aplicando XGBoost con una exactitud del 90%, una precisión del 91% y una predicción negativa del 58%, es el modelo aceptado para la implementación.

3.5.2 Revisar el proceso

Informe de la revisión del proceso

El proceso para la construcción de los modelos seleccionados fue llevado a cabo de acuerdo a la planificación redactada en la fase 1 Compresión del Negocio, se puso más énfasis en la limpieza de la data ya que hubo datos deficientes, esto se debe a las migraciones que han tenido por el cambio de Core financiero y también a las absorciones de otras instituciones financieras, que al tratar de integrar la data en un motor de base datos se suprimieron alguna información de las bases originales.

Para tratar de subsanar esta falta de información se llevaron a cabo técnicas de validación de data y limpieza de la misma escritas en la fase 3 Preparación de los datos.

3.5.3 Determinar los siguientes pasos a ejecutar

Listado de acciones posibles

Como se revisó en la anterior etapa se validó que modelo es el correcto y el que se acerca más al objetivo del proyecto.

Decisión razonada de cómo proceder

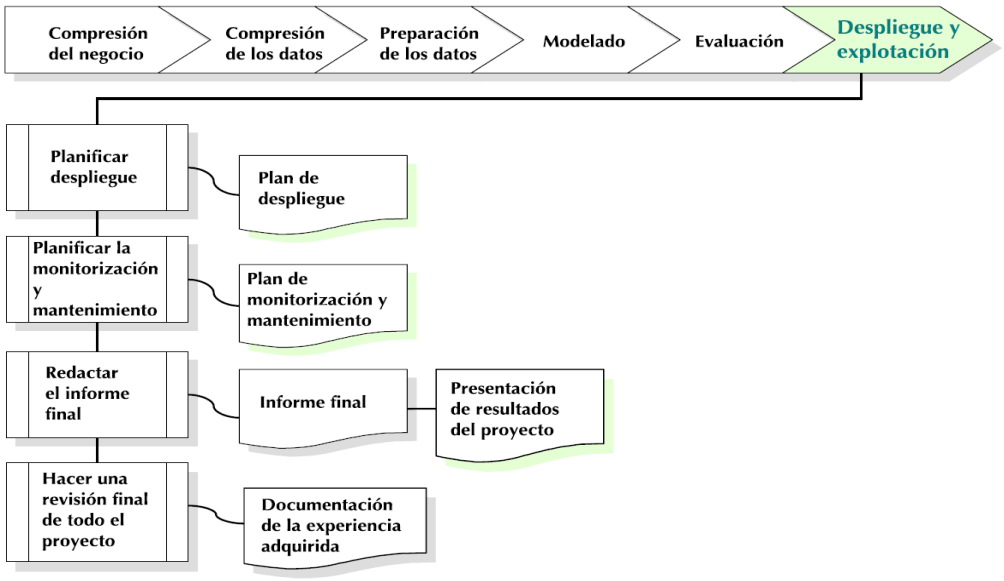
Obtenido ya el modelo que ayudará con los objetivos del Negocio, el siguiente paso del trabajo de investigación es llevar a cabo la fase de Implantación o Despliegue.

3.6 Fase 6: Implantación o Despliegue

En la Figura 47, se aprecia las tareas y documentos que se desarrollará en la fase de despliegue y explotación.

Figura 47

Metodología CRIPS-DM, Fase de Despliegue y Explotación



Nota: Adaptado de “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM” (p. 12), por C. García-Osorio, 2019, 10.13140/RG.2.2.34208.02566, CC BY 2.0.

3.6.1 Planificar despliegue

Plan de despliegue

Para poder implementar el modelo dentro de la institución financiera se crearon 4 fases, que ayudaron a los colaboradores a utilizar Machine Learning como medio de evaluación del riesgo crediticio:

Fase1: Luego de la creación del modelo y de la revisión de cuál es el más opcionado para cumplir con los objetivos del negocio, es necesario exportar el modelo es decir serializar el modelo entrenado para poder utilizar en un Api Web (Interfaz Web que facilita acceso a funciones de un determinado software).

Fase2: Se desarrolla una API web con Flask este framework ayuda a crear aplicaciones web con Python, mediante esta aplicación los colaboradores podrán

ingresar desde cualquier navegador web y podrán realizar la predicción del socio luego de ingresar algunos datos.

Fase3: En esta fase se realiza una capacitación de uso del API web a los colaboradores que trabajan en el área de créditos y operaciones siendo estos los encargados de generar las solicitudes de créditos y la aprobación del mismo.

Fase4: Como última fase se realizarán entrevistas o reuniones de trabajo con los responsables de los procesos operativos para ver los resultados que se obtuvieron al clasificar al socio mediante el modelo y el API web que se realizó utilizando Machine Learning.

3.6.2 Planificar la monitorización y mantenimiento

Plan de monitorización y mantenimiento

El monitoreo y mantenimiento de la implementación del modelo es uno de los pasos importantes para llevar a cabo una buena predicción del socio, se debe tomar en cuenta que existe un programa de actualización de datos, que cada cierto tiempo realiza la entidad financiera y con la afinación de los parámetros se podrá mejorar el proceso de clasificación del modelo.

Además, hay que tomar en cuenta que cada mes hay cancelación de créditos o pre-cancelaciones, estos datos históricos se van actualizando en la base de datos de la institución, como plan de monitoreo y mantenimiento se puede seguir los siguientes procesos:

- Selección y extracción de datos actualizados semestralmente, es decir un proceso de minería de datos.
- Generación del modelo con los nuevos datos sin olvidar que se necesita el 70% para entrenamiento y 30% para testeo.
- Exportación del modelo mediante el proceso de serialización y actualización en el API Web.

- Tener una bitácora de actualización de los modelos y guardar en versión cada modelo.

3.6.3 Redactar el informe final

Informe final

El presente proyecto consiste en el Desarrollo de un modelo predictivo para la evaluación del riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne.

En este sentido, el eje principal del proyecto es el desarrollo de un modelo predictivo para lo cual se empleará los métodos teóricos de análisis y síntesis; para estudiar los conceptos, analizar sus relaciones y sintetizarlos en un conjunto de criterios.

Se empleará el análisis exploratorio de datos mediante Data Mining, utilizando el ciclo de vida de la metodología CRISP-DM.

Haciendo un recuento de las fases de la metodología aplicada se realiza una síntesis, descrita en las siguientes etapas:

Etapa uno. Se realizó un análisis de forma general del negocio con el fin de entender el proceso que conlleva de forma manual la evaluación del riesgo crediticio en la entidad financiera, además se establece objetivos para el negocio y objetivos para el proceso de minería de datos.

Etapa dos. Se realiza una exploración de los datos con el objeto de identificar los problemas de calidad de datos y de identificar las primeras variables para el análisis de la data principal.

Etapa tres. Se realiza el proceso de limpieza con el fin de descubrir las principales variables que contribuirán en la generación de los modelos, para esta etapa se contó con la ayuda del jefe de operaciones que es la persona experta en el proceso de entrega de créditos.

Etapa cuatro. Se hace un análisis de las técnicas de modelado que se va emplear con la ayuda de Machine Learning para predecir si un cliente es bueno o malo, después del análisis se llegó a la conclusión que se va a realizar modelos clasificatorios para lo cual se escogió a Ramdon Forest, Regresión Logística y Redes Neuronales para la construcción de los modelos.

Además, Python fue escogido como lenguaje para el desarrollo de dichos modelos, por las librerías especializadas en el manejo de este tipo de proyectos. Una vez generados los modelos a cada uno se evaluó con métricas detalladas en la matriz de confusión.

Etapa cinco. Se realizó una evaluación de todos los modelos mediante una comparación con los valores obtenidos de las métricas de la anterior etapa, esto sumado a la verificación de que modelo cumple con todos los objetivos tanto del negocio como de minería de datos, obteniendo así la aceptación de un modelo que en este proyecto es Ramdon Forest.

Presentación de resultados del proyecto

Realizado todas estas etapas se presenta los resultados obtenidos de este trabajo de investigación además en el apartado Análisis y Exposición de resultados finales, se detalla lo que se obtuvo al implementar este proyecto.

Hacer una revisión final de todo el proyecto

Documentación de la experiencia adquirida

Como parte final de este proyecto es la revisión de todos los problemas que se encontraron en el camino de cada fase de la metodología CRISP-DM, con esto se podrá generar una bitácora de incidentes para tomar en cuanta en alguna otra etapa de este trabajo de investigación o trabajos futuros.

Como punto muy importarte que se debe tomar en cuenta es la información de la institución financiera, como se ha explicado en los apartados anteriores de este

trabajo la data en algunos casos no existe o hay información errónea, es decir se debe analizar la veracidad de la base de datos esto se debe a las migraciones y las absorciones que ha sufrido la entidad durante su vida institucional.

Otro punto muy importante es el análisis de las variables esto en general se debe trabajar con una o las personas con experiencia dentro de la institución ya que son ellas las que generan el proceso de forma manual y entienden muy bien la línea del negocio.

Capítulo IV

4 Discusión de resultados

4.1 Introducción

En este capítulo se explicará a más detalle los resultados que se obtuvieron durante el desarrollo y ejecución de los modelos, para el desarrollo de dichos modelos se aplicaron tres técnicas de Machine Learning: Random Forest, Regresión Logística y Redes Neuronales utilizando la librería Scikit-learn de Python.

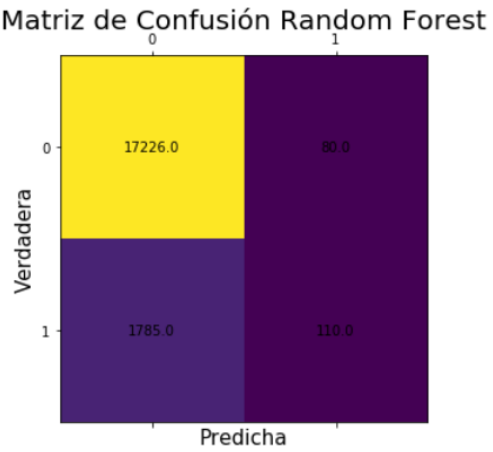
4.2 Validación de los resultados

En la aplicación de estos modelos se utilizó algoritmos de clasificación ya que el objetivo es predecir si el socio es o no sujeto a crédito evaluando el riesgo de crédito, mediante variables que fueron analizadas para el entrenamiento y testeo de los datos previamente extraídos de la base transaccional de la institución.

A continuación, se presenta los resultados de los modelos evaluados con los datos extraídos:

Random Forest

El primer modelo está desarrollado con XGBoots uno de los módulos de la librería Scikit-learn, recordando los datos de la matriz de confusión en la siguiente gráfica:



Indica que hay 17336 registros que fueron clasificados correctamente y que 1865 erróneamente, además se puede detallar lo siguiente:

17226 socios que están clasificados como sujetos a crédito.

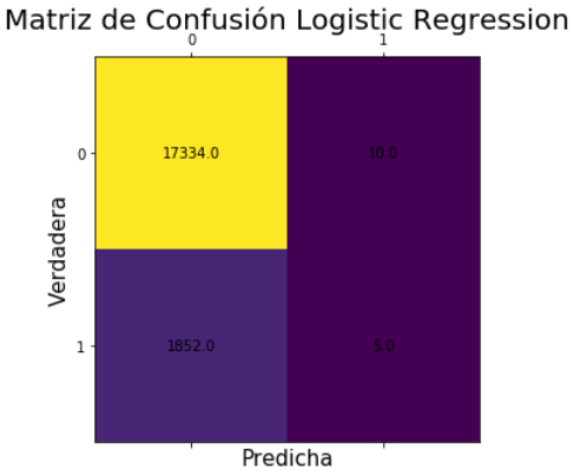
110 socios que fueron clasificados de forma correcta como no sujetos a crédito.

80 socios que fueron clasificados erróneamente como no sujetos a crédito.

1785 socios que fueron clasificados erróneamente como sujetos a crédito.

Regresión Logística

El segundo modelo desarrollado con LogisticRegression de igual manera un módulo de la librería de Scikit-learn, muestra los siguientes datos de la matriz de confusión:



Indica que hay 17339 registros que fueron clasificados correctamente y que 1862 erróneamente, además se puede detallar lo siguiente:

17334 socios que están clasificados como sujetos a crédito.

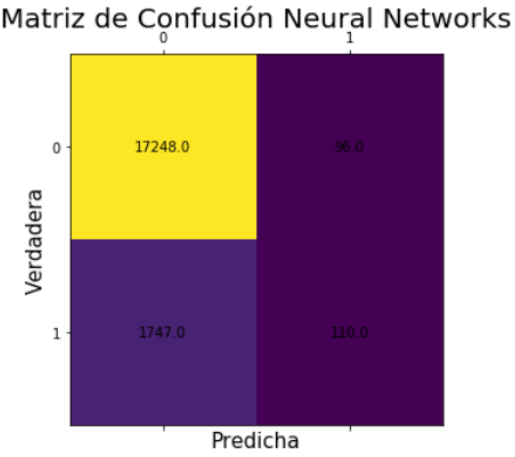
5 socios que fueron clasificados de forma correcta como no sujetos a crédito.

10 socios que fueron clasificados erróneamente como no sujetos a crédito.

1852 socios que fueron clasificados erróneamente como sujetos a crédito.

Redes Neuronales

El tercer modelo está desarrollado con MLPClassifier módulo de la librería de Scikit-learn, muestra los siguientes datos de la matriz de confusión:



Indica que hay 17358 registros que fueron clasificados correctamente y que 1843 erróneamente, además se puede detallar lo siguiente:

17248 socios que están clasificados como sujetos a crédito.

110 socios que fueron clasificados de forma correcta como no sujetos a crédito.

96 socios que fueron clasificados erróneamente como no sujetos a crédito.

1747 socios que fueron clasificados erróneamente como sujetos a crédito.

4.3 Exposición de resultados

Una vez empleada la metodología CRISP-DM para el análisis, desarrollo y evaluación de los modelos como se menciona en la fase 5 de dicha metodología, se llegó a la conclusión que el modelo con más eficacia es el Random Forest.

Además, para verificar si se dio solución al planteamiento del problema y la solvencia de la hipótesis se realizó un aplicativo web, en cual se ingresa los datos de las variables y con el procesamiento del modelo entrenado da como resultado si el socio es bueno (persona es sujeto a crédito) o malo (persona no sujeta a crédito) y con este análisis verificar si se optimizó el proceso de otorgamiento de crédito.

Para el análisis se realizó una entrevista a las personas que llevan a cabo el proceso de calificación de un crédito Ver Anexo 3, los resultados obtenidos de esta entrevista dan a conocer que el 95% de las personas afirman que la calificación de un socio mediante este modelo de clasificación ayudó optimizar y agilizar el proceso de entrega de un crédito al socio, en cambio el 5% comentan que no estaban tan seguros en aplicar esta forma para calificar a un crédito ya que los datos iniciales del socio no eran verídicos.

Capítulo V

5 Conclusiones y recomendaciones

5.1 Conclusiones

- Se acepta la Hipótesis **H0** planteada con un 95%, en el desarrollo del trabajo de investigación.
- Se verifica que el uso de las técnicas de clasificación binaria, son un método eficaz para la evaluación del riesgo crediticio en la Cooperativa de Ahorro y Crédito Virgen del Cisne.
- La elección de la metodología CRISP-DM, ayudó a generar objetivos que cumplan con las necesidades de la línea del negocio y también se enfoque a resolver la hipótesis.
- Para la elaboración de los modelos se realizó una investigación de las técnicas de Machine Learning, que están dentro del análisis predictivo y que se enfoquen en el riesgo crediticio, con este análisis se utilizaron tres modelos de tipo clasificatorio que son: Random Forest, Regresión Logística y Redes Neuronales.
- Con el desarrollo y evaluación de estos modelos de Machine Learning, se ratificó que el modelo de Random Forest utilizando el módulo de XGBoots, es el más acertado para predecir si un socio es o no sujeto a crédito.
- Mediante la evaluación de los modelos se generó patrones de desempeño de un socio, basado en la extracción de las variables importantes, esto puede ayudar al personal de operaciones a realizar una breve validación del socio en campo.

- Con la aplicación de Machine Learning para la generación del modelo más efectivo y el desarrollo de una Api Web (ingreso de datos mediante formulario y proceso de predicción mediante el modelo serializado), esto ayudará a agilizar el proceso de entrega de créditos y mejorar el servicio y atención al socio, a quienes se debe la institución.

5.2 Recomendaciones

- Analizar los resultados obtenidos en el entrenamiento y test de los modelos en conjunto con Gerencia y Subgerencia, con el objeto de mejorar las parametrizaciones y afinar el modelo de una manera más efectiva.
- Capacitar al personal de TI en temas de Minería de Datos e Inteligencia Artificial, con el propósito de predecir resultados que ayuden a mejorar no solo el servicio de crédito, sino también el servicio de captaciones.
- Trabajar conjuntamente con el área de operaciones, para el análisis de nuevas variables que ayuden a mejorar la precisión del modelo.

5.3 Investigaciones Futuras

- En el futuro será muy interesante realizar una investigación para el área de captaciones, con el propósito de predecir el comportamiento del socio cuando decide pre-cancelar una inversión.
- También se puede realizar un análisis con Machine Learning en el área de cumplimiento, verificando si existe o existirá lavado de activos mediante el comportamiento de las transacciones que realiza el socio.

6 Bibliografía

Asobanca. (2019). *ESTÁNDARES REGULATORIOS FINANCIEROS INTERNACIONALES* (Tecnico N.º 4). https://www.asobanca.org.ec/publicaciones/estudios-especiales/informe-técnico-estándares-regulatorios-financieros

Elizondo, A., & Altman, E. I. (2004). *Medición integral del riesgo de crédito*. Editorial Limusa.

Espinosa Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, *21*(3), 1-16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Fernández Hidalgo, D. L. (2013). *Diseño de un modelo de scoring de crédito para la cooperativa de ahorro y crédito Pujili ltda. Ubicada en el cantón Pujilí, provincia de Cotopaxi*. 138.

García, M. L. S., & García, M. J. S. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuadernos de Administración*, *23*(40), Article 40. https://doi.org/10.11144/Javeriana.cao23-40.mpmr

García Osorio, C. (2019). *Metodologías de desarrollo de proyectos de minería de datos—Una visión centrada en CRISP-DM*. https://doi.org/10.13140/RG.2.2.34208.02566

García Sánchez, M., & Sánchez Barradas, C. (2005). *Riesgo de crédito en México: Aplicación del modelo CreditMetrics* [Universidad de las Américas Puebla]. http://catarina.udlap.mx/u_dl_a/tales/documentos/laex/garcia_s_m/

Junta de Política y Regulación Monetaria y Financiera. (2015). *NOTA METODOLÓGICA SOBRE LAS ESTADÍSTICAS MONETARIAS Y FINANCIERAS: NUEVA SEGMENTACIÓN DE CRÉDITO*.

https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/IEMensual/m1967/nota_monetaria.pdf

Ltda, M., Cuenca, Gonzalo, C., & Cuenca, J. (2019). *Propuesta de modelo de machine learning para la evaluación de riesgo de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito La*.

Montes de Oca, J. (2015, julio 20). *Crédito*. Economipedia. Recuperado el 03 de febrero de 2021, de https://economipedia.com/definiciones/credito.html

Pinto Galindo, D. A. (2020, diciembre 10). *Diseño de un Modelo Predictivo de Fuga de Clientes Utilizando Algoritmos Machine Learning*. https://repositorio.ecci.edu.co/bitstream/handle/001/773/Dise%c3%b1o%20de%20un%20Modelo%20Predictivo%20de%20Fuga%20de%20Clientes.pdf?sequence=1&isAllowed=y

Rouhiainen, L. (2018). *INTELIGENCIA ARTIFICIAL 101 COSAS QUE DEBES SABER HOY SOBRE NUESTRO FUTURO* (Centro de Libros PAPF, SLU.). Planeta, S.A.

Salinas Pérez, A. C., & Chee Tse, J. Y. L. (2020). *MODELO DE MEDICIÓN DE RIESGO CREDITICIO EN ENTIDADES FINANCIERAS BASADO EN MINERÍA DE DATOS. CASO PRÁCTICO: CACPECO LTDA.* 218.

Ticona Carpio, C. J. (2016). Modelo de Predicción de la Morosidad en el otorgamiento de Crédito Financiero Aplicando Metodología CRISP-DM. *uancv http://repositorio.uancv.edu.pe/*. http://repositorio.uancv.edu.pe/handle/UANCV/743

Timón, E. (2017). *Análisis predictivo: Técnicas y modelos utilizados y aplicaciones del mismo—Herramientas Open Source que permiten su uso*. 65.

Toscano Palomo, G. N. (2019). *Modelo predictivo del comportamiento de la cartera crediticia para cooperativas de ahorro y crédito*. 101.

Wolters Kluwer. (2019). *Acuerdos de Basilea*.
https://guiasjuridicas.wolterskluwer.es/Content/Documento.aspx?params=H4sIAAAAAAAEAMtMSbF1jTAAASMjY1MztbLUouLM_DxbIwMDS0NDA1OQQGZapUt-ckhlQaptWmJOcSoAMIbOTDUAAAA=WKE

7 Anexos