Pontificia Universidad Católica del Perú
Facultad de Ciencias e Ingeniería

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DEL PERÚ

MEJORA DEL ACCESO AL FINANCIAMIENTO BANCARIO DE EMPRESAS MYPES, USANDO HERRAMIENTAS DE DATA MINING

Tesis para optar el Título de **Ingeniería Industrial**, que presentan los bachilleres

Alvaro Danilo Samaniego Osorio
José Felipe Viamonte Yucra

Profesor: Ing. Jonatán Edward Rojas Polo

Lima, Junio 2020

TEMA DE TESIS

PARA OPTAR	:	Título de Ingeniero Industrial
ALUMNO	:	**ALVARO DANILO SAMANIEGO OSORIO** **JOSÉ FELIPE VIAMONTE YUCRA**
CÓDIGO	:	2011 1050 8 2011 1140 3
PROPUESTO POR	:	Ing. Eduardo Rocca Espinoza.
ASESOR	:	Ing. Eduardo Rocca Espinoza. Ing. Jonatán Edward Rojas Polo.
TEMA	:	MEJORA DEL ACCESO AL FINANCIAMIENTO BANCARIO EN EMPRESAS MYPES, USANDO HERRAMIENTAS DE DATA MINING
N° TEMA	:	
FECHA	:	

JUSTIFICACIÓN:

En el Perú, las Micro y pequeñas empresas mejor conocidas como MYPEs, desempeñan un rol importante en el desarrollo económico, tanto a nivel local como nacional. Estos pequeños negocios emprendedores son considerados impulsores del crecimiento económico, debido a su aporte en la generación de empleo y la reducción de la pobreza (Okpara y Wynn, 2007). Según Produce[1], estas MYPEs han generado aproximadamente el 40% del PBI (Producto Bruto Interno) de nuestro país, de las cuales un 99.3% también contribuye con tributos derivados de su condición de formales.

Estas cifras motivan a un mayor desarrollo de las Microfinanzas[2] - préstamos dirigidos a personas o grupos con medios económicos reducidos que serían excluidos del sistema financiero tradicional – en el Perú, no solo por el potencial de rentabilidad de la colocación de créditos en el sector, sino también por los impactos colaterales generados en las sociedades locales. De acuerdo al estudio “Microscopio Global 2015” de *The Economist Intelligence Unit*, el sector MYPEs es uno de los ambientes más propicios para el desarrollo de este sector, ya que se cuenta con un enfoque adecuado para la aplicación de estrategias de inclusión financiera.

[1] Ministerio de la Producción (2014). Anuario Estadístico Industrial, Mipyme y Comercio Interno. http://www.produce.gob.pe/images/stories/Repositorio/estadistica/anuario/anuario-estadistico-Mype-2014.pdf

Sin embargo, las MYPEs continúan enfrentando diversos obstáculos que impiden su supervivencia en un mediano y largo plazo, tanto así que solo el 30% de ellas logran sobrevivir los dos primeros años[3].

Estos obstáculos son, generalmente, de índole estratégica, administrativa u operativa. La mayoría de las MYPEs, por ejemplo, no poseen un conocimiento adecuado de las herramientas de TI actuales ni de las ventajas del uso del canal digital para una mayor cobertura de sus productos/servicios, no poseen una estructura eficiente de costos, no se adhieren a un régimen tributario formal, entre otros. Si bien es cierto que algunas Mypes desarrollan un *know-how* en sus procesos productivos, ello no es garantía de un posterior crecimiento, ya que existen otros factores exógenos (por ejemplo, la globalización) que puede afectar su desarrollo.

La cantidad promedio de años en el mercado de una micro y pequeña empresa son 6.2 y 8.3 años respectivamente, de las cuales aproximadamente el 59.1% tiene una vida promedio de 5 años en el mercado. La tasa de mortalidad actual nos indica que de cada 100 Mypes existentes, 6 salen del mercado[4], principalmente por un escollo de índole estratégica que afecta al crecimiento y sostenibilidad de una empresa recién implementada: el acceso al crédito.

Para resolver esa problemática, se desarrollará un estudio exhaustivo acerca de las diferentes características de las MYPEs que han contribuido al rechazo o aceptación de un crédito bancario con el objeto de determinar un sistema eficiente de puntaje que pueda ser fácilmente interpretado por la MYPE utilizando una herramienta de gran potencial como lo es el Data Mining a través de modelos predictivos de gran fiabilidad.

Esta última herramienta es de común aplicación en Inteligencia de Negocios, enfocado en el manejo de las relaciones con los clientes (CRM); en Banca, discriminando entre clientes con potencial de *default*, en Terrorismo, utilizando *text mining* para rastrear a los culpables de los atentados del 11-S[5] e inclusive en Logística, determinando modelos analíticos de autoaprendizaje para ajustarse a la estacionalidad de la demanda y programar pedidos automáticamente y de mínimo costo.

OBJETIVO GENERAL:

A través del uso de indicadores de puntaje basados en herramientas de Data Mining, brindar una aproximación de la situación actual de la MYPE a partir de sus características, y de esa manera, identificar aspectos a mejorar para incrementar la probabilidad de acceso al financiamiento bancario.

[3] La Repúbllica. 70 % de las MYPEs fracasan en su negocio. 2010
https://larepublica.pe/economia/491334-70-de-las-mypes-fracasan-en-su-negocio
[4] Ministerio de la Producción (2014). Las Mipyme en cifras 2014.
http://www.produce.gob.pe/reMype/data/Mype2014.pdf
[5] Minería de Datos (s/a). "¿Qué es? ¿Para qué sirve?"
http://www.aprenderaprogramar.com/index.php?option=com_content&id=252:mineria-de-datos-data-mining-ique-es-ipara-que-sirve-1o-parte-dv00105a&Itemid=164

OBJETIVOS ESPECIFICOS:

- Conocer los fundamentos técnicos para clasificar a una MYPE en el Perú.
- Describir la situación actual de las MYPEs en el Perú.
- Conocer los fundamentos teóricos de la aplicación de un proyecto de Minería de Datos y entender los algoritmos de análisis de datos y su aplicación específica.
- Comparar la eficiencia de los algoritmos utilizados del Data Mining y comprender su interpretación a indicadores.
- Identificar un modelo de empresa MYPE para ejemplificar la gestión basada en indicadores y,
- Cuantificar el beneficio económico del uso de estas herramientas.

PUNTOS A TRATAR:

a. Marco Teórico.
Se explicarán los conceptos de composición de las MYPEs en nuestro país, describiendo sus características inherentes, las normas legales y regímenes tributarios a las que deben someterse, el carácter de formalidad de una empresa como tal, entre otros. También se presentarán los fundamentos teóricos de la Data Mining, los algoritmos de aplicación y el uso potencial de estos algoritmos.

b. Estudio de Casos.
Se proporcionará información sobre publicaciones y/o investigaciones relacionadas con el uso de herramientas del Data Mining aplicadas al sector de la micro y pequeña empresa en otros países.

c. Descripción de la situación actual.
Se presentará un estudio con mayor profundidad sobre la situación actual de la MYPE en términos de su productividad, desempeño y evolución en su aporte a la economía peruana.

d. Aplicación de la Minería de Datos.
Se utilizarán las técnicas de la Minería de Datos que mejor se puedan aplicar a la información disponible para segmentar a las MYPES de acuerdo a sus características. También se elaborarán dos modelos predictivos – regresión logística y redes neuronales artificiales – para comparar la eficiencia de su predicción y su viabilidad y facilidad de aplicación para una MYPE.

e. Evaluación Económica.
Se sustentará la aplicación de este proyecto de Data Mining mediante un ahorro cuantificable relativo al valor esperado del préstamo.

f. Conclusiones y recomendaciones.

ASESOR

RESUMEN

El presente trabajo tiene como objetivo dar a conocer una metodología simple para optimizar el acceso al financiamiento bancario para una MYPE a través del uso de herramientas de minería de datos que puedan plasmarse en un aplicativo móvil con una interface amigable para el usuario, que en este caso podría ser el gerente general, el gerente financiero, entre otros; sin demandar una inversión muy alta.

La herramienta de minería de datos que se aplicó fue una red neuronal con aprendizaje profundo, pues involucra más de una capa oculta – mayor cantidad de capas, mayor precisión – para a partir de variables disponibles en un set de datos, determinar el peso relativo de cada una de ellas y estimar la probabilidad de que una MYPE en particular pueda acceder a un crédito bancario. Se aplicó también otra herramienta conocida como regresión logística, sin embargo, por el potencial de aplicación del algoritmo anterior, se descartó la última opción.

En ambos casos se usó un *dataset* de un banco representativo de nuestro país, con historial de créditos aprobados o denegados para MYPEs de diferentes segmentos.

La practicidad del resultado del algoritmo de minería de datos permite que pueda convertirse fácilmente en un app para móviles que, a través de una simple interface de usuario, le permite a una MYPE conocer la probabilidad de acceso al financiamiento de forma personalizada. Esta información es de mucha utilidad para facilitar la toma de decisiones a nivel gerencial y a nivel estratégico (negociar con nuevos proveedores, con clientes, etc.)

Se estimó un beneficio estimado anual de S/1683 por el uso de este aplicativo respecto a no utilizarlo, en un escenario normal proyectado para 5 años en adelante. De la misma forma, se tuvo un VAN de S/3368 para un COK de 14.71%. Asimismo, para un WACC de 20.95% producto de una estructura de financiamiento 20% deuda y 80% aporte propio, el VAN calculado es de S/2360. En ambos escenarios el proyecto de implementación resulta económicamente viable.

Sintetizando, se tendrá un aplicativo móvil desarrollado a partir del algoritmo de minería de datos –redes neuronales– que permitirá a la MYPE tomar decisiones más acertadas.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Las MYPEs en el Perú representan un buen porcentaje de la participación económica en nuestro país, sin embargo, son muy propensas a fracasar durante sus primeros años de funcionamiento debido a diversos factores, sin embargo, uno de los más importantes es la dificultad de acceder a préstamos bancarios para poder seguir creciendo en el negocio.

La hipótesis que se pretende resolver es la factibilidad de elaborar, con el uso de herramientas de Data Mining, un algoritmo de aprendizaje automático que permita conocer y predecir – a partir de características particulares de cualquier MYPE – la probabilidad que tenga ésta de acceder a un crédito dadas sus condiciones actuales. Asimismo, en concordancia con el avance tecnológico actual, utilizar las respuestas de la aplicación de este algoritmo para diseñar un aplicativo móvil que permita que la MYPE pueda realizar este ejercicio periódicamente y de manera muy sencilla.

El presente trabajo comienza con el marco teórico, el cual brindará información y los conocimientos necesarios para entender la definición de un crédito a nivel de riesgos de crédito y, adicionalmente, fundamentos técnicos estadísticos para poder realizar correctamente la aplicación de una herramienta de Data Mining, enmarcada en una visión holística de un proyecto de minería de datos.

En el segundo capítulo se mostrarán dos casos reales donde se aplicaron herramientas de Data Mining a MYPEs a nivel mundial, demostrando que el uso de Data Mining permite encontrar *insights* vitales sobre su desempeño.

En el tercer capítulo se realizará una descripción exhaustiva de la situación actual de las MYPEs en el Perú, a nivel de productividad, competitividad y requisitos legales.

En el cuarto capítulo se realizará la ejecución del proyecto de minería de datos siguiendo la metodología indicada en el marco teórico y se hará una breve descripción de cómo interpretar los resultados en el aplicativo móvil propuesto.

Por último, en el quinto capítulo, se tratará la evaluación económica de la implementación de este aplicativo para una MYPE y se calcularán los indicadores de viabilidad del proyecto para tres diferentes escenarios, demostrando su viabilidad.

CAPÍTULO 1. MARCO TEÓRICO

En el presente capítulo se detalla la teoría a emplear para el desarrollo de la problemática planteada, dividida en dos partes: la descripción del riesgo financiero y la definición, importancia y pasos a seguir para el desarrollo de la minería de datos.

1.1. El Riesgo Financiero

Según la Real Academia Española (2016), el riesgo, proveniente del latín *risicare* que significa "atreverse", se define como una contingencia o proximidad de un daño o, en su forma más general, como cada una de las contingencias que pueden ser objeto de un contrato de seguro. Desde un punto de vista clásico, el riesgo puede ser detallado como "la probabilidad de que ocurra un evento adverso durante un periodo de tiempo delimitado", o incluso como "la medida numérica de la pérdida esperada asociada a un evento adverso". De esta manera, en la práctica de la evaluación del riesgo, a éste se le relaciona con dos elementos importantes: la **severidad** y la **frecuencia** del evento adverso, mediante la siguiente ecuación:

$$RIESGO = SEVERIDAD\ x\ FRECUENCIA$$

Por otro lado, desde un enfoque ampliamente aceptado en el sistema financiero, el riesgo es entendido como "el grado de incertidumbre asociado a una operación financiera o comercial" (BCRP, 2011). De esta manera, el riesgo estaría más relacionado directamente con la probabilidad que con la severidad de un evento. Este riesgo, sin embargo, no está siempre asociado con alguna repercusión negativa, pues podría tratarse de un resultado beneficioso para los implicados; sin embargo, en la práctica son asociados de manera peyorativa.

Dadas las anteriores definiciones y la práctica financiera actual, el **riesgo financiero** puede ser entendido como el nivel de incertidumbre de ocurrencia de un evento no esperado en operaciones financieras o comerciales, el cual genere pérdidas de valor en entidades con actividad financiera. En nuestro país, las entidades susceptibles principalmente son los bancos, AFP's y aseguradoras, entre otros participantes que están afectos a la Ley General del Sistema Financiero de la SBS (Ley N°26702). En vista que dichos eventos no esperados son de diversa índole, a lo largo de la evolución del sistema financiero, se les agrupó en grandes ramas y se les clasificó de acuerdo a los diferentes tipos de riesgo identificados. Los tipos de riesgo más recurrentes en la práctica bancaria son los siguientes (Samaniego, 2008):

- Riesgo de mercado

- Riesgo de crédito
- Riesgo operacional
- Riesgo de liquidez
- Riesgo país

En la labor bancaria, los riesgos son constituyentes de las actividades financieras diarias, ya que al no poseer la información completa de las contrapartes de dichas operaciones ni saber con exactitud lo que realmente ocurrirá en el futuro sobre el cumplimiento de los contratos financieros, o incluso no tener una certeza sobre el impacto de las fluctuaciones económicas, el banco estaría frente a diversos eventos que pueden producir pérdidas económicas en sus utilidades y su capital. Por esta razón, ante la posibilidad de dichos escenarios riesgosos, se ha resaltado la necesidad de implementar una política de Gestión de Riesgos por parte de dichas entidades y, como medida de supervisión obligatoria, por los reguladores del Sistema Financiero, con el objetivo de proteger tanto el capital de las entidades financieras, así como los demás involucrados en las operaciones financieras, tales como personas naturales y/o jurídicas, empresas, entre otros.

1.1.1. El Riesgo de Crédito

El crédito se define como “una operación económica en la que existe una promesa de pago con algún bien, servicio o dinero en el futuro” (BCRP, 2011). En otras palabras, el crédito se constituye cuando un acreedor le entrega dinero a un prestatario, quien incurrirá en la obligación de devolución de dicho monto más los conceptos monetarios generados (intereses) bajo una promesa de pago.

En la práctica bancaria, después de aprobarse la operación de crédito, surge la incertidumbre acerca de la promesa de pago de la contraparte. Si bien es cierto que ambas partes están vinculadas a un contrato, podría darse el caso que el prestatario incumpla en el pago de sus obligaciones, ya sea en términos de monto o plazo. Así, el deudor entra en mora y disminuye su calidad crediticia como sujeto de crédito. Posteriormente, el banco ante dicha duda puede plantearse diferentes alternativas ante dichos eventos, por ejemplo, ejecutar la garantía establecida en el contrato, notificar al deudor de la demora en su pago, etc., con el objetivo de no incurrir en pérdidas económicas ni afectar el capital de la entidad financiera.

Dado el ejemplo, el **Riesgo de Crédito**, según la SBS (Superintendencia de Banca, Seguros y AFP’s), se define como “la posibilidad de pérdidas por la incapacidad o falta de voluntad de los deudores, contrapartes, o terceros obligados, para cumplir

sus obligaciones contractuales registradas dentro o fuera del balance". Así mismo, como el crédito no solo depende de la promesa del acreedor, sino también del valor subyacente de los instrumentos que sostienen el crédito, tales como las garantías, el riesgo estaría determinado por:

- El Incumplimiento del contrato por la contraparte.
- La Pérdida del valor del activo en riesgo o de las garantías.

En vista que una de las actividades principales o *core* del negocio financiero consiste en el otorgamiento de créditos, este hecho corrobora que el Riesgo de Crédito es inherente a la actividad de financiamiento. De esta manera, las entidades financieras se enfocan como una función secundaria a la medición, control y gestión de los diferentes riesgos, en particular el Riesgo de Crédito.

1.1.2. Nuevo Acuerdo de Capital – Basilea II

El Riesgo de Crédito no constituye un tema reciente en el negocio financiero, pues este ha existido a lo largo de la historia del hombre, desde la forma más primitiva de los créditos existente hasta la actualidad. La diferencia radicaba en que las medidas para controlar dicho riesgo eran muy rudimentarias y no estaban sujetas a un juicio lógico. Así mismo, no existían entes de regulación financiera que se encarguen de supervisar dichas operaciones de préstamo.

Durante los siglos XIX y XX, el acelerado crecimiento económico de la humanidad impulsó una mayor actividad del sistema financiero, principalmente para el financiamiento de las naciones y de las primeras grandes compañías y/o empresas de la historia. Posteriormente, se reconocería el importante papel que desempeñaría la Banca, tanto como el motor de crecimiento de las grandes naciones como la causante de las primeras crisis sistémicas causadas por esta, tales como la de 1873, 1884, 1894, 1907 y una de las más importantes, la Crisis de 1929. Ante estos sucesos, se hizo imperativa la necesidad de establecer entidades de regulación y/o supervisión que velasen por la correcta gestión de las entidades financieras y la protección de los depositantes. Producto de ello, nace el Sistema de la Reserva Federal en EE.UU.

A lo largo del siglo XX, la globalización y la proliferación de la tecnología produjo que el mundo económico sea cada vez más complejo y dinámico, lo cual trajo consigo un mayor desarrollo de los temas de supervisión financiera (Samaniego, 2008). En dicho contexto, uno de los primeros logros alcanzados fue el establecimiento del Comité de Supervisión Bancaria para la realización del **Acuerdo de Basilea** en 1988. Dicho

comité inicialmente fue creado en 1974 bajo el nombre de Comité de Regulación y Prácticas Supervisoras Bancarias, el cual estaba conformado por los países de la G10. Este acuerdo conocido como **Basilea I** impuso las bases para la exigencia de los Requerimientos de Capital, los cuales funcionaban como una especie de "colchón" que amortiguase las pérdidas por Riesgo de Crédito; así mismo se establece la clasificación y ponderación de activos de acuerdo al riesgo crediticio. Cabe indicar que dicho comité no posee una jurisdicción **propia**; sin embargo, en la práctica regulatoria de diferentes países, sus recomendaciones siempre son tomadas como un referente para la implementación de políticas de supervisión bancaria.

Básicamente, los objetivos establecidos para este acuerdo eran preservar la estabilidad del sistema bancario e incentivar una competencia igualitaria. A pesar que resultó ser un buen avance en temas de regulación financiera, aún presentaba debilidades que podían ser aprovechadas por los entes del Sistema Financiero.

Para la determinación de los Requerimientos de Capital, se clasificaba a los riesgos de los activos financieros en cinco categorías de riesgo; sin embargo, la principal desventaja de esta clasificación era que carecía de algún sustento estadístico, más se basó en algo meramente cualitativo.

Tabla 1. Categorías de riesgo del Acuerdo de 1988

Pesos	Valores
0%	Emitidos por Estados o Bancos Centrales de los países de la OCDE[6]
10%	Emitidos por Administradores Públicos distintos al Estado. En el caso de la Unión Europea, activos emitidos por entidades crediticias especializadas en el descuento de papel público
20%	Operaciones interbancarias o bien con países no pertenecientes a la OCDE con duraciones menores al año
50%	Préstamos con garantías hipotecarias de vivienda
100%	El resto de las operaciones

Fuente: Samaniego (2008)

En lo referente al cálculo del Requerimiento Mínimo de Capital, este se realizaba con la ponderación indicada en la **Tabla 1**. Así el capital regulatorio debía constituirse al menos con el 8% de estos activos ponderados por riesgo. Este nivel mínimo de capital regulatorio fue establecido a partir de la opinión de expertos. Además, como se puede apreciar, la asignación de estos pesos por los diferentes valores no otorgaba una distinción entre activos adecuada. Así mismo, otro problema presente era la característica "estática" de estas ponderaciones, las cuales no incorporaban cambios en la calidad de los activos crediticios. Otro gran inconveniente del Acuerdo

[6] Organización para la Cooperación y el Desarrollo Económicos.

de Basilea I fue la omisión de incluir otros tipos de riesgos como parte de la Gestión de Riesgos.

De esta manera, todas estas debilidades presentadas conllevan a que se proponga y se establezca un Nuevo Acuerdo de Capital, el cual presentaría mejoras considerables e incluiría elementos omitidos en el Acuerdo de Basilea I.

Así nació el **Nuevo Acuerdo de Capital**, mejor conocido como **Basilea II**, cuyo documento definitivo se publica el 2004, muestra una normativa enfocada en lo siguiente:

- Ámbito de aplicación
- Pilar I: Requerimiento mínimo de capital
- Pilar II: Revisión supervisora
- Pilar III: Comportamiento de mercado

A diferencia del anterior acuerdo de Basilea I, los siguientes elementos estarían presentes en esta edición (Samaniego, 2008):

- Métodos de calificación interna
- Utilización de evaluación externa del crédito en el método estándar
- Técnicas de cobertura del riesgo de crédito
- Titulización de activos
- Tratamiento del riesgo operativo
- Examen supervisor
- Disciplina de mercado

Con esta nueva estructura, se manifestó una mayor eficacia en los objetivos propuestos de Basilea I, ya que se presentaba una gestión con mayor sensibilidad al riesgo, una mayor flexibilidad al adaptarse a las características propias (tamaño, sector desempeñado, concentración de créditos) de las entidades financieras y sus respectivos clientes. Así mismo, por primera vez, se incentiva a dichas entidades a contar con **metodologías internas** que permitan **medir el riesgo de crédito**.

La propuesta del Ámbito de aplicación nace a partir de la presencia de los bancos internacionales y la posible implementación de estas normas por bancos pequeños. Como ambos tipos de bancos en su mayoría pertenecen a grandes grupos empresariales, se hace hincapié en la identificación de los niveles de acción de la misma, ya que la quiebra o *default* de este banco puede traer consigo un riesgo de

contagio[7] al conjunto perteneciente. De esta manera, se propone que la normativa sea aplicada a los tres niveles de acción presentes:

- Nivel de grupo – Consolidación global
- Nivel de subgrupo – Subconsolidación
- Bancos internacionalmente activos

El objetivo de esta medida es mantener solidez tanto a nivel individual en la entidad financiera como a nivel general en el grupo empresarial y evitar que las pérdidas surgidas en un elemento del grupo contagien al resto.

En el Pilar I, a diferencia de Basilea I, se incorporan nuevos riesgos al cálculo del capital regulatorio, así se refuerza el establecimiento del Requerimiento Mínimo de Capital, el cual se puede interpretar con la siguiente ecuación:

$$\frac{Capital\ Regulatorio}{Activos\ ponderados\ por\ riesgo\ de\ crédito,\ de\ mercado\ y\ operacional} \geq 8\%$$

Así mismo, como se ha mencionado anteriormente, el Acuerdo de Basilea II incentiva a que los bancos se ajusten a las nuevas metodologías (Enfoque Estándar y Modelos Internos), los cuales incorporan estándares mínimos de calidad en la aplicación de estos, una mejor ponderación de los activos en riesgo, entre otros. Además, entre los nuevos riesgos incorporados, se resalta la relevancia del **Riesgo de concentración**, es decir, las probables pérdidas provenientes del ejercicio de otorgar créditos a un único cliente o el desenvolvimiento en un único sector en particular.

En el Pilar II, se realza el papel desarrollado por los entes reguladores y supervisores nacionales, así el objetivo de la supervisión no se limitará solamente a verificar el cumplimiento normativo de indicadores regulatorios de las entidades financieras, sino también a realizar un seguimiento prudente en la suficiencia del capital regulatorio e identificar posibles debilidades en la gestión del riesgo antes que estos se materialicen. De esta manera, la SBS, el regulador nacional bancario, está en plena facultad de exigir un mayor capital regulatorio según lo crea conveniente y este sea capaz de reflejar el riesgo asumido por la entidad financiera. Básicamente, el Pilar II del Nuevo Acuerdo comprende en total los siguientes 4 principios mencionados:

- Autoevaluación de la suficiencia de capital de la entidad financiera de acuerdo con su perfil de riesgo.

[7] Riesgo derivado por una institución bancaria de alto renombre, que afecta a otras instituciones bancarias de dimensión pequeña al afectar la estabilidad de sus operaciones financieras. Es uno de los principales *drivers* para el origen de crisis financieras en tiempos recientes (OECD, 2012).

- El organismo supervisor debe cerciorarse que el banco posea adecuados procedimientos de autoevaluación, monitoreo y cumplimiento de los ratios de capital regulatorios.
- El organismo supervisor exige el cumplimiento del capital regulatorio mínimo, así como un capital en exceso por encima del mínimo.
- Intervención oportuna y establecimiento de medidas correctivas antes que se produzca una caída del capital regulatorio por debajo del mínimo.

Por consiguiente, el examen supervisor rectifica que la tarea del supervisor esté dirigida a fomentar que las mejores prácticas bancarias sean las que prevalezcan a través de una valoración adecuada en la Gestión de riesgos y una evaluación permanente de la suficiencia de capital y las estrategias de planeamiento de este en escenarios normales y de estrés.

En el Pillar III, se exponen dos conceptos importantes: Transparencia y Disciplina de Mercado. El primero se basa en la idea que la transparencia de la información ayuda a los participantes involucrados en el Sistema Financiero, ya sean clientes, reguladores, inversionistas, entre otros; mientras que el segundo se justifica como consecuencia del primero, esto se puede apreciar en el hecho de que la información financiera brinda una idea general de los riesgos asumidos por los bancos y al ser conocido por el público, obliga a que el banco sea más prudente en sus actividades y realice una adecuada Gestión de Riesgos.

En general, el objetivo de este pilar es propiciar una correcta y adecuada conducta de mercado de las entidades financieras, a través de la publicación de su información financiera, con el fin de que los participantes estén siempre al tanto de las actividades y de las prácticas realizadas por estos. En sí, gran parte de la divulgación de dicha información es llevada por los reguladores nacionales; en nuestro caso, la SBS revisa y decide qué reportes y/o noticias estarán disponibles para el público en general. Básicamente, el conocimiento sobre la situación y decisiones bancarias permitirá que el público sea capaz de discernir sobre estos y determinarán su acercamiento o alejamiento de dichas entidades, ya que los participantes podrán “premiar” a los bancos más seguros y “castigar” a los más imprudentes en su Gestión de Riesgos (Samaniego, 2008).

1.1.3. Medición del Riesgo de Crédito – Modelos

Dentro del marco de la evaluación de riesgo de crédito, debe tenerse en cuenta que existen ciertos factores determinantes, los cuales forman parte del cálculo de las pérdidas generadas por este tipo de riesgo, dichos factores son:

- Probabilidad de incumplimiento o *default* (PD)
- Exposición al momento de incumplimiento (EAD)
- Severidad o Tasa de Recuperación (LGD)

Antes de definir estos tres elementos, es necesario mencionar que las pérdidas generadas por Riesgo de Crédito pueden ser divididas en dos: Pérdida Esperada (PE) y No Esperada (PNE). Para la definición de la primera será necesario conocer a dichos factores mencionados.

La **Probabilidad de Incumplimiento** (en inglés, *Probability of default, PD*) se define como la probabilidad de que la contraparte no pueda hacer frente a sus obligaciones contractuales en un determinado horizonte de tiempo -generalmente un año- la cual podría ser expresado a través de esta ecuación en uno de los modelos propuestos para su cálculo:

$$PD_t = \frac{Operaciones\ morosas\ en\ el\ año\ t\ de\ su\ vida}{Total\ de\ operaciones\ que\ han\ vivido\ hasta\ el\ año\ t}$$

La **Exposición** (en inglés, *Exposure at default, EAD*) se refiere al valor de todos los derechos con la contraparte en el momento del default, es decir, el valor monetario de la parte de la obligación contractual que no pudo ser pagada a tiempo.

La **Severidad** (en inglés *Loss Given Default, LGD*) se refiere al porcentaje de pérdida que, debido al incumplimiento y después de la ejecución de la recuperación, finalmente se produce. Este indicador, generalmente, contiene los costes de recuperación y el valor de ejecución de la garantía.

Dados estos tres elementos, la pérdida esperada por riesgo de crédito resulta como:

$$PE = PD * EAD * LGD$$

De esta manera, la Pérdida Esperada puede ser definida como la "Pérdida media en un determinado horizonte de tiempo derivado de los incumplimientos incurridos en una cartera de préstamos". Por otro lado, la Pérdida No Esperada (PNE) resulta de la diferencia entre las esperadas y las reales, es decir, es una medida de la volatilidad o desviación respecto de las pérdidas totales a un nivel de confianza.

Cabe señalar que en el marco de Basilea II, otro factor importante a tomarse en cuenta en el análisis cualitativo y cuantitativo de los riesgos es la **Maduración** o **Vencimiento efectivo (M)**, también conocido como Duración de Macaulay, la cual, a diferencia de la maduración tradicional, toma en consideración para su cálculo la

forma, cantidad y tiempo de devolución de un préstamo, ya que resulta diferente pagar dicha obligación en una única cuota o en un conjunto de estas (Guzmán, 2000). Por esta razón, se muestra que en el cálculo de la Duración o Maduración Efectiva para incluir este hecho se incluye una ponderación de los flujos de caja respecto a los períodos de pago:

$$Duración\ (M) = \frac{\sum_{t=1}^{n} \frac{t * Cf}{(1+i)^t}}{P}$$

Dónde:

- M: Maduración efectiva o Duración de un activo
- t: Número de período, siendo n la cantidad de períodos total.
- i: Tasa efectiva del período
- Cf: Valor del flujo de caja en el período t
- P: Valor del activo

1.1.3.1. Método Estándar (External-Rating Based)

En el marco del Acuerdo de Basilea II, el Método Estándar para la Medición del Riesgo de Crédito, a diferencia del Acuerdo de 1988 (Basilea I), incluye el uso del *rating* como herramienta para la calificación de emisiones de deuda o préstamos con el objetivo de brindar una mejor ponderación de los activos en riesgo (APR's). Con ello se obtuvo una mejor clasificación de los créditos, de acuerdo a la calidad crediticia del emisor de deuda, el tipo de emisor, las características del crédito y el contrato de por medio, entre otros detalles. De esta manera, el Acuerdo de Basilea II resalta la importancia del rol de las agencias de calificación crediticia externas en los temas de regulación; entre las más conocidas a nivel mundial, se cuenta con Standard & Poor's, Moody's y Fitch Rating (BBVA, 2015). El trabajo de dichas agencias externas, en breves palabras, consiste en otorgar una calificación de crédito a partir de un análisis general de los emisores de deuda (o prestatarios) y el entorno económico que enfrentan, dicha calificación expresa, a través de una secuencia de letras, números y signos, la capacidad de cumplimiento de pagos del préstamo de dicha entidad prestataria, o incluso, puede estar asociada únicamente a la emisión de deuda en sí, cuyo análisis considera las características del crédito, las condiciones de contrato, entre otros. Así mismo, solamente las agencias que estén designadas como ECAI[8] son reconocidas oficialmente para realizar este tipo de trabajos.

[8] ECAI: Institución de Evaluación Crediticia Externa, en inglés *External Credit Assessment Institutions*.

Así, con el uso de estas calificaciones o *ratings*, el Banco puede asignar porcentajes más adecuados para los diferentes activos ponderados por riesgo, tal como se puede apreciar en la tabla siguiente:

Tabla 2. Coeficientes propuestos en Basilea II

Sectores	**Calificaciones crediticias**						
	AAA a AA-	AA+ a A-	BBB+ a BBB-	BB+ a BB-	B+ a B-	Por debajo de B-	No clasificado
Soberano	0%	20%	50%	100%	100%	150%	100%
Gobiernos Regionales y Locales	Igual que las entidades de crédito						
Empresas públicas	Igual que las entidades de crédito						
Entidades de crédito							
Opción 1	20%	50%	50%	100%	100%	150%	50%
Opción 2	20%	50%	100%	100%	100%	150%	100%
Empresas	20%	50%	100%	100%	150%	150%	100%

Opción1: Ponderación basada en la calificación de un organismo externo de evaluación.
Opción 2: Ponderación basada en la calificación asignada al país en la que está establecida la entidad más un escalón.

Fuente: Dominguez (2003)

Una vez obtenidas las calificaciones de crédito y por consiguiente los porcentajes asignados, se calcula el total de los Activos Ponderados por Riesgo y se procede a hallar el capital regulatorio, el cual debe consistir de al menos el 8% de los APR's, según los establecimientos de Basilea. Así mismo, con las nuevas modificaciones realizadas en el Nuevo Acuerdo, al incluir diversas técnicas de cobertura de riesgo, tales como las garantías o colaterales, la exposición o cuantía de préstamo en riesgo de crédito podría reducirse, lo cual conllevaría a una reducción del capital regulatorio por dicho activo.

Si bien es cierto que la aplicación del Método Estándar ofrece cálculos rápidos y sencillos en la Medición del Riesgo de Crédito, es criticado por este mismo hecho y su dependencia en las agencias externas, lo cual podría propiciar que los bancos no realicen sus propios análisis y se expongan más allá de lo debido. Además, las nuevas modificaciones al Método Estándar no eliminaron por completo algunos errores recurrentes de Basilea I, ya que diferentes estudios empíricos mostraron que, para algunas calificaciones crediticias, las pérdidas inesperadas no eran cubiertas debidamente por el capital regulatorio (Domínguez, 2003).

1.1.3.2. Métodos Internos (Internal-Rating Based)

La aprobación de uso de métodos internos para las entidades financieras resultó ser una novedad del Nuevo Acuerdo de Capital, ya que incentivar el desarrollo de estos origina una mejor comprensión cuantitativa y cualitativa del riesgo asumido por los

bancos, lo cual podría ser usado en la toma de decisiones de dicha entidad. De esta manera, el comité introduce nuevos parámetros que deberán ser usados para el cálculo del Riesgo de Crédito, así como un conjunto de requisitos mínimos que se deberá cumplir para el uso de los métodos internos, para lo cual los supervisores nacionales serán los encargados de evaluar la pertinencia de su uso.

- **IRB Básico:** En una primera etapa, la entidad financiera tiene por obligación hallar la PD asociada a las diferentes exposiciones, según la línea de negocios que ésta atienda (soberanos, empresas, bancos, entre otros); mientras que, el resto de las variables, como la EAD y la LGD, serán proporcionadas por el ente regulador y determinarán su correspondencia de acuerdo a las características de riesgo asumido por la entidad.
- **IRB Avanzado:** En una etapa avanzada del desarrollo de algún modelo interno, las entidades financieras deberán estimar todos los parámetros para el cálculo del Riesgo de Crédito, incluyendo la EAD y la LGD. La aprobación de uso de dichos modelos está sujeto al cumplimiento de requerimientos mínimos establecidos por el Comité de Basilea y el regulador nacional debe asegurar una constante supervisión de este.

En general, la implementación del modelo interno tiene como objetivo reducir la dependencia de las entidades financieras respecto de las agencias de calificación externa como en el Método Estándar y contar con metodologías que buscaban sensibilizar las cargas de capital respecto al riesgo asumido. De esta manera, el incentivo para la migración hacia métodos más avanzados y sofisticados se basaba en el ahorro significativo en el requerimiento mínimo de capital (Poggi, 2005) y una mejor gestión de los riesgos asumidos.

En cuanto al cálculo del capital regulatorio para cubrir las pérdidas inesperadas, este se obtiene a través de la suma de los requerimientos de las categorías de exposición, los cuales surgen con la estimación de los diferentes parámetros mencionados. Así, el tratamiento interno de las garantías se ve expresado al menos en uno de ellos.

$$Pérdida\ Esperada = PD * LGD * EAD$$

Así mismo, cuando la entidad financiera esté autorizada por el regulador para la incorporación de modelos internos, esta tiene por obligación llevar un historial de la información de sus exposiciones y asignar un *rating* interno de acuerdo al nivel de calidad crediticio. Por consiguiente, con dicha información se obtendrá medias

históricas de los principales parámetros con el fin de adaptar el modelo a través del tiempo y apoyar en la gestión crediticia del banco.

1.1.4. Marco Normativo Peruano

Dentro del marco normativo peruano, la Superintendencia de Banca, Seguros y AFP's y el Banco Central de Reserva del Perú (BCRP) son los principales encargados de velar por la estabilidad del Sistema Financiero Peruano, a través de su respectiva supervisión y regulación de las entidades que la conforman. Por un lado, el BCRP tiene como principal objetivo preservar la estabilidad monetaria en nuestro país, cuyo meta se expresa en un 2.0% de inflación acumulada anual[9], el cual lo consigue a través de diferentes mecanismos de transmisión de políticas. Así mismo, entre sus funciones se destacan la regulación de la moneda y el crédito en el Sistema Financiero, administración de las reservas, emisión de billetes y monedas, entre otros.

Por otro lado, la SBS tiene como finalidad defender los intereses del público a través de la regulación y supervisión del cumplimiento de las normas legales, reglamentarias y estatuarias con el fin de preservar la solidez económica y financiera de las entidades que proveen servicios financieros, principalmente los bancos, las aseguradoras y los fondos privados de pensiones, así mismo se podrían incluir cajas municipales y rurales, microfinancieras, algunas cooperativas de ahorro y crédito, entre otros que estén bajo su control (Resolución N°1906 – 2016).

Acerca del Riesgo de Crédito en nuestro país, las principales resoluciones dadas por la SBS que influyen en este son las siguientes:

- Resolución N° 11356 – 2008: Reglamento para la Evaluación y Clasificación del deudor y la exigencia de provisiones.
- Resolución N° 14354 – 2009: Reglamento para el Requerimiento de Patrimonio Efectivo por Riesgo de Crédito.
- Resolución N° 3780 – 2011: Reglamento de Gestión de Riesgo de Crédito
- Resolución N° 8425 – 2011: Reglamento para el Requerimiento de Patrimonio Efectivo Adicional.

Para el desarrollo de la presente investigación, se debe tener en cuenta la definición de créditos a Micro y Pequeñas Empresas (MYPEs), la cual puede variar según el país o región. Para el caso peruano, la SBS da las siguientes definiciones:

[9] Índice actualizado al 2016 en el Portal Web del BCRP.

- **Créditos a Pequeñas Empresas:** Son aquellos que financian actividades de producción, comercialización o prestación de servicios, otorgados a personas naturales o jurídicas, cuyo endeudamiento total en el sistema financiero (excluyendo los créditos hipotecarios para vivienda) se encuentra en los S/. 20,000 y los S/. 300,000 durante los 6 últimos meses.
- **Créditos a Microempresas:** Son aquellos que financian actividades de producción, comercialización o prestación de servicios, otorgados a personas naturales o jurídicas, cuyo endeudamiento total en el sistema financiero (excluyendo créditos hipotecarios para vivienda) es no mayor a S/. 20,000 durante los 6 últimos meses.

En nuestro país, de acuerdo a la Resolución N° 11356 – 2008, se establece que existen 5 categorías de clasificación crediticia del deudor en la cartera de créditos para la Banca Minorista[10], la cual se basa en los días de atraso, y la Banca No Minorista[11], que se trata de una evaluación subjetiva:

Tabla 3. Clasificación del deudor – Resolución SBS N°11356 - 2008

Clasificación del deudor	**Días de Atraso**	
	Pequeña, Microempresa y Consumo	Hipotecario
Normal	Atraso hasta 8 días	Atraso hasta 30 días
CPP[12]	De 9 a 30 días	De 31 a 60 días
Deficiente	De 31 a 60 días	De 61 a 120 días
Dudoso	De 61 a 120 días	De 121 a 365 días
Pérdida	Más de 120 días	Más de 365 días

Fuente: SBS (2016)

En el Sistema Financiero Peruano, por normativa, la SBS exige provisiones para créditos incobrables, el cual se trata de un proceso que permite ajustar el valor (restando) de los saldos de los créditos a su valor recuperable. Estos reflejan la situación en la que el cliente o deudor crediticio ya no se encuentra en capacidad de pagar la promesa de pago con fiabilidad, también son conocidos como **Provisión por Riesgo de Crédito** (SBS, 2016). De esta manera, dadas la anterior clasificación del deudor y el tipo de crédito en riesgo, así como la valuación de las garantías que respaldan dicha operación, se procede con la exigencia y cálculo de las provisiones, las cuales pueden ser de dos tipos:

- **Provisiones genéricas:** Aquellas que se constituyen, de manera preventiva, sobre créditos directos y exposiciones equivalentes a riesgo crediticio de créditos indirectos bajo la clasificación de deudor Normal. A su vez, estas

[10] Banca compuesta por la Pequeña Empresa, Microempresa, el segmento Consumo e Hipotecario.
[11] Banca dirigida a créditos corporativos, medianas y grandes empresas.
[12] CPP: Con Problemas Potenciales

pueden ser obligatorias, exigidas de acuerdo a la SBS, y voluntarias, adicionales de acuerdo a la decisión de la Gerencia.

- **Provisiones específicas:** Aquellas que se constituyen sobre créditos directos y exposiciones equivalentes a riesgo crediticio de créditos indirectos bajo una clasificación de deudor más riesgosa que Normal (desde CPP a Pérdida).

En el capítulo III de la Resolución N° 11356 – 2008, se trata el tema de la exigencia de provisiones en la que se especifica las diferentes tasas mínimas obligatorias determinadas por la SBS, así como condiciones que las entidades financieras deben cumplir en su tratamiento. Una de las más resaltantes es el tratamiento de las **provisiones genéricas pro-cíclicas**, la cual surge a partir de la característica inherente de "prociclicidad" o amplificación de ciclos económicos del sistema financiero, ya que la actividad crediticia se relaciona de manera directa con el desempeño del PBI, así dichas provisiones se constituyen como un elemento de cobertura por riesgo de crédito en los periodos de auge (ASBANC, 2014). La activación de estas, de acuerdo con la regla Pro-cíclica, sucede cuando el promedio de tasa de crecimiento anual del PBI de los últimos 30 meses es mayor o el promedio de los últimos 12 meses supere en 2 puntos porcentuales al de un año antes. En otras palabras, la exigencia de estas provisiones se constituye como un respaldo o "colchón" recaudado en tiempos de auge para poder enfrentar mayores posibles pérdidas en tiempos de recesión.

En el tratamiento general de las provisiones, la SBS debe asegurarse que las entidades financieras resguarden las provisiones calculadas dentro de su contabilidad financiera y asegurarse que estén por encima del mínimo requerido, ya que la retención de estas provisiones tiene como fin cubrir la posible pérdida esperada en la cartera de créditos. Esta operación se ve reflejada en el tratamiento de créditos con más de 90 días de atraso, con la cual, la empresa puede obtener una mejor estimación de la pérdida esperada, teniendo en cuenta la coyuntura actual y la verdadera situación que envuelve a los crédito atrasados. Así, en caso dicha estimación sea mayor que el tratamiento general de provisiones realizado, la entidad financiera tendrá por obligación constituir un mayor monto de provisiones específicas. Posteriormente, en caso haya evidencia real y comprobable de la pérdida por riesgo crediticio, la entidad financiera procede con el castigo de créditos incobrables, las cuales se registran de acuerdo a las normas contables vigentes de la SBS[13].

[13] Disponibles en el Manual de Contabilidad SBS.

De esta forma, ante la supervisión bancaria peruana y el marco de Basilea II, las provisiones y el capital regulatorio funcionan como instrumentos para la cobertura ante el riesgo de crédito: las provisiones estarán destinadas a cubrir las pérdidas esperadas por deterioro en la cartera de créditos; mientras que, el capital regulatorio cubrirá las pérdidas inesperadas del mismo. Por esta razón, a las entidades financieras se les exige modelos internos sofisticados de medición del riesgo de crédito con el fin que puedan afrontar la volatilidad de las pérdidas totales. Tal como se puede apreciar en el siguiente gráfico, la simulación de un escenario pesimista (al 99,99% de nivel de confianza) nos permite cubrir la posible pérdida total de un escenario de alta incertidumbre, tal como lo fue la Crisis Económica del 2008.

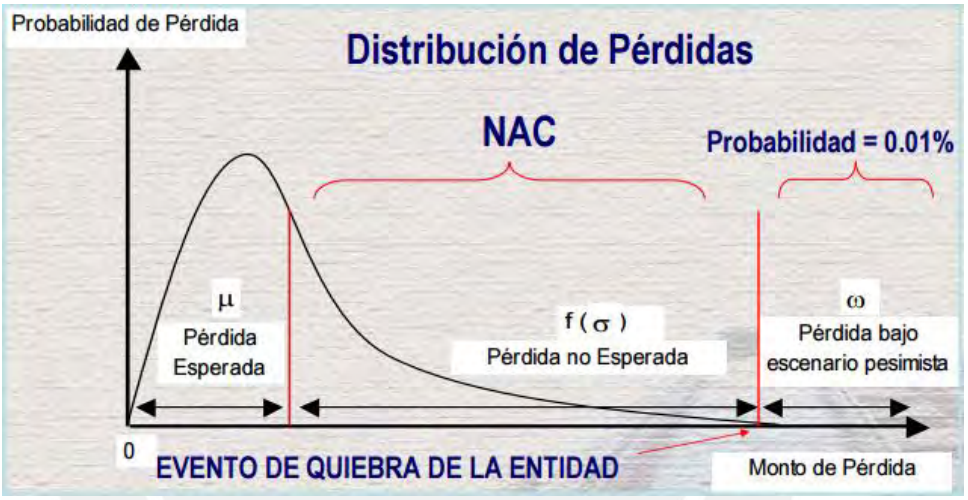


Figura 1. Distribución de Pérdidas
Fuente: Domínguez (2003)

En cuanto al tratamiento del **Requerimiento de Patrimonio Efectivo** por Riesgo de Crédito, anteriormente, se ha mencionado de forma general las novedades implementadas de acuerdo a Basilea II, debido a que dicho comité no posee autoridad sobre los países miembros, estos son libres de decidir si seguir o no sus sugerencias. De esta manera, de acuerdo a la realidad de nuestro país, la SBS decide publicar la Resolución N° 14354 – 2009 en la que se encuentra lo referente a tratamiento de riesgo crediticio. Acerca del método estándar, lo resaltante es que el requerimiento no es del 8% respecto del APR tal como establece Basilea II, sino 10% para el caso peruano[14]; mientras que acerca de los modelos internos, se incorporan los requisitos mínimos para la implementación de estos, tanto para el básico como el avanzado. Posteriormente, la SBS publica la Resolución N° 8425 – 2011, en la que adhiere las nuevas sugerencias dadas por el Comité de Supervisión Bancaria de

[14] Resolución N° 14354 – 2009: Capítulo II, Artículo 7.

Basilea, las cuales surgen a partir de lo acontecido durante la Crisis Económica del 2008, con el fin de promover mayor seguridad y solvencia en el sistema financiero y adaptarse al caso peruano. De esta manera, se incluyen nuevas pautas para el reforzamiento del Patrimonio Efectivo Adicional por ciclo económico para el Método Estándar e Internos, lo que implicaría reconocer sus efectos sobre la solvencia de las entidades financieras.

Al mismo tiempo, con el fin de promover que las entidades financieras susceptibles a riesgo de crédito mejoren en la gestión de dicho riesgo, la SBS dispone de la Resolución N° 3780 – 2011 en la que se enfatiza conceptos relacionados íntegramente con la Administración del Riesgo Crediticio, tales como las funciones de la Unidad de Riesgos de Crédito, las políticas del seguimiento, entre otros. De esta manera, esta resolución sirve como una guía de cumplimiento de los requerimientos mínimos de una adecuada gestión.

1.2. La Minería de Datos

1.2.1. Definición de la Minería de Datos

En el mundo donde vivimos actualmente existen enormes cantidades de datos se recogen continuamente producto del avance tecnológico en el uso de las redes y la globalización. Datos sobre transacciones de ventas, inventarios almacenados, descripciones de productos, promociones, perfiles de clientes, entre otros, son tan masivos que difícilmente cabrían – en términos físicos – en una oficina.

Toda esta revolución tecnológica se origina producto de la necesidad humana de tomar decisiones cada vez más rápido y de forma eficiente.

Sin embargo, los datos[15] por si mismos son meramente descriptivos y por ende, no son relevantes para la toma de decisiones, por lo cual es necesario analizar dichos datos para obtener información[16] que si pueda ser relevante para tomar una decisión.

Es por ello que con ayuda de soporte computacional – pues hacerlo manualmente es prácticamente imposible – estos datos se convertirán en información útil que permita generar conocimiento importante para tomar buenas decisiones.

Todo este proceso de llevar los datos a información y luego a conocimiento recibe el nombre de **Mineria de Datos** (en inglés*, Data Mining*), un término ampliamente

[15] Representación simbólica (números, etc.) de hechos o detalles de una característica de una entidad.
[16] Conjunto de datos organizados y estrucutrados que brindan conocimiento sobre alguna entidad.

utilizado a nivel global y derivado de la equivalencia con la actividad minera. De manera formal se le conoce como KDD (*Knowledge Discovery from Data).*

Figura 2. Minería de datos.
Fuente: Han (2006)

Según Makhabel (2015), la minería de datos comprende “el descubrimiento de un modelo basado en datos que permite hallar conocimiento no esperado de la información pero que es útil, válido y comprensible”. Esta “minería” puede explicarse como un sistema de procesos iterativos (Han, 2012):

1) Limpieza de datos
2) Integración de las fuentes de datos
3) Selección de datos relevantes para el análisis
4) Transformación de los datos
5) Minería de los datos (técnicas/algoritmos de detección de patrones)
6) Evaluación e interpretación de patrones y hallazgos encontrados
7) Presentación formal del conocimiento adquirido al usuario

Los pasos (1) al (4) son usualmente esquematizados en la etapa de pre procesamiento de los datos, pero esto puede variar pues el preprocesamiento comprende todo lo necesario para que los datos estén aptos para ser utilizados en el paso (5), que comprende en sí mismo la aplicación de algoritmo especializado de acuerdo al motivo de la investigación.

Si bien es cierto que el volumen de datos se ha incrementado exponencialmente, las nuevas tecnologías hacen que manejar dichas cantidades de datos no sea un problema. Según Han (2012), la minería de datos se puede aplicar tanto con sistemas

de bases de datos, almacenes de datos, bases de datos on-line (transacciones) e incluso desde algunos servidores de internet.

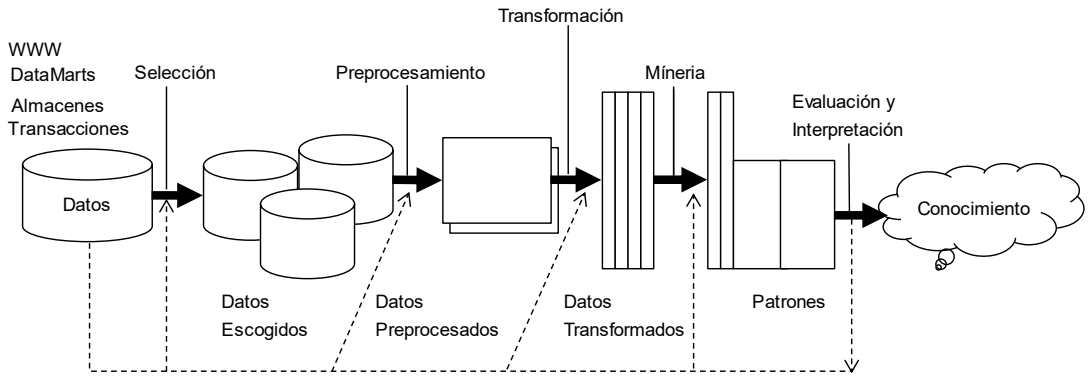


Figura 3. Macroproceso de la minería de datos
Elaboración propia.

Las funcionalidades logradas con la minería de datos pueden dividirse según el resultado de cada funcionalidad en descriptivas y predictivas. Las funcionalidades descriptivas están relacionadas a la manera en la que se pueden caracterizar o resumir los datos, la forma en cómo pueden discriminarse y agruparse los datos y como pueden encontrarse valores atípicos o extraños.

Tabla 4. Funcionalidades de la minería de datos

Función	Tipo	Descripción Breve
Descriptivo	Caracterización	Medidas de resumen de cada atributo en términos de la variable objetivo[17] (histogramas, gráficas 2D, 3D, etc.)
	Discriminación	Medida de comparación de cada atributo en términos de múltiples clases de otra variable objetivo
	Valores Extremos	Datos clasificados como “extremos” a partir de métricas de distancia “extrema” basada en la distribución de los datos
Predictivo	Patrones frecuentes	Comportamientos, secuencias que se repiten entre registros[18] o arreglos (reglas de asociación y correlación)
	Clasificación	Construcción de un modelo predictivo que distinga y describa las clases de una variable objetivo a partir de algunos atributos. Puede ser supervisada[19] (arboles de decisión, redes neuronales, regresión logística) o no supervisada (Naive Bayes, Dendrogramas)
	Regresión	Construcción de un modelo predictivo, matemático, que estime el valor numérico de una variable objetivo a partir de una combinación lineal de los valores de algunos atributos que sean estadísticamente significativos.

Fuente: Han (2012).
Elaboración propia.

[17] Llamada también *target*, es el atributo que se busca describir o predecir.
[18] Un registro en una base de datos, es una colección única de atributos. Un atributo es, pues, una descripción única de una característica de un registro.
[19] Supervisada, cuando se conocen *a priori* las clases a las que pertenece cada registro y no supervisada, cuando no se conocen *a priori* las clases pero se generan con un algoritmo *a posteriori.*

Se puede concluir que la minería de datos posee una dependencia natural con la rama estadística. No obstante, incorpora también a otras múltiples disciplinas.

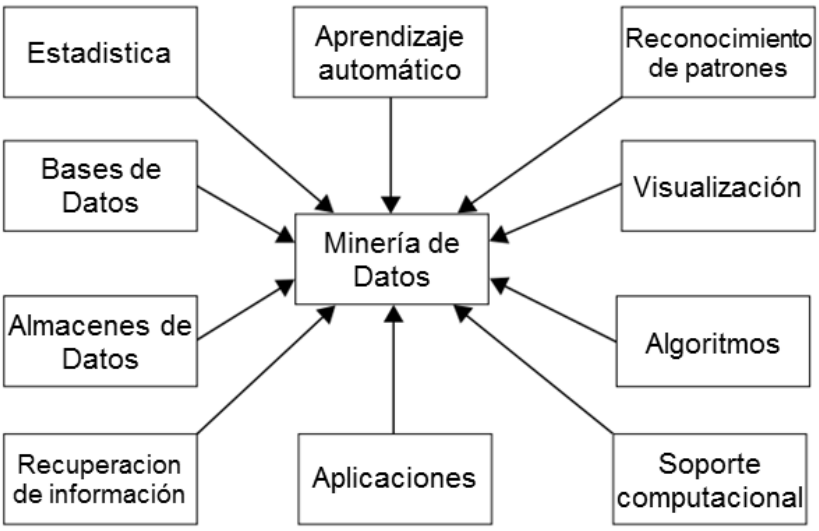


Figura 4. Aporte integral de otras disciplinas a la minería de datos.
Fuente: Han (2012)
Elaboración propia.

1.2.2. Importancia de la Minería de Datos

Como bien decía Dan Schulman, CEO de Paypal, “el mayor impedimento del éxito futuro de una compañía son sus éxitos pasados”, indicado la constante necesidad de adelantarse a los avances tecnológicos, los cuales son parte de la estrategia competitiva de toda empresa que quiera sobrevivir en esta época.

Adaptar los procesos de negocio a las necesidades tecnológicas actuales es lo que se conoce como “transformación digital”, la cual es, ahora, una realidad tangible. Hoy en día prácticamente todas las personas de toda edad poseen por lo menos un equipo tecnológico cuyas funcionalidades son muy variadas: entretenimiento, publicidad, negocios electrónicos, viajes, redes sociales; pero el común denominador es la interacción inmediata: conectados en todo momento.

Esta interacción consecuentemente genera una inmensa cantidad de datos a tal nivel que, según McKinzey (2016) “solo el 90% de los datos digitales en este mundo han sido creados en los últimos 2 años”.

Por esa razón, los datos se están volviendo el activo más valioso de toda organización.

Esta situación motiva la competencia entre compañías. Las organizaciones suelen competir en base a un mejoramiento continuo de sus procesos, tanto internos como externos, para lo cual buscan construir ventajas competitivas que hagan que la compañía aumente de valor; en particular, una de las ventajas competitivas del futuro sopesará en la capacidad de poder transformar la información generada a partir de sus diversos procesos de negocio en conocimiento y este a su vez debe tangibilizarse en estrategias que mejoren la rentabilidad de la compañía.

Resulta ser de particular importancia para la gestión de los procesos en el sector bancario, que en los últimos años ha crecido exponencialmente, producto de una mayor diversificación de portafolios y especialización de productos, lo cual ha llevado al aumento de préstamos otorgados para el sector consumo y para las micro y pequeñas empresas, que en el Perú representan el 99.0% de empresas existentes y contribuyentes (INEI, 2013)

Es por ello que el flujo de información concerniente a las características de sus clientes – desde el punto de vista del banco – deberá ser gestionado por aplicaciones y/o herramientas apropiadas de IT de manera que se pueda extraer conocimiento y patrones de clientes para ofrecer productos financieros que logren una gestión de los mismos que minimice la probabilidad de pérdida.

Estas herramientas ya se aplican en la creación de modelos cualitativos y cuantitativos de *creditscoring*[20], que en la mayoría de casos se traduce en puntos que permiten jerarquizar a los clientes de mayor riesgo y de menor riesgo.

Sin embargo, estos modelos de *creditscoring* quedan en manos del banco por motivos completamente razonables, y pocos modelos toman en consideración un análisis inverso: encontrar las características que deberían tener las empresas que quieren acceder a un préstamo para maximizar su probabilidad de obtenerlo, a partir de –pero no completamente– las condiciones que define un banco representativo del país para otorgar un préstamo.

Todas estas características han motivado el creciente deseo por utilizar esta integración de disciplinas para lograr un mejor entendimiento de los procesos clasificatorios en el servicio bancario (en cuanto a riesgo crediticio) como pilar fundamental para el funcionamiento de su negocio y como aporte sustancial a la

[20] *Credit Score*: Es un modelo de clasificación crediticia que sirve para categorizar a los solicitantes de crédito cuyas peticiones se rechazarán o aceptarán, en función a características cuantitativas (nivel de ingresos, edad) como cualitativas (sexo, estado civil, entre otros).

necesidad que tienen las empresas (usualmente micro y pequeñas empresas) de acceder a un crédito bancario para lograr crecer en el sector y así aportar al desarrollo del país.

1.2.3. El proyecto de Minería de Datos

Es preciso diferenciar las etapas del **proceso** de minería de datos con el **proyecto** de minería de datos. El proyecto de Minería de Datos es el procedimiento estructurado de aplicar las etapas del proceso de minería de datos en un proyecto asociado a las necesidades de conocimiento que se requieren.

Como en todo proyecto, se deben definir lineamientos necesarios para que los objetivos logrados con el proceso de minería de datos coincidan con lo requerido en el proyecto.

Según Makhabel (2015), hay 2 enfoques sistemáticos para la realización del proyecto de Minería de Datos, que son:

- Proceso estandarizado multi-sectorial para la Minería de Datos (en inglés, CRISP-DM)
- Selección, Exploración, Modificación, Modelado y Evaluación (en inglés, SEMMA)

El primero de ellos es el enfoque más utilizado a nivel mundial y consta de 6 etapas. Es un enfoque secuencial pero con capacidad de flexibilizarse y retornar a pasos previos si ocurren nuevos descubrimientos o cambios en el proyecto.

Tabla 5. Etapas del CRISP-DM.

Etapas	Hitos a establecer
Entendimiento del negocio	Objetivos del negocio, evaluación de la situación actual, metas del proyecto de minería de datos y plan de ejecución.
Comprensión de la información	Evaluación de la integridad de los datos. Incluye la recolección, descripción, exploración y evaluación de la calidad de los datos.
Preprocesamiento de la información	Limpieza, selección y transformación de la data para ser aplicada en el método o algoritmo de minería de datos acorde al objetivo del proyecto.
Modelamiento	Uso de técnicas de clusterización para visualizar la segmentación de los datos y aplicación de la técnica apropiada de minería de datos.
Evaluación del modelo	Verificación que los resultados del modelo se alinean con el modelo de negocio y evaluación de la performance del modelo a nivel matemático. Es posible que en esta etapa surjan nuevas necesidades del negocio.
Implementación	Incorporación de los resultados del modelo validado y del conocimiento adquirido en los procesos de negocio de la compañía.

Fuente: Makhabel (2015)
Elaboración propia.

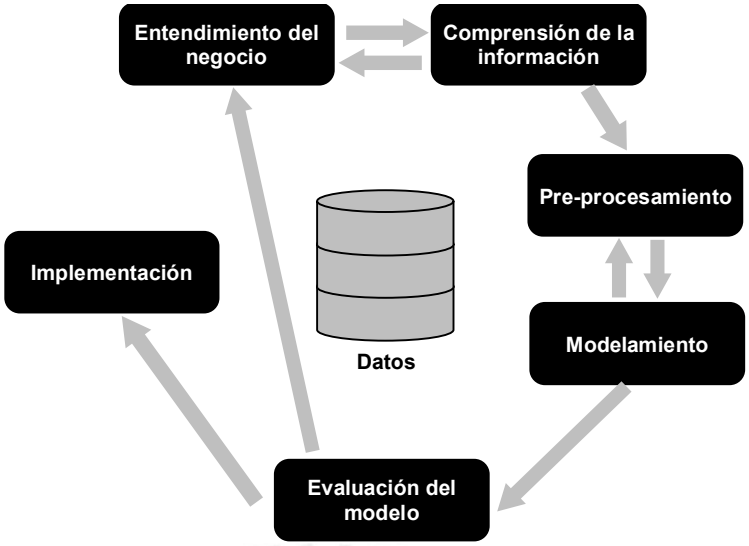


Figura 5. Flujo del proyecto de minería de datos (CRISP-DM).
Elaboración propia.

El segundo enfoque fue desarrollado por el Instituto de Análisis de Sistemas Estadísticos[21]. Es un resumen de las mejores prácticas que se han aplicado en proyectos de minería de datos previos y que los analistas pueden aplicar sin importar el tipo de industria. Funciona como una guía de pasos para la implementación de soluciones de minería de datos, a comparación del enfoque CRISP-DM, que es una metodología integral que funciona en todo tipo de proyecto.

Por lo último descrito, se utilizará el enfoque CRISP-DM de manera simplificada, puesto que algunos pasos por el tamaño del proyecto no se realizarán.

1.2.4. Etapas del proyecto de Minería de Datos

1.2.4.1. Conocimiento del negocio

La primera fase de esta etapa comienza con la **determinación de los objetivos del negocio**, en los cuales se debe entender lo que busca el negocio y como se relaciona con el proceso de minería de datos desde una perspectiva empresarial. En palabras simples, esta fase consta de construir las inquietudes que el negocio busca responder con el proyecto de minería de datos, sujeta a las restricciones y limitaciones que actualmente posee la compañía.

La segunda fase de esta etapa corresponde a la **evaluación de la situación actual** que implica una investigación más profunda acerca de todos los recursos,

[21] El *SAS Institute* (en inglés) es una empresa estadounidense fabricante de soluciones de tecnología de información que hoy por hoy es una de las más grandes proveedoras de soluciones de TI a nivel mundial.

limitaciones, supuestos y otros factores que deben considerarse para afinar los objetivos del proyecto y encaminar el plan de acción.

Tabla 6. Factores para determinar la situación actual.

Factores	Descripción Breve
Inventario de recursos	Recursos disponibles para el proyecto, tanto personal (expertos de negocio, en minería de datos, en TI y soporte técnico, etc.) como informáticos (software/hardware, almacenamiento de datos).
Requerimientos, Limitaciones y Supuestos	Requerimientos del proyecto (cronogramas, marco legal y de seguridad; supuestos y simplificaciones consideradas y limitaciones del proyecto, tanto físicas (recursos disponibles) como tecnológicas (manejo de información).
Riesgos y contingencias	Riesgos o contingencias que puedan retrasar el proyecto, así como planes de contingencia para afrontar dichas situaciones.
Terminología	Glosario de terminología relevante al proyecto de minería de datos.
Análisis costo-beneficio	Evaluación económico-financiera de la necesidad del proyecto.

Elaboración propia.

La tercera fase de esta etapa corresponde a la **determinación de los objetivos de la Minería de Datos** que implica traducir los objetivos del proceso de minería de datos con los objetivos del proyecto, así como también métricas de eficiencia mínimas que deben lograrse con los resultados obtenidos en el proyecto.

Tabla 7. Construcción de objetivos del proceso de minería de datos

Objetivo del negocio	Objetivos del proceso de minería de datos
Incrementar ventas por catálogo de los clientes actuales	Hallar patrones de consumo de los clientes actuales para conocer posibles compras cruzadas.
Minimizar las perdidas por impago de créditos de consumo para una tienda de retail	Construir un modelo predictivo de regresión logística para estimar la probabilidad de impago de un cliente de retail.

Elaboración propia.

La última fase de esta etapa corresponde a la **descripción del plan de acción**, que describe el plan previsto para los logros de los objetivos del proyecto para que consecuentemente se puedan alcanzar los objetivos de negocio. El plan de acción debe incluir una lista de las etapas a ser ejecutadas en el proyecto, junto con la duración de las mismas, los recursos necesarios, las entradas, salidas y dependencias, así como los riesgos que pueden afectar al cronograma del proyecto. Se sugiere que sea un documento dinámico que pueda ser modificado en función al avance en el proyecto y a los cambios presentados. Además, debe definirse un marco general acerca de las herramientas a utilizar.

1.2.4.2. Comprensión de la información

La primera fase de esta etapa comienza con la **recolección inicial de los datos**, a partir de la información descrita en la lista de recursos. Por lo general los datos son recolectados a través de múltiples fuentes y de múltiples bases de datos, por lo tanto,

es fundamental considerar que este conjunto de datos sea posible de integrar. Aquí es necesario establecer un listado de las fuentes empleadas y la forma en cómo han sido extraídas.

La segunda fase de esta etapa comienza con la **descripción de los datos**, en la cual se trata de examinar profundamente a los valores de los atributos ubicados en todos los registros que se encuentran en el subconjunto de datos seleccionado en el paso previo. Asimismo, se verifican si los formatos son adecuados para trabajar.

Se analiza el contenido de las variables (atributos) y en base a ello se determina su tipo a nivel estadístico. Pueden ser **variables cualitativas**, cuyo contenido indica una cualidad o característica de un registro; y **variables cuantitativas**, las cuales tienen un contenido que por su naturaleza debe estar expresada en números. Cada una de estas variables tiene una **escala de medición** que define si los valores de una misma variable pueden ser comparables o no.

Tabla 8. Tipos de variables estadísticas.

Variable	Escala de Medición	Descripción Breve	Ejemplo
Cualitativa	Nominal	Nombres o símbolos que representan una clase o categoría. Solo pueden compararse mas no establecer una jerarquía.	Estado Civil, Sexo
Cualitativa	Ordinal	Nombres o categorías que pueden ordenarse entre sí y pueden compararse por niveles.	NSE
Cuantitativa	Discreta	Valores numéricos que por su naturaleza deben expresarse en números enteros o discretos.	Número de hijos
Cuantitativa	Continua	Valores numéricos que por su naturaleza deben expresarse en números reales o continuos.	Edad, Talla

Fuente: Adaptado de Han (2012).

Tabla 9. Escala de mediciones de las variables estadísticas.

Escala	Descripción Breve
Nominal	La medida de la variable corresponde a un valor que es interpretado como una categoría o clase. En términos estadísticos, solo pueden realizarse las operaciones de = y ≠.
Ordinal	Son categorías que pueden ordenarse entre sí y pueden compararse por niveles (Ej: NSE). Pueden realizarse las operaciones de =, ≠, < y >.
De intervalo	Indican diferencias entre atributos en una escala de rangos máximos y mínimos. Solo representa una medida de orden, por tanto, no tiene un cero absoluto. Pueden realizarse las operaciones de +, -, =, ≠, < y >.
De razón	Indican diferencias absolutas entre atributos. Los valores de los atributos pueden ordenarse y se pueden calcular medidas de tendencia y desviación. Pueden realizarse las operaciones de *, ÷, +, -, =, ≠, < y >.

Fuente: Adaptado de Han (2012).

Para cada tipo de variable existen indicadores estadísticos apropiados. Sin embargo, la mayoría de medidas estadísticas[22] para la descripción de un conjunto de datos tienen como objetivo conocer su valor central y su nivel de dispersión. Por ello las medidas presentadas se aplican mejor a variables del tipo cuantitativas. Estas medidas, así como sus características, se presentan a continuación:

Tabla 10. Medidas de tendencia central.

Medidas	Descripción Breve
Media muestral (poblacional)	Valor promedio de los valores del atributo. Sensible a valores extremos. Calculado para una muestra o para una población.
Mediana	Valor mínimo tal que el 50% de los datos de un atributo son menores o iguales a dicho valor. Robusto frente a valores extremos.
Moda	Valor con mayor frecuencia dentro de una variable.
Percentil k-ésimo	Valor mínimo tal que el k% de los datos de un atributo, son menores o iguales a dicho valor.
Cuartil k-ésimo	Es en esencia, un percentil notable. El Q1, Q2 y Q3 representan los percentiles 25, 50 y 75% de la variable

Fuente: Adaptado de Han (2012).

Cabe destacar que el cálculo de dichos valores difiere si el estudio se realiza con **datos agrupados en intervalos** o **datos sin intervalos**. Los datos se agrupan en intervalos cuando existen muchos valores numéricos muestrales discretos y diferentes, por lo que conviene agruparlos en intervalos para resumir mejor la información.

Para las medidas de dispersión de los datos sucede algo similar. Solo se considerarán las medidas de dispersión para variables con datos cuantitativos no agrupados.

Tabla 11. Medidas de dispersión de los datos.

Medida de Dispersión	Descripción Breve
Rango	Diferencia entre el valor máximo y el valor mínimo en el atributo.
Rango intercuartil	Diferencia entre el tercer y el primer cuartil. Expresa una medida de la amplitud de los valores en la distribución.
Desviación estándar (muestral)	Mide la desviación de los valores con respecto a su media muestral. Utilizada para datos a partir de una muestra como estimación de la desviación estándar poblacional.
Coeficiente de Asimetría de Pearson	Mide la diferencia relativa entre la media muestral y la mediana para estimar la forma de la distribución de los datos. Dependiendo del signo se infiere la asimetría (izquierda o derecha) o simetría de la distribución.

Fuente: Adaptado de Han (2012).

Finalmente, es de igual importancia conocer si existen relaciones a nivel estadístico entre las variables dentro de la base de datos en función a las variables contrastadas.

[22] Llamados también **estadísticos**, pues son tomados a partir de una porción de estudio de una población (muestra). Cuando se analiza toda la población, las medidas se denominan **parámetros**.

La Tabla 12[23] explica las pruebas y/o cálculos estadísticos necesarios para establecer relaciones significativas entre variables:

Tabla 12. Pruebas/medidas de asociación de variables

Variable X	**Variable Y**	
	Cualitativa	Cuantitativa
Cualitativa	• Prueba Chi-cuadrado • Prueba de independencia	• Análisis de factores • Diseño de experimentos
Cuantitativa	• Análisis de factores • Diseño de experimentos	• Coeficiente de correlación • Coeficiente de determinación

Elaboración propia.

La tercera fase de esta etapa consiste en la **exploración de la información**, en la cual se trata de visualizar los patrones que deberían existir luego de la aplicación del proceso de minería de datos, utilizando herramientas de visualización (una, dos, tres o más dimensiones) a partir de la descripción estadística de los datos. Las herramientas de visualización dependerán entonces del tipo y las variables a mirar.

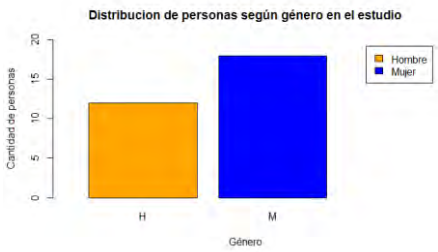


Gráfico de barras: Representación gráfica de un atributo cualitativo en forma de barras verticales. Miden la frecuencia de los valores observados (de forma absoluta o proporcional)

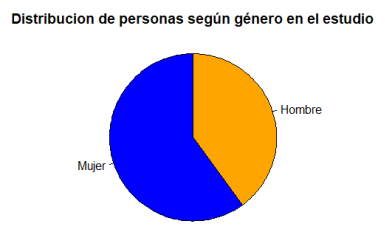


Gráfico de pastel: Representación gráfica de un atributo cualitativo en forma de pastel, con ángulos proporcionales al porcentaje observado para cada clase del atributo.

Figura 6. Gráficos utilizados para atributos cualitativos.
Elaboración propia.

Para un atributo de tipo **cuantitativo discreto** se utiliza el siguiente gráfico:

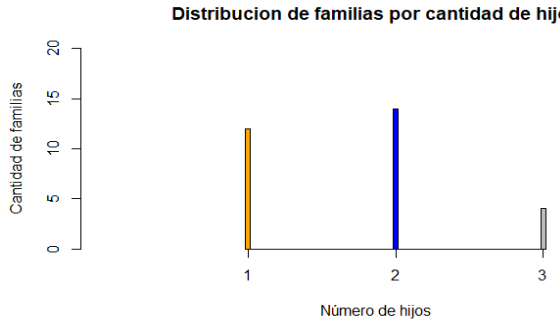
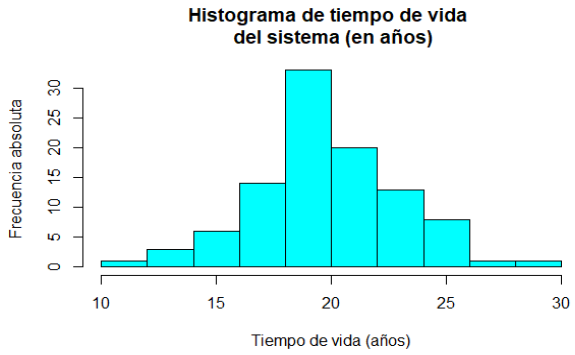


Gráfico de bastones: Representación gráfica de un atributo cuantitativo discreto en forma de bastones verticales. Miden la frecuencia de los valores observados.

Figura 7. Gráfico de bastones
Elaboración propia.

[23] Mayor detalle puede encontrarse en Montgomery (2012).

Finalmente, para un atributo de tipo **cuantitativo continuo** se utilizan los siguientes gráficos (a partir de la tabla de frecuencias con intervalos):



Histograma: Representación gráfica de la distribución de un atributo cuantitativo continúo agrupado en intervalos. Es utilizado para analizar simetrías.

Figura 8. Histograma.
Elaboración propia.

Los anteriores gráficos son utilizados frecuentemente para visualizar la asimetría de la distribución y su dispersión, sin embargo, a menudo es utilizado este tipo de grafico tanto en variables cuantitativas discretas como en cuantitativas continuas para tener una visualización gráfica de la composición de los datos (y cuan cerca o lejos están de sus medidas de tendencia central):

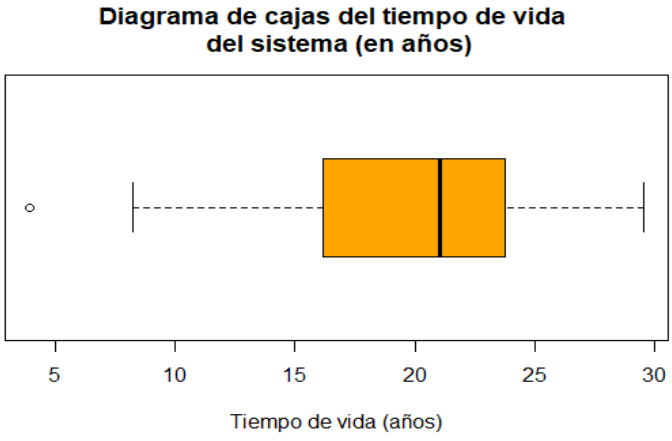


Diagrama de cajas: Es un gráfico en forma de caja que brinda información acerca de los cuartiles de una variable determinada, sus valores máximos, mininos y atípicos, y el nivel de simetría y dispersión de la distribución.

Figura 9. Diagrama de cajas
Elaboración propia.

Por otra parte, las herramientas de uso frecuente para comparar dos atributos en simultáneo y hasta tres atributos, es el diagrama de dispersión.

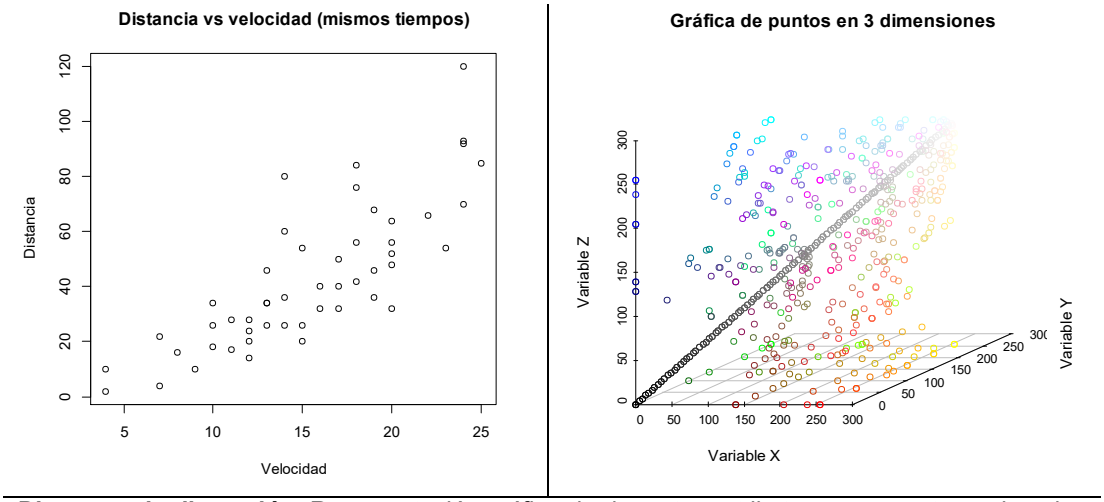


Diagrama de dispersión: Representación gráfica de dos o tres atributos como puntos ordenados dentro de un espacio cartesiano. A la izquierda, un diagrama de dispersión en 2 dimensiones. A la derecha, un diagrama de dispersión en 3 dimensiones. Muy útil para encontrar patrones o tendencias o agrupamientos entre dos atributos numéricos.

Figura 10. Diagramas de dispersión en 2 y 3 dimensiones.
Elaboración propia.

Si es necesario analizar más de 3 atributos en simultáneo, lo más conveniente y lógico consiste en analizar las variables 2 a 2, utilizando lo presentado anteriormente.

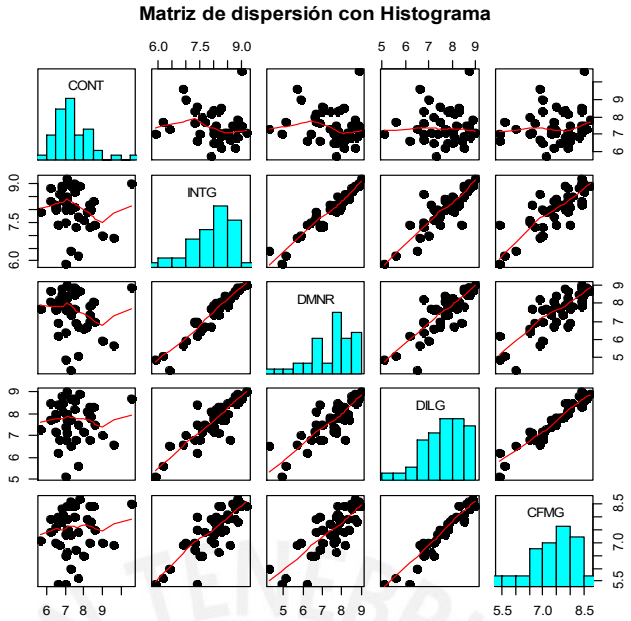
Finalmente, luego de realizar la visualización, se deben documentar los hallazgos encontrados, así como también los diagramas o gráficas que requieren examinación adicional, por tener valores muy sesgados o extremos, ya sea por criterios estadísticos o por criterios de agrupación de clases (más de 2 variables).

La última fase de esta etapa consiste en la **verificación de la calidad de la información**, en la cual se examina la calidad de los datos en términos de los siguientes objetivos (Makhabel, 2015):

- Precisión: Los datos deben haberse registrado correctamente.
- Integridad: Los datos deben ser relevantes para el proyecto.
- Singularidad: Los datos no deben ser redundantes.
- Actualidad: los datos no deben ser tan antiguos.
- Consistencia: Los datos deben ser coherentes con su definición.

Todo error de calidad deberá ser registrado y subsanado en la medida de lo posible para poder proceder a la etapa de preprcesamiento de la información.

Figura 11. Matriz de gráficas de variables en múltiples dimensiones.
Elaboración propia.



Matriz de dispersión con gráfica individual: Representación gráfica de múltiples atributos enfrentados 2 a 2. Los recuadros de la diagonal inferior corresponden a gráficas individuales (histograma, cajas, etc.) de cada variable.

1.2.4.3. Preprocesamiento de la información

En muchos casos, la base de datos seleccionada para la minería de datos debe ser "mejorada", "preparada" o "limpiada" para cumplir con las observaciones rescatadas en la fase de verificación de la calidad y para preparar los datos para el ingreso a modelos o algoritmos. Los problemas típicos encontrados son los siguientes:

Datos incompletos:	Valores faltantes en los atributos
	Cada registro puede carecer de atributos de interés
	Solo totalizaciones o resúmenes de los atributos.
Ruidos	Errores en el ingreso de los datos
	Valores anomalos en la distribución de los datos
	Valores atipicos que afectan a la distribución
Inconsistencias	Discrepancias en codigos o nombres
	Los datos no guardan relación con el atributo y viceversa
	Valores de registros redundantes

Figura 12. Problemas comunes encontrados en las bases de datos.
Fuente: Torgo (2011).

A este tratamiento de la información se le conoce como **preprocesamiento de la información,** que consta de 5 fases (Makhabel 2015).

La primera fase consiste en la **Limpieza de datos**, que consiste en: completar los valores faltantes, suavizar el nivel de ruido de los datos e identificar valores anómalos, y posteriormente corregir las inconsistencias presentes en la base de datos seleccionada. Algunas de las medidas más utilizadas para completar valores faltantes se muestran en la siguiente tabla, sin embargo, cabe resaltar que existe una

vasta cantidad de técnicas para ser aplicadas, que difieren en gran proporción en la complejidad de su utilización.

Tabla 13. Principales medidas de imputación de datos faltantes

Metodología	Ventajas	Desventajas
Eliminación del registro	✓ Elimina el registro que contiene valores faltantes en uno o más atributos. Útil si la base de datos es muy grande.	✓ Perdida de información relevante si la base de datos es pequeña
Llenado manual	✓ De fácil aplicación y sencillo.	✓ Alto sesgo, no recomendado.
Usar una constante global	✓ Simplifica enormemente el llenado de datos faltantes ya que no discrimina las características de cada atributo.	✓ Puede generar sesgo por no incluir el efecto de las características de otros atributos.
Usar una medida de tendencia central	✓ Imputa eficientemente distribuciones simétricas y no simétricas y toma en cuenta las características de cada atributo. Se usa la media, mediana o moda según la distribución y el tipo de atributo con valor faltante.	✓ Requiere un nivel de ejecución computacional mínimo de acuerdo al volumen de datos.
Usar una medida de distancia	✓ A partir de la técnica de los K-vecinos más cercanos. Considera el efecto de los demás atributos. Buena precisión.	✓ Requiere un nivel de ejecución computacional considerable, el cual puede ser costoso o infectable si la base de datos es muy grande.
Usar el valor más probable	✓ A partir de algoritmos de clasificación, tales como arboles de decisión o Naive Bayes, aprende de su clasificación automáticamente.	✓ Puede tener un sesgo considerable si dentro del atributo existe un valor que se repite múltiples veces.
Usar un valor estimado	✓ A partir de técnicas de regresión, brinda una estimación matemática a partir de otros atributos. Buena precisión.	✓ Puede verse muy sesgado a valores numericos extremos. Si la base de datos es muy grande el nivel de procesamiento computacional será muy grande.

Fuente: Adaptado de Makhabel (2015).

Por otro lado, algunas de las técnicas más conocidas para identificar valores atípicos presentes en los datos son:

- **A partir de una técnica de regresión**: A partir de un ajuste eficiente para los datos, es posible reemplazar un valor anómalo u *outlier* con la estimación de la regresión de forma similar a la imputación de datos faltantes. Sin embargo, esto solo es necesario si el *outlier* causa una alteración significativa de los principales indicadores estadísticos de una distribución. De lo contrario contribuye a la variabilidad del modelo.

- **A partir de una técnica de clasificación**: A partir de una regla de clasificación, se puede identificar valores atípicos que no pertenezcan a la misma distribución de los datos “normales”. Si un registro no pertenece a ninguno de los grupos generados mediante clasificación, entonces es un *outlier* y debe ser tratado como tal.

Con respecto a la corrección de las inconsistencias, algunos de los métodos más conocidos son:

- **Relevancia del registro**: Revisar el subconjunto de datos seleccionado y analizar si a partir de las características de los valores de los atributos en el registro se debe considerar el registro (si los valores para los atributos guardan sentido con su denominación).

- **Corrección de coherencia en los datos**: Eliminar los registros duplicados y corregir los errores en los datos en función de la descripción de cada atributo.

La segunda fase de esta etapa consiste en la **integración de los datos**. Es muy probable que el subconjunto de datos seleccionado haya provenido de múltiples fuentes de información. Por tanto, para incrementar la eficacia de la aplicación de la minería de datos, es necesario integrar estas múltiples fuentes en una estructura de datos coherente, buscando lograr los siguientes objetivos:

- Datos homogéneos, con la misma clave[24]
- Mismas definiciones o descripciones de los atributos
- Datos tomados en un mismo lapso de tiempo (sincronización)
- Evitar los datos heredados de un viejo sistema
- Factores sociológicos (límites para la adquisición de datos)

Algunas estrategias para conseguir lo anterior son:

- Identificar las relaciones entre entidades, utilizando técnicas de emparejamiento y analizar entre bases de datos las variables que corresponden a un mismo atributo, de manera que puedan eliminarse.

- Detectar los registros duplicados y eliminar las redundancias entre valores dentro de los atributos.

La tercera fase de esta etapa consiste en la **transformación de los datos**, en el cual, se convierte los datos a un formato apropiado para el uso de las herramientas de la minería de datos[25]. Esto involucra según Han (2012), lo siguiente:

1) El suavizado del ruido de los datos, explicado anteriormente.
2) La construcción de atributos a partir de otros.

[24] También llamada **variable llave**, indica que la variable que identifica al registro (ej. código del producto) de una base de datos a integrar, debe ser la misma para todas las bases a integrar.
[25] En algunas aplicaciones, como, por ejemplo, en el uso de Redes Neuronales Artificiales (ANR), se necesita que los datos estén transformados en un rango de entrada reescalado entre -1 y 1.

3) La totalización, que consiste en aplicar operaciones de agregación o suma para mostrar resúmenes de los datos por cada atributo.
4) La normalización, básicamente escalamiento de datos.
5) La discretización, en donde los valores de atributos continuos son agrupados en intervalos numéricos o intervalos conceptuales.

Los puntos 1), 2) y 3) son determinados en fases anteriores. Los puntos a cubrir y por ende de mayor importancia corresponden a la **normalización** y la **discretización** de variables.

Tabla 14. Normalización y discretización de atributos

Tipo	Descripción Breve	Fórmula
Normalización Min-max	Preserva las relaciones entre los valores originales usando una transformación lineal basada en los valores más altos y bajos.	$v' = \frac{v - min_A}{max_A - min_A}(max'_A - min'_A) + min'_A$
Normalización Z-Score	Los valores son estandarizados en función a la media y la desviación estándar del atributo.	$v'_i = \frac{v_i - \bar{v}}{\sigma_v}$
Normalización Decimal	Se mueve el punto decimal el mayor valor del atributo tal que al dividirlo resulte ser menor que 1.	$v'_i = \frac{v_i}{10^j}$, $j \in Z^+ / j \rightarrow \max(\lvert v'_i \rvert < 1)$
Discretización por histogramas	Los valores numéricos de un atributo se agrupan en intervalos, normalmente del mismo ancho.	$C_k = I_k = [l_k, u_k]$ $X \in C_k \leftrightarrow X \subset I_k$
Discretización por clústers (agrupación)	Los datos son agrupados de acuerdo al método de clusterización (valores comunes)	$C_k = Centroide(a_j), \quad k = \{1 \dots T\}$ $a_j \; son \; T \; valores \; elegidos \; aleatoriamente.$ $X \in C_k \leftrightarrow d(X, C_k) = \min\big(d(X, C_k)\big)$

Fuente: Adaptado de Makhabel (2015)

La última fase de esta etapa consiste en la **reducción de la dimensión de los datos** la cual es necesaria cuando se tiene una inmensa cantidad de atributos dentro del subconjunto de datos seleccionados. El objetivo de la reducción de la dimensión es reducir el tamaño de la matriz de información seleccionada a una dimensión mucho más pequeña, perdiendo una mínima variabilidad.

Tabla 15. Métodos de reducción de dimensionalidad

Metodología	Descripción y características
Análisis de Componentes Principales (ACP)	Busca representar de la mejor manera a la información (en términos de la variabilidad de los datos originales) a partir de una composición lineal de atributos correlacionados y no correlacionados para minimizar el MCE de la de la composición lineal.
Selección iterativa de atributos	Reduce el tamaño del subconjunto de datos al remover o eliminar atributos de forma secuencial. Según el orden y selección de atributos la selección puede ser: • **Hacia adelante:** Empieza con 0 atributos y agrega uno a uno, desde el primer atributo, únicamente para aquellos atributos que aumenten el ajuste a los datos originales. • **Hacia atrás:** Empieza con todos los atributos y elimina uno a uno, desde el último atributo, únicamente para aquellos atributos que aumenten el ajuste a los datos originales. • **Hacia adelante/atrás:** Es una mezcla de ambas técnicas previas. • **Inducción por arboles de decisión:** El algoritmo de árboles de decisión selecciona el mejor atributo para dividir la base de datos en sub-clases. Los atributos utilizados para la división son considerados los más relevantes.

Fuente: Adaptado de Makhabel (2015).

A continuación, se muestra un resumen de la etapa de preprocesamiento:

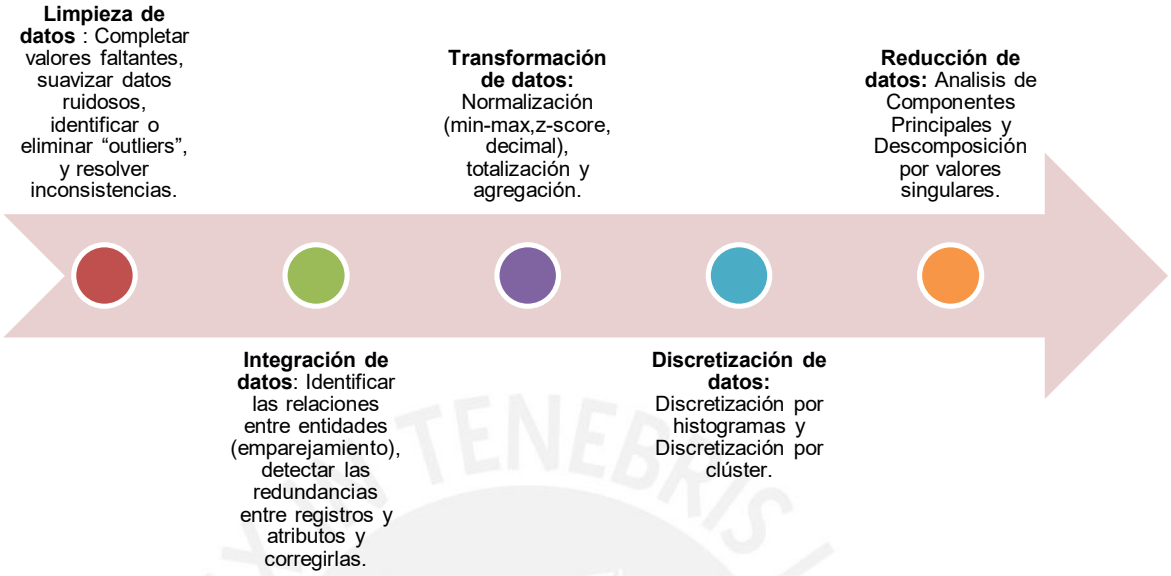


Figura 13. Preprocesamiento de la información.
Elaboración propia.

1.2.4.4. Modelamiento de la información

Una vez preparada la entrada de datos, es posible definir los modelos para obtener el conocimiento.

La primera fase de esta etapa consiste en la **selección de la técnica de modelado**, en la cual se aplican las técnicas descritas en el plan de acción acorde con los objetivos del proceso de minería de datos.

Estas técnicas de modelado se clasifican en técnicas de **aprendizaje supervisado y aprendizaje no supervisado**. Según Berry y Linoff (2004), las técnicas de aprendizaje supervisado se utilizan a partir de la existencia de un variable objetivo en el modelo, que se quiere estimar para nuevos registros. En cambio, las técnicas de aprendizaje no supervisado no tienen supuestos acerca de cuál es la variable objetivo. Estas últimas son útiles para hallar patrones y explicar relaciones entre variables que permitan identificar grupos o asociaciones.

Dentro del aprendizaje supervisado, se incluyen las herramientas de clasificación, estimación y predicción, mientras que, en el aprendizaje no supervisado, se incluyen las técnicas de clusterización, las máquinas de vectores de soporte (SVM) y las reglas de asociación, las cuales no se detallarán al no aplicarse al presente estudio. Algunas de estas herramientas pueden ser usadas en ambos aprendizajes.

- **Árbol de clasificación ID5**: Herramienta de clasificación y de predicción que se utiliza principalmente para dividir una gran colección de registros en conjuntos sucesivamente más pequeños, aplicando reglas de decisión simples. Con cada división sucesiva, los miembros de los conjuntos resultantes se vuelven más similares.

Tabla 16. Pseudocódigo: Árbol de clasificación

Algoritmo 1: Árbol de Clasificación ID5

```
Variables SubsetInicial (ANT,AT), Atributo Target (AT), Atributos No Target (ANT_i), Valores de
Atributo No Target (V_j).
Inicio
   Crear raíz de árbol de clasificación
   Leer SubsetInicial (Atributos No Target, Atributo Target)
(*) Para cada ANT_i del Subset
      Calcular Función RatioGanancia (ANT_i,AT)
      Si RatioGanancia(ANT_k,AT) es máximo entonces
        Asignar ANT_k como nodo del árbol
        Para cada V_j del atributo ANT_k
          Añadir una rama a raíz con ANT_k = V_j
          Establecer subsets para instancias con ANT_k = V_j
          Si Entropia(AT) dentro de Subset_j = 0 entonces
            Terminar árbol
          Caso contrario
            Repetir (*) para Subset_j hasta que Entropia(AT) = 0
Fin
```

Fuente: Adaptado de Sempere (s/f)

- **Red Neuronal Artificial (RNA)**: Es una herramienta de clasificación y de predicción caracterizado por ser uno de los modelos con mayor flexibilidad. Hace uso de modelos matemáticos dentro de la red para "simular" la interpretación de diversas entradas de variables dentro del nodo de procesamiento, de manera que la respuesta de la interpretación sea cada vez mejor conforme se haya entrenado mejor a la red. Cada nivel de interpretación se conoce como capa, por lo que existen redes neuronales de 1 o más capas.

Tabla 17. Componentes de una red neuronal artificial.

Componente	Descripción Breve
Entradas	Las entradas representan los valores de los atributos que se seleccionan para entrenar a la red neuronal. Usualmente, se sugiere escoger solo las variables que determinen mejor (de forma intuitiva) la salida deseada.
Función de Combinación	La función de combinación, que se da en cada nodo de la red neuronal, combina todos las entradas usualmente como una suma ponderada, donde cada atributo tiene su propio peso w_i. En este paso se sugiere reescalar las entradas de cada atributo al rango de salida de la función de transferencia.
Función de transferencia	Se encarga de reescalar la función de combinación al rango de salida de la función de transferencia. Típicamente es la función sigmoidal o la función Logit.
Salida	Es el resultado de la aplicación de la función de transferencia, el cual puede representar otra entrada hacia otro nodo de la red neuronal o puede ser la salida final. En caso sea la salida final, de ser necesario se debe reescalar al rango real de la salida, en función a la variable objetivo definida.

Fuente: Berry (2004)

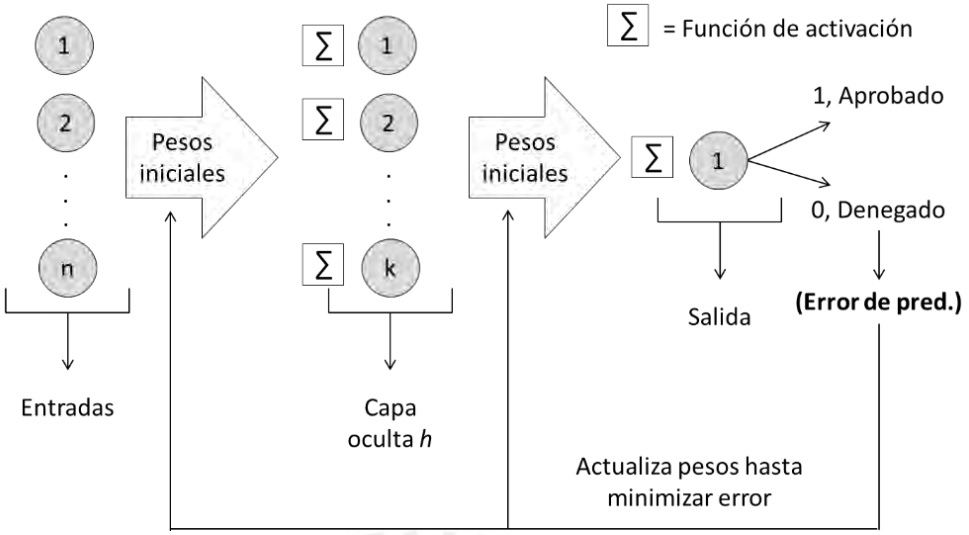


Figura 14. Funcionamiento y estructura de una red neuronal artificial.
Fuente: Berry (2004) (Adaptado)

- **Clasificador Bayesiano Simple (Naive Bayes):** Es una herramienta de clasificación caracterizada por ser uno de los modelos de más fácil aprendizaje. Usa probabilidades calculadas *a priori* con datos de entrenamiento para determinar las probabilidades de pertenencia a cada clase. El nuevo registro se clasificará en la clase que tenga una mayor probabilidad de pertenencia condicional, como resultado del Teorema de Bayes aplicado a los valores de los atributos del registro. Es de mucha utilidad cuando se dispone de una base de datos extensa.

Tabla 18. Pseudocódigo: Naive Bayes

Algoritmo 2: Naive Bayes

Supuesto: Las variables del conjunto de datos son independientes.
Variables Data Entrenamiento D{($x_{ij} \in X_j$, $X_j \in A$, $y_k \in C$)}, Vector a clasificar D'($x_{ij}' \in X_j'$,$X_j' \in A$), C {1..k}

Inicio
Para cada j en A
  Para cada i en X
    Calcular $p(x_{ij}) = n_{ij}/n$
    Para cada clase k **en** C
      Calcular $p(x_{ij}/y_k)= (n_{ij}/n)/(n_k/n)$
    Fin Para cada
  Fin Para cada
Fin Para cada
Para cada clase k en C dentro de D'
  Hacer $p(x_{ij}'/ y_k) = p(x_{ij}/y_k)$
  Calcular $p(y_k/A) = p(y_k/X_j') = \Pi\, p(y_k/x_{ij}') \propto \Pi\, (p(x_{ij}'/ y_k) * p(y_k))$ (Teorema de Bayes)
Fin Para cada
Hacer Clase(D') = $\text{argmax}_k(\Pi\, (p(x_{ij}'/ y_k) * p(y_k))$
Fin

Elaboración Propia.

- **Máquinas de Vectores de Soporte (SVM)**: Es una herramienta de clasificación basada en un método de optimización matemática. Se crea un

hiperplano que maximice la separación entre grupos de registros notablemente diferentes. La ecuación del plano óptimo que divide con la máxima separación a ambos grupos representa un problema de optimización que la mayoría de paquetes estadísticos computacionales realiza con facilidad. Permite separar la mayoría de conjuntos de datos, pero a costa de extensa capacidad computacional.

Tabla 19. Formulación de un hiperplano de separación con SVM

Algoritmo 3: Maquinas de vectores de soporte (SVM)
Variables w_i (parámetros hiperplano genérico), b, α_i (parámetros hiperplano optimo), vectores de soporte A(i). Inicio **Hallar** ecuación del hiperplano, $P = \Sigma w_i x_i$ **Aplicar kernel**[26] a la ecuación P' = kernel(P) para linearizar el hiperplano **Hallar** ecuación del hiperplano de máximo margen usando los puntos linearizados (o vectores de soporte) en la ecuación: $$P = b + \sum_{i=1}^{n} \alpha_i y_i A(i) * A$$ Fin

Fuente: González (s/f)

- **Clusterización por K-medias**: Es una metodología para dividir un grupo de instancias en términos de una "distancia" entre los registros. Cada clúster es representado por su centroide. El algoritmo clasifica a una nueva instancia al clúster cuyo centroide esté más cerca del valor del registro.

Tabla 20. Pseudocódigo: Clasificación K-medias

Algoritmo 4: K-medias

Variables k (número de clusters), D (vector de valores numéricos)
 	
Inicio
Tomar k valores de D aleatoriamente
Asignar $Grupo_k$= k
Para cada Grupo
  Calcular centroide
Fin Para cada
Para cada i **en** D-k no escogidos
  Asignar i al $Grupo_k$ más cercano vía centroide
  Calcular centroide $Grupo_k$ asignado
Fin Para cada
Fin

Fuente: Adaptado de Han (2012)
Elaboración propia.

- **Clusterización por conglomerados**: Es una extensión de la herramienta de clusterización que se utiliza cuando no se define con anterioridad la cantidad de clústers que se espera encontrar. También llamado dendrograma, realiza

[26] Un ***kernel*** es una función de transformación que permite convertir un registro de N dimensiones en uno con N+1 dimensiones, de manera que se puedan separar grupos a partir del hiperplano con mayor facilidad.

una aplicación sucesiva del algoritmo de k-medias (k=2) para determinar los clústers definidos de mayor a menor “distancia”.

Figura 15. Clusterización por conglomerados o dendrograma.
Elaboración propia.

- **Regresión**: Es una herramienta de predicción que permite estimar a partir de valores de atributos de un nuevo registro, un valor estimado de la variable objetivo. Puede utilizarse cuando se desea estimar valores objetivos de atributos continuos (regresión lineal múltiple) o cuando se desea estimar valores binarios de clasificación (regresión logística).

La segunda fase de esta etapa consiste en el **diseño de la elaboración del modelo**, en el cual se definen los mecanismos para evaluar y validar el modelo. Dependiendo del tipo de técnica utilizada (aprendizaje supervisado o no supervisado), se seleccionará un subconjunto adicional para el entrenamiento, la prueba y la validación del modelo.

La tercera fase de esta etapa consiste en la **construcción del modelo**, donde se debe ejecutar la herramienta utilizada para la construcción del modelo o modelos, identificando los parámetros utilizados en ellos e interpretando los resultados obtenidos a partir de los modelos. Asimismo, se debe documentar cualquier dificultad encontrada en esta fase.

Finalmente, en la última fase de esta etapa se debe **evaluar el modelo,** interpretando los resultados obtenidos de acuerdo a los objetivos de la minería de datos que se deseaban cumplir. Los modelos deben ser ordenados en función a medidas de precisión estadísticas y los recursos utilizados para la construcción de cada modelo, de manera que se identifique de forma definitiva al mejor modelo.

1.2.4.5. Evaluación del proyecto

La primera fase de esta etapa consiste en la **evaluación de resultados.** Aquí se evalúa la medida en que la precisión y la exactitud de las predicciones o clasificaciones cumplen con los objetivos propuestos de la minería de datos y guardan relación con los objetivos del negocio. Además, se realiza la aplicación del modelo en un sistema real, para determinar su desempeño con relación a las restricciones de tecnología, tiempo y presupuesto. No todos los resultados deben guardar relación con el negocio, sin embargo, son fuente adicional de conocimiento. Por lo tanto, en esta fase se declara si el modelo es aprobado o no.

La segunda fase de esta etapa consiste en la **revisión general del proceso.** Aquí se debe reflexionar acerca de lo hecho en el proceso de Minería de Datos y preguntarse si algún factor importante no ha sido tomado en cuenta en el análisis. Por ejemplo, analizar si las variables que se encuentran en el modelo final estarán disponibles para futuros análisis.

La última fase de esta etapa consiste en la **decisión de implementación.** Según los resultados del proceso de Minería de Datos, se aprobará la implementación del proyecto en los sistemas de información actuales, o se deberán realizar algunos ajustes a los resultados del proceso asociadas a las restricciones de presupuesto actual.

1.2.4.6. Implementación del modelo

Si la implementación del modelo ha sido aprobada, se desarrollará esta etapa.

La primera fase de esta etapa consiste en la elaboración de un **plan de implementación.** El equipo de proyecto -junto con los responsables del negocio- desarrolla una estrategia para la implementación del modelo. Si se identifica un procedimiento general para crear el modelo, este procedimiento será documentado en el plan de implementación.

La segunda fase de esta etapa consiste en el **mantenimiento y monitoreo continuo** de los cambios implementados. Como en todo proyecto, un plan de monitoreo es vital para evitar periodos innecesarios de uso incorrecto de los resultados de la minería de datos. Así, se debe documentar el plan de monitoreo y mantenimiento en relación a la implementación del nuevo modelo, incluyendo la estrategia a adoptar y los pasos necesarios para ejecutar la misma.

Finalmente, la última fase de esta etapa (y de todo el proceso) es la elaboración de un **reporte final** de proyecto. Se redactará un informe final que describa un resumen del proyecto y de los resultados obtenidos. Así mismo, se preparará una presentación final de los resultados de la aplicación de la minería de datos. Por último, se resumirán descubrimientos importantes obtenidos durante el proyecto, así como también los errores y aciertos en búsqueda de la mejor aplicación de las herramientas de la minería de datos, en la parte de recomendaciones.

CAPÍTULO 2. ESTUDIO DE CASOS

En la actualidad, el uso de la minería de Datos es sumamente importante para optimizar la gestión del riesgo crediticio, puesto que una mejor precisión en la predicción de la probabilidad de *default* (o impago) de un cliente en particular (MYPEs peruanas) permite una mejora en la asignación crediticia, tanto en plazos como en montos involucrados en los préstamos. A continuación, se detallarán cuatro casos:

- El primero hace hincapié en las ventajas de una modelación hibrida en un modelo de clasificación crediticia.
- El segundo prioriza la determinación de variables que son significativas en la probabilidad de pertenecer al sector informal, en una ciudad determinada.
- El tercero realiza un estudio de la evaluación crediticia o *credit scoring* gestionada a partir de la tecnología *blockchain* actualmente utilizada para las transacciones financieras en línea, como ventaja sustancial para la consolidación de información en el sistema bancario y la creación de modelos de mucha mayor efectividad y potencial.
- El último realiza una comparativa de modelos de regresión logísticos versus modelos de redes neuronales artificiales y la eficacia de cada modelo a nivel global y a nivel especifico, buscando aquel que minimice la posibilidad de captar clientes que sean malos pagadores.

2.1. Caso 1: Modelo híbrido de selección de variables y clasificadores de autoaprendizaje para el *creditscoring*, en un banco hindú.

2.1.1. Descripción del problema

El caso de estudio se centra en una institución financiera hindú, que busca mejorar la gestión del riesgo crediticio de sus clientes a partir de una mejor oferta de servicios diferenciados para cada tipo de cliente. Con el fin de lograr un mejor control del riesgo, los bancos hacen uso del *creditscoring*. Sin embargo, en la mayoría de los casos no existe una comparación de modelos de *creditscoring* a utilizar y se selecciona un modelo acorde con la experiencia de los analistas de riesgo. Por tanto, el banco en cuestión busca seleccionar el modelo más preciso dentro de los diferentes modelos propuestos para mejorar la gestión del riesgo, en particular, para mejorar la asignación de créditos.

2.1.2. Metodología

Con el objetivo de dar una alternativa a este problema, Koutanaei *et al* (2015) desarrolla un análisis de los modelos propuestos para la asignación siguiendo el flujograma que se muestra a continuación:

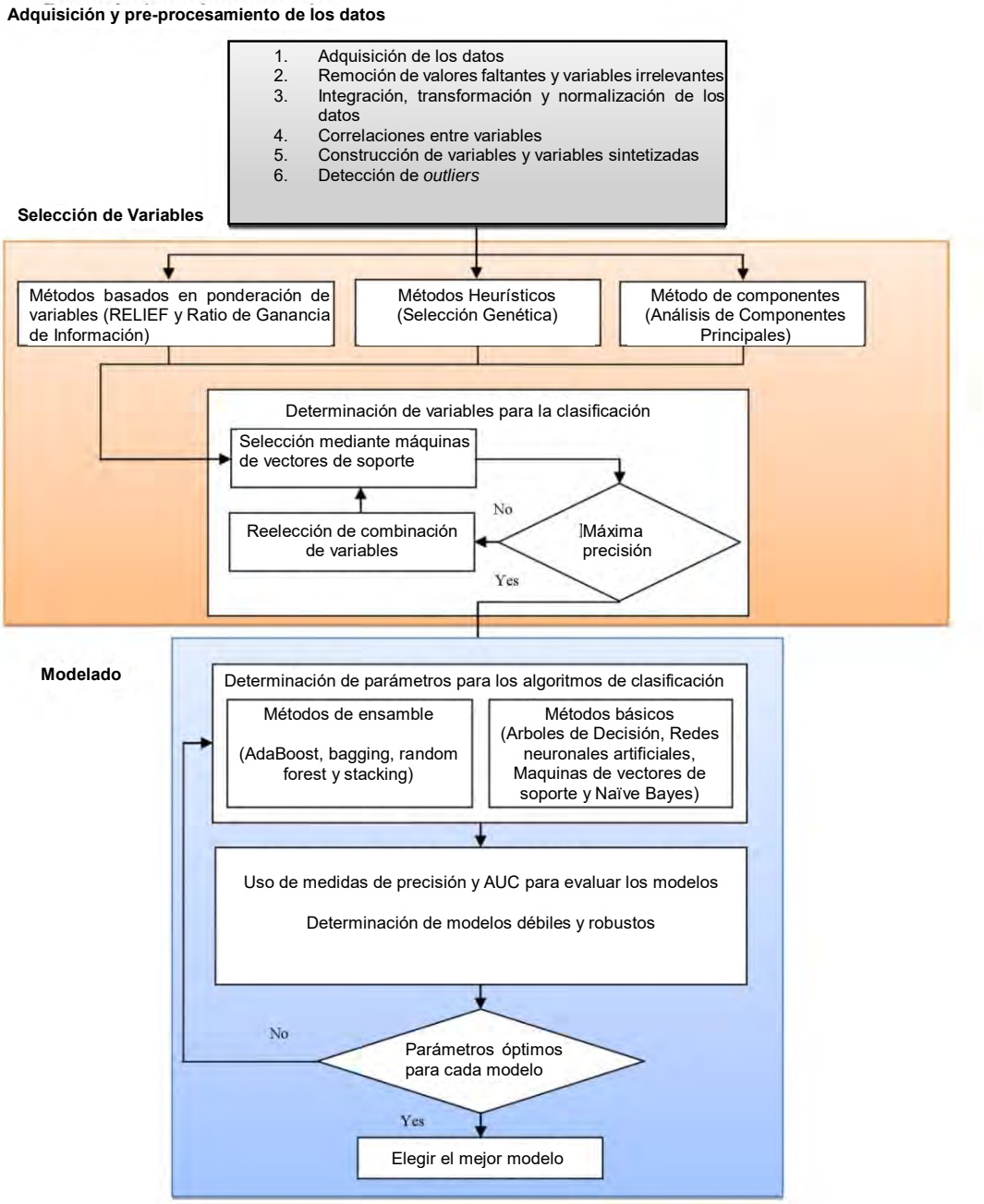


Figura 16. Diagrama de bloques de la generación de modelos propuestos.
Fuente: Adaptado de Koutanaei et al (2015)

Como se puede apreciar, este flujograma guarda una estrecha relación con la metodología de proyectos de minería de datos conocido como CRISP-DM.

2.1.3. Aplicación de la minería de datos

Se inició con la recopilación de la información y el preprocesamiento de los datos. Posteriormente, se realizó una selección de variables a partir del aporte de la inclusión de cada variable de acuerdo a los algoritmos siguientes:

Tabla 21. Algoritmos de selección de variables.

Método	Algoritmos por tipo de selección	Descripción Breve
Métodos heurísticos	Selección Genética	Tomar y combinar sets de variables aleatoriamente hasta encontrar el set de mejor acercamiento.
Método por pesos ponderados	RELIEF	Únicamente para objetivos binarios. Permite realizar un ranking de las variables que más poder de clasificación tienen, usando un efecto que pondera la cercanía entre 2 registros de una misma clase y 1 de una clase y el otro de la otra clase, para mismos atributos.
	Ratio de Ganancia de Información	Consiste en construir un árbol de clasificación donde los atributos que salen de las ramas van de mayor a menor ratio de ganancia de información. Define los atributos relevantes que pueden representar con gran precisión a un subgrupo de la información.
Método de componentes	Análisis de Componentes Principales	Modelo estadístico que determina las variables relevantes del modelo a través de un hiperplano de ajuste que minimice los errores cuadráticos obtenidos por el ajuste. Los componentes principales se obtienen a partir de una composición lineal de variables correlacionadas y no correlacionadas.

Fuente: Koutanaei et al. (2015)

Los parámetros principales obtenidos a partir de estos métodos son contrastados usando SVM para encontrar la mejor combinación de variables para cada método, en términos de la precisión obtenida. Posteriormente, en la etapa de modelado, se utilizaron métodos de clasificación básicos (arboles de clasificación, RNA, SVM, Naive Bayes) y métodos de clasificación por ensamble, que significa una extensión de la clasificación de un registro a partir de diferentes modelos de clasificación analizados conjuntamente. Estos métodos se muestran a continuación:

Tabla 22. Métodos de ensamble.

Método	Descripción Breve
Bagging	En términos simples, genera “m” modelos **de un mismo tipo** a partir de diferentes sets de datos aleatorios, se prueban los “m” modelos y se pondera el resultado de cada uno para mejorar la precisión de la clasificación.
Adaptive Boosting	Desarrollado por Freund y Shapire (1996), construye un modelo de clasificación “apilado” en función a un conjunto de clasificadores base. De manera secuencial, toma una muestra aleatoria de la base de datos disponible y calcula el nivel de error presente al adicionar en el modelo un clasificador base, medido como: $E(f) = e^{-y(x)f(x)}$ La ponderación del peso de cada clasificador base estará en función de la reducción del error lograda.
Stacking	En términos simples, es la predicción de la clasificación a partir de la predicción de modelos **de diferente tipo**, pero con el mismo objetivo.
Random forest	En términos simples, es un algoritmo que genera múltiples árboles de decisión con “k” variables para cada uno (escogidas aleatoriamente) y pondera los resultados de todos los árboles para dar una decisión más precisa. En problemas de clasificación, brinda como clase predicha la que se repite más veces entre todos los árboles.

Fuente: Koutanaei et al. (2015)

2.1.4. Solución del problema y resultados obtenidos

Se obtuvieron los resultados concernientes a los algoritmos simples de clasificación y a los métodos de ensamble, que se observan en la figura siguiente como medida porcentual de la precisión y del AUC:

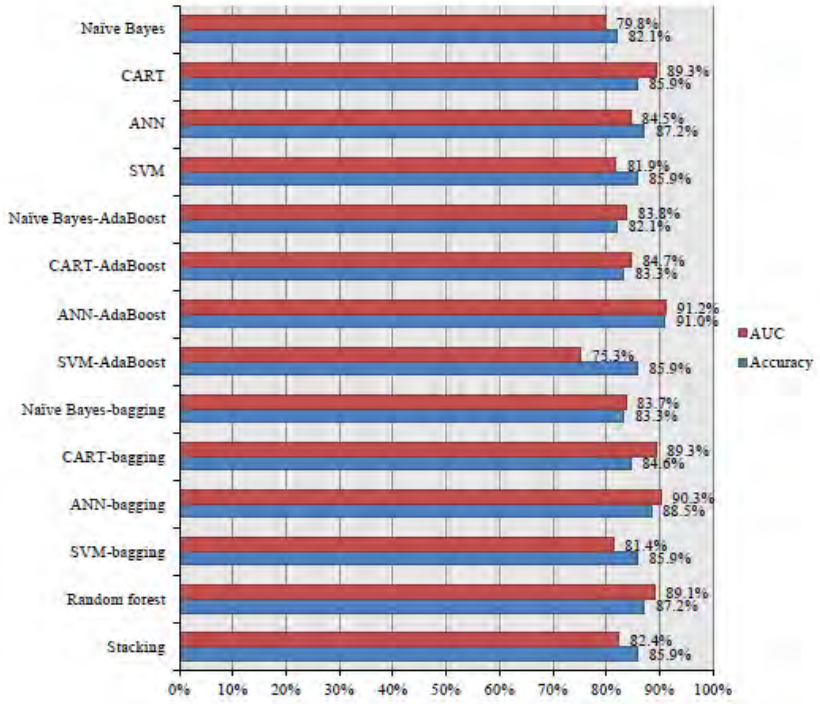


Figura 17. Comparación de eficiencia de algoritmos de clasificación.
Fuente: Koutanaei et al (2015)
Elaboración propia.

En general, el modelo de clasificación basado en RNA-AdaBoost fue el que mejor desempeño logro en términos de los indicadores propuestos anteriormente, mientras que los que lograron los peores desempeños fueron los algoritmos SVM-AdaBoost y Naive Bayes. Por lo tanto, queda demostrado que un modelo hibrido de clasificación logra mejores resultados que modelos analizados por separado.

2.1.5. Conclusiones

El uso de modelos de clasificación híbridos, a partir de algoritmos de ensamble, es un modelo que en definitiva puede ser usado para problemas que involucren la creación de modelos de *creditscoring*, pues no solo logra una mayor precisión y un alto nivel de confianza, sino que además reduce el sesgo producido por los supuestos de algunos modelos de clasificación.

2.2. Caso 2: Determinación de los niveles y factores de informalidad para MYPEs usando regresión logística, en Lahore, Pakistán.

2.2.1. Descripción del problema

El caso de estudio se centra en el análisis de los motivos por los cuales los emprendedores o micro-empresarios tienen la necesidad de pertenecer al sector de informalidad ya sea voluntariamente o de forma obligada. Contrario a muchos otros estudios acerca de la situación de las micro-empresas en la informalidad, esta investigación analiza la hipótesis de que efectivamente, existen diferentes grados de informalidad y prueba esta hipótesis determinando aquellos factores que determinan su nivel de formalidad de manera significativa y su pertenencia a un cierto nivel de informalidad.

2.2.2. Metodología

Se condujo una encuesta cara-a-cara a 300 micro empresarios ubicados en la ciudad de Lahore entre Octubre del 2012 a Enero del 2013, utilizando un muestreo de variación máxima[27] en lugar de un muestreo aleatorio simple para la selección de participantes, con el fin de evitar a la población inaccesible. Se hizo una partición estratificada en 7 zonas según el nivel de ingreso promedio de la localidad: alto nivel de ingreso, mediano nivel de ingreso y bajo nivel de ingreso. Dentro de cada zona, se realizó un muestreo estratificado para seleccionar la cantidad de participantes a escoger por cada localidad de la zona.

Las variables que se consideraron para el estudio se establecieron de mayor grado de generalidad de la informalidad hasta los motivos por los cuales son informales. De acuerdo a Hussmans (2005), para construir un índice de informalidad se debe conocer: (1) el estatus legal de la microempresa, (2) su estado de contribución de impuestos y (3) si mantienen un registro de sus actividades en sus libros contables.

Se analizaron también las características que presentan los microempresarios por nivel de informalidad.

[27] Técnica de muestreo no probabilística que busca capturar una amplia gama de perspectivas acerca de una condición de interés. Las tendencias de la población estudiada estarán incluidas dentro de la amplia gama de perspectivas analizadas (tanto los extremos como las tendencias centrales). Sin embargo, esta técnica solo sirve de manera descriptiva.

Tabla 23. Niveles de informalidad establecidos a partir del índice de formalidad.

Nivel de informalidad	Estatus legal	Registrado como contribuyente	Cuentas contables formales	% por opción	Clasificación	% por clasificación
Totalmente formal	Bueno	Sí	Sí	6.7	3	6.7
Bajo nivel de informalidad					2	30.2
Opción 1	Malo	Sí	Sí	0		
Opción 2	Bueno	Sí	No	30.2		
Opción 3	Bueno	No	Sí	0		
Alto nivel de informalidad					1	33.6
Opción 1	Malo	No	Sí	0		
Opción 2	Malo	Sí	No	1.0		
Opción 3	Bueno	No	No	32.6		
Totalmente informal	Malo	No	No	29.5	0	29.5

Fuente: Williams et al. (2015)

Tabla 24. Características de los microempresarios por nivel de informalidad, en %.

Nivel de informalidad	Totalmente informal	Alto nivel de informalidad	Bajo nivel de informalidad	Totalmente formal
De todos los encuestados	29.5	33.6	30.2	6.7
Encuestadas mujeres	1.2	1.0	0.0	10.0
Edad				
15-24	13.8	8.1	3.4	0.0
25-35	48.3	55.6	39.3	55.0
36-64	36.8	35.4	55.1	35.0
65+	1.2	1.0	2.3	10.0
Nivel educativo culminado				
Sin educación	26.1	17.0	5.6	0.0
Primaria	30.7	31.0	24.4	5.0
Secundaria	34.1	40.0	45.6	30.0
Diploma	2.3	6.0	16.7	25.0
Universidad	6.8	6.0	7.8	40.0
Ingreso neto				
Es fuente principal de ingresos	80.5	76.5	79.8	65.0
< 20,000	54.6	21.0	4.4	5.0
20,000 - 29,999	28.4	38.0	33.3	15.0
30,000 - 39,999	4.6	11.0	33.3	30.0
40,000 - 49,999	3.4	13.0	11.1	15.0
> 50,000	5.7	9	14.4	35.0
No reporta	3.4	8	3.3	0.0
No es fuente principal de ingresos	19.5	23.5	20.2	35.0
Familiares que aportan al ingreso				
Ninguno	53.5	49.5	52.8	21.1
Uno	18.6	17.2	13.5	21.1
Dos	17.4	17.2	18.0	21.1
Tres a más	10.5	16.2	15.7	36.8
Sector				
Retail	31.8	22.2	40.5	70.0
Manufactura	29.7	31.0	41.6	25.0
Comida instantánea	38.5	47.0	18.0	5.0

Fuente: Williams et al. (2015)

Por último, se estudiaron también los principales motivos por los cuales los microempresarios operan en la informalidad, mostrados en la tabla siguiente:

Tabla 25. Motivos principales para operar en la informalidad, por nivel de informalidad, en %.

Nivel de informalidad	Totalmente informal	Alto nivel de informalidad	Bajo nivel de informalidad	Totalmente formal
Factores de exclusión				
No encontró un trabajo regular	53.4	32.0	22.2	5.0
Necesitaba un ingreso adicional	5.7	7.0	4.4	10.0
Razones racionales				
Regulaciones poco claras	20.5	28.0	11.1	0.0
Impuestos altos	22.7	29.0	4.4	0.0
Corrupción del sector público	21.6	32.0	11.1	0.0
Resentimiento	26.1	43.0	22.2	5.0
Falta de interés y conocimiento	39.8	26.0	17.8	0.0
Entorno institucional				
Ser informal es muy riesgoso	3.6	1.0	13.3	20.0
Es normal operar siendo informal	44.3	37.0	21.1	20.0

Fuente: Williams et al. (2015)

Con esta información, y junto con otras características de la microempresa, se formularon modelos de regresión logística para determinar los factores que influyen en la determinación de los niveles de informalidad.

2.2.3. Aplicación de la minería de Datos

Se establecieron cuatro modelos de regresión logística a partir de: (1) las características de los empresarios informales, (2) las características de la microempresa, (3) las razones racionales para operar en la informalidad y (4) el impacto del entorno institucional formal e informal en la región paquistaní. Para cada factor, se tienen los siguientes sub-niveles:

- **Modelo 1:** Incorpora las características de los empresarios informales:
    - Sexo (Masculino y Femenino)
    - Edad (15-24,25-39,40-64 y mayores de 65 años)
    - Nivel educativo culminado (sin educación, educación primaria, secundaria, diploma o universidad)
    - Ingresos netos en términos de la moneda local (menor a 20000, 20000-29999,30000-39999,40000-49999, mayor a 50000, no reportados o el empresario no representa la principal fuente de ingresos)
    - Ingresos por otros familiares si es que el empresario no representa la principal fuente de ingresos (por uno, dos, 3 o más familiares)
    - Sector económico (Retail, Manufactura o Comida instantánea)

- **Modelo 2:** Incorpora las variables del modelo 1 y las características de la microempresa:
    - Contabilidad propia (si el empresario lleva las cuentas contables o si lo hacen trabajadores contratados para ello)
    - Sector económico (Retail, Manufactura o Comida instantánea)
    - Local de negocio (si el local de negocio ya existía o no)
    - Emprendimiento del negocio (si el empresario comenzó el negocio solo, con ayuda de otros participantes o con ayuda de familiares)
    - Cuenta bancaria (si el empresario cuenta con una cuenta bancaria asociada o no)
    - Tamaño del negocio (uno, dos, tres, cuatro, cinco a diez trabajadores)
    - Fuentes de financiamiento (por medio de parientes o familiares, amigos o vecinos, de forma propia, con crédito de proveedores o con pagos adelantados de los clientes).
    - Historial de préstamos (si el empresario reporta haber aplicado para un préstamo bancario o no)
    - Trayectoria de la empresa (en años, y en años al cuadrado, indicando una medida del tiempo que perdura la empresa desde su creación).
    - Factores que llevan a iniciar en la informalidad (no poder conseguir trabajo o necesitar un ingreso adicional)
    - Actitudes de emprendimiento (no es más rentable que un trabajo regular, no es más rentable pero prefiero ser mi propio jefe, es más rentable pero no prefiero ser mi propio jefe, es más rentable y prefiero ser mi propio jefe)
- **Modelo 3:** Incorpora las variables del modelo 2 y los principales razones racionales por las que el empresario opera informalmente.
    - Resentimiento (afirmativa si la respuesta es que “es informal porque el estado no hace nada por la gente, así que para que seguir la ley”, negativa en caso contrario)

- Impuestos altos (afirmativa si la respuesta es que “es informal porque los impuestos a pagar son muy altos”, negativa en caso contrario)
- Regulaciones poco claras (afirmativa si la respuesta es que “es informal porque el proceso de “formalidad” es muy complicado”, negativa en caso contrario)
- Corrupción del sector público (afirmativa si la respuesta es que “es informal porque el sistema formal es corrupto”, negativa en caso contrario)
- Falta de interés y conocimiento (afirmativa si la respuesta es que “es informal porque no sé si debo registrarme”, negativa en caso contrario)

- **Modelo 4:** Incorpora las variables del modelo 3 y el impacto del entorno institucional formal e informal en la región paquistaní:
  - Ética sobre la informalidad (es muy aceptable ser informal en Pakistán, es poco aceptable ser informal en Pakistán y no es para nada aceptable ser informal en Pakistán)
  - Riesgo de ser informal (es muy riesgoso ser informal en Pakistán, es poco riesgoso ser informal en Pakistán y no es para nada riesgoso ser informal en Pakistán)

Los resultados estadísticos de los cuatro modelos se presentan en la sección siguiente.

2.2.4. Solución del problema y resultados obtenidos

Realizados los modelos anteriores, el análisis de regresión logística concluye que el modelo 4 es que mayor explicación realiza sobre las características de la informalidad empresarial, producto de la inclusión de los variables de informalidad del entorno que influyen fuertemente en la probabilidad de que una empresa pertenezca a un determinado nivel de informalidad. Asimismo, se tiene un reporte del peso relativo de cada variable en el mejor modelo (modelo 4) haciendo uso de los *odds ratio* correspondientes.

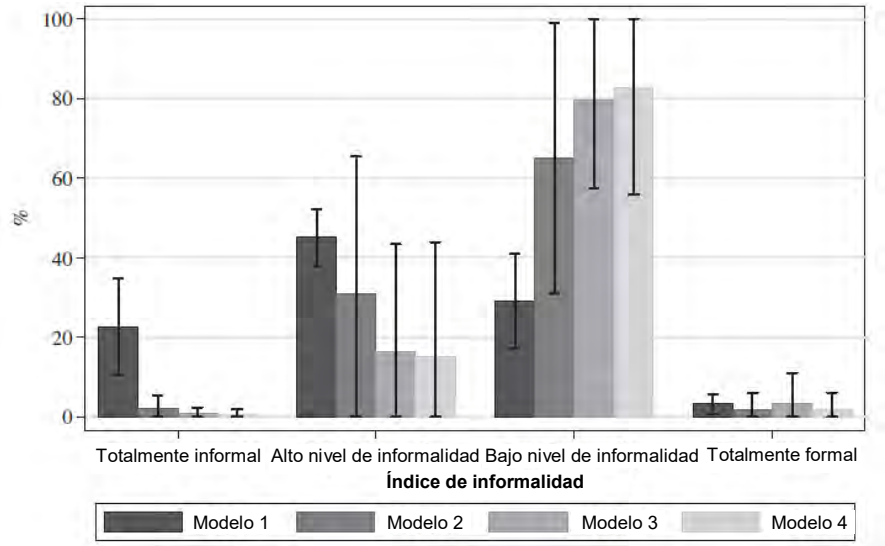


Figura 18. Comparación de probabilidades de pertenencia a la informalidad según modelos e índice de informalidad.
Fuente: Williams et al. (2015)

Tabla 26. *Odds ratio* de variables significativas del modelo 4.

Nivel de informalidad	Totalmente informal	Alto nivel de informalidad	Bajo nivel de informalidad	Totalmente formal
Características de los empresarios				
Encuestadas mujeres				
Edad (15-24)				
25-35	- 7.4	- 11.5	18.8	0.1
40-64				
65+	-9.1	- 66.7	- 41.1	34.7
Nivel educativo (Sin educación)				
Primaria	- 5.5	- 11.9	17.3	0.1
Secundaria	- 6.1	- 11.3	17.4	0.1
Diploma	- 8.3	- 38.6	46.3	0.6
Universidad	- 8.0	- 35.5	43.0	0.5
Ingreso neto (< 20,000)				
20,000 - 29,999	- 9.9	- 22.9	32.6	0.2
30,000 - 39,999	- 9.5	- 34.4	43.4	0.5
40,000 - 49,999	- 5.6	- 17.6	23.0	0.2
> 50,000	- 6.6	- 22.8	29.1	0.2
Características de la empresa				
Contabilidad propia	51.2	10.7	- 61.4	- 0.5
Sector económico (Retail)				
Manufactura	12.5	10.3	- 22.7	- 0.1
Local de empresa es nuevo	8.6	5.7	- 14.2	- 0.0
Posee cuenta en banco	- 14.8	- 55.0	68.3	1.6
Financiamiento (familiares)				
Amigos/Vecinos	- 7.2	- 41.1	47.5	0.7
Propia	- 12.0	-7.5	19.4	0.0
Crédito de proveedores	- 9.0	- 45.8	53.9	0.9
Trayectoria	- 1.0	- 1.5	2.5	0.0
Trayectoria (cuad.)	0.0	0.0	- 0.0	- 0.0
Razones para operar de informal				
Regulaciones poco claras	8.4	6.3	- 14.6	- 0.00
Corrupción del sector público	5.6	32.0	- 11.2	- 0.0
Falta de interés y conocimiento	6.1	6.3	- 12.3	- 0.0

Entorno institucional				
Riesgo de ser informal (alto)				
Riesgo de ser informal (bajo)	11.7	20.7	- 32.3	- 0.2
Es bueno ser informal				
Es algo bueno ser informal	- 10.5	- 14.3	24.6	0.1
No es bueno ser informal	- 6.9	- 22.3	29.0	0.2
Observaciones		259		
χ^2		203.13		
Prob > χ^2		0.00		
Pseudo R^2		0.49		

Fuente: Williams et al. (2015)

2.2.5. Conclusiones

El uso de la regresión logística como herramienta de la minería de datos permite establecer de manera categórica y con sustento estadístico que existen diferentes clasificaciones o niveles de informalidad asociados al sector denominado como informal. Por tanto, es erróneo clasificar únicamente a las empresas como formales e informales, sino debe describirse a cada una de las empresas con mayor profundidad. Por otro lado, se han revelado las principales características de los empresarios informales que están fuertemente relacionados con estos niveles de informalidad y a su vez, los factores que están relacionados con el nivel de formalidad de estas microempresas. Esto es útil para impulsar campañas de formalidad –ya sea por los agentes financieros o por el gobierno nacional – que tengan como objetivo mejorar las condiciones para que una MYPE en ese país pueda ser formal.

2.3. Caso 3: Un estudio de la evaluación crediticia de las PYMEs basada en la tecnología Blockchain

2.3.1. Descripción del problema

El caso de estudio se centra en el sustento de las ventajas que el uso de la tecnología *blockchain* – actualmente usada como una cadena de transacciones financieras digitales 100% segura con el uso de procesos criptográficos– puede brindar aplicándose al proceso de consolidación de data bancaria (transacciones a nivel bancario, principalmente) una forma confidencial sin poner en riesgo su integridad, a prueba de todo tipo de manipulaciones y sobre todo con la capacidad de hacerle seguimiento de manera privada y segura. De esa manera, la creación de modelos de evaluación de perfiles crediticios logra tener un mejor potencial puesto que los perfiles de un mismo cliente consiguen ser únicos a nivel de todo el sistema bancario y el estudio sustenta esta afirmación con el desarrollo de un modelo de regresión logística con la creación de una base de datos obtenida de múltiples entidades bancarias.

2.3.2. Metodología

Se tomaron 150 copias de muestras de data a través de entrevistas directas a las empresas de la clasificación mencionada líneas arriba. Se definió como clientes buenos o malos de acuerdo a (1) su capacidad para abonar los sueldos de los trabajadores, los préstamos solicitados a clientes y otras deudas y (2) su capacidad para cumplir en las condiciones de pago establecidas en sus contratos. De esta manera, la data se dividió en 90 casos buenos y 60 casos malos.

Se seleccionaron 38 variables características iniciales, de las cuales 4 correspondían a características relativas a la empresa, 16 a características relativas al estado financiero de la empresa y 18 a características propias del dueño de la compañía, a saber, que en pequeñas empresas la capacidad de pago suele estar fuertemente ligada a la capacidad crediticia del dueño de la empresa.

Para evitar repetición de información, se hizo un análisis de correlación entre las variables con un punto de quiebre de 0.65, eliminando 6 variables.

2.3.3. Aplicación de la minería de Datos

Se realizó un pre procesamiento a las variables resultantes previo a la aplicación del modelo. Se utilizó un modelo de regresión logística con optimización hacia adelante y hacia atrás, para finalmente al modelo optimizado incluir variables basadas en decisión experta, logrando una mejor explicación de los resultados del modelo.

2.3.4. Solución del problema y resultados obtenidos

Las variables finales resultantes, los indicadores de precisión y la curva OC se muestran a continuación:

Tabla 27. Variables seleccionadas del modelo final.

Variables	Coeficiente	S.E	G. Libertad	Significancia	Exp(B)
ind	-0.68	0.53	1	0.2	0.507
cr	0.051	0.193	1	0.794	1.052
dtar	-0.089	0.132	1	0.501	0.915
roa1	7.546	2.657	1	0.005	1893.15
roe	1.022	0.738	1	0.166	2.779
cat	0.12	0.14	1	0.391	1.127
ta	-0.122	0.15	1	0.418	0.885
crae	-0.099	0.063	1	0.116	0.906
db1	-0.087	0.044	1	0.049	0.917
dn1	0.00001	0	1	0.007	1.000
hy	-0.41	0.149	1	0.006	0.664
cren1	-0.069	0.092	1	0.449	0.933
Constante	2.879	1.757	1	0.101	17.796

Fuente: Zhang et al. (2019)

Tabla 28. Indicadores de precisión del modelo, valor de corte 0.5

Modelo	-2*log	R^2	X_2	G. Libertad	Significancia
1	43.086	0.681	1.211	8	0.997

Fuente: Zhang et al. (2019)

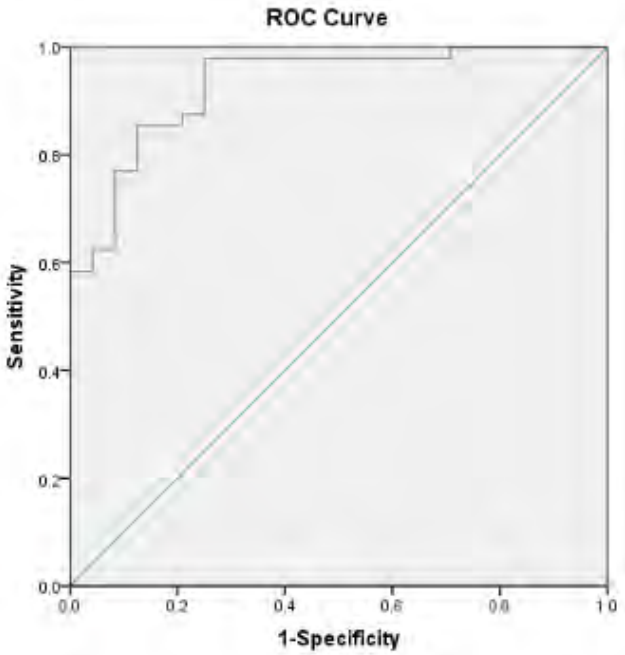


Figura 19. Curva OC del modelo final
Fuente: Zhang et al. (2019)

2.3.5. Conclusiones

El estudio se basa en la consolidación de la reportaría de informes de evaluación de crédito de múltiples bancos a múltiples empresas. Siguiendo este supuesto, los bancos comerciales y las instituciones financieras crearían una base de datos compartida para incluir los registros de préstamos y el estado crediticio de los propietarios, lo que garantizaría la fiabilidad e integridad de las fuentes de datos y mejoraría el sistema de evaluación crediticio. Sobre esta base, el modelo logístico modificado tiene una gran mejora en precisión y usabilidad. Basados en este supuesto, el modelo logra buenos resultados en el objetivo principal, es decir, identificar clientes buenos y malos y predicciones. Sin embargo, el modelo también tiene algunos problemas. En la realidad, la participación universal y el intercambio de datos no pueden realizarse en su totalidad. En el futuro, se podrían considerar más plataformas de datos para expandir las fuentes de datos en todos los aspectos, y se puede establecer un sistema de intercambio de datos basado en *blockchain* para expandir los escenarios de aplicación del modelo de crédito.

2.4. Caso 4: Modelos de manejo del riesgo de crédito bancario realizados por bancos comerciales de Jordania con el uso de redes neuronales

2.4.1. Descripción del problema

El caso de estudio se centra en el rechazo generalizado de las instituciones bancarias de Jordania para con el uso de herramientas de minería de datos como soporte a la toma de decisiones relativas a evaluaciones crediticias. El estudio busca como respuesta a ello, proponer dos modelos de evaluación de riesgo crediticio basadas en dichas herramientas que tengan como objetivo mejorar la gestión del riesgo crediticio, así como también reducir el tiempo invertido en el análisis manual y el costo de dicha gestión. Los modelos utilizados en este estudio fueron los modelos de regresión logística y el modelo de redes neuronales artificiales, cuyos resultados serán comparados para indicar cuál de estos modelos tiene mayor efectividad para identificar a los clientes con potencial de caer en *default*.

2.4.2. Metodología

La data colectada consistió en 492 casos obtenidos como muestras de tanto solicitudes de crédito aceptadas, como denegadas, de diferentes bancos comerciales de Jordania, guardando para cada banco las confidencialidades del caso de forma que no se vea afectada dicha entidad. De esta muestra, 292 (59.3%) aplicaciones fueron aceptadas y 200 (40.3%) fueron rechazadas.

El número total de variables recopiladas fueron 13; 7 de ellas fueron variables de escala, mientras que el complemento fueron variables categóricas.

2.4.3. Aplicación de la minería de Datos

Se realizó un pre procesamiento a las variables recopiladas, transformando las variables categóricas en valores numéricos de forma que pueda ser utilizada tanto por el modelo logístico como por el modelo de redes neuronal artificiales. Las variables de escala fueron estandarizadas utilizando la opción del programa estadístico utilizado (SPSS) para el desarrollo del análisis del presente estudio. La variable dependiente también fue codificada, indicando como 1 cuando la solicitud de crédito ha sido aprobada, y 0 cuando haya sido rechazada.

Como se mencionó anteriormente, se realizaron 2 modelos: el modelo de redes neuronales RBF, donde los errores de cada iteración permiten el ajuste del modelo,

hacia delante; y el modelo de regresión logística con optimización del modelo hacia adelante en cada paso. Para ambos modelos, se consideraron 440 casos seleccionados aleatoriamente para entrenamiento y 52 casos para prueba.

2.4.4. Solución del problema y resultados obtenidos

Los resultados para el modelo de regresión logística con optimización hacia adelante confirman que solo 7 variables fueron significativas para la construcción del modelo, Las variables finales se muestran a continuación:

Tabla 29. Variables seleccionadas del modelo final, regresión logística

Variables	Coeficiente	S.E	G. Libertad	Significancia	Exp(B)
Proposito de crédito					
Proposito de crédito 1, LP1	-1.89	0.759	1	0.013	0.15
Proposito de crédito 1, LP2	-1.21	0.544	1	0.028	0.30
Tipo de compañía, CT	-3.54	1.143	1	0.002	0.03
Garante, GU	-2.58	0.081	1	0.003	0.08
Ratio de deuda	-0.968	1.028	1	0	0.38
Duracion en meses	-0.01	0.004	1	0.014	0.99
Tasa de interés	-33.66	11.907	1	0.005	0.00
Ingresos totales	-1.28	0.546	1	0.019	0.28
Constante	14.47	2.563	1	0	1,924,159.87

Fuente: Becket, H y Kamel, S. (2014)

Para el modelo de redes neuronales artificiales como es de esperar se han considerado las 12 variables, luego de las iteraciones correspondientes se obtuvieron los pesos que minimizan la suma de los errores, y estos pesos son tabulados en una tabla con una medida de importancia estadística, calculada a partir del cambio relativo de la variable independiente versus el resultado obtenido en la variable dependiente u objetivo. La tabla descrita se muestra a continuación:

Tabla 30. Variables seleccionadas del modelo final, red neuronal artificial

Variables	Importancia	Importancia normalizada
Genero, G	4.3	28.3
Propósito de crédito, LP	4.5	29.6
Tipo de compañía, CT	5.2	33.7
Garante, GU	4.2	27.6
Nacionalidad, N	4.6	30.0
Edad, A	11.1	72.8
Ratio de deuda, DPR	15.3	100.0
Monto del préstamo, LA	11.8	77.3
Ingresos totales, TI	10.6	69.1
Periodo en meses con actual empleador, PE	9.3	61.0
Duración en meses, DM	10.9	71.1
Tasa de interés, IR	8.3	54.3

Fuente: Becket, H y Kamel, S. (2014)

Para comparar la efectividad de ambos modelos, se ha resumido de acuerdo a los errores tipo I y tipo II, tanto en el subconjunto de datos de prueba y el de entrenamiento, de acuerdo a la siguiente tabla:

Tabla 31. Comparación de eficiencia y precisión de los modelos de RL y RN

Modelo	Muestra	% Clasificación	% Precisión total	Error Tipo I (%)	Error Tipo II (%)
LR	Entrenamiento	84.8	85.4	11.6	20.4
	Prueba	90.4		3	21.1
RBF	Entrenamiento	80.9	81.5	23.9	12.2
	Prueba	86.5		12.1	15.8

Fuente: Becket, H y Kamel, S. (2014)

Como se puede apreciar, si bien el modelo de redes neuronales posee una menor precisión general, tiene mayor potencia para detectar a aquellos que tengan una tendencia a caer en *default*, pues tiene un menor error tipo II.

2.4.5. Conclusiones

Ambos modelos demuestran brindar poderosos resultados para el mismo objetivo que es mejorar la gestión del riesgo de crédito. A pesar que el modelo de regresión logística tuvo un mejor performance a nivel general, el modelo de regresión lineal es más poderoso en descubrir potenciales malos pagadores. Sin embargo, la decisión final de cual modelo utilizar la define siempre la institución financiera que desarrolle estos modelos, ponderando los costos estimados de perder a potenciales buenos pagadores versus aceptar a potenciales malos pagadores y la decisión final, a pesar del rigor matemático, esta correlacionada a una evaluación de estos posibles escenarios.

CAPÍTULO 3. DIAGNÓSTICO DE LA SITUACIÓN ACTUAL

3.1. Características generales de las MYPEs en el Perú

De acuerdo a Ley N° 28015 – 2003 de la Promoción y Formalización de la MYPE, La Micro y Pequeña Empresa es la unidad económica constituida por una persona natural o jurídica, bajo cualquier forma de organización o gestión empresarial contemplada en la legislación vigente, que tiene como objeto desarrollar actividades de extracción, transformación, producción, comercialización de bienes o prestación de servicios. En nuestro país, el desarrollo económico proviene en gran parte de estas, ya que representan más del 99% de las unidades productivas formales (Muñoz, 2013), así mismo generan aproximadamente el 40% del PBI nacional (Alzamora, 2015). De esta manera, se puede colocar en evidencia el rol que desempeñan las MYPEs tanto a nivel económico como institucional, ya que son las unidades que mayor generación de empleo a nivel de América Latina (OIT, 2015); sin embargo, también presentan deficiencias que limitan el crecimiento de estas, tales como la informalidad, el bajo nivel de productividad, la dificultad de acceso al crédito, entre otros obstáculos.

Para definir a las MYPEs en el contexto peruano al cual se refiere esta tesis, se brindará dos enfoques acerca de sus características, es decir, desde la perspectiva del TUO de la Ley de Impulso al Desarrollo Productivo y al Crecimiento Empresarial [28] y la Normativa SBS[29].

Tabla 32. Definición de MYPE de acuerdo a Produce y la SBS.

Tipo de empresa	Produce	SBS[30]
Microempresa	Ventas anuales hasta por un monto de 150 UIT[31]	Endeudamiento crediticio total en el sistema financiero menor a S/.20,000
Pequeña Empresa	Ventas anuales entre un monto de 150 y 1700 UIT	Endeudamiento crediticio total en el sistema financiero entre S/.20,000 y S/.300,000

Fuente: Produce (2013) y SBS (2008)

Cabe mencionar que con la aprobación del TUO mencionado, este decretaba la Derogatoria del Decreto Supremo N° 007 – 2008 – TR[32]. Así a partir de ello y la

[28] Decreto Supremo N° 013 – 2013 – PRODUCE.
[29] Resolución N° 11356 – 2008.
[30] Endeudamiento crediticio total durante los 6 últimos meses, no incluye el préstamo hipotecario.
[31] Unidad Impositiva Tributaria 2016: S/. 3,950 nuevos soles.
[32] Texto Único Ordenado de la Ley de Promoción de la Competitividad, Formalización y Desarrollo de la Micro y Pequeña Empresa y del Acceso al Empleo Decente, Ley MYPE.

entrada en vigencia de la Ley N° 30056[33], el número de trabajadores ya no se constituye como una característica diferenciadora entre los tipos de empresa; sin embargo, aquellas empresas que surgieron bajo el decreto derogado, todavía pueden regirse bajo dicha característica.

3.2. Las MYPEs en Latinoamérica

La situación actual de las MYPEs en sus diferentes aspectos no es un caso que solo concierne al Perú, sino también a gran parte de Latinoamérica, e incluso el Caribe, ya que se puede apreciar muchas similitudes de estas en los países que conforman este gran bloque geográfico. No obstante, cabe resaltar que los estudios sobre las MYPEs (pymes o mipymes en otros países), en particular en esta región, resultan ser complicados debido a la alta presencia de la informalidad y la falta de estandarización en la definición de una mype, ya sea por ventas anuales, número de trabajadores u otros factores. Estos dos hechos provocan que se obtenga datos sesgados y complica la posibilidad de hacer una comparación adecuada (Urmeneta, 2016).

A nivel de Latinoamérica, la presencia de la micro y pequeña empresa es considerable a nivel porcentual, aproximadamente representan el 99% de las empresas formales, así mismo proveen gran parte de los puestos de trabajo al emplear aproximadamente al 67% del total de la población ocupada (CEPAL, 2016). De esta manera, con dichas estadísticas debe tenerse en cuenta que estas representan actores clave en la generación del empleo y el fortalecimiento de las economías locales.

De acuerdo a un informe de CEPAL (2012), en Latinoamérica, las MYPEs se caracterizan por su gran heterogeneidad en aspectos como el acceso al mercado, las tecnologías de información y el capital humano, es decir, se trata de un grupo con una alta dispersión en sus características. A partir de este hecho, existen obstáculos para la posibilidad de generar políticas de fomento mype que beneficien a todas por igual, ya que no cubrirían las necesidades específicas de algunas. Por ello, reducir la brecha empresarial dentro de las MYPEs debería formar parte de las reformas estructurales necesarias dentro de los organismos institucionales latinoamericanos.

La heterogeneidad presente en las MYPEs repercute en varios aspectos concernientes a su desarrollo y/o potencial de crecimiento, principalmente en su productividad y su capacidad de exportación (CEPAL, 2012), lo cual produce que

[33] Ley que modifica diversas leyes para la facilitar la inversión, impulsar el desarrollo productivo y el crecimiento empresarial.

exista un vasto grupo con baja productividad y solo un sector reducido para la alta, cuyas probabilidades de sobrevivencia son más altas que las primeras. De este modo, realizando una comparación de productividad de una mype contra una gran empresa, en promedio, la productividad de esta llega hasta 33 veces la de la microempresa y 6 veces en el caso de las pequeñas empresas, lo cual comparado contra Europa, nos deja en una situación de mucha desventaja, ya que la misma relación en promedio en los países europeos oscila entre 1.3 y 2.4 veces. A través de dichas estadísticas, se puede apreciar porque la contribución de la mype en el PBI es relativamente baja. Por el lado de la exportación, en América Latina cerca del 10% de las MYPEs exportan una fracción de su producción; mientras que, en Europa, alrededor del 40% de las MYPEs realizan exportaciones. Añadiendo otras debilidades, la MYPEs latinoamericana se caracteriza por otorgar un bajo valor agregado a sus productos, debido a que exporta en gran mayoría materia prima, y poseer una mano de obra deficiente. En gran parte, la combinación de todos estos factores genera que el trabajo de la MYPEs en América Latina sea calificado como un estrato concentrado de muy baja productividad relativa con marcadas asimetrías entre los segmentos empresariales y la de los trabajadores (CEPAL, 2012). A partir de este hecho, en Latinoamérica, se pudo identificar que gran parte de las MYPEs de baja productividad, que frecuentemente se encuentran bajo el régimen de la informalidad, se tratan de empresas basadas en el autoempleo y la supervivencia, ya que no poseen el acceso a niveles adecuados de los principales recursos para su crecimiento: capital humano, financiamiento, capacidad de exportación, entre otros. Así mismo, el poco valor agregado de su producción se explica por los reducidos niveles de requerimientos técnicos y humanos desarrollados en sus actividades operativas.

3.2.1. Definición de las MYPEs

De acuerdo a Cardozo (2012), las características más comunes para la definición de mype por país, instituciones gubernamentales o estudios empíricos en América Latina son el volumen de ventas anuales, el número de empleados y el nivel de activos de la empresa, así mismo entre otros criterios identificados fueron patrimonio neto, tecnología, situación jurídica, ventas brutas anuales, volumen de exportación, etc. En general, el autor manifiesta que el 90% de los países[34] utiliza como criterio el

[34] El autor considera los siguientes países de América Latina: Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, República Dominicana, Uruguay y Venezuela.

número de trabajadores; el 60%, las ventas anuales; un 35%, los activos de la empresa y menos del 10% a los criterios de ventas brutas anuales y patrimonio neto.

Si bien es cierto que se evidencia el uso de criterios similares para la definición de una mype en América Latina, los límites en los intervalos de estos resultan ser heterogéneos en los diferentes países que conforman la región. Dicha situación resulta ser contraria a la mostrada por la Unión Europea en la que se encuentran con criterios estandarizados en cuanto a límites de ocupación, ventas anuales y nivel de activos, dichos intervalos pueden ser comparadas a continuación:

Tabla 33. Límites de número de trabajadores en países de Latinoamérica y la UE.

	Micro	Pequeña	Mediana	PYME
UE	0 - 10	10 - 50	50 - 250	10 - 250
Argentina	0 - 5			Industria: 10 - 200 Comercio y Servicios: 5 - 100
Brasil	Industria: 0 - 20 Comercio y Servicios: 0 - 10 Sector informal: 0 - 5	Industria: 20 - 100 Comercio y Servicios: 10 - 50	Industria: 100 - 500 Comercio y Servicios: 50 - 100	Industria: 20 - 500 Comercio y Servicios: 10 - 100
Chile	0 - 10	10 - 50	50 - 200	10 - 200
Ecuador	0 - 10	10 - 50	50 - 100	10 - 100
México	0 - 10	Industria: 10 - 50 Comercio y Servicios: 10 - 30	Industria: 50 - 250 Comercio y Servicios: 30 - 100	Industria: 10 - 250 Comercio y Servicios: 10 - 100
Perú	0 - 10	10 - 50	50 - 250	10 - 250
Uruguay	0 - 5	5 -20	20 - 100	5 - 100

Fuente: OIT (2009).

Para el caso del criterio de número de trabajadores, puede apreciarse que hay diferencias presentes, pero que no son tan significativas, podría acotarse que dicho criterio es uno de los de mayor homogeneidad respecto a la UE. Así mismo, cabe resaltar la diferencia presentada en el caso de Argentina, Brasil y México, los cuales incluyen un sub-criterio sectorial, esto puede resultar de uso conveniente debido a los diferentes niveles de productividad en que se desarrollan estos sectores.

Por otro lado, respecto al criterio de ventas anuales, resulta que este presenta una alta heterogeneidad, ninguno de los países coincide de manera exacta en los límites inferiores y superiores. Así mismo, se puede notar que los límites establecidos dentro la UE están muy por encima de los demás países latinoamericanos, lo cual trae como consecuencia que la “verdadera percepción” de una mediana empresa en la UE correspondería de mejor manera a la de una grande empresa en América Latina. Así, para realizar la comparación estadística entre América Latina y la Unión Europea para el caso de las MYPEs, se debe considerar que las diferencias presentes se explican por las distintas realidades de cada región en aspectos económicos,

sociales, educativos, entre otros. Aunque el uso de este criterio es uno de los más extendidos en diferentes países, se advierte que, al menos en el caso de Latinoamérica, la información monetaria puede presentar problemas de confiabilidad en aquellas empresas de menor tamaño (OIT, 2009).

Tabla 34. Límites ventas anuales en países de Latinoamérica y UE, miles US$[35] .

	Micro	Pequeña	Mediana	PYME
UE	0 - 2800	2800 - 14000	14000 - 70000	2800 - 70000
Argentina	Industria: 0 - 329 Comercio: 0 - 486 Servicios: 0 - 123	Industria: 329 - 1973 Comercio: 486 - 2919 Servicios: 123 - 885	Industria: 1973 - 15780 Comercio: 2919 - 23354 Servicios: 885 - 5902	Industria: 329 - 15780 Comercio: 486 - 23354 Servicios: 123 - 5902
Chile	0 - 95	95 - 394	394 – 3975	95 - 3975
Ecuador	0 -100	100 - 1000	1000 - 5000	100 - 5000
México	Industria: 0 - 500 Comercio: 0 - 1000 Servicios: 0 - 250			Industria: 500 - 24000 Comercio: 1000 - 48000 Servicios: 250 - 12000
Perú	0 - 178	178 - 1000	1000 - 13794	178 - 13794
Uruguay	0 - 60	60 - 180	180 - 5000	60 - 5000

Fuente: OIT (2009).

Tabla 35. Límites ventas activos en países de Latinoamérica y UE, miles US$.

	Micro	Pequeña	Mediana	PYME
UE	0 - 2800	2800 - 14000	14000 - 60000	2800 - 60000
Ecuador	0 - 20	20 - 750	750 - 4000	20 - 4000
Uruguay	0 - 20	20 - 50	50 - 350	20 - 350

Fuente: OIT (2009).

Finalmente, de acuerdo al criterio del límite de los activos, se puede identificar el mismo problema de la alta heterogeneidad, así como una reducida participación de países que utilicen dicho criterio.

En general, estas disparidades y/o heterogeneidades presentes son las que dificultan y condicionan el diseño de las políticas para las MYPEs, es decir, debido a la diversidad empresarial de estas, los gobiernos de turnos no pueden simplemente implementar una "receta común", sino, a partir de los diferentes potenciales de crecimiento en productividad y empleo, así como las restricciones propias del sector y el tamaño de las empresas, se deben proponer y crear diversas políticas y estrategias que beneficien a las MYPEs (CEPAL, 2012).

3.2.2. Participación de las MYPEs en América Latina

[35] Tipo de cambio al 05.07.2009

En cuanto a información referente a las MYPEs, América Latina cuenta con desfases en la recopilación de su información o, lo que es peor, no cuenta con ninguna información sobre estas. En general, como se ha dicho anteriormente, el problema se debe a la alta informalidad presente, así mismo también se debe a la falta de consenso entre las instituciones públicas de un mismo país. Por ejemplo, en el caso de Perú, el INEI, la SUNAT y Produce utilizan diferentes fuentes para el registro y recopilación estadística de las MYPEs, pero cabe resaltar que estas diferencias no son tan significativas, e incluso mantienen una tendencia similar, lo cual hace que la información sea confiable y su uso viable para la comparación estadística. Sin embargo, esto no sucede a lo largo de los demás países latinoamericanos, en los cuales la información está desactualizada por diversas razones. En cambio, para el caso de la Unión Europea, este territorio cuenta con un recopilatorio común de información estadística de las MYPEs, el Observatorio Eurostat trabaja a través de instancias especializadas usando una metodología homogénea y reúne antecedentes de sus 27 estados miembros (OIT, 2009), lo cual contrasta con el problema de la información y la heterogeneidad de América Latina.

Tanto en América Latina como en Europa, se mantiene la tendencia que las MYPEs poseen la mayor participación por estrato empresarial, de las cuales las microempresas conforman aproximadamente entre el 80% a 90% de participación, a excepción de Argentina y Chile. Luego, el conglomerado de pequeñas y medianas empresas (también conocido como pymes) conforma en promedio menos del 10% del estrato empresarial. Por otro lado, las grandes empresas constituyen menos del 2% del número total. Si bien es cierto, hay algunas diferencias notables con ciertos países, estas se pueden explicar por la diferencia en la definición de la mype o por el desarrollo económico del país. Generalmente, estos porcentajes de participación se mantienen estables a lo largo del tiempo, excepto cuando ocurren cambios estructurales tales como crisis económicas, cambios en la política económica, o incluso cambios en la definición de mype.

También, cabe resaltar que la relación existente entre el porcentaje de microempresas y grandes empresas en la gran mayoría de países resulta ser inversamente proporcional, es decir, en la medida que haya menor concentración de las primeras, se presentará unos cuantos puntos porcentuales más de las grandes empresas. De esta manera, la composición de las empresas por su tamaño se puede resumir en la tabla siguiente.

Tabla 36. Participación porcentual del número de empresas por tamaño, 2011

País	Empresas		
	Micro	Pyme	Grande
Argentina	69,7	28,4	1,9
Brasil	90,1	9,3	0,6
Chile	78,3	20,3	1,4
Colombia	96,4	3,5	0,1
Ecuador	95,4	4,4	0,2
El Salvador	91,2	8,4	0,4
México	95,5	4,3	0,2
Perú	94,5	4,9	0,6
Uruguay	83,4	16,1	0,5
UE (25)	92,0	7,8	0,2
Alemania	82,0	17,5	0,4
Bélgica	93,7	6,2	0,1
España	94,0	5,9	0,1
Francia	94,7	5,1	0,1
Italia	95,0	4,9	0,1
República Checa	96,0	3,8	0,1
Reino Unido	89,7	10,0	0,3

Fuente: CEPAL (2016)

Estas estadísticas obtenidas por CEPAL provienen de la información manejada por instituciones públicas, la cual solo considera a aquellas empresas registradas y reguladas por el marco laboral, es decir, formales. De esta manera, la "real" participación de la mype en una economía nacional podría ser mucho mayor a la presentada, ya que esta unidad económica generalmente está asociada a negocios de subsistencia y baja productividad (OIT, 2015), los cuales están inmersos en la informalidad y por lo tanto no se posee registros de ellos. Por el lado del empleo, se conserva una tendencia a que las MYPEs en América Latina y la Unión Europea concentren gran parte del empleo total de los países, las microempresas poseen alrededor, en algunos casos, hasta el 50% de la participación del empleo. Para el caso del conglomerado mipyme (mediana, pequeña y micro empresa) se puede observar que puede llegar a incluir entre el 60% al 80% de la participación total, exceptuando los casos de Argentina y Brasil para Latinoamérica, y Reino Unido para el caso europeo. En general, la mype se constituye como la unidad económica de mayor creación de empleo en la región explicada en su alta tasa de creación, lo cual a su vez está representado en su concentración en el número de empresas y en el empleo; sin embargo, persiste el problema de que la gran mayoría de estos negocios se basan en el uso de recursos (humanos, tecnológicos, etc.) deficientes que trae consigo un bajo nivel de productividad, lo cual tiene como consecuencia que exista una alta tasa de salida para dicho estrato empresarial. En sí, la elevada concentración de las MYPEs en América Latina y sus características asociadas son las que se traducen directamente en las desigualdades o brechas del sector laboral, ya que al tratarse de sectores de baja productividad, se presentan bajos salarios, condiciones laborales deplorables y un bajo acceso a la protección para una proporción amplia

de la fuerza laboral, lo cual contrasta totalmente de su extremo opuesto en el que una proporción minoritaria goza de altos niveles de productividad (OIT, 2015).

Tabla 37. Participación porcentual del empleo en empresas por tamaño, 2011

País	Empleo		
	Micro	Pyme	Grande
Argentina	11,5	39,6	48,9
Brasil	13,7	28,3	58,0
Chile	44,1	30,9	25,0
Colombia	50,6	30,3	19,1
Ecuador	47,3	29,8	22,9
El Salvador	37,8	27,7	34,6
México	45,7	23,6	30,8
Perú	48,5	19,2	32,4
Uruguay	24,1	43,1	32,8
UE (25)	31,5	38,3	30,2
Alemania	19,5	44,0	36,5
Bélgica	34,8	38,4	26,8
España	41,5	35,1	23,4
Francia	31,8	35,1	33,1
Italia	48,5	33,4	18,1
República Checa	32,8	37,6	29,6
Reino Unido	19,8	37,0	43,2

Fuente: CEPAL (2016)

3.2.3. Participación en las exportaciones

La mype en América Latina se caracteriza por poseer un bajo nivel de internacionalización, es decir, está limitado a satisfacer a la demanda interna nacional, lo cual se expresa en su bajo volumen de ventas en las empresas que participan en las exportaciones. De acuerdo a CEPAL (2012), este hecho se fundamenta en la propia estructura de las exportaciones de la región, la que en gran medida se focaliza hacia los recursos naturales y sus derivados, ósea actividades primarias de extracción, como la minería, la pesca, la tala, entre otros. De por sí, este ámbito se encuentra dominado por las grandes empresas por los altos requerimientos de inversión necesarios, lo cual origina que se reste oportunidades a las pequeñas, e incluso a las medianas empresas, ya que no cumplen con dichos estándares necesarios, lo cual representa la principal restricción para el acceso a los mercados externos. Como consecuencia de ello, la gran mayoría de MYPEs se limitan únicamente a abastecer el propio mercado local y/o demanda interna, lo cual genera que este tipo de empresas posean una elevada dependencia de la situación de la economía nacional. Así, durante periodos de recesión, se puede apreciar que una cantidad considerable de MYPEs quiebran debido a la desaceleración del consumo.

Tabla 38. Participación porcentual de las empresas en las exportaciones, 2010

	Micro	Pequeñas	Medianas	Grandes
Argentina	0,3	1,6	6,5	91,6
Brasil	0,1	0,9	9,5	82,9
Chile	-	0,4	1,5	97,9
España	11,1	13,3	22,6	47,1
Italia	9	19	28	44
Alemania	8	12	18	62
Francia	17	10	15	58

Fuente: CEPAL (2012)

Como se puede apreciar en la tabla anterior, en los países latinoamericanos, el impacto de la mype en las exportaciones resulta insignificante ante la participación de las grandes empresas que sobrepasa el 80%; mientras que en Europa, no se presenta dicha tendencia a la concentración, lo cual se muestra con un impacto de la mype relativamente mayor con participaciones entre el 20% y el 30%, así como un impacto de las grandes empresas que resulta ser hasta 20% menos respecto a la latinoamericana. Este hecho conlleva a la existencia de una gran brecha en los países de América Latina, en la que las grandes empresas pueden representar más del 85% del valor de las exportaciones y, al mismo tiempo, tratarse del menos del 10% del total de las empresas exportadoras.

De acuerdo a Urmeneta (2016), otra manera de demostrar el problema de la concentración es a través del análisis de Pareto para el percentil 1, en el cual puede verificarse que el 1% de las empresas exportadoras pueden llegar a concentrar hasta el 73.5% del valor de las exportaciones en promedio en América Latina, lo cual contrasta con el 47.0% en el mismo indicador para la Unión Europea. Por otro lado, en la revisión por países, las más altas concentraciones están en los países de Colombia, Paraguay y Venezuela, en los que dicho parámetro representa el 83.6%, 89.7% y 98.2%, respectivamente para el año 2013.

3.2.4. Productividad

Como se ha mencionado anteriormente, la falta de una definición estándar de la mype a lo largo de los países latinoamericanos ocasiona que la comparación entre estas sea complicada. Por ejemplo, algunos errores comunes son sobreestimar la cantidad de MYPEs, no tomar en cuenta las características específicas sectoriales, entre otros. A pesar de este hecho sí puede afirmarse que existen tendencias y patrones relativos a la productividad que describen a la gran mayoría de MYPEs en Latinoamérica, en promedio, las pequeñas empresas poseen un 16% a 36% de la productividad de las grandes empresas, lo cual discrepa totalmente del contexto europeo, en el que las pequeñas empresas alcanzan entre 63% y 75% de la productividad de las grandes empresas. En la siguiente tabla, se puede apreciar

mejor dichas estadísticas, en la cual se coloca como base que la productividad de la gran empresa es de un 100%:

Tabla 39. Productividad relativa en países de Latinoamérica y la UE

	Micro	Pequeñas	Medianas	Grandes
Argentina	24	36	47	100
Brasil	10	27	40	100
Chile	3	26	46	100
México	16	35	60	100
Perú	6	16	50	100
Alemania	67	70	83	100
España	46	63	77	100
Francia	71	75	80	100
Italia	42	64	82	100

Fuente: Indicadores productivos 2010 – CEPAL (2014)

En general, se puede ver que la brecha existente entre la micro y pequeña empresa en Latinoamérica es mucho más acentuada que en Europa. Así mismo, las brechas entre la mediana y la grande empresa son más reducidas para Europa que para Latinoamérica. En general, estas diferencias están estrechamente relacionadas con la estructura productiva y la distribución sectorial (CEPAL, 2012). Por otro lado, respecto a estas estadísticas, un posible sesgo presente en ella se puede deber a la falta de una definición estándar de la mype según características en los países latinoamericanos, lo cual complica no solo las comparaciones, sino también la clasificación en sí de dichas empresas debido a la heterogeneidad presente.

3.2.5. El clima de negocios en América Latina

De acuerdo a CEPAL (2012), el clima de negocios y la estructura productiva que se basa en factores de acceso al financiamiento, a las tecnologías, recursos humanos y la existencia de sistemas articulados de producción, entre otros, son factores que afectan al desempeño de la mype. En la medida que el Gobierno establezca políticas públicas que incidan en el clima de negocios, las MYPEs enfrentarán menos dificultades para su desarrollo.

3.3. Las MYPEs en el Perú

En nuestro país, el rol desarrollado por las Mypes es de gran relevancia debido a que se constituyen como unidades empresariales generadoras de empleo, así mismo se reconoce su gran aporte al desarrollo socioeconómico de las zonas en donde se ubican (Avolio, 2011). De la misma manera, el Perú es reconocido como uno de los entornos más propicios para el desarrollo de las microfinanzas (Portal de Microfinanzas, 2015), el cual es uno de los principales componentes clave para el financiamiento de las MYPEs y el despliegue de estrategias innovadoras dirigidas a

la inclusión financiera de poblaciones excluidas del Sistema Financiero tradicional, generalmente segmentos de ingresos medios–bajos (Ríos, 2015). Si bien es cierto que, actualmente, el sistema de las microfinanzas en nuestro país está considerado como uno de los más avanzados del mundo; por el lado de las MYPEs, estas aún presentan grandes retos a superar, entre las que se puede destacar su bajo nivel de productividad y la falta de herramientas de TI (Campodónico, 2016).

A continuación, se presentará algunas de las principales estadísticas y aspectos que representan al sector Mype en nuestro país y su respectivo impacto.

3.3.1. Aporte a la economía nacional

En los últimos años, el Perú ha presentado altas tasas de crecimiento de PBI en la región latinoamericana, durante el periodo 2001 – 2010, el PBI nacional ha crecido a un promedio de 6.3%, así se destacó como uno de los países de más rápido crecimiento dentro de la región. Así mismo, se presentó un fuerte crecimiento del empleo y los ingresos que conllevaron a una reducción drástica de los índices de pobreza de 55.6% hasta un 21.8% en la población entre el 2005 y 2015 (Banco Mundial, 2016). Este crecimiento sostenido de la economía peruana está apoyado, principalmente, en el entorno favorable de la economía mundial que se basaba en la mayor demanda e incremento de los precios internacionales de los minerales y de las exportaciones no tradicionales (Produce, 2010). Así, tomando como referente el PBI del Perú, este ha crecido en los últimos años de la siguiente manera:

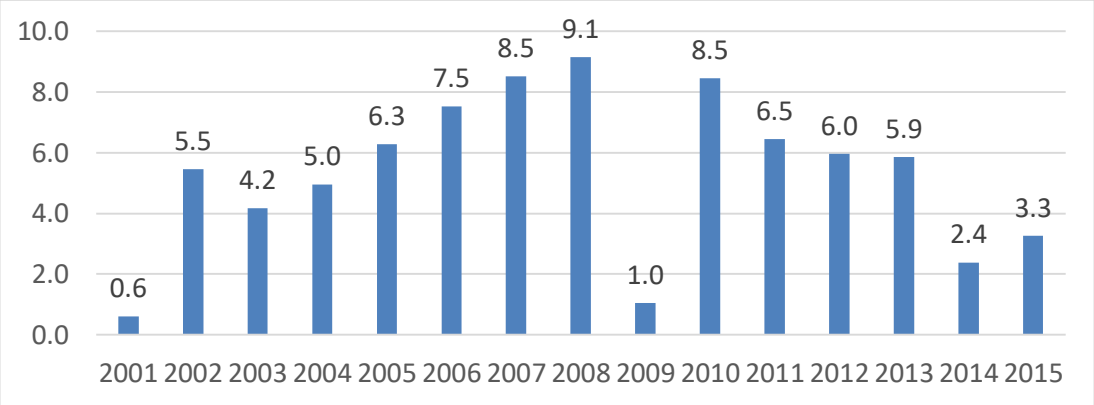


Figura 20. PBI 2001 – 2015 (Variación Porcentual Real).
Fuente: Cuadros Anuales Históricos – BCRP (2016)

De acuerdo a Wennekers y Thurik (2001), existen estudios econométricos que evidencian que la actividad empresarial es un factor determinante para el crecimiento económico, el cual a su vez está asociado en forma positiva con la tasa de creación de empresas. Sin embargo, no solo dependerá de ello, sino también de aquellos que

introduzcan innovaciones al mercado y amplíen el ámbito de los negocios (Produce, 2010). A partir de ello, se debe resaltar la existencia de una relación recíproca entre la entrada de las nuevas MYPEs al mercado y su supervivencia a lo largo del tiempo y el crecimiento económico y laboral de una determinada región; debido a que la Mype al constituirse como ofertante de productos nacionales, con el paso del tiempo, también se convierte en generadora de empleo. Posteriormente, se sigue con los empleados de las MYPEs al contribuir al consumo de otros productos nacionales. De esta manera, se constituye un "círculo" que empieza con la creación de estos pequeños negocios y que "gira" siempre y cuando los negocios se encuentren activos.

A partir de estas proposiciones y estudios sobre la coyuntura actual del sector, el Gobierno a través de sus diferentes entidades, tales como el Ministerio de Trabajo y Promoción del Empleo (MTPE) y el Ministerio de la Producción (Produce), entre otros, se encuentra en la obligación de fomentar programas que constantemente contribuyan a la constitución y crecimiento de las pequeños negocios. En general, dichos programas están dirigidos a la promoción de la formalización de las MYPEs peruanas, la capacitación de sus trabajadores, el acceso a los créditos asequibles, entre otras materias que tengan como finalidad aumentar la competitividad de las MYPEs ante un escenario cada vez más globalizado.

En nuestro país, aproximadamente el 99% de las empresas formales están conformadas por las micro (94%) y pequeñas empresas (5%), lo cual coloca en evidencia la importancia de estas en la generación de PIB nacional. Así mismo, el crecimiento mostrado en la variación porcentual anual de las MYPEs 2010-2014, se mantuvo en promedio por encima del 7%. Si bien es cierto que dicho segmento se caracteriza por tener ventas anuales relativamente bajas respecto a las grandes empresas, las MYPEs realizan un gran aporte por la alta concentración y/o volumen de estas. Así, en el año 2014, aproximadamente el 40% del PBI fue generado por las MYPEs (Gestión, 2014) y dicho porcentaje ha logrado permanecer constantemente entre un rango cercano.

3.3.2. Evolución del número de MYPEs, por estrato empresarial

De esta manera, respecto al crecimiento del PBI nacional y la participación de las MYPEs en la economía nacional, en los últimos años, la promulgación de diversas leyes dirigidas a las MYPEs, enfatizaban reglamentos especiales acerca de su funcionamiento, el régimen tributario aplicado, los deberes y derechos de los empleados, entre otros temas. Dichas facilidades redujeron las barreras de entrada

para la constitución de la micro y pequeña empresa formal en nuestro país, así durante el IV trimestre del 2015, de acuerdo al DCEE[36], se crearon aproximadamente 62581 empresas; sin embargo, al mismo tiempo, se dieron de baja a 45393 empresas (SEMANAeconómica.com, 2016). Como se puede apreciar, las empresas en general, especialmente las MYPEs, están superando el obstáculo mencionado, pero aún persisten problemas que dificultan su crecimiento sostenido, de ahí, que se presenten un elevado número de bajas. De esta manera, a lo largo de los últimos años, con la entrada y salida de las empresas, al final de cada año, se obtuvieron saldos positivos, lo cual revela una tendencia creciente del número total de empresas y que se presente un porcentaje de concentración alto y estable de MYPEs en el Perú, tal como indica la siguiente tabla:

Tabla 40. Evolución de las empresas formales, por estrato empresarial

Año	Microempresa	Pequeña empresa	Mediana empresa	Grande empresa	N° MYPEs	Total
2009	1,074,235	50,637	1,885	5,487	1,124,872	1,132,244
2010	1,138,091	55,589	2,031	6,342	1,193,680	1,202,053
2011	1,221,343	61,171	2,325	7,285	1,282,514	1,292,124
2012	1,270,009	68,243	2,451	7,908	1,338,252	1,348,611
2013	1,439,778	70,708	2,520	8,306	1,510,486	1,521,312
2014	1,518,284	71,313	2,635	8,388	1,589,597	1,600,620

Fuente: Las Mipyme en cifras 2014 – Produce (2015)

Enfocándonos en la evolución de las MYPEs, en los últimos años, el porcentaje representado por estos sobre el total de empresas formales siempre ha sido mayor al 99%. Para una diferenciación más específica entre micro y pequeña empresa, los respectivos porcentajes se mantuvieron alrededor de 95% y 4%, respectivamente.

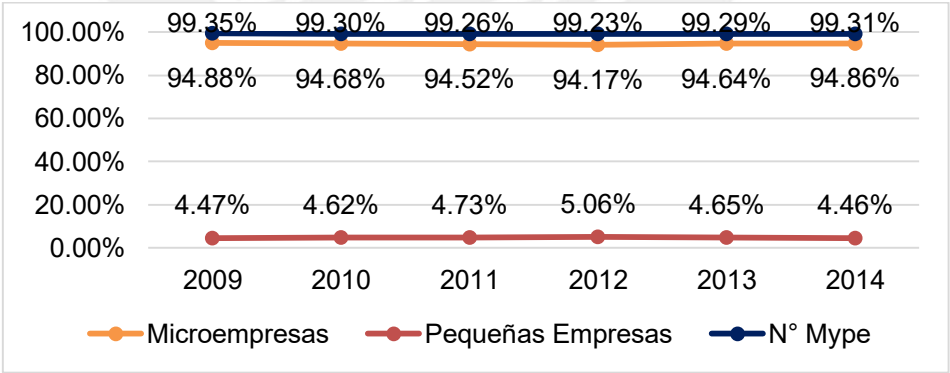


Figura 21. Evolución de empresas formales, por estrato empresarial, en %
Fuente: Las Mipyme en cifras 2014 - Produce (2015)

[36] Directorio Central de Empresas y Establecimientos del Instituto Nacional de Estadística e Informática (INEI)

3.3.3. Distribución de la PEA ocupada, por estrato empresarial

Otro aspecto resaltante de las Mype en la economía de un país es su función como fuente generadora de empleo, principalmente en los países de América Latina (OIT, 2015). En Perú alrededor del 60% de PEA[37] se encuentra empleada por las Mipyme[38] (Produce, 2015), mostrando una alta concentración en las empresas con menor cantidad de trabajadores. Así, según la clasificación del tamaño empresarial, la distribución de la PEA empleada muestra que el 48.2% de ésta trabaja en empresas de 2 a 10 trabajadores; mientras que el 20.3% deciden por el "autoempleo".

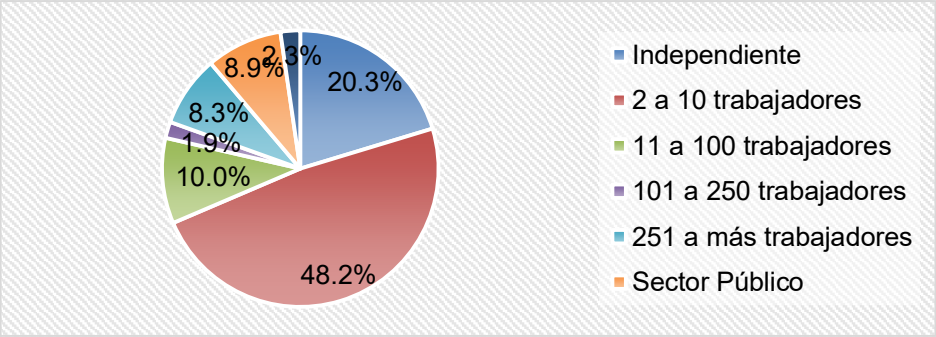


Figura 22. Distribución de la PEA, según tamaño empresarial, 2014
Fuente: Las Mipyme en cifras 2014 - Produce (2015)

De esta manera, de acuerdo a las cifras recopiladas por Produce, en nuestro país, alrededor del 60% de la PEA ocupada, se encuentra empleada en Mypes formales, lo cual realza una evidente dependencia de la generación del empleo con la creación de los pequeños negocios en forma de Mype. Este hecho está expresado en las estadísticas que muestran una tendencia al aumento del número de trabajadores en Mype, pero cuya participación total porcentual mostró una disminución promedio anual de menos de 1% entre el periodo 2004 - 2011. Así, en el 2004, esta población representaba el 63.34% de la PEA ocupada y para el 2011, la misma bordeaba el 59.59% de la PEA ocupada. Dicha insignificante disminución porcentual no resulta alarmante para el sector Mype por el momento, pero, coloca en evidencia el nivel de dinamismo que lo rodea y que sus características asociadas a la baja productividad e informalidad persistente de estas podrían hacer que, en el futuro, los nuevos miembros de la PEA estén más dispuestos a trabajar en una Gran Empresa que en una Mype, lo cual podría tener un impacto severo en el nivel de empresarialidad de nuestro país.

[37] Población Económicamente Activa
[38] Mediana, micro y pequeña empresa

Tabla 41. Perú: Distribución de la PEA ocupada, según estrato empresarial

Año	PEA ocupada	PEA ocupada en Mype	% PEA ocupada en Mype
2004	13,059,798	8,272,482	63.34%
2005	13,120,443	8,254,553	62.91%
2006	13,682,993	8,543,987	62.44%
2007	14,197,152	8,682,326	61.16%
2008	14,459,188	8,742,388	60.46%
2009	14,757,684	8,923,127	60.46%
2010	15,089,871	9,085,879	60.21%
2011	15,307,326	9,121,940	59.59%
2012	15,541,484	9,132,395	58.76%

Fuente: Mype 2011 – Produce (2012)

3.3.4. Evolución del número de Mipymes, por sector económico

En el Perú, los diferentes estudios sobre las estadísticas de las empresas formales, especialmente Mype y/o Mipyme, acerca de la clasificación por sector económico poseen como consenso que son 7 los grandes sectores que agrupan a estas. De acuerdo a la clasificación CIIU, estos serían los sectores analizados:

- Agropecuario
- Pesca
- Minería
- Manufactura
- Construcción
- Comercio
- Servicios

En cuanto a la creación de Mipyme en el Perú, durante el periodo mencionado, se ha observado que los sectores que presentan saldos positivos más altos[39] son los de Comercio, Servicios y Manufactura, los cuales concentran alrededor del 94% de las Mipyme formales, dicha composición bordea 45%, 39% y 10% de su respectivo sector económico (Produce, 2014). Así mismo, al realizar la comparación, según el estrato comercial y la actividad económica, se replica el comportamiento apreciado a nivel macroeconómico: el número de microempresas en un determinado sector económico representa entre el 90% y 95%; mientras que la pequeña empresa, alrededor del 5% de la cantidad total de Mipyme formales.

[39] Respecto de la entrada y salida de empresas de dicho sector económico.

Tabla 42. Evolución de Mipyme formales, por sector económico

Sector económico	Año 2010	2011	2012	2013	2014
Agropecuario	22,202	22,597	22,298	24,131	23,879
Comercio	547,651	601,930	630,193	694,358	720,299
Construcción	31,898	39,327	39,662	47,378	49,150
Manufactura	121,242	129,189	131,731	144,506	145,499
Minería	6,375	6,955	8,793	9,620	13,530
Pesca	3,493	3,813	3,437	3,656	3,539
Servicios	462,850	481,028	504,589	589,357	636,336
Total	1,195,711	1,284,839	1,340,703	1,513,006	1,592,232

Fuente: Las Mipyme en cifras 2013 – Produce (2014)

Acerca de la importancia económica, se reconoce que los sectores que los sectores mencionados (Comercio, Servicios y Manufactura) aportan alrededor del 70% del PBI nacional, de acuerdo a las Estadísticas del BCRP. De hecho, el más destacado de ellos es el sector Servicios que, en los últimos años, realiza un aporte que bordea el 47% del PBI y existe una tendencia explícita hacia un mayor desarrollo de dicho sector. Luego, le sigue el sector Manufactura con un aporte promedio del 15%, dicho sector resulta ser más dinámico que los otros debido a que presenta variaciones más marcadas. Actualmente, este sector está pasando por un periodo de recesión, así en el 2015, su variación porcentual del PBI fue de -1.7% y aún continuará en recesión (LaRepública.pe, 2016). Además, el sector Comercio aporta el 11% del PBI nacional y presenta un crecimiento moderado, Por otro lado, debe agregarse que el sector Minería aporta alrededor de 12% al PBI, a pesar que el número de Mipyme en este sector conforma menos del 1% del total de estas.

Como en el caso de la minería, el hecho que un sector con un porcentaje bajo de empresas aporte una cantidad considerable al PBI se basa en la existencia de una alta concentración de empresas de bajo nivel de ventas anuales en los diferentes sectores, lo cual, si bien es cierto, representaría un gran aporte por volumen, no llega a equipararse contra la ganancia obtenida de las grandes empresas que se apoya en las ventajas de la economía de escalas y el acceso al mercado internacional.

3.3.5. El dinamismo empresarial de las Mipyme

Como se había mencionado anteriormente, el desarrollo económico de las naciones está directamente relacionado con la creación de nuevas empresas, a partir de ello, se ha hecho necesario estudiar las características asociadas a sus procesos de creación y desaparición de estas, principalmente con las Mipyme. En el Perú, las Mipyme representan un segmento empresarial muy dinámico en la que los eventos de entrada, crecimiento, declive y salida y/o cambio de giro de negocio son frecuentes (Produce, 2014). Así, para la evaluación del dinamismo que envuelve a las Mipyme,

usualmente se recurre a los indicadores de cantidad de años en el mercado, la tasa de creación, la tasa de salida y la matriz de transición, entre otras variables.

Acerca del tiempo de vida en el mercado de las Mipyme, se determinó que el 59.1% de las Mipyme tiene como máximo 5 años en el mercado (Produce, 2015). Particularmente, durante los años 2013 y 2014, las estadísticas de Produce mostraron la tendencia de que la mediana sea menor a la media en el análisis de la cantidad de años en el mercado, lo cual indica que se presenta una distribución marcadamente asimétrica que ocasiona una alta concentración de empresas en el rango inferior, a partir de ello, se genera que el tiempo de vida promedio en el mercado sea muy bajo para la Mipyme.

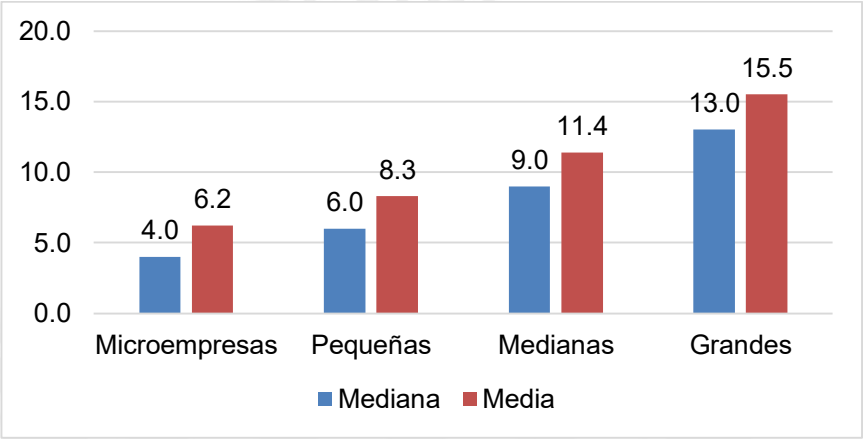


Figura 23. Promedio de años en el mercado, según estrato empresarial.
Fuente: Las Mipyme en cifras 2014 - Produce (2015)

Como se puede apreciar en la gráfica, existe una relación directa entre el promedio de años en el mercado y el tamaño de la empresa, lo cual podría explicarse a través de la alta tasa de mortalidad de las Mipyme comparado contra la de las grandes. Dicha tendencia se mantiene a lo largo de todos los sectores económicos; sin embargo, como la situación de las empresas dependen en gran medida de la del entorno económico, algunos sectores presentarán características heterogéneas respecto a los otros, como puede ser un mayor promedio de vida, una mayor brecha, según el estrato empresarial, entre otras que dependerán de las características estudiadas. Por ejemplo, el sector manufactura revelaría una mayor probabilidad de supervivencia hasta alcanzar la categoría de Gran empresa, de ahí, le seguirían los sectores Comercio y Servicios.

Tabla 43. Años promedio en el mercado

Sector económico	Microempresa	Pequeña	Mediana	Grande
Agropecuario	4.0	6.0	7.0	12.0
Pesca	3.0	3.0	8.0	10.0
Minería	1.0	4.0	5.5	12.0
Manufactura	4.0	7.0	10.0	17.0
Construcción	2.0	4.0	8.0	10.0
Comercio	4.0	6.0	9.0	12.0
Servicios	4.0	6.0	9.0	12.0

Fuente: Mipyme en cifras 2014 – Produce (2015)

Así mismo, la relación existente entre el tiempo de vida en el mercado de una empresa y el estrato empresarial perteneciente se puede reflejar al clasificarlos de acuerdo a un rango de edad. De esta manera, en los últimos años, se ha presentado una mayor concentración de las Mipyme en el intervalo de hasta 5 años, principalmente en las microempresas que conforman el 47.6% y el 59.7% para los años 2013 y 2014, respectivamente; mientras que, para los intervalos de mayor rango de edad, el porcentaje se reduce considerablemente. Por otro lado, cabe notar que para las grandes empresas, la situación resulta totalmente distinta, existe un mayor porcentaje de estas en los intervalos más altos, como el de mayor a 16 años que bordea el 38.3% en el año 2013 y 39.6% en el año 2014.

Tabla 44. Empresas formales por estrato empresarial, según rango de edad, 2013

	Hasta 5 años	De 6 a 10 años	De 11 a 15 años	De 16 a 21 años
Microempresas	47.6%	21.0%	13.3%	18.1%
Pequeñas	46.0%	25.4%	12.3%	16.3%
Medianas	30.0%	27.7%	16.5%	25.9%
Grandes	21.1%	22.5%	18.0%	38.3%

Fuente: Mipyme en cifras 2013 – Produce (2014)

Tabla 45. Empresas formales por estrato empresarial, según rango de edad, 2014

	< 5 años	De 6 a 10 años	De 11 a 15 años	De 16 a 20 años	> 20 años
Microempresas	59.7%	19.4%	9.1%	6.9%	4.9%
Pequeñas	47.0%	25.9%	11.8%	7.6%	7.8%
Medianas	30.1%	27.1%	17.1%	11.5%	14.2%
Grandes	20.5%	22.5%	17.3%	16.6%	23.0%

Fuente: Mipyme en cifras 2014 – Produce (2015)

CAPÍTULO 4. APLICACIÓN DE LA MINERÍA DE DATOS

En este capítulo se detallará la aplicación de cada una de las etapas en un proyecto de minería de datos según la metodología CRISP-DM a una base de datos que contiene información de préstamos evaluados para empresas MYPEs de un banco representativo de nuestro país, algunos de ellos aceptados o denegados.

4.1. Conocimiento del negocio

4.1.1. Objetivos del negocio

Siendo nuestro cliente del proyecto una MYPE – que representa para nuestros fines a la institución que requiere el proyecto de minería de datos – se debe analizar primero cuáles son sus necesidades o cuestionamientos acerca del proceso de solicitud de un préstamo bancario, utilizando tanto criterios internos como externos (Gestión, 2018ª):

- ¿El giro de negocio influye en la decisión?
- ¿La rotación de mis productos/servicios puede influir en la decisión?
- ¿Mi nivel de ventas/costos mensuales pueden influir en la decisión?
- Si tengo poco tiempo en el negocio y es el primer préstamo que realizo, ¿es más probable que el préstamo sea denegado?
- ¿Es más probable obtener un préstamo en soles o en dólares?
- ¿A partir de que monto es más probable que sea aceptado un préstamo?
- ¿La garantía contra préstamo influye en la decisión?
- Si no labora en Lima, ¿es más probable que mi préstamo sea denegado?
- ¿Cuáles son los factores que más influyen a la aceptación de un préstamo?

4.1.2. Objetivos del proyecto de minería de datos

De acuerdo a las necesidades de las MYPEs, se plantean objetivos del proyecto restringidos a la base de datos que se tiene para el proyecto:

- Encontrar relación entre la aceptación de un préstamo y las características de la MYPE,
- Verificar la hipótesis si el monto aprobado depende de la moneda emitida.
- Desestimar la hipótesis de que la probabilidad de aceptación depende del lugar de emisión del préstamo.
- Hallar factores significativos que influyan en la aprobación de un préstamo bancario.

4.2. Comprensión de la información

La base de datos cuenta con 37 campos y 31161 registros e incluye información sobre solicitudes de préstamos a MYPEs para el año 2016.

Tabla 46. Campos de la base de datos.

Campo	Descripción
CODMES	Mes y año, concatenados
MES	Mes en el que la solicitud fue atendida
TIPO_CAMPANIA	Nombre de la campaña
ESTADO	Aceptado (ACE) o denegado (DEN)
MONEDA	Tipo de moneda indicada en la solicitud del préstamo
MTOAPROBADO	Monto solicitado, en la moneda indicada en la solicitud
MONTODOLARES	Monto solicitado, en dólares
MONTOSOLES	Monto solicitado, en soles
TASA	Tasa del préstamo (en porcentaje)
DESSECTORECONOMICO	Sector económico de la empresa
CODVENDEDOR	ID del asesor del préstamo
CODAPROBADOR	ID del aprobador del préstamo
FLGGAR	Indicador de tenencia de garantía
DESCAMPANIA	Clasificación préstamo según campaña
TIPESTCONTRATOSOLICITUDMIC	“S” si en el contrato se establece la garantía, “N” caso cont.
FLGTIPPER	Tipo de persona indicada en la solicitud
MTOCAMBIOALNUEVOSOL	Tipo de cambio en la fecha del préstamo
MARCA	Variable de segmentación del banco
ZONA	Ubicación: Nivel 1, agrupación banco
AREA	Ubicación: Nivel 2, agrupación banco
REGION	Ubicación: Nivel 3, agrupación banco
CODOFI	Ubicación: Nivel 4, agrupación banco
DESCODDEPARTAMENTO	Departamento
FLG_CAMPAÑA	Indicador del tipo de campaña según estacionalidad
CANALVENTA	Canal de venta
CANALEVALUACION	Canal de evaluación
TIPESTJUSTIFICACIÓN	ID del tipo de justificación para solicitudes denegadas
DESJUSTIFICACIÓN	Justificación: Nivel 3, para solicitudes denegadas
DESJUS_AGRUPACIÓN	Justificación: Nivel 2, para solicitudes denegadas
GRUPOJUSTIFICACIÓN	Justificación: Nivel 1, para solicitudes denegadas
APROB_CANALEVALUACION	Indicador único del estado y el canal de aprobación
CODCLAVECIC	Identificador del préstamo
FLG_DESEMBOLSO	"S" si el préstamo fue desembolsado, "N" caso contrario
FLG_APROBACION	"S" si el préstamo fue aprobado, "N" caso contrario
FLG_APROBACION_RBM	"S" si el préstamo fue aprobado por RBM, "N" caso contrario
PD_CLIENTE	Probabilidad de default de la empresa
SEG_COM	Segmento comercial según banco
SEGMENTO	Segmento según clasificación de riesgo ente supervisor

Elaboración Propia.

Completando la descripción estadística de los campos con el tipo, la escala, la cantidad de valores únicos y la cantidad de valores faltantes en la tabla siguiente:

Tabla 47. Descripción estadística de la base de datos.

Campo	Tipo	Escala	%Faltantes
CODMES	Cualitativa	Ordinal	0
MES	Cuantitativa Discreta	Nominal	0
TIPO_CAMPANA	Cualitativa	Nominal	0
ESTADO	Cualitativa	Nominal	0
MONEDA	Cualitativa	Nominal	0
MTOAPROBADO	Cuantitativa Continua	De razón	0
MONTODOLARES	Cuantitativa Continua	De razón	0
MONTOSOLES	Cuantitativa Continua	De razón	0
TASA	Cuantitativa Continua	De razón	0

DESSECTORECONOMICO	Cualitativa	Nominal	0
CODVENDEDOR	Cualitativa	Nominal	0
CODAPROBADOR	Cualitativa	Nominal	0
FLGGAR	Cualitativa	Nominal	24
DESCAMPANIA	Cualitativa	Nominal	0
TIPESTCONTRATOSOLICITUDMIC	Cualitativa	Nominal	5
FLGTIPPER	Cualitativa	Nominal	13
MTOCAMBIOALNUEVOSOL	Cuantitativa Continua	De razón	0
MARCA	Cualitativa	Nominal	0
ZONA	Cualitativa	Ordinal	0
AREA	Cualitativa	Ordinal	0
REGION	Cualitativa	Ordinal	0
CODOFI	Cualitativa	Nominal	0
DESCODDEPARTAMENTO	Cualitativa	Nominal	0
FLG_CAMPAÑA	Cualitativa	Nominal	0
CANALVENTA	Cualitativa	Nominal	0
CANALEVALUACION	Cualitativa	Nominal	0
TIPESTJUSTIFICACIÓN	Cualitativa	Nominal	0
DESJUSTIFICACIÓN	Cualitativa	Nominal	82
DESJUS_AGRUPACIÓN	Cualitativa	Nominal	82
GRUPOJUSTIFICACIÓN	Cualitativa	Nominal	82
APROB_CANALEVALUACION	Cualitativa	Nominal	0
CODCLAVECIC	Cualitativa	Nominal	0
FLG_DESEMBOLSO	Cualitativa	Nominal	0
FLG_APROBACION	Cualitativa	Nominal	0
FLG_APROBACION_RBM	Cualitativa	Nominal	0
PD_CLIENTE	Cuantitativa Continua	De razón	10
SEG_COM	Cualitativa	Ordinal	10
SEGMENTO	Cualitativa	Ordinal	10

Elaboración Propia.

Finalmente, se aplicarán tablas de enfrentamiento entre cada una de las variables y la variable que se quiere predecir, en este caso la variable “ESTADO”.

Estos resultados pueden visualizarse en el **Anexo 01**.

4.3. Preprocesamiento de la información

4.3.1. Limpieza de datos

Iniciamos con la limpieza de la base de datos tal y como se ha obtenido originalmente y en función al porcentaje de valores faltantes y criterios de importancia de la variable, se definirá un método de imputación. La tabla siguiente resume el procedimiento:

Tabla 48. Criterios de limpieza de datos.

Campo	% Faltantes	¿Quitar espacios?	¿Unificar a mayus?	¿Imputar?	¿Método?	Imputación
CODMES	0	1	0	0	--	No tiene
MES	0	1	0	0	--	No tiene
TIPO_CAMPANA	0	1	1	0	--	No tiene
ESTADO	0	1	1	0	--	No tiene
MONEDA	0	1	1	0	--	No tiene
MTOAPROBADO	0	1	0	0	--	No tiene
MONTODOLARES	0	1	0	0	--	No tiene
MONTOSOLES	0	1	0	0	--	No tiene

TASA	0	1	0	0	--	No tiene
DESSECTORECONOMICO	0	1	1	0	Valor	'No aplica'
CODVENDEDOR	0	1	1	0	--	No tiene
CODAPROBADOR	0	1	1	0	--	No tiene
FLGGAR	24	1	1	1	Valor	'No aplica'
DESCAMPANIA	0	1	1	0	--	No tiene
TIPESTCONTRATOSOLICITUDMIC	5	1	1	1	--	No tiene
FLGTIPPER	13	1	1	1	--	No tiene
MTOCAMBIOALNUEVOSOL	0	1	0	0	--	No tiene
MARCA	0	1	1	0	--	No tiene
ZONA	0	1	1	0	--	No tiene
AREA	0	1	1	0	--	No tiene
REGION	0	1	1	0	--	No tiene
CODOFI	0	1	1	0	--	No tiene
DESCODDEPARTAMENTO	0	1	1	0	--	No tiene
FLG_CAMPAÑA	0	1	1	0	--	No tiene
CANALVENTA	0	1	1	0	--	No tiene
CANALEVALUACION	0	1	1	0	--	No tiene
TIPESTJUSTIFICACIÓN	0	1	1	0	--	No tiene
DESJUSTIFICACIÓN	82	1	1	1	Valor	'No aplica'
DESJUS_AGRUPACIÓN	82	1	1	1	Valor	'No aplica'
GRUPOJUSTIFICACIÓN	82	1	1	1	Valor	'No aplica'
APROB_CANALEVALUACION	0	1	1	0	--	No tiene
CODCLAVECIC	0	1	1	0	--	No tiene
FLG_DESEMBOLSO	0	1	1	0	--	No tiene
FLG_APROBACION	0	1	1	0	--	No tiene
FLG_APROBACION_RBM	0	1	1	0	--	No tiene
PD_CLIENTE	10	1	0	1	Media	~SEGMENTO
SEG_COM	10	1	1	1	Valor	'Sin clasificación'
SEGMENTO	10	1	1	1	Valor	'Sin clasificación'

Elaboración Propia.

Asimismo, el código utilizado en R para dicho fin se muestra en el **Anexo 02**.

4.3.2. Transformación de variables

De acuerdo a la información descriptiva de las variables obtenida previamente y en función a criterios del modelo, se eliminarán o crearán nuevas variables:

Tabla 49. Criterios de transformación o eliminación de variables.

Campo	Acción[40]	Comentarios
CODMES	(3)	El año es igual para todos los registros. Dato recogido en la variable MES
MES	(1)	Correcto
TIPO_CAMPANA	(2)	DIA DE LA MADRE = DMA, ESCOLAR = ESC, FIESTAS PATRIAS = FPA, NAVIDAD = NAV
ESTADO	(2)	ESTADO: 'ACE' = 1 y 'DEN' = 0
MONEDA	(1)	Correcto
MTOAPROBADO	(1)	Correcto
MONTODOLARES	(1)	Correcto
MONTOSOLES	(1)	Correcto
TASA	(1)	Correcto
DESSECTORECONOMICO	(1)	Correcto
CODVENDEDOR	(4)	Nueva variable: FLG_VEND_APR
CODAPROBADOR		1 si la categoría del vendedor es igual a la del aprobador, 0 caso contrario.
FLGGAR	(1)	Correcto
DESCAMPANIA	(3)	No aporta al conjunto de variables

[40] (1) Mantener. Dejar la variable intacta.
(2) Modificar. Modificar algunos valores de la variable según comentarios.
(3) Eliminar. Eliminar la variable.
(4) Crear y eliminar. Usar una o más variables para crear otras nuevas, luego eliminar las antiguas.

TIPESTCONTRATOSOLICITUDMIC	(3)	No aporta al conjunto de variables
FLGTIPPER	(1)	Correcto
MTOCAMBIOALNUEVOSOL	(3)	No aporta al conjunto de variables
MARCA	(3)	No aporta al conjunto de variables
ZONA AREA REGION CODOFI DESCODDEPARTAMENTO	(4)	Crear: FLGTIPLOCAL: 'A' = Agen. y 'S'= Suc. ZONA_FINAL: Si DEPARTAMENTO <> 'Lima' tomar DEPARTAMENTO, caso contrario agrupar según criterio de agrupación
FLG_CAMPAÑA	(3)	No aporta al conjunto de variables
CANALVENTA	(2)	Funcionario sin cartera: FSC, Funcionario con cartera: FCC, Otros: OTR
CANALEVALUACION	(1)	Correcto
TIPESTJUSTIFICACION	(3)	Solo tienen datos cuando ESTADO = 'DEN', no aporta al modelo
DESJUSTIFICACION	(3)	Solo tienen datos cuando ESTADO = 'DEN', no aporta al modelo
DESJUS_AGRUPACION	(3)	Solo tienen datos cuando ESTADO = 'DEN', no aporta al modelo
GRUPOJUSTIFICACION	(3)	Solo tienen datos cuando ESTADO = 'DEN', no aporta al modelo
APROB_CANALEVALUACION	(3)	Similar a la variable ESTADO. No aporta al modelo.
CODCLAVECIC	(1)	Identificador del cliente
FLG_DESEMBOLSO	(3)	El préstamo ya es aprobado, puede estar pendiente el desembolso
FLG_APROBACION	(3)	Campos resultantes luego de la aplicación de la solicitud. No son útiles
FLG_APROBACION_RBM	(3)	Campos resultantes luego de la aplicación de la solicitud. No son útiles
PD_CLIENTE	(3)	Son datos internos del banco
SEG_COM	(2)	SEG_COM: 1 = S1, 2 = S2, 3 = S3, 4 = S4, SIN CLASIFICACION = SC
SEGMENTO	(2)	SEGMENTO: 1. A1 = A1, … , DEFAULT = F, INACTIVO = I, SIN CLASIFICACION = S

Elaboración Propia.

El código utilizado en R se encuentra en el **Anexo 03** debido a razones de espacio.

4.3.3. Eliminación de valores extraños o *outliers*

De acuerdo con la distribución de clientes según el anexo 01, se puede apreciar que aproximadamente un 10% de los registros corresponden a un único cliente, y examinando detenidamente a dicho cliente, se pudo notar que se trataba de un cliente que no se comporta como la distribución regular de las otras MYPEs (registraban diferentes sectores económicos y solo gestiones de préstamos bancarios en el mes 2). Con el fin de no afectar el modelamiento y sesgar los resultados a este mes en particular, se decidió eliminar todos los registros correspondientes a ese cliente.

```
[…]
#Paso 5. Eliminamos registros que puedan causar sesgo.
DB_Banco<-DB_Banco[CODCLAVECIC!="999999",]
DB_Banco<-DB_Banco[!is.na(DB_Banco$ESTADO),]
[…]
```

Figura 24. Código en R para eliminación de valores extraños.
Elaboración propia

Ahora que la base ha sido tratada y limpiada, es momento de hacer uso de técnicas de minería de datos para realizar un modelo apropiado.

4.4. Modelamiento de la información

4.4.1. Selección de técnica de modelado

Para definir la técnica de modelado es necesario analizar las características de la variable objetivo o la que se busca predecir. Esta variable se denomina “ESTADO” y está compuesta por valores dicotómicos 0 y 1. Corresponde, por tanto, utilizar modelos matemáticos de aprendizaje supervisado que permitan estimar dichos valores. Según Han (2012), se podrían utilizar:

- Regresión logística
- Redes neuronales artificiales

Se hará uso de ambas metodologías para luego compararlas y determinar aquella que sea más efectiva en términos de precisión y procesamiento computacional.

Sin embargo, previo al desarrollo computacional del modelo, se evaluó la distribución de valores positivos versus aquellos negativos, resultando en una relación de 4:1 entre sí. Para no afectar las ejecuciones de ambos modelos, se decidió tomar una sub-muestra de la data ya pre-procesada, de manera que la proporción sea equitativa.

4.4.2. Aplicación de técnicas de modelado

Regresión Logística

Utilizando el código que se indica en el **Anexo 04** correspondiente al modelamiento, y la muestra de entrenamiento, se obtienen los resultados del modelo:

```
Call:
glm(formula = ESTADO ~ ., family = binomial, data = DB_Banco[s_train,
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8227  -1.1526  -0.2782   1.1000   2.5962

Coefficients: (10 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.177e+00  2.323e+00   1.798 0.072178 .
MES                 -8.862e-02  2.564e-02  -3.457 0.000546 ***
TASA                -2.095e-03  3.553e-03  -0.590 0.555475
MTOAPROBADO          5.780e-06  1.850e-05   0.312 0.754775
MONTODOLARES        -1.799e-04  9.635e-05  -1.868 0.061829 .
MONTOSOLES           4.576e-05  3.414e-05   1.340 0.180136
```

```
TIPO_CAMPANA_g1        -5.003e-01  1.022e-01  -4.894 9.87e-07 ***
TIPO_CAMPANA_g2         2.586e-01  9.314e-02   2.776 0.005496 **
TIPO_CAMPANA_g3         2.018e-01  1.354e-01   1.490 0.136135
TIPO_CAMPANA_g4                NA         NA      NA       NA
DESSECTORECONOMICO_g1 -1.534e+00  1.103e+00  -1.391 0.164317
DESSECTORECONOMICO_g2 -1.487e+00  1.101e+00  -1.350 0.176970
DESSECTORECONOMICO_g3 -1.595e+00  1.102e+00  -1.448 0.147628
DESSECTORECONOMICO_g4 -1.770e+00  1.156e+00  -1.531 0.125760
DESSECTORECONOMICO_g5          NA         NA      NA       NA
MONEDA_g1               1.740e-03  1.672e+00   0.001 0.999169
MONEDA_g2                      NA         NA      NA       NA
SEG_COM_g1             -6.861e-02  6.873e-01  -0.100 0.920479
SEG_COM_g2              2.810e-02  6.860e-01   0.041 0.967333
SEG_COM_g3             -3.098e-02  6.864e-01  -0.045 0.964003
SEG_COM_g4              1.840e-01  6.953e-01   0.265 0.791281
SEG_COM_g5                     NA         NA      NA       NA
SEGMENTO_g1             6.338e-01  4.597e-01   1.379 0.168015
SEGMENTO_g2             6.569e-01  4.627e-01   1.420 0.155704
SEGMENTO_g3             7.424e-01  4.614e-01   1.609 0.107595
SEGMENTO_g4             6.831e-01  4.648e-01   1.470 0.141616
SEGMENTO_g5             3.024e-01  4.749e-01   0.637 0.524194
SEGMENTO_g6             6.582e-01  5.293e-01   1.244 0.213671
SEGMENTO_g7             4.772e-01  4.916e-01   0.971 0.331671
SEGMENTO_g8                    NA         NA      NA       NA
SEGMENTO_g9                    NA         NA      NA       NA
FLGGAR_g1              -2.572e+00  2.418e-01 -10.634  < 2e-16 ***
FLGGAR_g2               2.784e-01  7.991e-02   3.484 0.000494 ***
FLGGAR_g3                      NA         NA      NA       NA
FLGTIPPER_g1           -2.429e+00  2.527e-01  -9.612  < 2e-16 ***
FLGTIPPER_g2           -3.032e+00  2.553e-01 -11.879  < 2e-16 ***
FLGTIPPER_g3                   NA         NA      NA       NA
FLGTIPLOCAL_g1          5.418e-02  5.818e-02   0.931 0.351726
FLGTIPLOCAL_g2                 NA         NA      NA       NA
ZONA_FINAL_g1          -7.861e-02  1.021e+00  -0.077 0.938644
ZONA_FINAL_g2           2.910e-01  1.025e+00   0.284 0.776467
ZONA_FINAL_g3           1.851e-01  1.026e+00   0.180 0.856910
ZONA_FINAL_g4           1.640e-01  1.023e+00   0.160 0.872664
ZONA_FINAL_g5           9.036e-02  1.063e+00   0.085 0.932260
ZONA_FINAL_g6           3.159e-01  1.025e+00   0.308 0.758069
ZONA_FINAL_g7           8.805e-01  1.073e+00   0.821 0.411795
ZONA_FINAL_g8                  NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9645.7  on 6957  degrees of freedom
Residual deviance: 9173.0  on 6921  degrees of freedom
AIC: 9247

Number of Fisher Scoring iterations: 5
```

Figura 25. Reporte de regresión logística, iteración inicial.
Elaboración Propia

Utilizamos el algoritmo de optimización hacia atrás para conseguir un modelo reducido con un poder predictivo similar, basado en el indicador AIC, teniendo en cuenta las correlaciones entre variables:

```
Call:
glm(formula = ESTADO ~ MES + TIPO_CAMPANA_g1 + TIPO_CAMPANA_g2 +
    TIPO_CAMPANA_g3 + DESSECTORECONOMICO_g1 + DESSECTORECONOMICO_g2 +
    DESSECTORECONOMICO_g3 + DESSECTORECONOMICO_g4 + SEG_COM_g2 +
    SEG_COM_g4 + SEGMENTO_g1 + SEGMENTO_g2 + SEGMENTO_g3 + SEGMENTO_g4 +
    FLGGAR_g1 + FLGGAR_g2 + FLGTIPPER_g1 + FLGTIPPER_g2 + ZONA_FINAL_g2 +
```

```
    ZONA_FINAL_g3 + ZONA_FINAL_g4 + ZONA_FINAL_g6 + ZONA_FINAL_g7,
    family = binomial, data = DB_Banco[s_train, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8137  -1.1549  -0.2782   1.1051   2.5850

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.29217    1.13881   3.769 0.000164 ***
MES                  -0.09581    0.02523  -3.798 0.000146 ***
TIPO_CAMPANA_g1      -0.41497    0.09232  -4.495 6.95e-06 ***
TIPO_CAMPANA_g2       0.27289    0.09252   2.950 0.003183 **
TIPO_CAMPANA_g3       0.27542    0.12550   2.195 0.028193 *
DESSECTORECONOMICO_g1 -1.58911    1.10419  -1.439 0.150106
DESSECTORECONOMICO_g2 -1.53336    1.10274  -1.391 0.164377
DESSECTORECONOMICO_g3 -1.62890    1.10323  -1.476 0.139814
DESSECTORECONOMICO_g4 -1.84383    1.15707  -1.594 0.111039
SEG_COM_g2            0.08770    0.05908   1.485 0.137667
SEG_COM_g4            0.22954    0.13424   1.710 0.087273 .
SEGMENTO_g1           0.24450    0.10149   2.409 0.015994 *
SEGMENTO_g2           0.27765    0.11550   2.404 0.016220 *
SEGMENTO_g3           0.36679    0.10917   3.360 0.000780 ***
SEGMENTO_g4           0.30968    0.12206   2.537 0.011174 *
FLGGAR_g1            -2.58303    0.24162 -10.691  < 2e-16 ***
FLGGAR_g2             0.21670    0.07163   3.025 0.002483 **
FLGTIPPER_g1         -2.37297    0.25123  -9.446  < 2e-16 ***
FLGTIPPER_g2         -2.99384    0.25386 -11.793  < 2e-16 ***
ZONA_FINAL_g2         0.34396    0.09812   3.506 0.000456 ***
ZONA_FINAL_g3         0.23012    0.11339   2.029 0.042414 *
ZONA_FINAL_g4         0.21954    0.08287   2.649 0.008071 **
ZONA_FINAL_g6         0.36611    0.10515   3.482 0.000498 ***
ZONA_FINAL_g7         0.97420    0.32894   2.962 0.003060 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9645.7  on 6957  degrees of freedom
Residual deviance: 9186.0  on 6934  degrees of freedom
AIC: 9234

Number of Fisher Scoring iterations: 5
```

Figura 26. Reporte de regresión logística optimizado según AIC.
Elaboración Propia

Analizando los resultados, determinamos que existen variables que no aportan al modelo por: no tener un valor-p mayor a 0.05 o estar correlacionadas con otras variables. Por tanto, se replantea el modelamiento eliminando las variables:

- SEG_COM
- SEGMENTO

Finalmente, se obtiene los resultados del modelo de regresión final:

```
Call:
glm(formula = ESTADO ~ ., family = binomial, data = DB_Banco[s_train,
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8018  -1.1539  -0.2856   1.1072   2.5120
```

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.11174    0.27503  11.314  < 2e-16 ***
MES                   -0.09452    0.02515  -3.758 0.000171 ***
TIPO_CAMPANA_g1       -0.41926    0.09213  -4.551 5.35e-06 ***
TIPO_CAMPANA_g2        0.30168    0.09199   3.279 0.001040 **
TIPO_CAMPANA_g3        0.29650    0.12501   2.372 0.017699 *
DESSECTORECONOMICO_g1  0.03103    0.07868   0.394 0.693290
DESSECTORECONOMICO_g2  0.09336    0.05854   1.595 0.110768
DESSECTORECONOMICO_g4 -0.20991    0.35293  -0.595 0.552001
DESSECTORECONOMICO_g5  1.64829    1.10245   1.495 0.134886
FLGGAR_g1             -2.75424    0.24672 -11.163  < 2e-16 ***
FLGGAR_g3             -0.16394    0.06568  -2.496 0.012560 *
FLGTIPPER_g1          -2.35988    0.25086  -9.407  < 2e-16 ***
FLGTIPPER_g2          -2.99656    0.25375 -11.809  < 2e-16 ***
ZONA_FINAL_g2          0.35285    0.09767   3.613 0.000303 ***
ZONA_FINAL_g3          0.22726    0.11327   2.006 0.044811 *
ZONA_FINAL_g4          0.21009    0.08263   2.542 0.011007 *
ZONA_FINAL_g5          0.11860    0.29553   0.401 0.688197
ZONA_FINAL_g6          0.37141    0.10490   3.541 0.000399 ***
ZONA_FINAL_g7          0.97806    0.32880   2.975 0.002934 **
ZONA_FINAL_g8          0.06200    1.01591   0.061 0.951337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9645.7  on 6957  degrees of freedom
Residual deviance: 9202.2  on 6938  degrees of freedom
AIC: 9242.2

Number of Fisher Scoring iterations: 5
```

Figura 27. Reporte de regresión logística, modelo final.
Elaboración Propia

La ecuación de la recta de regresión que corresponde a este modelo es:

$$\ln\left(\frac{\pi}{1-\pi}\right) = 3.11 - 0.09x_1 - 0.42x_2 + 0.30x_3 + 0.30x_4 + 0.03x_5 + 0.09x_6 - 0.20x_7 + 1.65x_8 - 2.75x_9 - 0.16x_{10} - 2.36x_{11} - 2.99x_{12} + 0.35x_{13} + 0.23x_{14} + 0.21x_{15} + 0.12x_{16} + 0.37x_{17} + 0.98x_{18} + 0.06x_{19}$$

Donde:

x_1: Mes, en formato numérico.

x_2: 1 cuando TIPO_CAMPANA = “DMA”, 0 caso contrario.

x_3: 1 cuando TIPO_CAMPANA = “ESC”, 0 caso contrario.

x_4: 1 cuando TIPO_CAMPANA = “FPA”, 0 caso contrario.

x_5: 1 cuando DESSECTORECONOMICO = “COMERCIO”, 0 caso contrario.

x_6: 1 cuando DESSECTORECONOMICO = “INDUSTRIA”, 0 caso contrario.

x_7: 1 cuando DESSECTORECONOMICO = “OTROS”, 0 caso contrario.

x_8: 1 cuando DESSECTORECONOMICO = “SERVICIO”, 0 caso contrario.

x_9: 1 cuando FLGGAR = “N”, 0 caso contrario.

x_{10}: 1 cuando FLGGAR = “S”, 0 caso contrario.

x_{11}: 1 cuando FLGTIPPER = “J”, 0 caso contrario.

x_{12}: 1 cuando FLGTIPPER = “N”, 0 caso contrario.

x_{13}: 1 cuando ZONAFINAL = “LIMA ESTE”, 0 caso contrario.

x_{14}: 1 cuando ZONAFINAL = “LIMA MODERNA”, 0 caso contrario.

x_{15}: 1 cuando ZONAFINAL = “LIMA NORTE”, 0 caso contrario.

x_{16}: 1 cuando ZONAFINAL = “LIMA PROVINCIAS”, 0 caso contrario.

x_{17}: 1 cuando ZONAFINAL = “LIMA SUR”, 0 caso contrario.

x_{18}: 1 cuando ZONAFINAL = “LIMA TOP”, 0 caso contrario.

x_{19}: 1 cuando ZONAFINAL = “LIMA PROVINCIAS”, 0 caso contrario.

Regresión Logística: Métricas

Dado que la regresión logística devuelve como resultado de predicción valores de probabilidad, se necesita establecer un punto de corte o *cutoff* que determina el límite para catalogar a un determinado registro como 0 o como 1.

Para un modelo de regresión logística balanceado[41], el punto de corte típico es de 0.5, sin embargo, para modelos de regresión logística no balanceados, el punto de corte depende fuertemente del objetivo de uso del modelo y cuan riesgoso es catalogar erróneamente 1 como 0 o viceversa.

Considerando estos puntos, y tomando en cuenta que para la construcción del modelo se seleccionó aleatoriamente una sub-muestra balanceada, analizaremos la eficiencia del modelo utilizando las métricas más conocidas para los modelos de regresión logística: precisión, sensibilidad, especificidad, y recall, así como también

[41] Hace referencia a que aproximadamente el 50% de los valores de la variable objetivo son 1 y 0.

el indicador GINI, tanto en la muestra de entrenamiento como en la muestra de prueba:

Tabla 50. Indicadores para distintos valores de cutoff, entrenamiento

Cutoff	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
Precisión	50%	53%	53%	53%	57%	59%	55%	51%	50%	50%	50%
Sensibilidad	100%	100%	99%	99%	91%	66%	22%	1%	0%	0%	0%
Especificidad	0%	6%	7%	7%	22%	52%	87%	99%	100%	100%	100%
Score F1	66%	68%	43%	68%	68%	61%	33%	3%	0%	0%	0%

Elaboración Propia.

Tabla 51. Indicadores para distintos valores de cutoff, prueba

Cutoff	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
Precisión	50%	54%	55%	84%	56%	59%	53%	49%	49%	49%	49%
Sensibilidad	100%	100%	100%	96%	89%	66%	22%	1%	0%	0%	0%
Especificidad	0%	7%	8%	25%	21%	51%	84%	99%	100%	100%	100%
Score F1	66%	63%	69%	40%	40%	62%	33%	2%	0%	0%	0%

Elaboración Propia.

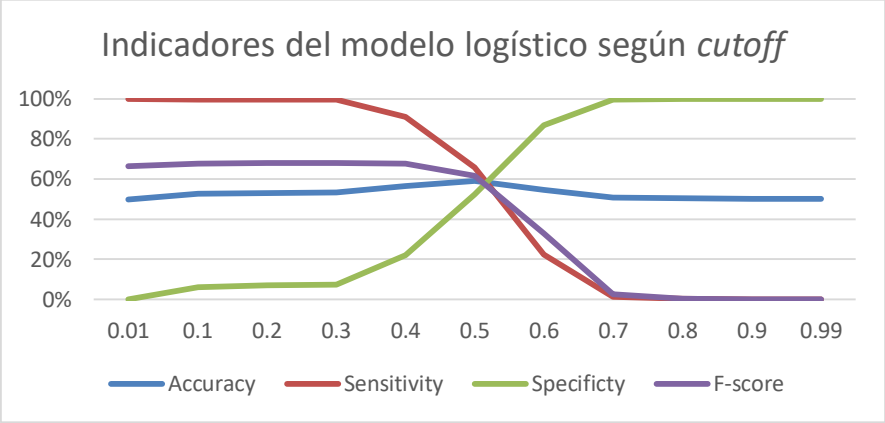


Figura 28. Sensibilidad de indicadores del modelo logístico, entrenamiento
Elaboración Propia.

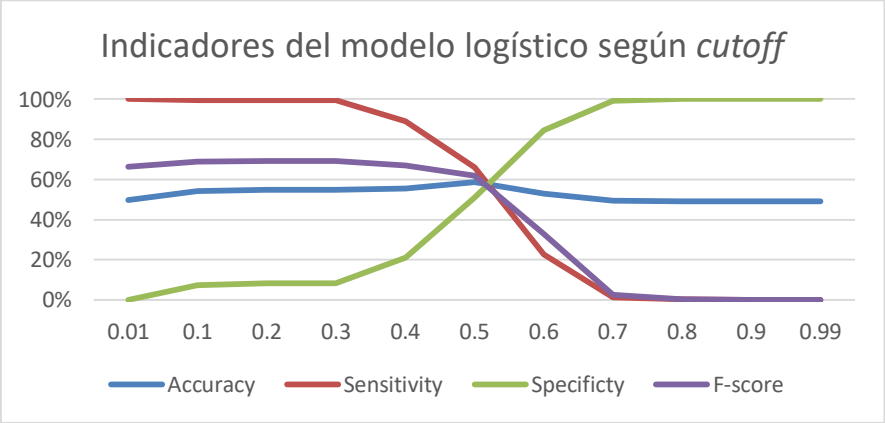


Figura 29. Sensibilidad de indicadores del modelo logístico, prueba
Elaboración Propia.

Para determinar un valor "óptimo" del punto de corte en el modelo logístico, existen 2 opciones que se pueden evaluar:

- Tomar el punto de corte que tenga las mejores métricas en general para todos los indicadores de recesión logística o,
- Tomar un punto que tenga en cuenta la mínima penalización total por los errores tipo I y tipo II.

Se debe analizar cuál de los errores resultaría más perjudicial para el objetivo del modelo, que es facilitar a una MYPE el acceso al financiamiento bancario.

- Cometer el error tipo I: Significa que bajo el modelo matemático la MYPE no tendrá acceso al préstamo cuando en la realidad si podría acceder al préstamo. La pérdida del préstamo representa un costo de oportunidad al no poder concretar el negocio para el cual ha suscrito un préstamo, sin embargo, dependiendo de la complejidad de los factores que hacen que la MYPE en el momento de la solicitud no pueda acceder a un préstamo, podría subsanarlos rápidamente e intentar con otra solicitud probablemente poco tiempo después.
- Cometer el error tipo II: Significa que bajo el modelo matemático la MYPE si tendrá acceso al préstamo cuando en la realidad no accedería al financiamiento. Es probable que, dada dicha información, la MYPE busque adelantar las gestiones para concretar eficientemente el negocio (contratar proveedores, comprar materiales, contratar mano de obra eventual, contratar transportistas, asegurar pedidos a clientes, etc.) con la premisa de que el dinero del préstamo "ya está en su poder". Sin embargo, de acuerdo al modelo, la MYPE no recibiría dicho préstamo y perdería no solo la oportunidad de negocio, sino también importantes relaciones con sus *stakeholders*, lo cual es fundamental para todas las MYPEs en crecimiento.

Según las condiciones de la muestra, se asume sin perdida de generalidad que la sub-muestra balanceada es representativa, por lo que el cutoff se mantiene en 0.5, dado que con este valor se consigue maximizar las métricas de sensibilidad y especificidad.

Tabla 52. Indicadores finales del modelo de regresión logística.

Indicadores	Entrenamiento	Prueba
Exactitud	58.98%	58.62%
Sensibilidad	65.65%	66.10%
Especificidad	52.38%	50.88%
Score F1	61.44%	61.90%
AUC	62.32%	61.15%
GINI	24.64%	22.29%

Elaboración Propia.

Redes Neuronales Artificiales

Una de las condiciones necesarias para que el algoritmo de redes neuronales pueda ejecutarse es que se los datos se encuentren normalizados, esto es, que aquellas variables con valores numéricos sean normalizadas a un rango entre 0 y 1, y que las variables categóricas sean discretizadas con indicadores binarios.

Usando el código del **Anexo 07** correspondiente al nuevo procesamiento de las variables y al modelamiento de la red neuronal, se tienen los siguientes resultados:

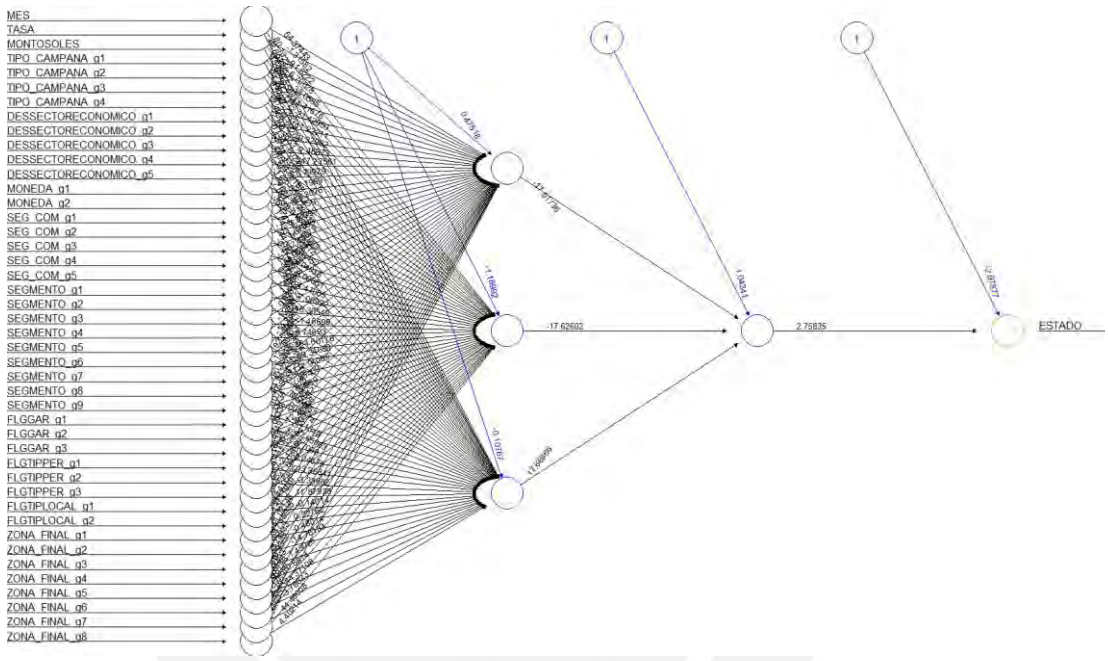


Figura 30. Red neuronal artificial de 2 capas, data de entrenamiento.
Elaboración Propia.

Se utilizará el mismo *cutoff* obtenido en el análisis del modelo de regresión logística.

Tabla 53. Indicadores finales del modelo de red neuronal artificial.

Indicadores	Entrenamiento	Prueba
Exactitud	60.05%	57.47%
Sensibilidad	70.67%	69.38%
Especificidad	49.51%	45.15%
Score F1	63.78%	62.40%
AUC	65.78%	61.90%
GINI	31.57%	23.80%

Elaboración Propia.

Comparación de modelos

Los modelos fueron ejecutados usando una PC con 16 GB de memoria RAM, sistema operativo Windows 10 e interface R-Studio bajo compilador C++. En ambos modelos

el ratio de entrenamiento a prueba fue de 80 contra 20, seleccionados aleatoriamente.

En general, ambos modelos producen resultados similares para un mismo *cutoff*. Si bien es cierto la red neuronal logra destacar en algunos puntos sus indicadores, su ejecución toma un tiempo relativamente mayor respecto del modelo de regresión logística,

Aún así, el método heurístico de optimización que utiliza este algoritmo le brinda la capacidad de trabajar con otros tipos de datos (de diferente magnitud) y su salida – vista como una "ponderación" de la influencia de otras variables – brinda una fácil conversión a una interpretación de resultados traducible en una escala de 0 a 1. Así mismo, es sujeto a trabajar directamente con una muestra no balanceada, puesto que los pesos se autocorregirán tanto en función a la muestra como a la optimización del *output* del modelo.

Por esa razón el modelo de redes neuronales resulta más efectivo.

4.4.3. Implementación de resultados de minería de datos

Utilizando la información sobre los pesos de cada atributo dentro del modelo de red neuronal artificial, se puede construir una interfaz gráfica a través de una app en móviles. El usuario – en este caso el gerente de la MYPE – puede introducir las características propias de su empresa que coincidan con las variables que el modelo de red neuronal utiliza para estimar los resultados.

Figura 31. UI para reporte de probabilidad de acceso.
Elaboración Propia.

Figura 32. Reporte de microfinanciamiento, parte 1.
Elaboración Propia.

Figura 33. Reporte de microfinanciamiento, parte 2.
Elaboración Propia.

CAPÍTULO 5. EVALUACIÓN ECONÓMICA

En este capítulo se calculará el costo del desarrollo del modelo completo incluyendo la aplicación construida a partir de los resultados del algoritmo de minería de datos y su precio aproximado de venta netamente para el recupero de la inversión, mas no para fines de lucro pues el objetivo también es apoyar a la MYPE. La proyección se tomará para los próximos 5 años considerando que el ahorro se calcula con respecto a un escenario de incertidumbre donde la MYPE según su punto de vista tendría igualdad de probabilidades de acceder a un crédito.

5.1. Costo de implementación

Según WageIndicator Foundation (2017) el salario medio mensual de un programador de aplicaciones es de S/: 3033, y se estima un tiempo de desarrollo de 1 mes. El software R y su interface R-studio son gratuitos y por ende no se contabilizan en el flujo de inversión. Adicionalmente se incluye la licencia de entrada a Google Play Store, pues el aplicativo se orienta a sistemas operativos instalados en equipos móviles de bajo costo. Los costos de operación – entiéndase consumo eléctrico, viáticos, etc. – se incluyen también en la **Tabla 50**:

Tabla 54. Inversión en modelado y aplicación móvil.

Costos	Total por concepto
Desarrollador	S/3,033.00
Licencia Google Play	S/81.40
Consumo operativo	S/54.13
Total	**S/3,189.88**

Elaboración Propia.

Dado que la finalidad del proyecto es apoyar a las MYPEs en lograr el financiamiento bancario, se determina un precio de uso del servicio sin margen de ganancia a partir de aplicaciones con funcionalidades similares o complementarias a la propuesta del proyecto.

Tabla 55. Aplicaciones similares.

Aplicaciones Similares	Precio (S/.)
Mi Score Equifax ® - Consulta otras empresas	6.66
Sentinel PyME – Plan 30 Consulta Unitaria	5.00
Promedio	**5.83**

Elaboración Propia.

También existen aplicaciones gratuitas como la aplicación de la SBS que permiten visualizar y exportar la información. A pesar de la existencia de estas aplicaciones, la aplicación móvil elaborada busca reemplazar a estas aplicaciones brindando un reporte de acceso al crédito personalizado por cada MYPE. Para recuperar la inversión, considerando el precio promedio sin margen, el número de usuarios registrados y pagadores mínimo para lograr el punto de equilibrio se muestra a continuación, asumiendo que por cada MYPE solo se tendrá un usuario – el gerente general usualmente –.

$$\frac{S/.\ 3189.88}{S/.\ 5.83} = 548\ usuarios$$

5.2. Costo de implementación para la MYPE

El costo de implementación para una MYPE incluye, adicional al valor del aplicativo, una capacitación en temas de administración de riesgos y *creditscoring*, de manera que el gerente de la microempresa pueda comprender los conceptos detrás del cálculo del indicador de predicción de acceso al financiamiento bancario. El costo de esta capacitación es de S/. 1800, por 24 horas de capacitación para una persona.

De ser necesario, se puede imprimir con un costo de S/. 10 el manual de usuario.

Tabla 56. Inversión para la MYPE.

Concepto de Inversión	Costos
Aplicación	S/. 5.83
Capacitación Scoring de Crédito	S/. 1800.00
Materiales de Implementación - Manual	S/. 10.00
Total	**S/. 1815.83**

Elaboración Propia.

Para determinar el concepto de beneficios del proyecto de minería de datos, se debe analizar las diferencias entre el importe recibido estimado por el préstamo sin acceso al reporte del aplicativo versus el importe recibido estimado con la información del aplicativo. Se examinaron 3 diferentes escenarios sobre las posibilidades del impacto de este cálculo. En un escenario normal, se asume que la probabilidad de acceso al crédito sin la información del aplicativo es de 50%; en uno pesimista; de 60% y en uno optimista; de 40%. Tomando como referencia una MYPE del rubro servicios, la probabilidad de acceso al crédito con el modelo es de 83.67%.

Tabla 57. Ahorro potencial con el uso del aplicativo móvil.

Escenarios	Prob. acceso al crédito, sin aplicativo	Prob. acceso al crédito, con aplicativo	Beneficio potencial
Optimista	40%	83.65%	43.65%
Normal	50%	83.65%	33.65%
Pesimista	60%	83.65%	22.65%

Elaboración Propia.

Considerando esta información, el flujo económico para la inversión realizada por esta MYPE, asumiendo un préstamo promedio de capital de trabajo de S/. 5,000 cada año durante 5 años, es el siguiente:

Tabla 58. Flujo económico y financiero del proyecto por escenario.

Optimista	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Ahorro		S/2,183	S/2,183	S/2,183	S/2,183	S/2,183
Egreso	S/1,816					
Flujo Económico	-S/1,816	S/2,183	S/2,183	S/2,183	S/2,183	S/2,183
Financiamiento	S/363					
Amortización		-S/27.04	-S/40.83	-S/61.65	-S/93.09	-S/140.56
Intereses		-S/185.21	-S/171.43	-S/150.60	-S/119.16	-S/71.69
Flujo Financiero	-S/1,453	S/1,970	S/1,970	S/1,970	S/1,970	S/1,970
Normal	**Año 0**	**Año 1**	**Año 2**	**Año 3**	**Año 4**	**Año 5**
Ahorro		S/1,683	S/1,683	S/1,683	S/1,683	S/1,683
Egreso	S/1,816					
Flujo Económico	-S/1,816	S/1,683	S/1,683	S/1,683	S/1,683	S/1,683
Financiamiento	S/363					
Amortización		-S/27.04	-S/40.83	-S/61.65	-S/93.09	-S/140.56
Intereses		-S/185.21	-S/171.43	-S/150.60	-S/119.16	-S/71.69
Flujo Financiero	-S/1,453	S/1,470	S/1,470	S/1,470	S/1,470	S/1,470
Pesimista	**Año 0**	**Año 1**	**Año 2**	**Año 3**	**Año 4**	**Año 5**
Ahorro		S/1,183	S/1,183	S/1,183	S/1,183	S/1,183
Egreso	S/1,816					
Flujo Económico	-S/1,816	S/1,183	S/1,183	S/1,183	S/1,183	S/1,183
Financiamiento	S/363					
Amortización		-S/27.04	-S/40.83	-S/61.65	-S/93.09	-S/140.56
Intereses		-S/185.21	-S/171.43	-S/150.60	-S/119.16	-S/71.69
Flujo Financiero	-S/1,453	S/970	S/970	S/970	S/970	S/970

Elaboración Propia.

Se incluyó un flujo financiero en la tabla anterior si es que la MYPE desease realizar un micro préstamo con el banco para cubrir parte de la inversión. Para determinar la tasa de interés de la deuda se realizó una simulación en base al monto solicitado con bancos y cajas resultando una tasa promedio del alrededor del 51% anual, de acuerdo a Gestión (2018[b]).

Asimismo, la modalidad de financiamiento se reparte entre 20% financiamiento con bancos y 80% capital propio.

Para determinar la viabilidad del proyecto se calculó el costo de oportunidad del capital (COK) exclusivamente con aporte propio y el costo ponderado de capital

(WACC) de darse la modalidad de financiamiento entre banca y aporte propio, usando las formulas siguientes:

$$COK = Tasa\ libre\ de\ riesgo + Prima\ de\ riesgo\ x\ Beta\ apalancado + Riesgo\ Pais$$

$$WACC = \%\ Aporte\ Propio\ x\ COK + \%\ Financiado\ Deuda\ x\ Tasa\ Interés\ x\ (1 - I.Renta\ MYPE)$$

Para el cálculo del Beta apalancado se debe considerar el Beta desapalancado según la industria y apalancarlo según la estructura de financiamiento de la MYPE.

$$Beta\ Apalancado = \left(1 + \frac{\%\ Financiado\ Deuda}{\%\ Aporte\ Propio}\ x\ (1 - I.Renta\ MYPE)\right) x\ Beta\ Desapalancado$$

El valor del Beta desapalancado por rubro de empresa se extrajo de la página web de Aswath Damodarán (s/f). En la **Tabla 55** se muestra el Beta desapalancado por rubros de empresas que correspondan a actividades de servicio.

Tabla 59. Beta desapalancado por rubro de empresa.

Nombre de la industria	Beta desapalancado
Repuestos de Autos	0.85
Servicios de negocios y consumo	0.97
Equipos y servicios de oficina	0.99
Restaurante	0.68
Retail	0.85
Servicios de software (aplicaciones)	0.98
Promedio	**0.89**

Elaboración Propia.

Con esta información se procede a calcular el Beta apalancado, COK y WACC:

Tabla 60. Beta apalancado para la estructura de financiamiento.

Concepto	Valor
Tasa libre de riesgo	9.35 pp
Prima de riesgo	3.81 pp
Beta desapalancado	0.89
Riesgo Pais	1.21 pp
% Financiado Deuda / % Aporte Propio	2/8
Impuesto a la renta MYPE	10%
Beta apalancado	**1.09**

Fuente: Gestión, Trading Economics.

Tabla 61. COK y WACC.

Concepto	Valor
Beta apalancado	1.09
Costo deuda	51%
COK	**14.71%**
WACC	**20.95%**

Fuente: Gestión, Trading Economics.

La Tasa interna de Retorno (TIR) se calcula para cada escenario de acuerdo a la estructura de financiamiento escogida. Para cada alternativa, se muestran las TIRs:

Tabla 62. TIRs para cada escenario y estrategia de financiamiento.

Escenarios	COK	WACC
Escenario optimista	118%	134%
Escenario normal	89%	98%
Escenario pesimista	59%	59%

Elaboración Propia.

Por último, analizamos la sensibilidad del proyecto según la estructura de financiamiento:

Tabla 63. Sensibilidad del VAN del proyecto

Estructura de Financiamiento	100% aporte propio			20 % deuda 80 % aporte propio		
Tasa Descuento	VAN optimista	VAN normal	VAN pesimista	VAN optimista	VAN normal	VAN pesimista
0%	S/9,097	S/6,597	S/4,097	S/8,399	S/5,899	S/3,399
10%	S/5,871	S/4,147	S/2,424	S/5,469	S/3,746	S/2,023
14.71%	S/4,839	S/3,368	S/1,896	S/4,531	S/3,060	S/1,589
20%	S/3,926	S/2,680	S/1,434	S/3,700	S/2,454	S/1,207
20.95%	S/3,784	S/2,573	S/1,362	S/3,570	S/2,360	S/1,149
30%	S/2,692	S/1,755	S/819	S/2,574	S/1,637	S/700
40%	S/1,876	S/1,149	S/422	S/1,827	S/1,100	S/373
50%	S/1,316	S/737	S/158	S/1,313	S/734	S/155
60%	S/922	S/451	-S/21	S/949	S/478	S/6
70%	S/637	S/246	-S/144	S/685	S/294	-S/97
80%	S/427	S/98	-S/231	S/489	S/160	-S/169
90%	S/269	-S/12	-S/292	S/341	S/61	-S/220
100%	S/149	-S/93	-S/335	S/228	-S/14	-S/256

Elaboración Propia.

Se observa que bajo todos los escenarios y en las 2 estructuras de financiamiento, el proyecto es rentable, sin embargo, es sensible a la tasa de interés de la deuda contraída, si se escogiese dicha modalidad de financiamiento.

CAPÍTULO 6. CONCLUSIONES Y RECOMENDACIONES

Finalmente, en este capítulo se detallarán las conclusiones y recomendaciones producto de la realización del presente trabajo de investigación. Asimismo, se brindarán sugerencias que permitan ampliar la investigación más adelante.

6.1. Conclusiones

- El proyecto de minería de datos resultó ser rentable al obtener un VAN estimado para un escenario normal de S/3368 con financiamiento completamente con fondos propios (COK) y S/2360 con una estructura de financiamiento 20% deuda y 80% recursos propios (WACC), mientras que para el escenario pesimista estos indicadores apenas se redujeron a la mitad. La diferencia entre los VAN de cada estructura de financiamiento se ve fuertemente impactada por el alto costo de deuda que existe actualmente para las empresas MYPE.

- Este proyecto estima un beneficio esperado de un 33.7% con ayuda de la información obtenida a partir del aplicativo móvil basado en el algoritmo de redes neuronales, lo cual multiplica la probabilidad para que una MYPE pueda acceder al financiamiento bancario hasta en 1.6 veces, en un escenario normal.

- Otras alternativas de algoritmos de minería de datos que pudieron haberse utilizado para el mismo propósito fueron los arboles de decisión o los clasificadores ingenuos de Bayes. Sin embargo, y a pesar de que el tiempo de entrenamiento es mucho mayor a comparación de los algoritmos antes mencionados, se escogió utilizar las redes neuronales por la flexibilidad de reconocer patrones con data de diferente calidad (con ruido, distorsionada o mal escalada) y su capacidad para adaptarse muy fácilmente a nuevos ingresos de data calibrando únicamente los pesos de cada una de las variables con dicha nueva información.

- La finalidad de este trabajo de investigación fue servir como base para el desarrollo de futuras aplicaciones móviles – sostenidas con cualquier herramienta de minería de datos que aplique a la situación – capaces de lograr que el desempeño de la MYPE en el Perú mejore continuamente en todos sus aspectos (productividad, calidad, manejo estratégico y financiero, entre otros) y demostrar que la minería de datos forma parte ya de la revolución tecnológica actual.

6.2. Recomendaciones

- Para ayudar a la comprensión de los resultados del aplicativo móvil, sería útil utilizar un semáforo de probabilidad de acceso al crédito, pues el apoyo visual le permitiría conocer rápidamente que si es verde tiene una buena probabilidad y si es rojo tiene una probabilidad muy baja. Asimismo, incluir de manera anónima los resultados sobre las predicciones de otras compañías, de manera que la MYPE pueda compararse y pueda determinar que variable o variables hacen que la competencia tenga mejores o peores probabilidades medias de acceso al crédito.

- Para un mejor desempeño del aplicativo móvil, se recomienda por lo menos contar con un *Smartphone* con 1GB de memoria RAM y 8GB de memoria ROM. A mayor memoria RAM el tiempo de aprendizaje del algoritmo de minería de datos se reducirá y con ello se presentarán los resultados más rápidamente.

- Se recomendaría incrementar las fuentes de datos, pues si bien es cierto se trabajó con información de un banco representativo del Perú, sería muy importante contar con información de otros bancos y también de cajas municipales o rurales, debido a que la naturaleza de operación de estas últimas es diferente a la de un banco tradicional, y ello podría enriquecer los resultados del algoritmo de minería de datos.

- Dentro del marco del uso de herramientas de minería de datos, la MYPE debería comenzar a analizar cuáles de sus procesos de operación pueden ser mejorados con el uso de estas herramientas, para así adelantarse a los avances tecnológicos y lograr una ventaja con respecto a su competencia.

BIBLIOGRAFÍA

- **LIBROS Y DOCUMENTOS DE TRABAJO**

BERRY, M. y LINOFF, G.
2004 *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management.* Publicaciones Wiley: 2da Ed. Indianápolis, IN.

BIGUS, J.
1996 *Data Mining with Neural Networks: Solving business problems*. McGraw-Hill: Estados Unidos.

COMISIÓN ECONÓMICA PARA AMÉRICA LATINA Y EL CARIBE (CEPAL)
2012 *Perspectivas económicas de América Latina 2013: Políticas de pymes para el cambio estructural.* Santiago, Chile: Autor.

2014 *La hora de la igualdad: Brechas por cerrar, caminos por abrir.* Santiago, Chile: Autor.

2016 *Estudio Económico de América Latina y el Caribe: La Agenda 2030 para el Desarrollo Sostenible y los desafíos del financiamiento para el desarrollo.* Santiago, Chile: Autor.

HAN, J, KAMBER, M y PEI, J.
2012 *Data Mining: Concepts and Techniques.* Publicaciones Morgan Kaufmann: 3ra Ed. Massachusetts, MA.

INSTITUTO NACIONAL DE ESTADISTICA E INFORMATICA (INEI)
2013 *Perú: Estructura Empresarial, 2013.* Lima, Perú: Autor.

MAKHABLEL, B.
2014 *Learning Data Mining with R.* Publicaciones Packt. Birmingham, Inglaterra.

MINISTERIO DE LA PRODUCCIÓN
2010 *Estadísticas de la Micro y Pequeña Empresa 2009.* Lima, Perú: Autor.

2013 *Mipyme 2012: Estadísticas de la Micro, Pequeña y Mediana Empresa*. Lima, Perú: Autor.

2014 *Las Mipyme en cifras 2013*. Lima, Perú: Autor.

2015 *Las Mipyme en cifras 2014*. Lima, Perú: Autor.

ORGANIZACIÓN INTERNACIONAL DEL TRABAJO (OIT)
2015 *Panorama Laboral Temático: Pequeñas Empresas, grandes brechas. Empleo y condiciones de trabajo en las MYPE de América Latina y el Caribe.* Lima, Perú: Autor.

SAMANIEGO, R.
2008 *El Riesgo de Crédito en el marco de acuerdo de Basilea II.* Publicaciones Delta. Madrid, España.

SUPERINTENDENCIA DE BANCA, SEGUROS Y AFP (SBS)
2016 *Banca.* XVII Programa de Extensión. Lima, Perú. Autor.

2016 *Contabilidad.* XVII Programa de Extensión. Lima, Perú. Autor.

2016 *Riesgo de Crédito.* XVII Programa de Extensión. Lima, Perú. Autor.

TORGO, L.
2014 *Data Mining with R. Learning with Case Studies*. Publicaciones Chapman & Hall/CRC Press. Florida, FL.

URMENETA, R.
2016 *Dinámica de las empresas exportadoras en América Latina: el aporte de las pymes.* Editorial CEPAL. Santiago, Chile.

- **PAPERS**

AVOLIO, B., MESONES, A. y ROCA, E.
2011 Factores que limitan el crecimiento de las micro y pequeñas empresas en el Perú (MYPES). *Strategia 22*(22), 70–80.

BLANCO, G. y JOSÉ, G.
2000 Duración: un concepto de la matemática financiera. *Faces UNMDP 7*(7), 117-126. Recuperado de http://eco.mdp.edu.ar/cendocu/repositorio/FACES_n7_117-126.pdf

BEKHET, H. y KAMEL, S.
2014 Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance 4*, 20-28. Recuperado de https://doi.org/10.1016/j.rdf.2014.03.002

HUSSMANNS, R
2004 Measuring the informal economy: From employment in the informal sector to informal employment. *Bureau of Statistics 53*(53), 3-18. Recuperado de https://www.ilo.org/wcmsp5/groups/public/---dgreports/---integration/documents/publication/wcms_079142.pdf

KOUTANAEI, F., SAJEDI, H. y KHANBABAEI, M.
2015 A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services 21*(21), 12-23. Recuperado de https://www.sciencedirect.com/science/article/pii/S0969698915300060

MUÑOZ, Jorge y otros
2013 Analizando el endeudamiento de las Micro y Pequeñas Empresas. *Revista La Moneda (156)*, 19-24. Recuperado de http://www.bcrp.gob.pe/docs/Publicaciones/Revista-Moneda/moneda-156/moneda-156.pdf

ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICO
2012 Financial Contagion in the Era of Globalised Banking?. *OECD Economics Policy Notes 14* (14), 2-8. Recuperado de https://www.oecd.org/eco/monetary/50556019.pdf

POGGI, J., LUY, M., VÁSQUEZ, S. y BASTANTE, E.

2005 El Nuevo Acuerdo de Capital en economías emergentes: un estudio para el caso peruano. *Revista de Temas Financieros SBS 2*(1), 187-222. Recuperado de http://www.sbs.gob.pe/Portals/0/jer/EDIPUB_VOLUMEN2/10POGGI.pdf

THURIK, R. y WENNEKERS, S.
2001 A note on Entrepreneurship, Small Business and Economic Growth. *Erasmus Research Institute of Management Report Series* (n/d), 1-8. Recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1265430

VATSA, K
2004 Risk, vulnerability, and asset-based approach to disaster risk management. *International Journal of Sociology and Social Policy 24 (*10), 1-48. Recuperado de http://dx.doi.org/10.1108/01443330410791055

WILLIAMS, C., SHAHID, M. y MARTINEZ, A.
2015 Determinants of the Level of Informality of Informal Micro-Enterprises: Some Evidence from de City of Lahore, Pakistan. *World Development 84*, 312-325. Recuperado de https://doi.org/10.1016/j.worlddev.2015.09.003

ZHANG, J., LYU, T. y LI, R.
2019 A Study on SMIE Credit Evaluation Model Based on Blockchain Technology. *Procedia CIRP 83*, 616-623. Recuperado de https://doi.org/10.1016/j.procir.2019.05.003

- **DIAPOSITIVAS**

BOTÍA, J.
2009 *Preprocesado de Datos* [diapositiva]. Universidad de Murcia: Departamento de Ingeniería de la Información y las Comunicaciones. Murcia, España.

GONZÁLEZ, C.
s/f *SVM: Maquinas de vectores de soporte* [diapositiva]. Universidad de Valladolid: Departamento de Informática. Valladolid, España.

SEMPERE, J.
s/f *Aprendizaje de árboles de decisión* [diapositiva]. Valencia: Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.

- **PÁGINAS WEB Y DOCUMENTOS VIRTUALES**

ALZAMORA, M.
2015 (17 de noviembre). MYPES aportan el 40% del PBI. Recuperado de https://asep.pe/index.php/mypes-aportan-el-40-del-pbi/

ASOCIACIÒN DE BANCOS DEL PERÙ (ASBANC)
2014 (12 de diciembre) Regla procíclica: ¿Qué significa y qué implicancias tiene su reciente desactivación? Recuperado de https://www.asbanc.com.pe/Publicaciones/ASBANC%20SEMANAL%20N%C2%BA%20131.pdf

BANCO BILBAO VIZCAYA ARGENTARIA (BBVA)
2017 (02 de mayo) ¿Qué son las agencias de 'rating' o agencias de calificación? Recuperado de https://www.bbva.com/es/las-agencias-calificacion-rating/

BANCO CENTRAL DE RESERVA DEL PERÚ (BCRP)
2011 (31 de marzo) Glosario de Términos Económicos. Recuperado de http://www.bcrp.gob.pe/docs/Publicaciones/Glosario/Glosario-BCRP.pdf
2016 (n.d.) Cuadros Mensuales Históricos. Recuperado de http://www.bcrp.gob.pe/estadisticas/cuadros-mensuales-historicos.html

2016 (n.d.) Preguntas Frecuentes. Recuperado de http://www.bcrp.gob.pe/sobre-el-bcrp/preguntas-frecuentes.html

BANCO MUNDIAL
2016 (n.d.) Perú: Panorama General. Recuperado de http://www.bancomundial.org/es/country/peru/overview#1

BASEL COMMITTEE ON BANKING SUPERVISION
2001 Pillar 2 – Supervisory Review Process. Recuperado de http://www.bis.org/publ/bcbs189.pdf

DOMÍNGUEZ, M, MIRANDA, F, PALLAS, J y PERAZA, C.
2003 (n.d.) La medición del riesgo de crédito y el nuevo acuerdo de capital del Comité de Basilea. Recuperado de: http://www.uv.es/asepuma/XI/31.pdf

DAMODARÁN, A.
s/f (n.d.) Betas by Sector. Recuperado de http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/Betas.html

PORTAL DE MICROFINANZAS
2015 (n.d.) Perú. Recuperado de http://www.microfinancegateway.org/es/pa%C3%ADs/per%C3%BA

WAGEINDICATOR FOUNDATION
2017 (12 de diciembre) Salary Check. Recuperado de https://wageindicator.org/

SUPERINTENDENCIA DE BANCA, SEGUROS Y AFP
2006 (01 de octubre) Basilea II: El Nuevo Acuerdo de Capital. Recuperado de http://www.sbs.gob.pe/Portals/0/jer/REGUL_PROYIMP_BASIL_FUNSBS/BasileaII-Introduccion-JPoggi-MLuy.pdf

- **NOTICIAS Y ARTICULOS ESPECIALIZADOS**

CAMPODÓNICO, H.
2016 (10 de enero) ¿Sólo las MYPES salvarán al Perú?. *La República*. Recuperado de http://larepublica.pe/impresa/opinion/732722-solo-las-Mypes-salvaran-al-peru

GESTIÓN
2014 (15 de mayo) Participación de las Mypes en el PBI sigue en descenso, alertó la SNI". *Gestión*. Recuperado de http://gestion.pe/economia/sni-necesario-mejorar-competitividad-y-productividad-Mypes-2097330

2018a (01 de octubre) ¿Qué evalúa un banco para otorgar un préstamo a un emprendedor? *Gestión.* Recuperado de https://gestion.pe/tu-dinero/evalua-banco-otorgar-pr%C3%A9stamo-emprendedor-110033

2018b (08 de junio) Créditos personales: ¿Cuánto debe ser la tasa de interés para ser considerada 'baja'?. *Gestión.* Recuperado de https://gestion.pe/tu-dinero/finanzas-personales/creditos-personales-debe-tasa-interes-considerada-baja-143555

HENKE, N., LIBARIKIAN,A. y WISEMAN,B.
2016 (31 de octubre) Straight talk about big data. *McKinzey*. Recuperado de http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/straight-talk-about-big-data

LA REPÚBLICA
2016 (19 de marzo) BCR estima que 2016 será un mal año para el sector manufactura. *La República*. Recuperado de http://larepublica.pe/impresa/economia/749619-bcr-estima-que-2016-sera-un-mal-ano-para-el-sector-manufactura

RÍOS, M.
2015 Las microfinancieras y su rol descentralizador. *Gestión*. Recuperado de http://gestion.pe/mercados/microfinancieras-y-su-rol-descentralizador-2138997

SEMANA ECONÓMICA
2016 (13 de marzo) INEI: 62,581 empresas se crearon en el cuarto trimestre del 2015. *SEMANAeconomica.com*. Recuperado de http://semanaeconomica.com/article/economia/macroeconomia/182920-inei-se-crearon-62581-empresas-se-crearon-en-el-cuarto-trimestre-del-2015/

- **DECRETOS, LEYES Y RESOLUCIONES**

MINISTERIO DE LA PRODUCCIÓN
2013 Decreto Supremo N° 013 – 2013: Texto Único Ordenado de la Ley de Impulso al Desarrollo Productivo y al Crecimiento Empresarial.

SUPERINTENDENCIA DE BANCA, SEGUROS Y AFP (SBS)
2008 Resolución N°11356 – 2008: Reglamento para la Evaluación y Clasificación del deudor y la exigencia de provisiones.

2009 Ley N°26702: Ley General del Sistema Financiero y del Sistema de Seguros y Orgánica de la Superintendencia de Banca y Seguros.

2009 Resolución N° 14354 – 2009: Reglamento para el Requerimiento de Patrimonio Efectivo por Riesgo de Crédito.

2011 Resolución N° 3780 – 2011: Reglamento de Gestión de Riesgo de Crédito.
2011 Resolución N° 8425 – 2011: Reglamento para el Requerimiento de Patrimonio Efectivo Adicional.

2016 Resolución N° 1906 – 2016: Aprueban Estructura Orgánica de la SBS.

ANEXOS

Anexo 01. Distribución de campos versus variable objetivo

Variable: CODMES

Únicos	9
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
201608	4,777	15.3%	3,849	928	80.6%	19.4%
201603	4,003	12.8%	3,318	685	82.9%	17.1%
201605	3,334	10.7%	2,832	502	84.9%	15.1%
201609	3,290	10.6%	2,622	668	79.7%	20.3%
201606	3,271	10.5%	2,768	503	84.6%	15.4%
201607	3,270	10.5%	2,724	546	83.3%	16.7%
201604	3,219	10.3%	2,662	557	82.7%	17.3%
201601	3,069	9.8%	2,539	530	82.7%	17.3%
201602	2,927	9.4%	2,352	575	80.4%	19.6%

Variable: MES

Únicos	9
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
8	4,777	15.3%	3,849	928	80.6%	19.4%
3	4,003	12.8%	3,318	685	82.9%	17.1%
5	3,334	10.7%	2,832	502	84.9%	15.1%
9	3,290	10.6%	2,622	668	79.7%	20.3%
6	3,271	10.5%	2,768	503	84.6%	15.4%
7	3,270	10.5%	2,724	546	83.3%	16.7%
4	3,219	10.3%	2,662	557	82.7%	17.3%
1	3,069	9.8%	2,539	530	82.7%	17.3%
2	2,927	9.4%	2,352	575	80.4%	19.6%

Variable: TIPO_CAMPANA

Únicos	4
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
NAVIDAD	9,833	31.6%	7,943	1,890	80.8%	19.2%
ESCOLAR	7,485	24.0%	6,056	1,429	80.9%	19.1%
DIA DE LA MADRE	7,165	23.0%	5,976	1,189	83.4%	16.6%
FIESTAS PATRIAS	6,677	21.4%	5,691	986	85.2%	14.8%

Variable: ESTADO

Únicos	2
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
ACE	25,666	82.4%	25,666	0	100.0%	0.0%
DEN	5,494	17.6%	0	5,494	0.0%	100.0%

Variable: MONEDA

Únicos	2
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
DOLARES	12	0.0%	7	5	58.3%	41.7%
SOLES	31,148	100.0%	25,659	5,489	82.4%	17.6%

Variable: MTOAPROBADO

Únicos	1916
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
60000	5,064	16.3%	4,162	902	82.2%	17.8%
90000	3,083	9.9%	1	543	0.2%	99.8%
30000	1,778	5.7%	1,443	335	81.2%	18.8%
150000	1,510	4.8%	1,220	290	80.8%	19.2%
50000	1,361	4.4%	1,146	215	84.2%	15.8%
40000	783	2.5%	646	137	82.5%	17.5%
45000	758	2.4%	612	146	80.7%	19.3%
70000	623	2.0%	518	105	83.1%	16.9%
20000	537	1.7%	453	84	84.4%	15.6%
48000	490	1.6%	404	86	82.4%	17.6%
72000	470	1.5%	383	87	81.5%	18.5%
10000	351	1.1%	296	55	84.3%	15.7%
80000	345	1.1%	271	74	78.6%	21.4%
64000	332	1.1%	280	52	84.3%	15.7%
120000	320	1.0%	277	43	86.6%	13.4%
100000	312	1.0%	251	61	80.4%	19.6%
Otros (<1%)	13,043	41.9%	10,764	2,279	82.5%	17.5%

Variable: MONTODOLARES

Únicos	5204
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
18039.69	738	2.4%	612	126	82.9%	17.1%
26533.02	635	2.0%	525	110	82.7%	17.3%
18331.81	609	2.0%	503	106	82.6%	17.4%
17793.59	595	1.9%	492	103	82.7%	17.3%
18242.63	589	1.9%	480	109	81.5%	18.5%
17883.76	586	1.9%	489	97	83.4%	16.6%
17688.68	580	1.9%	468	112	80.7%	19.3%
17291.07	529	1.7%	441	88	83.4%	16.6%
17026.11	501	1.6%	399	102	79.6%	20.4%
44221.7	418	1.3%	351	67	84.0%	16.0%
26470.59	410	1.3%	335	75	81.7%	18.3%
27059.53	359	1.2%	286	73	79.7%	20.3%
26690.39	357	1.1%	299	58	83.8%	16.2%
17647.06	337	1.1%	278	59	82.5%	17.5%
Otros (<1%)	23,917	76.8%	19,708	4,209	82.4%	17.6%

Variable: MONTOSOLES

Únicos	1923
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
60000	5,064	16.3%	4,162	902	82.2%	17.8%
90000	3,083	9.9%	2,540	543	82.4%	17.6%
30000	1,777	5.7%	1,443	334	81.2%	18.8%
150000	1,510	4.8%	1,220	290	80.8%	19.2%
50000	1,361	4.4%	1,146	215	84.2%	15.8%
40000	783	2.5%	646	137	82.5%	17.5%
45000	757	2.4%	611	146	80.7%	19.3%
70000	623	2.0%	518	105	83.1%	16.9%
20000	537	1.7%	453	84	84.4%	15.6%
48000	490	1.6%	404	86	82.4%	17.6%
72000	469	1.5%	383	86	81.7%	18.3%
10000	351	1.1%	296	55	84.3%	15.7%
80000	345	1.1%	271	74	78.6%	21.4%
64000	332	1.1%	280	52	84.3%	15.7%
120000	320	1.0%	277	43	86.6%	13.4%
100000	312	1.0%	251	61	80.4%	19.6%
Otros (<1%)	13,046	41.9%	10,765	2,281	82.5%	17.5%

Variable: TASA

Únicos	132
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
20	2,116	6.8%	1,712	404	80.9%	19.1%
18	1,839	5.9%	1,512	327	82.2%	17.8%
19	1,755	5.6%	1,462	293	83.3%	16.7%
22	1,618	5.2%	1,344	274	83.1%	16.9%
23	1,591	5.1%	1,316	275	82.7%	17.3%
17	1,555	5.0%	1,281	274	82.4%	17.6%
15	1,544	5.0%	1,280	264	82.9%	17.1%
21	1,515	4.9%	1,255	260	82.8%	17.2%
24	1,510	4.8%	1,229	281	81.4%	18.6%
27	1,494	4.8%	1,240	254	83.0%	17.0%
26	1,483	4.8%	1,220	263	82.3%	17.7%
25	1,398	4.5%	1,155	243	82.6%	17.4%
16	1,228	3.9%	1,030	198	83.9%	16.1%
28	1,148	3.7%	946	202	82.4%	17.6%
29	922	3.0%	761	161	82.5%	17.5%
30	830	2.7%	686	144	82.7%	17.3%
32	753	2.4%	620	133	82.3%	17.7%
31	602	1.9%	496	106	82.4%	17.6%
33	587	1.9%	498	89	84.8%	15.2%
35	579	1.9%	464	115	80.1%	19.9%
34	509	1.6%	422	87	82.9%	17.1%
14	424	1.4%	334	90	78.8%	21.2%
38	408	1.3%	341	67	83.6%	16.4%
36	386	1.2%	323	63	83.7%	16.3%
Otros (<1%)	3,366	10.8%	2,739	627	81.4%	18.6%

Variable: DESSECTORECONOMICO

Únicos	4
Faltantes	142

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
COMERCIO	17,611	56.8%	14,596	3,015	82.9%	17.1%
SERVICIO	8,498	27.4%	6,908	1,590	81.3%	18.7%
INDUSTRIA	4,882	15.7%	4,032	850	82.6%	17.4%
OTROS	27	0.1%	25	2	92.6%	7.4%

Variable: CODVENDEDOR

Únicos	924
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
Otros (<1%)	31,160	100.0%	25,666	5,494	82.4%	17.6%

Variable: CODAPROBADOR

Únicos	631
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
S08757	6,624	21.3%	5,527	1,097	83.4%	16.6%
S08756	6,440	20.7%	5,436	1,004	84.4%	15.6%
U21889	5,254	16.9%	4,487	767	85.4%	14.6%
U19068	4,207	13.5%	3,688	519	87.7%	12.3%
S13533	3,207	10.3%	2,789	418	87.0%	13.0%
E21269	2,541	8.2%	2,109	432	83.0%	17.0%
Otros (<1%)	2,887	9.3%	1,630	1,257	56.5%	43.5%

Variable: FLGGAR

Únicos	2
Faltantes	7,411

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
N	18,365	77.3%	15,343	3,022	83.5%	16.5%
S	5,384	22.7%	4,522	862	84.0%	16.0%

Variable: DESCAMPANIA

Únicos	8
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
APROB. NAVIDAD	9,777	31.4%	7,924	1,853	81.0%	19.0%
APROBADOS CAMPANA ESCOLAR	7,392	23.7%	6,045	1,347	81.8%	18.2%
APROBADOS DIA DE LA MADRE	7,063	22.7%	5,949	1,114	84.2%	15.8%
APROBADOS FIESTAS PATRIAS	6,649	21.3%	5,677	972	85.4%	14.6%
CAMPAÑA DIA DE LA MADRE	102	0.3%	27	75	26.5%	73.5%
CAMPANA ESCOLAR	93	0.3%	11	82		
CAMPAÑA DE NAVIDAD	56	0.2%	19	37	33.9%	66.1%

CAMPAÑA 28 DE JULIO	28	0.1%	14	14	50.0%	50.0%

Variable: TIPESTCONTRATOSOLICITUDMIC

Únicos	3
Faltantes	1461

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
S	22,810	76.8%	22,810	0	100.0%	0.0%
N	5,514	18.6%	20	5,494	0.4%	99.6%
X	1,375	4.6%	1,375	0	100.0%	0.0%

Variable: MTOCAMBIOALNUEVOSOL

Únicos	9
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
3.392	4,777	15.3%	3,849	928	80.6%	19.4%
3.326	4,003	12.8%	3,318	685	82.9%	17.1%
3.372	3,334	10.7%	2,832	502	84.9%	15.1%
3.4	3,290	10.6%	2,622	668	79.7%	20.3%
3.289	3,271	10.5%	2,768	503	84.6%	15.4%
3.355	3,270	10.5%	2,724	546	83.3%	16.7%
3.273	3,219	10.3%	2,662	557	82.7%	17.3%
3.47	3,069	9.8%	2,539	530	82.7%	17.3%
3.524	2,927	9.4%	2,352	575	80.4%	19.6%

Variable: MARCA

Únicos	2
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
ANTIGUO	28,694	96.6%	23,621	5,073	82.3%	17.7%
NUEVO	2,466	8.3%	2,045	421	82.9%	17.1%

Variable: AREA

Únicos	6
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
PROV 1	10,938	35.1%	9,098	1,840	83.2%	16.8%
PROV 2	10,092	32.4%	8,151	1,941	80.8%	19.2%
LIMA 1	5,819	18.7%	4,876	943	83.8%	16.2%
LIMA 3	2,427	7.8%	1,999	428	82.4%	17.6%
LIMA 2	1,882	6.0%	1,541	341	81.9%	18.1%
SIN AREA	2	0.0%	1	1	50.0%	50.0%

Variable: REGION

Únicos	28
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
PROV 1-ORIENTE	3,351	10.8%	2,746	605	81.9%	18.1%
PROV 2-SUR 3	2,726	8.7%	2,231	495	81.8%	18.2%
PROV 1-NORTE 2	2,271	7.3%	1,912	359	84.2%	15.8%
PROV 1-NORTE 1	2,182	7.0%	1,812	370	83.0%	17.0%
PROV 2-CENTRO	1,925	6.2%	1,584	341	82.3%	17.7%
PROV 2-SUR 1	1,808	5.8%	1,464	344	81.0%	19.0%

LIMA 1-REGION 3	1,678	5.4%	1,397	281	83.3%	16.7%
PROV 2-SUR CHICO	1,649	5.3%	1,333	316	80.8%	19.2%
PROV 2-SUR 2	1,647	5.3%	1,284	363	78.0%	22.0%
PROV 1-NORTE 3	1,563	5.0%	1,278	285	81.8%	18.2%
LIMA 3-REGION 4	1,450	4.7%	1,212	238	83.6%	16.4%
PROV 1-NORTE CHICO	1,258	4.0%	1,086	172	86.3%	13.7%
LIMA 1-REGION 2	1,224	3.9%	1,027	197	83.9%	16.1%
LIMA 1-REGION 1	1,141	3.7%	940	201	82.4%	17.6%
LIMA 1-REGION 5	906	2.9%	778	128	85.9%	14.1%
LIMA 1-REGION 4	870	2.8%	734	136	84.4%	15.6%
LIMA 2-REGION 1	860	2.8%	726	134	84.4%	15.6%
LIMA 2-REGION 2	522	1.7%	398	124	76.2%	23.8%
LIMA 3-REGION 5	409	1.3%	316	93	77.3%	22.7%
PROV 2-SURCHICO	337	1.1%	255	82	75.7%	24.3%
Otros (<1%)	1,383	4.4%	1,153	230	83.4%	16.6%

Variable: ZONA

Únicos	3
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
PROV	20,805	66.8%	17,030	3,775	81.9%	18.1%
LIMA	10,353	33.2%	8,635	1,718	83.4%	16.6%
SIN ZONA	2	0.0%	1	1	50.0%	50.0%

Variable: DESCODDEPARTAMENTO

Únicos	25
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
LIMA	10,408	33.4%	8,678	1,730	83.4%	16.6%
AREQUIPA	2,032	6.5%	1,649	383	81.2%	18.8%
LA LIBERTAD	1,985	6.4%	1,696	289	85.4%	14.6%
JUNIN	1,741	5.6%	1,424	317	81.8%	18.2%
LAMBAYEQUE	1,329	4.3%	1,110	219	83.5%	16.5%
CUZCO	1,256	4.0%	969	287	77.1%	22.9%
PIURA	1,241	4.0%	1,027	214	82.8%	17.2%
TACNA	1,217	3.9%	1,014	203	83.3%	16.7%
ICA	1,076	3.5%	870	206	80.9%	19.1%
PUNO	1,066	3.4%	853	213	80.0%	20.0%
HUANUCO	1,059	3.4%	874	185	82.5%	17.5%
SAN MARTIN	1,018	3.3%	868	150	85.3%	14.7%
ANCASH	924	3.0%	787	137	85.2%	14.8%
CAJAMARCA	880	2.8%	700	180	79.5%	20.5%
AYACUCHO	762	2.4%	595	167	78.1%	21.9%
UCAYALI	703	2.3%	575	128	81.8%	18.2%
CALLAO	517	1.7%	426	91	82.4%	17.6%
LORETO	485	1.6%	371	114	76.5%	23.5%
TUMBES	322	1.0%	251	71	78.0%	22.0%
AMAZONAS	259	0.8%	218	41	84.2%	15.8%
APURIMAC	247	0.8%	199	48	80.6%	19.4%
MOQUEGUA	220	0.7%	179	41	81.4%	18.6%
CERRO DE PASCO	191	0.6%	155	36	81.2%	18.8%
MADRE DE DIOS	143	0.5%	115	28	80.4%	19.6%
HUANCAVELICA	79	0.3%	63	16	79.7%	20.3%

Variable: CODOFI

Únicos	270
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
SUC AYACUCHO	516	1.7%	394	122	76.4%	23.6%
AG.AVIACION - LIMA	485	1.6%	397	88	81.9%	18.1%
SUC HUANUCO	460	1.5%	366	94	79.6%	20.4%
SUC CUZCO	403	1.3%	310	93	76.9%	23.1%
SUC LIMA	382	1.2%	327	55	85.6%	14.4%
SUC TACNA	370	1.2%	323	47	87.3%	12.7%
AG.PRIMAVERA - TRUJILLO	369	1.2%	302	67	81.8%	18.2%
SUC TARAPOTO	360	1.2%	307	53	85.3%	14.7%
AG.LEONARDO ORTIZ - CHICLAYO	341	1.1%	273	68	80.1%	19.9%
AG.SAN ROMAN - JULIACA	338	1.1%	265	73	78.4%	21.6%
AG.TORRE AMERICA - LIMA	328	1.1%	273	55	83.2%	16.8%
Otros (<1%)	26808	85.0%	22129	4679	82.5%	17.5%

Variable: FLG_CAMPAÑA

Únicos	1
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
Otros (<1%)	31,160	43.0%	25,666	5,494	82.4%	17.6%

Variable: CANALVENTA

Únicos	3
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
Funcionario con cartera	26,558	85.2%	21,906	4,652	82.5%	17.5%
Funcionario sin cartera	3,742	12.0%	3,050	692	81.5%	18.5%
Otros	860	2.8%	710	150	82.6%	17.4%

Variable: CANALEVALUACION

Únicos	6
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
Canal Simplificado	28,402	91.1%	24,145	4,257	85.0%	15.0%
Funcionario	1,375	4.4%	404	971	29.4%	70.6%
Canal Integral	1,062	3.4%	908	154	85.5%	14.5%
Otros	231	0.7%	130	101	56.3%	43.7%
GDA	86	0.3%	79	7	91.9%	8.1%
Analista de Campo	4	0.0%	0	4	0.0%	100.0%

Variable: TIPESTJUSTIFICACION

Únicos	6
Faltantes	2

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
BREC	23,684	76.0%	23,684	0	100.0%	0.0%
SAPE	1,008	3.2%	1,008	0	100.0%	0.0%
P45O	987	3.2%	0	987	0.0%	100.0%

P37G	938	3.0%	0	938	0.0%	100.0%
P39P	696	2.2%	0	696	0.0%	100.0%
P36M	583	1.9%	0	583	0.0%	100.0%
ABXC	437	1.4%	437	0	100.0%	0.0%
P35C	405	1.3%	0	405	0.0%	100.0%
P40M	379	1.2%	0	379	0.0%	100.0%
APBD	352	1.1%	352	0	100.0%	0.0%
P38P	320	1.0%	0	320	0.0%	100.0%
Otros (<1%)	1,369	4.4%	183	1,186	13.4%	86.6%

Variable: DESJUSTIFICACION

Únicos	65
Faltantes	25,666

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
PYME45 INFORMACIÓN BASICA INCOMPLETA	987	18.0%	0	987	0.0%	100.0%
PYME37 NO CUMPLE FILTRO POLITICA DE GARANTIAS	938	17.1%	0	938	0.0%	100.0%
PYME39 NO CUMPLE FILTROS PODERES	696	12.7%	0	696	0.0%	100.0%
PYME36 NO CUMPLE FILTRO MORA Y SOBREGIROS VIGENTES	583	10.6%	0	583	0.0%	100.0%
PYME35 NO CUMPLE FILTRO CUENTA CORRIENTE AHORRO	405	7.4%	0	405	0.0%	100.0%
PYME40 NO CUMPLE MAS DE UN FILTRO CANAL SIMPLIFICADO	379	6.9%	0	379	0.0%	100.0%
PYME38 NO CUMPLE FILTRO POSICION DEUDORA	320	5.8%	0	320	0.0%	100.0%
PYME10 ERROR DE DIGITACION MIC	290	5.3%	0	290	0.0%	100.0%
PYME33 OTRAS JUSTIFICACIONES	243	4.4%	0	243	0.0%	100.0%
PYME02 CLIENTE DESESTIMO POR TIEMPO DE PROCESO	86	1.6%	0	86	0.0%	100.0%
PYME08 DOCUMENTACION OBLIGATORIA INCOMPLETA	64	1.2%	0	64	0.0%	100.0%
Otros (<1%)	503	9.2%	0	503	0.0%	100.0%

Variable: DESJUS_AGRUPACION

Únicos	4
Faltantes	25666

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
Despliegue Operativo	4,833	88.0%	0	4,833	0.0%	100.0%
Otros Motivos	290	5.3%	0	290	0.0%	100.0%
Cliente	249	4.5%	0	249	0.0%	100.0%
Política	120	2.2%	0	120	0.0%	100.0%
Score	2	0.0%	0	2	0.0%	100.0%

Variable: GRUPOJUSTIFICACION

Únicos	3
Faltantes	25666

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
Despliegue	4,833	88.0%	0	4,833	0.0%	100.0%
Cliente,Otros	539	9.8%	0	539	0.0%	100.0%
Política,Score,Modelado	122	2.2%	0	122	0.0%	100.0%

Variable: APROB_CANALEVALUACION

Únicos	7
Faltantes	7

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
ACE	25,666	82.4%	25,666	0	100.0%	0.0%
DEN-CRE2	4,331	13.9%	0	4,331	0.0%	100.0%
DEN-FDN2	500	1.6%	0	500	0.0%	100.0%
DEN-FDN1	453	1.5%	0	453	0.0%	100.0%
DEN-FDN3	119	0.4%	0	119	0.0%	100.0%
DEN-CRE1	81	0.3%	0	81	0.0%	100.0%
DEN-CRE3	3	0.0%	0	3	0.0%	100.0%

Variable: CODCLAVECIC

Únicos	17202
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
999999	2,927	9.4%	2,352	575	80.4%	19.6%
Otros (<1%)	28,233	90.6%	23,314	4,919	82.6%	17.4%

Variable: FLG_DESEMBOLSO

Únicos	2
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
S	22,810	73.2%	22,810	0	100.0%	0.0%
N	8,350	26.8%	2,856	5,494	34.2%	65.8%

Variable: FLG_APROBACION

Únicos	2
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
S	30,846	99.0%	25,666	5,180	83.2%	16.8%
N	314	1.0%	0	314	0.0%	100.0%

Variable: FLG_APROBACION_RBM

Únicos	2
Faltantes	0

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
S	30,584	98.2%	25,666	4,918	83.9%	16.1%
N	576	1.8%	0	576	0.0%	100.0%

Variable: PD_CLIENTE

Únicos	23555
Faltantes	3069

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
0.01% a 0.91%	6,327	22.5%	5,237	1,090	82.8%	17.2%
0.91% a 1.81%	4,973	17.7%	4,097	876	82.4%	17.6%
1.81% a 2.71%	3,359	12.0%	2,768	591	82.4%	17.6%
2.71% a 3.61%	2,640	9.4%	2,164	476	82.0%	18.0%
3.61% a 4..52%	2,035	7.2%	1,688	347	82.9%	17.1%
4.52% a 5.42%	1,717	6.1%	1,414	303	82.4%	17.6%
5.42% a 6.32%	1,454	5.2%	1,206	248	82.9%	17.1%
6.32% a 7.22%	1,159	4.1%	967	192	83.4%	16.6%
7.22% a 8.12%	835	3.0%	679	156	81.3%	18.7%
8.12% a 9.03%	737	2.6%	595	142	80.7%	19.3%
9.03% a 9.93%	580	2.1%	486	94	83.8%	16.2%
9.93% a 10.80%	391	1.4%	329	62	84.1%	15.9%
10.80% a 11.70%	330	1.2%	268	62	81.2%	18.8%
11.70% a 12.60%	286	1.0%	228	58	79.7%	20.3%
Mayor a 12.60%	1,268	4.5%	1,035	233	81.6%	18.4%

Variable: SEG_COM

Únicos	4
Faltantes	2976

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
2	9,915	35.2%	8,206	1,709	82.8%	17.2%
1	8,648	30.7%	7,258	1,390	83.9%	16.1%
3	8,561	30.4%	6,940	1,621	81.1%	18.9%
4	1,060	3.8%	882	178	83.2%	16.8%

Variable: SEGMENTO

Únicos	8
Faltantes	2972

			Cross: ESTADO			
Valor	Freq	Porc	ACE	DEN	% ACE	% DEN
1. A1	13,855	49.2%	11,405	2,450	82.3%	17.7%
2. A2	5,692	20.2%	4,778	914	83.9%	16.1%
INACTIVO	4,090	14.5%	3,411	679	83.4%	16.6%
3. B1	2,659	9.4%	2,213	446	83.2%	16.8%
4. B2	1,053	3.7%	821	232	78.0%	22.0%
5. C	513	1.8%	408	105	79.5%	20.5%
6. D	282	1.0%	228	54	80.9%	19.1%
DEFAULT	44	0.2%	26	18	59.1%	40.9%

Anexo 02. Limpieza de variables, código en R

```
#Paso 1. Eliminar espacios en variables----
EliminarVacios<-c('CODMES','MES','TIPO_CAMPANA',
                  'ESTADO','MONEDA','MTOAPROBADO',
                  'MONTODOLARES','MONTOSOLES','TASA',
                  'DESSECTORECONOMICO','CODVENDEDOR',
                  'CODAPROBADOR','FLGGAR','DESCAMPANIA',
                  'TIPESTCONTRATOSOLICITUDMIC',
                  'FLGTIPPER','MTOCAMBIOALNUEVOSOL','MARCA',
                  'AREA','REGION','ZONA','DESCODDEPARTAMENTO',
                  'CODOFI','FLG_CAMPAÑA','CANALVENTA',
                  'CANALEVALUACION','TIPESTJUSTIFICACION',
                  'DESJUSTIFICACION','DESJUS_AGRUPACION',
                  'GRUPOJUSTIFICACION','APROB_CANALEVALUACION',
                  'CODCLAVECIC','FLG_DESEMBOLSO','FLG_APROBACION',
                  'FLG_APROBACION_RBM','PD_CLIENTE','SEG_COM',
                  'SEGMENTO')
for (i in EliminarVacios) {
  DB_Banco[,i] <- trimws(DB_Banco[,i])
  rm(i)
  }
#Paso 2. Unificar a mayúsculas.----
UnifMayuscula<-c('TIPO_CAMPANA','ESTADO','MONEDA',
                  'DESSECTORECONOMICO','CODVENDEDOR',
                  'CODAPROBADOR','FLGGAR','DESCAMPANIA',
                  'TIPESTCONTRATOSOLICITUDMIC','FLGTIPPER',
                  'MARCA','ZONA','AREA','REGION','CODOFI',
                  'DESCODDEPARTAMENTO','FLG_CAMPAÑA',
                  'CANALVENTA','CANALEVALUACION',
                  'TIPESTJUSTIFICACION','DESJUSTIFICACION',
                  'DESJUS_AGRUPACION','GRUPOJUSTIFICACION',
                  'APROB_CANALEVALUACION','CODCLAVECIC',
                  'FLG_DESEMBOLSO','FLG_APROBACION',
                  'FLG_APROBACION_RBM','SEG_COM','SEGMENTO')

for (i in UnifMayuscula) {
  DB_Banco[,i] <- toupper(DB_Banco[,i])
  rm(i)
}
#Paso 3. Imputar datos categóricos----
Inputar<- c('FLGGAR','FLGTIPPER','DESSECTORECONOMICO',
            'DESJUSTIFICACION','DESJUS_AGRUPACION',
            'GRUPOJUSTIFICACION','SEG_COM','SEGMENTO')
Inputacion<- c('NO APLICA')

for (i in 1:length(Inputar)) {
  eval(parse(text=paste0('DB_Banco$',Inputar[i],'<-
replace(DB_Banco$',Inputar[i],',is.na(DB_Banco$',Inputar[i],"),'",Inputaci
on,"')")))
}
rm(Inputar)
rm(Inputacion)
#Paso 4. Imputar datos numéricos----
DB_Banco$PD_CLIENTE<-as.numeric(DB_Banco$PD_CLIENTE)
Segmentos <- unique(DB_Banco$SEGMENTO)
for (j in Segmentos){
  DB_Banco$PD_CLIENTE[is.na(DB_Banco$PD_CLIENTE) & DB_Banco$SEGMENTO == j]
<-
    mean(DB_Banco$PD_CLIENTE[!is.na(DB_Banco$PD_CLIENTE) &
DB_Banco$SEGMENTO == j],na.rm=T)
}
DB_Banco$PD_CLIENTE[is.na(DB_Banco$PD_CLIENTE)] <-
median(DB_Banco$PD_CLIENTE,na.rm=T)

rm(i,j,EliminarVacios,Segmentos,UnifMayuscula)
```

Anexo 03. Transformación de variables, código en R

```
#Paso 1. Eliminación de registros----
#Filtro N°1: Para una misma transacción (mismo codigo mes, mismo mes,
misma campana, mismo monto aprobado,
#misma p.default, mismo segmento), en caso de haber una denegacion
antecedente, se mantendrá y se eliminará la aprobación.
DB_Banco$Indicador_Transaccion =
as.numeric(duplicated(DB_Banco[,c("CODMES","MES","TIPO_CAMPANA","MTOAPROBA
DO","PD_CLIENTE","SEGMENTO")])) #5827 registros
DB_Banco<-DB_Banco[DB_Banco$Indicador_Transaccion==0,]
DB_Banco$Indicador_Transaccion=NULL
#Paso 2. Ordenamiento de variables----
DB_Banco <- DB_Banco[,c("CODCLAVECIC","CODMES","MES",
"TIPO_CAMPANA","DESCAMPANIA","FLG_CAMPAÑA","DESSECTORECONOMICO",
"MARCA","AREA","REGION","ZONA","DESCODDEPARTAMENTO","CODOFI","ESTADO",
"TASA","MONEDA","MTOAPROBADO","MONTODOLARES","MONTOSOLES",
"MTOCAMBIOALNUEVOSOL","SEG_COM","SEGMENTO","PD_CLIENTE","FLGGAR","FLGTIPPE
R",
"TIPESTCONTRATOSOLICITUDMIC","TIPESTJUSTIFICACION","DESJUSTIFICACION",
"DESJUS_AGRUPACION","GRUPOJUSTIFICACION","APROB_CANALEVALUACION",
"FLG_DESEMBOLSO","FLG_APROBACION","FLG_APROBACION_RBM","CODVENDEDOR",
"CODAPROBADOR","CANALVENTA","CANALEVALUACION")]
#Paso 3. Eliminación de variables----
#Filtro N°2: Eliminar todas las variables de aprobación/negación excepto
DESJUSTIFICACION y realizar clusterización personalizada.
Filtro2<-
c('TIPESTJUSTIFICACION','DESJUS_AGRUPACION','GRUPOJUSTIFICACION','APROB_CA
NALEVALUACION','FLG_DESEMBOLSO','FLG_APROBACION','FLG_APROBACION_RBM')
DB_Banco<- DB_Banco[,!names(DB_Banco)%in%Filtro2]
#Filtro N°3: Eliminar variables que no se utilizarán.
Filtro3<-
c('CODMES','DESCAMPANIA','TIPESTCONTRATOSOLICITUDMIC','MTOCAMBIOALNUEVOSOL
','MARCA','FLG_CAMPAÑA','PD_CLIENTE','CODVENDEDOR','CODAPROBADOR',
            'CANALVENTA','CANALEVALUACION')
DB_Banco<- DB_Banco[,!names(DB_Banco)%in%Filtro3]
#Paso 4. Creación/actualización de variables según catálogo----
HomologacionZona<-as.data.frame(read_xlsx('Homologacion.xlsx',sheet =
"ZonaFinal"))
DB_Banco[,"FLGTIPLOCAL"]<-
HomologacionZona[match(DB_Banco$CODOFI,HomologacionZona$CODOFI),"FLGTIPLOC
AL"]
DB_Banco[,"ZONA_FINAL"]<-
HomologacionZona[match(DB_Banco$CODOFI,HomologacionZona$CODOFI),"ZONA_FINA
L"]
DB_Banco<-
DB_Banco[,!names(DB_Banco)%in%c('ZONA','AREA','REGION','CODOFI','DESCODDEP
ARTAMENTO')]
rm(HomologacionZona)
HomologacionJustificacion<-
as.data.frame(read_xlsx('Homologacion.xlsx',sheet = "Justificacion"))
DB_Banco[,"JUSTIFICACION"]<-
HomologacionJustificacion[match(DB_Banco$DESJUSTIFICACION,HomologacionJust
ificacion$DESJUSTIFICACION),"DESJUSTIF_H"]
DB_Banco<- DB_Banco[,!names(DB_Banco)%in%c('DESJUSTIFICACION')]
rm(HomologacionJustificacion)
attach(DB_Banco)
DB_Banco[ESTADO=="ACE","ESTADO"] <- 1
DB_Banco[ESTADO=="DEN","ESTADO"] <- 0
DB_Banco[,"ESTADO"] <- as.numeric(DB_Banco[,"ESTADO"])
DB_Banco[TIPO_CAMPANA=="DIA DE LA MADRE","TIPO_CAMPANA"]<-"DMA"
DB_Banco[TIPO_CAMPANA=="ESCOLAR","TIPO_CAMPANA"]<-"ESC"
DB_Banco[TIPO_CAMPANA=="FIESTAS PATRIAS","TIPO_CAMPANA"]<-"FPA"
DB_Banco[TIPO_CAMPANA=="NAVIDAD","TIPO_CAMPANA"]<-"NAV"
DB_Banco[CANALVENTA=="FUNCIONARIO CON CARTERA" &
!is.na(CANALVENTA),"CANALVENTA"] <- "FCC"
DB_Banco[CANALVENTA=="FUNCIONARIO SIN CARTERA" &
!is.na(CANALVENTA),"CANALVENTA"] <- "FSC"
```

```
DB_Banco[CANALVENTA=="OTROS" & !is.na(CANALVENTA) |
is.na(CANALVENTA),"CANALVENTA"] <- "OTR"
DB_Banco[SEG_COM=="1","SEG_COM"] <- "S1"
DB_Banco[SEG_COM=="2","SEG_COM"] <- "S2"
DB_Banco[SEG_COM=="3","SEG_COM"] <- "S3"
DB_Banco[SEG_COM=="4","SEG_COM"] <- "S4"
DB_Banco[SEG_COM=="SIN CLASIFICACION","SEG_COM"] <- "SC"
DB_Banco[SEGMENTO=="1. A1","SEGMENTO"] <- "A1"
DB_Banco[SEGMENTO=="2. A2","SEGMENTO"] <- "A2"
DB_Banco[SEGMENTO=="3. B1","SEGMENTO"] <- "B1"
DB_Banco[SEGMENTO=="4. B2","SEGMENTO"] <- "B2"
DB_Banco[SEGMENTO=="5. C","SEGMENTO"] <- "C"
DB_Banco[SEGMENTO=="6. D","SEGMENTO"] <- "D"
DB_Banco[SEGMENTO=="DEFAULT","SEGMENTO"] <- "F"
DB_Banco[SEGMENTO=="INACTIVO","SEGMENTO"] <- "I"
DB_Banco[SEGMENTO=="SIN CLASIFICACION","SEGMENTO"] <- "S"
#Paso 5. Eliminamos registros que puedan causar sesgo.
DB_Banco<-DB_Banco[DB_Banco$CODCLAVECIC!="999999",]
DB_Banco<-DB_Banco[!is.na(DB_Banco$ESTADO),]
DB_Banco=DB_Banco[,!names(DB_Banco)%in%c("CODCLAVECIC")]
#Paso 6. Correccion de formatos finales.
DB_Banco[,"MES"]<-as.numeric(DB_Banco[,"MES"])
DB_Banco[,"MTOAPROBADO"]<-as.numeric(DB_Banco[,"MTOAPROBADO"])
DB_Banco[,"MONTOSOLES"]<-as.numeric(DB_Banco[,"MONTOSOLES"])
DB_Banco[,"MONTODOLARES"]<-as.numeric(DB_Banco[,"MONTODOLARES"])
DB_Banco[,"TASA"]<-as.numeric(DB_Banco[,"TASA"])
DB_Banco[,"ESTADO"]<-as.numeric(DB_Banco[,"ESTADO"])
detach(DB_Banco)
```

Anexo 04. Modelo de regresión logística inicial, código en R

```
[…]
#Paso 1: Selección aleatoria de casos----
set.seed(1234)
ratio<-0.8
s_train<-sample(seq(nrow(DB_Banco)), size = floor(nrow(DB_Banco) * ratio),
replace = F)
s_test<-seq(nrow(DB_Banco))[!seq(nrow(DB_Banco))%in%s_train]
[…]
#Paso 3: Generación del modelo inicial e indicadores----
model<-glm(ESTADO ~ .,data = DB_Banco[s_train,], family = binomial)
p<-predict(model,type="response")
pr<-prediction(p,DB_Banco[s_train,"ESTADO"])
PerfROC<-performance(pr,"tpr","fpr")
PerfGINI<-performance(pr,"auc")
GINI<-abs(1-2*(unlist(PerfGINI@y.values)))
plot(PerfROC)
```

Anexo 05. Modelo de regresión logística optimizado según AIC, código en R

```
[…]
#Paso 4: Optimización del modelo final según AIC e indicadores-----
model_opt<-glm(ESTADO ~ .,data = DB_Banco[s_train,], family = binomial)
model_opt<-stepAIC(object = model_opt,direction = "backward")
p_opt<-predict(model_opt,type="response")
pr_opt<-prediction(p_opt,DB_Banco[s_train,"ESTADO"])
PerfROC_opt<-performance(pr_opt,"tpr","fpr")
PerfGINI_opt<-performance(pr_opt,"auc")
GINI_opt<-abs(1-2*(unlist(PerfGINI_opt@y.values)))
plot(PerfROC_opt)
```

Anexo 06. Modelo final, regresión logística (entrenamiento y prueba), código en R

```
[…]
#Paso 5: Ajuste de captura de variables del modelo optimizado-----
[…]
model_end<-glm(ESTADO ~ .,data = DB_Banco[s_train,], family = binomial)
p_end<-predict(model_end,type="response")
```

```
pr_end<-prediction(p_end,DB_Banco[s_train,"ESTADO"])
PerfROC_end<-performance(pr_end,"tpr","fpr")
PerfGINI_end<-performance(pr_end,"auc")
GINI_end<-abs(1-2*(unlist(PerfGINI_end@y.values)))
plot(PerfROC_end)
```

Anexo 07. Red neuronal artificial, código en R

```
#Función de normalización----
norm.01<-function(x){
  x <- (x - min(x))/(max(x) - min(x))
  return (x)
}
#Paso 1: Eliminamos variables correlacionadas-----
DB_Banco_nn =
DB_Banco_nn[,!names(DB_Banco_nn)%in%c("MONTODOLARES","MTOAPROBADO")]

#Paso 2: Normalización de variables-----
DB_Banco_nn = as.data.frame(sapply(DB_Banco_nn,norm.01))

[…]
#Paso 1: Entrenamiento----
n <- names(DB_Banco_nn)
f <- as.formula(paste("ESTADO ~", paste(n[!n %in% c('ESTADO')], collapse =
" + ")))
nn <-
neuralnet(f,data=DB_Banco_nn[s_train,],hidden=c(3,1),linear.output=F,
                lifesign = "full",threshold = 0.05,algorithm =
'rprop+',learningrate = 0.002,stepmax = 200000)

#Paso 2: Entrenamiento, indicadores----
output.train=compute(nn,DB_Banco_nn[s_train,-
which(names(DB_Banco_nn)=="ESTADO")])
output.train.pr<-
ROCR::prediction(unlist(output.train$net.result),DB_Banco_nn[s_train,"ESTA
DO"])
output.train.perf.GINI<-ROCR::performance(output.train.pr,"auc")
output.train.perf.ROC<-ROCR::performance(output.train.pr,"tpr","fpr")
GINI_nn_train<-abs(1-2*(unlist(output.train.perf.GINI@y.values)))

cutoff = 0.5 # según distribución de la data
y=as.numeric(as.character(unlist(output.train.pr@labels)))
y_hat=as.numeric(unlist(output.train.pr@predictions)>cutoff)
tp_v=as.numeric(y_hat==1 & y==1)
fp_v=as.numeric(y_hat==1 & y==0)
tn_v=as.numeric(y_hat==0 & y==0)
fn_v=as.numeric(y_hat==0 & y==1)

tp = sum(tp_v,na.rm = T)
tn = sum(tn_v,na.rm = T)
fp = sum(fp_v,na.rm = T)
fn = sum(fn_v,na.rm = T)
npos = tp+fn
nneg = fp+tn

sens_nn = tp/(tp+fn)
spec_nn = tn/(fp+tn)
acc_nn = (tp + tn)/(npos+nneg)
prec_nn = tp/(tp+fp)
f1_nn = 2*tp/(2*tp+fp+fn)

#Prueba: indicadores----
output.test=compute(nn,DB_Banco_nn[s_test,-
which(names(DB_Banco_nn)=="ESTADO")])
```

```
output.test.pr<-
ROCR::prediction(unlist(output.test$net.result),DB_Banco_nn[s_test,"ESTADO
"])
output.test.perf.GINI<-ROCR::performance(output.test.pr,"auc")
output.test.perf.ROC<-ROCR::performance(output.test.pr,"tpr","fpr")
GINI_nn_test<-abs(1-2*(unlist(output.test.perf.GINI@y.values)))

cutoff = 0.5 # según distribución de la data
y=as.numeric(as.character(unlist(output.test.pr@labels)))
y_hat=as.numeric(unlist(output.test.pr@predictions)>cutoff)
tp_v=as.numeric(y_hat==1 & y==1)
fp_v=as.numeric(y_hat==1 & y==0)
tn_v=as.numeric(y_hat==0 & y==0)
fn_v=as.numeric(y_hat==0 & y==1)

tp = sum(tp_v,na.rm = T)
tn = sum(tn_v,na.rm = T)
fp = sum(fp_v,na.rm = T)
fn = sum(fn_v,na.rm = T)
npos = tp+fn
nneg = fp+tn

sens_nn_test = tp/(tp+fn)
spec_nn_test = tn/(fp+tn)
acc_nn_test = (tp + tn)/(npos+nneg)
prec_nn_test = tp/(tp+fp)
f1_nn_test = 2*tp/(2*tp+fp+fn)
```