TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

TESIS

MODELO PREDICTIVO DE ANÁLISIS DE RIESGO CREDITICIO USANDO MACHINE LEARNING EN UNA ENTIDAD DEL SECTOR MICROFINANCIERO

PRESENTADA POR:

MIGUEL ROMILIO ACEITUNO ROJO

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PUNO, PERÚ

2019

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

TESIS

MODELO PREDICTIVO DE ANÁLISIS DE RIESGO CREDITICIO USANDO MACHINE LEARNING EN UNA ENTIDAD DEL SECTOR MICROFINANCIERO

PRESENTADA POR:

MIGUEL ROMILIO ACEITUNO ROJO

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

APROBADA POR EL SIGUIENTE JURADO:

PRESIDENTE

.....
Dr. LEONEL COYLA IDME

PRIMER MIEMBRO

.....
Dr. ELVIS AUGUSTO ALIAGA PAYEHUANCA

SEGUNDO MIEMBRO

.....
Dra. GUINA GUADALUPE SOTOMAYOR ALZAMORA

ASESOR DE TESIS

.....
Dr. HENRY IVAN CONDORI ALEJO

Puno, 03 de diciembre de 2019

ÁREA: Ciencias de la Computación.
TEMA: Formalización extraordinaria.
LÍNEA: Inteligencia Artificial.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

DEDICATORIA

Con mucho afecto para mis seres queridos:

mis hijos Anthony y Steve y

mi esposa Donia Alizandra Ruelas Acero,

mi padre Miguel Aceituno Huacani (Q.E.P.D.)

mi madre Ana María Rojo Mamani

y mi hermana Edith Roxana Aceituno Rojo.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

AGRADECIMIENTOS

A la Universidad Nacional del Altiplano – Puno, por la formación de posgrado que he recibido.

A los miembros del jurado: Dr. Leonel Coyla Idme, Dr. Elvis Augusto Aliaga Payehuanca, Dra. Guina Guadalupe Sotomayor Alzamora y Dr. Henry Iván Condori Alejo; por sus recomendaciones y sugerencias durante el desarrollo de la presente investigación.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

ÍNDICE GENERAL

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

ÍNDICE DE TABLAS

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

ÍNDICE DE FIGURAS

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

ÍNDICE DE ANEXOS

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

RESUMEN

Los microcréditos constituyen un componente importante en el desarrollo de la economía rural del país, éstos en su mayoría son otorgados por entidades microfinancieras, que tratan con índices altos de riesgo, estos son controlados por medio de personal especializado que realiza la evaluación y verificación de los clientes que solicitan estos microcréditos. La presente investigación propone un modelo de predicción de riesgo crediticio a partir de la evaluación de distintos modelos de *Machine Learning* para determinar el modelo que presenta el mejor nivel de asertividad en el otorgamiento de microcréditos. Para tal efecto se ha determinado las principales variables que intervienen en el proceso, seguidamente, se realizó el entrenamiento de seis modelos de *Machine Learning,* se logró determinar el nivel de asertividad en base a las métricas *Accuracy, Precision, Recall, F1 Score* y *AUC* para luego ser comparados, resultando ser el más asertivo *Artificial Neural Network* en comparación a Regresión Logística, *Random Forest*, *Support Vector Machine*, *Decision Tree* y *k-Nearest Neighbors,* de lo determinado, se afirma que con el modelo más asertivo se puede reducir riesgo crediticio al mejorar el nivel de asertividad en el otorgamiento microcrédito en base a las variables determinadas.

Palabras clave:

Machine Learning, Microfinanzas, Microcrédito, Riesgo crediticio, Sector rural.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

ABSTRACT

The microcredits constitute an important component in the development of the rural economy of the country, these are mostly granted by microfinance entities, which deal with high rates of risk, these are controlled by specialized personnel who perform the evaluation and verification of the clients requesting these microcredits. The present investigation proposes a credit risk prediction model based on the evaluation of different Machine Learning models to determine the model that presents the best level of assertiveness in the granting of microcredits. For this purpose, the main variables involved in the process have been determined, then the training of six Machine Learning models was carried out, the assertiveness level was determined based on the Accuracy, Precision, Recall, F1 Score and AUC metrics to then be compared, proving to be the most assertive Artificial Neural Network in comparison to Logistic Regression, Random Forest, Support Vector Machine, Decision Tree and k-Nearest Neighbors, it is stated that with the most assertive model risk can be reduced credit to improve the level of assertiveness in the microcredit grant based on the determined variables.

Keywords:

Credit Risk , Machine Learning, Microcredit, Microfinance, Rural Sector.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

INTRODUCCIÓN

Los microcréditos son una fuente principal de financiamiento formal para las diferentes actividades económicas en el país, así como en el sector rural, el cual impulsa la economía del país, este sector se dedica principalmente a la agricultura y ganadería para lo cual requiere de fuentes de financiamiento a través de distintos medios tales como los microcréditos que ofrecen las entidades microfinancieras, pero las mismas observan un alto nivel de riesgo en este sector debido a la falta o escaza información financiera de las personas y a su vez la no disposición de activos de calidad que garanticen los microcréditos, motivo por el cual el sector rural tiene acceso limitado a este servicio (Wendel & Harvey, 2006).

El otorgamiento de microcréditos sigue un proceso determinado, que es realizado por un personal especializado conocido como asesor de negocios o analista de créditos que tienen diferentes niveles en razón a la experiencia de los mismos, quienes utilizan la información para determinar la voluntad y capacidad de pago del cliente, procedimiento que mejoraron usando métodos computacionales, entre ellos los que usan la IA como *Machine Learning* y *Deep Learning*.

Machine Learning se volvió popular en las últimas décadas, el cual brinda habilidad de que un computador aprenda de la experiencia (Goodfellow et al., 2016; M. Jordan & Mitchell, 2015), el cual denota un proceso de estimar un modelo del mundo real a partir de un conjunto de datos (Wang & Tao, 2008).

La presente investigación tiene el objetivo de determinar el mejor modelo de *Machine Learning* que permita mejorar el nivel de asertividad en el otorgamiento de microcréditos, para lo cual se realizó la determinación de las variables a través de revisión de la literatura relacionadas al tema de la investigación y el análisis del proceso de otorgamiento de microcréditos de la entidad, seguidamente se realizó la preparación de datos, para el entrenamiento de los modelos de *Machine Learning*, siendo éstos evaluados a través de métricas tales como *Accuracy, Precision, Recall, F1 Score* y *AUC ROC* que permitió determinar el nivel de asertividad de los modelos.

Como resultado se obtuvo la identificación de un total de 34 variables independientes para el entrenamiento de los modelos a fin de determinar si un crédito debe ser otorgado o no, así mismo se determinó que el modelo más asertivo es *Artificial Neural Networks* con un nivel de asertividad 93.72 % en comparación con los modelos Regresión Logística

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

86.07 %, *Random Forest* 66.35 %, *Support Vector Machine* 84.44 %, *Decision Tree* 88.80% y *k-Nearest Neighbors* 65.98 %, los cuales también ofrecen un nivel alto de asertividad.

La presente investigación se ha estructurado en cuatro capítulos, en el Capítulo I se desarrolla la revisión bibliográfica expresada en el marco teórico y los antecedentes de investigación; en el Capítulo II se presenta la identificación y formulación del problema, la justificación y los objetivos de la investigación; en el Capítulo III se presenta los materiales y métodos que se utilizó en la investigación, también se describe la población y muestra considerada; en el Capítulo IV se exponen los resultados obtenidos y finalmente se expone las conclusiones y recomendaciones a las que se llegó con la investigación.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1 Marco Teórico

1.1.1 El sistema financiero peruano

El sistema financiero tiene un rol importante en el desarrollo y crecimiento de la economía, permitiendo expandir la frontera de producción y alcanzar mayores niveles de utilidad, es decir, mejorar el nivel del bienestar social; el sistema financiero peruano tiene diversos tipos de entidades que cumplen determinados requisitos y características que son definidos por el ente regulador denominado Superintendencia de Bancos y Seguros (SBS) (Lizarzaburu, 2014).

1.1.2 Crédito

El crédito es una herramienta que permite a acceder a un capital a través de diferentes tipos de productos, para incrementar la capacidad de producción o adquirir equipos que faciliten el trabajo de personas naturales o jurídicas, vale decir que el crédito fue usado para reactivar la economía en los siglos XIX y XX (Morales & Morales, 2014).

Guajardo (1991) define crédito como una estrategia de un determinado valor actual que puede ser dinero, producto o servicio, sobre la base de una confianza a cambio de un valor equivalente esperado en un futuro, teniendo la posibilidad de un interés pactado. Así mismo la SBS lo define como la suma de los créditos directos e indirectos, donde los directos son representados por cualquier producto otorgado a la persona que implique la entrega de una suma de dinero a la misma, y los indirectos son representados por la suma aquellos créditos que la persona avala o garantiza, las cartas fianza y las líneas de crédito no utilizadas (SBS, 2008).

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Otra definición especifica crédito como una herramienta que permite que una determinada persona acceda a un determinado capital que será reintegrado en un determinado tiempo con un interés pactado, para asegurar el reintegro del crédito otorgado se hace uso de las garantías, dado que en caso que el cliente no cumpliese el pago del crédito en el tiempo pactado se inicia con la ejecución de las mismas (Morales & Morales, 2014).

1.1.3 Tipos de créditos

La SBS (2008) define ocho tipos de créditos: créditos corporativos, créditos a grandes empresas, créditos a medianas empresas, créditos a pequeñas empresas, créditos a microempresas, crédito revolvente, créditos consumo no revolvente y créditos hipotecarios para empresas. Los créditos a microempresas buscan la financiación de actividades de producción, comercialización o prestación de servicios (a las cuales pueden acceder las personas naturales o jurídicas) en estas, la suma de sus créditos directos e indirectos no debe superar el monto de 20,000.00 soles en los tres últimos meses, si la persona supera este monto el tipo de crédito se reclasifica.

1.1.4 Clasificación crediticia

Así mismo, la SBS (2008) clasifica a las personas en cinco categorías de acuerdo al atraso en el pago de sus créditos.

- **Categoría normal:** Personas que cumplen sus pagos con una tolerancia de hasta 8 días calendarios de atraso.
- **Categoría con Problemas Potenciales:** Personas pagan sus créditos con un atraso entre 9 a 30 días calendarios.
- **Categoría Deficiente:** Personas pagan sus créditos con un atraso entre 31 a 60 días calendarios.
- **Categoría Dudoso:** Personas pagan sus créditos con un atraso entre 61 a 120 días calendarios.
- **Categoría Pérdida:** Personas pagan sus créditos con un atraso con más de 121 días calendarios.

1.1.5 Otorgamiento de créditos

La decisión de otorgar un crédito está ligada a una investigación conocida como evaluación de crédito, esta evaluación previa determina si la persona que solicita el

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

préstamo lo va pagar, realizado por un personal especializado conocido como analista de créditos, que identifica la voluntad y capacidad de pago de acuerdo a sus antecedentes crediticios (Morales & Morales, 2014).

La evaluación de crédito es un procedimiento para determinar la capacidad de pago de la persona, en esta se analizan sus antecedentes de los créditos otorgados en la central de riesgo de su país, los estados financieros del mismo, así como las garantías que la persona otorga (Chavez, 2017).

Uno de los criterios de evaluación está determinado por la capacidad de pago, el cual es definido en razón de su flujo de caja, obligaciones, patrimonio neto, calificaciones en el sistema financiero, así como sus antecedentes crediticios, y otros indicadores dependiendo el tipo de crédito (SBS, 2008).

Según Morales & Morales (2014) el proceso que sigue el otorgamiento de crédito se inicia con la investigación seguida del análisis y la aceptación del cliente como se muestra en la Figura 1, en esta sección se explica cada uno de los sub procesos que componen este proceso.

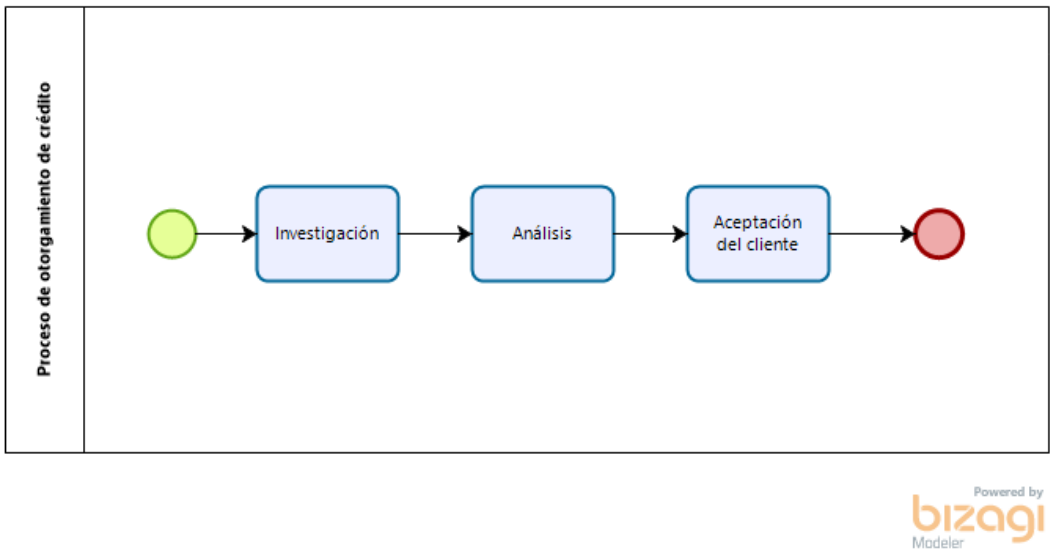


Figura 1. Proceso general de otorgamiento de crédito

Fuente: Adaptado de Morales & Morales (2014)

En el sub proceso de **investigación** se indaga sobre la capacidad financiera y los antecedentes financieros del cliente a fin de acceder a un crédito; para lo cual se investiga los datos de los créditos otorgados, referencias de riesgos, información de

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

otros proveedores y estados financieros del solicitante, en la que los autores resaltan los siguientes datos:

- Registro de pagos, permite predecir el habito de pago basado en su historial crediticio, es uno de los factores más importantes.
- Ingreso, se debe realizar una evaluación de acuerdo a sus ingresos y la fuente de los mismos además de sus obligaciones, lo que determinará la posibilidad de pago de acuerdo al periodo que solicita el crédito.
- Empleo, datos sobre la fuente principal de sus ingresos, la antigüedad en el mismo, y el puesto que ocupa.
- Residencia, se indaga en el lugar donde las personas residen, en la cual debería tener una antigüedad recomendada de 5 años, si fuera menor debería indagarse su anterior residencia y los motivos de cambios del mismo.
- Estado civil, se indaga el dato debido a que puede este afecta en sus ingresos y a su vez en la actitud de la persona respecto a una determinada obligación.
- Edad, es investigado debido a que existen patrones de pagos de acuerdo a la edad de la persona, dado que un joven puede que logre tener los ingresos suficientes para cubrir sus obligaciones, o una persona mayor, puede que tenga una jubilación, pero tiene la probabilidad de tener otros gastos como salud u otros.
- Referencias y reputación, se evalúa el carácter de lo indicado sea correcto y confiable.

En el **análisis** se realiza el estudio de los datos recabados en la investigación para determinar si el crédito es otorgado o no, así mismo se analizan los hábitos de pagos y la capacidad de compromiso de pago, lo cual está contenido en el informe de crédito y estados financieros, donde se logra analizar el capital propio en relación al crédito que se desea otorgar.

Finalmente, la **aceptación de cliente** consiste en definir si el crédito será otorgado o no, en el caso que sea otorgado se define las condiciones del crédito tales como el monto, plazo entre otros.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

El detalle del otorgamiento de crédito es mostrado en la Figura 2, este se basa en el procedimiento general antes expuesto, donde las entidades microfinancieras adoptan este procedimiento planteado por los autores debido a que este se adecua mejor al otorgamiento de microcréditos; cada entidad modifica según sus necesidades, considerando sus productos, políticas y metodología de trabajo para evaluar a la persona solicitante y determinar si la solicitud de crédito se acepta o se niega.

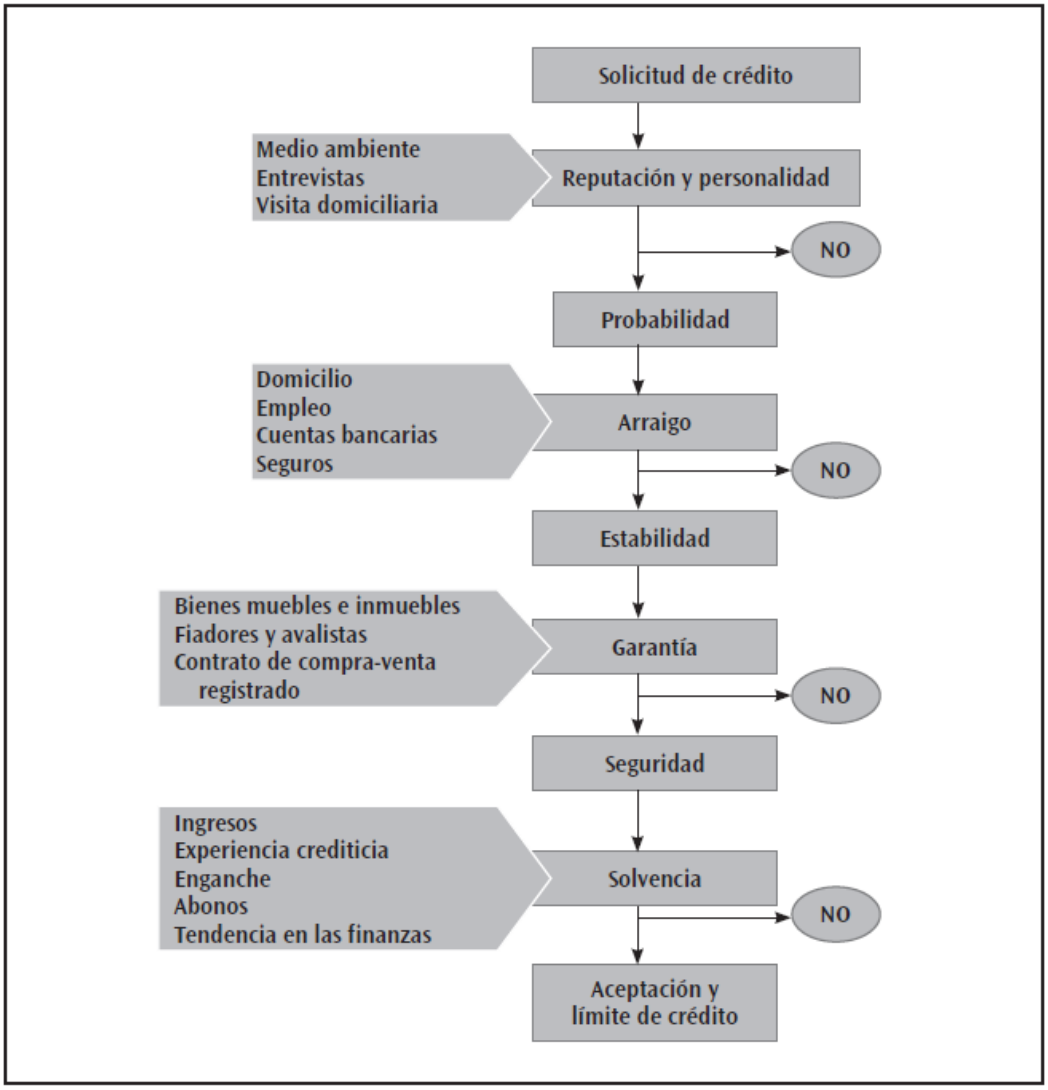


Figura 2. Proceso de análisis detallado para el otorgamiento de un crédito

Fuente: Morales & Morales (2014)

1.1.6 Riesgo crediticio

Indica la probabilidad del incumplimiento de las obligaciones crediticias que generarían perdidas a las entidades, es uno de los principales riesgos en sector financiero dado que afectan directamente a la rentabilidad y solidez de estas, es por

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

ello que toda entidad dedicada a las finanzas o microfinanzas debe contar con un sistema o metodología de gestión de riesgos para poder tener una forma de medir el riesgo crediticio, donde una que más acogen las entidades microfinancieras son las cinco “C” del crédito (Berger et al., 2003; Jiménez, 2016).

Para Morales & Morales (2014) en esta metodología contempla las características del perfil del cliente, situación del negocio y la situación de la industria, aquí se consideran criterios tales como Conducta, Capacidad de pago, Capacidad de endeudamiento, Capacidad de pago proyectada y Condiciones macroeconómicas como se muestra en la Figura 3.

Conducta: Evalúa la conducta del cliente, donde se determinan la calidad moral del mismo realizando un análisis cualitativo del riesgo del cliente, y a su vez se evalúa el grado de evidencia en la información, experiencia en pago en los últimos 24 meses, el conocimiento y experiencia del cliente en la entidad y en otras entidades, esto puede ser consultado en la Superintendencia de Bancos y Seguros. Así mismo se analiza las demandas administrativas y judiciales si la tuviera, también se evalúa el tipo de administración y estructura organizacional con lo cual se determina la capacidad que tiene el cliente para optimizar sus operaciones, la toma de decisiones, asertividad y visión empresarial.

Capacidad de pago: Evalúa si el cliente tuvo los recursos suficientes para el pago de sus deberes, analizando dos factores de riesgo: la operación histórica que evalúa la ventas y utilidades, y el flujo neto histórico que implica la capacidad de generar recursos como dinero desde su negocio o actividad económica.

Capacidad de endeudamiento: Mide la solidez de la estructura financiera del cliente analizando factores como liquidez, apalancamiento y la rentabilidad y eficiencia, estos son calculados con una serie de fórmulas específicas.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

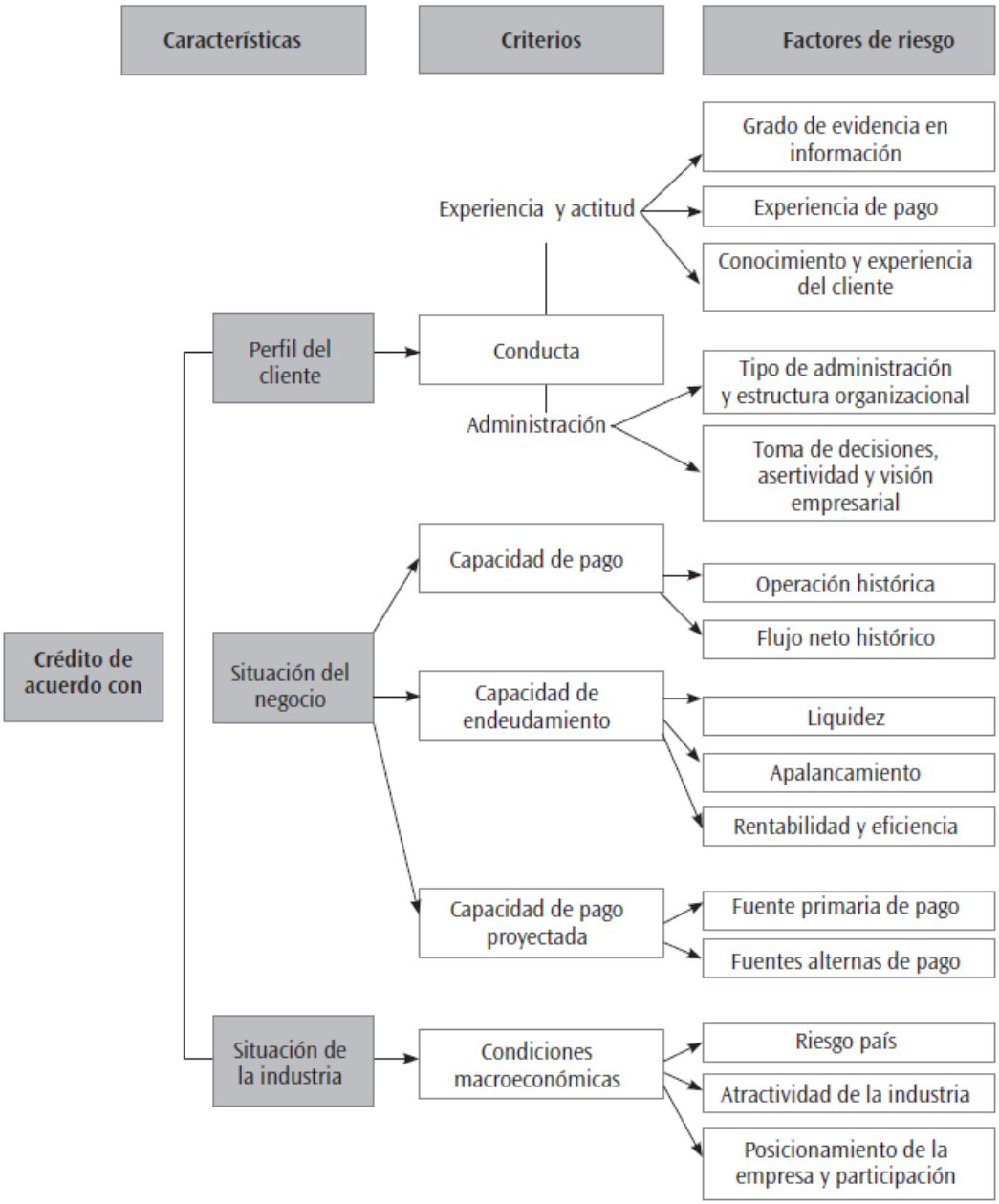


Figura 3. Las cinco "C" del crédito

Fuente: Morales & Morales (2014)

Condiciones macroeconómicas: Determina el comportamiento de la actividad económica en su conjunto, observando los factores como el riesgo del país que está directamente relacionado a su situación política y económica, atractividad de la industria y el posicionamiento de la empresa y participación.

Capacidad de pago proyectada: Analiza la capacidad de generar dinero a futuro para cumplir con sus obligaciones financieras, incluyendo factores como fuente primaria de pago, los cuales están expresados en el balance del cliente y las fuentes alternas de pago, que son recursos adicionales diferente fuente a la primaria de ingreso.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

1.1.7 Microcrédito

Las entidades microfinancieras otorgan a las personas un conglomerado de productos tales como: producto contra depósito, producto agropecuario, producto hipotecario, entre otros. El producto agropecuario es un microcrédito que se otorga a personas naturales o jurídicas que financian las actividades agropecuarias, este microcrédito puede ser otorgados en periodos mensuales, bimestrales, trimestrales, cuatrimestrales y semestrales, que no pueden exceder los doce meses (CRACLASA, 2019).

1.1.8 *Machine Learning*

Machine Learning es la habilidad de que un computador aprenda de la experiencia (Goodfellow et al., 2019; Jordan & Mitchell, 2015), este aprendizaje se da a través de un programa de computadora en el cual, dada la experiencia *E* con respecto a una clase de tarea *T*, usando medidas de desempeño tipo *P*, si su desempeño de la tarea *T* medida por *P* mejora la experiencia en *E* (Mitchell, 1997), donde la experiencia está basada en los datos generados de una determinada tarea o actividad realizada y esta es evaluada o medida a través de una determinada métrica para observar si esta mejora en la realización de la tarea definida. Esto permite abordar tareas complejas con un nivel de competencia igual o mejor que un humano experto (Luger, 2013).

Machine Learning tiene un efecto substancial en diversas áreas de la tecnología y la ciencia (Jordan & Mitchell, 2015), además viene siendo aplicado en el área de las finanzas (Tesén, 2017) y microfinanzas no es la excepción (Kalayci et al., 2018). Que se identifican dos tipos de aprendizajes que son el aprendizaje supervisado (Addo et al., 2018), aprendizaje no supervisado (Turkson et al., 2016). Se debe considerar que en finanzas se viene utilizando con más frecuencia el aprendizaje supervisado debido a la información histórica que tienen las entidades.

Kotsiantis (2007) sostiene que el aprendizaje supervisado es la búsqueda de algoritmos que razonan a partir de instancias suministradas externamente, para producir hipótesis generales que luego hacen predicciones sobre instancias futuras. Baştanlar & Özuysal (2014), mencionan que las técnicas de aprendizaje supervisado construyen modelos predictivos al aprender de una gran cantidad de ejemplos de entrenamiento, donde cada ejemplo de entrenamiento tiene una etiqueta que indica su salida. Sin embargo, Zhou (2018) considera que, aunque las técnicas actuales han logrado un gran éxito, cuando se tiene una sólida información de supervisión, es difícil obtener información de

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

supervisión sólida, como etiquetas de verdad completa debido al alto costo del proceso de etiquetado de datos. Los algoritmos de aprendizaje supervisado incluyen regresión lineal, regresión logística, *Random Forest, Support Vector Machine Neural Network Artificial, Decision Tree, k-Nearest Neighbors* entre otros.

1.1.9 Modelos de *Machine Learning*

1.1.9.1 Regresión Logística

Según Ng (2018) la regresión logística es un algoritmo de clasificación de aprendizaje supervisado en la que establece una relación de variables independientes representada por el conjunto $X = \{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(i)}, \dots, x^{(m)}\}$ y una variable dependiente conocida como y a través de la siguiente ecuación:

$$z^{(i)} = w^T x^{(i)} + b$$

$$a^{(i)} = \sigma(z^{(i)})$$

$$L(a^{(i)}, y^{(i)}) = -y^{(i)} \log(a^{(i)}) - (1 - y^{(i)}) \log(1 - a^{(i)})$$

$$J = \frac{1}{m} \sum_{i=1}^{m} L(a^{(i)}, y^{(i)})$$

Donde:

$x^{(i)}$: $Iésimo\ ejemplo\ de\ entrenamiento.$
$y^{(i)}$: $Etiqueta\ de\ salida\ el\ iésimo\ ejemplo\ de\ entrenamiento.$
W: $El\ peso\ vectorial.$
b: $Bias, variable\ que\ apoya\ en\ generalización\ del\ modelo$
$a^{(i)}$: $Predicción$
σ: $Función\ sigmoide$
L: $Función\ error$
J: $Función\ costo$
m: $Número\ de\ ejemplos\ de\ entrenamiento\ en\ el\ dataset$

La regresión logística es uno de los modelos de clasificación más conocidos y utilizados en análisis de riesgo crediticio, como los trabajos de Millán & Caicedo (2018), Addo et al. (2018), Kalayci et al., (2018), Arango & Restrepo (2017), Valencia (2017), Kruppa et al. (2013) entre otros.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

1.1.9.2 Support Vector Machine (SVM)

Support Vector Machine es un algoritmo de aprendizaje supervisado, identificado como clasificador binario, lo que implica que las etiquetas de clasificación deben ser 0 o 1, Verdadero o Falso, Azul o Verde, entre otros, este algoritmo utiliza un enfoque diferente al probabilístico que es usado en otros algoritmos de *Machine Learning*, permitiendo razonar de una forma geométrica basándose en productos internos y proyecciones (Deisenroth et al., 2019). A su vez indican que es un poderoso método de aprendizaje automático en la clasificación de datos (Liang et al., 2016). En general, la idea principal de *SVM* es determinar el hiperplano de separación que maximiza el margen entre dos clases de datos de entrenamiento, según la teoría de la optimización, dicho hiperplano óptimo se especifica mediante el vector de peso *w* y el sesgo *b*, tomando en cuenta los valores de entrada *x* (Figura 4), que son soluciones del problema de optimización restringida (Nguyen, 2016).

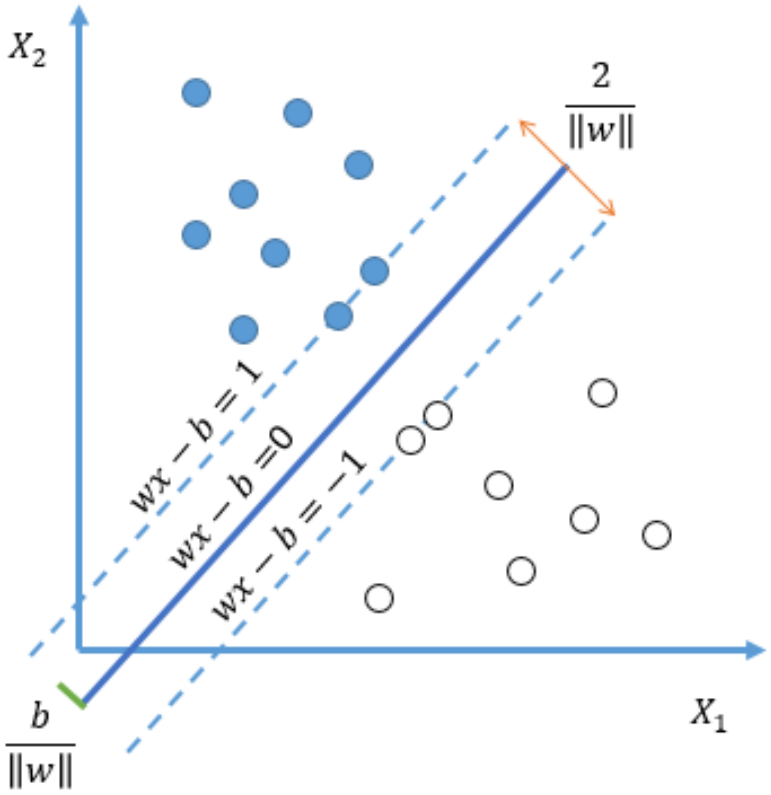


Figura 4. Clasificación a través de *Support Vector Machine*

Fuente: Adaptado de Nguyen (2016)

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

SVM fue utilizado en diversas investigaciones relacionadas al tema de riesgo crediticio tales como: Chakraborty & Joseph (2017), (Kalayci et al., 2018), Flores & Ramon (2014), Turkson et al. (2016) entre otros.

1.1.9.3 Artificial Neural Network (ANN)

ANN es un modelo computacional inspirado en la biología, consiste en elementos de procesamiento (llamados neuronas) y conexiones entre ellos con coeficientes (pesos) unidos a las conexiones. Estas conexiones constituyen la estructura neuronal y se unen a esta estructura los algoritmos de entrenamiento y recuperación. *ANN* se denominan modelos conexionistas debido a las conexiones que se encuentran entre las neuronas (Shanmuganathan, 2016).

Ng (2018) presenta el modelo integral de una *ANN* (Figura 5) donde se integra todos los componentes de dicho modelo como la inicialización de variables, el algoritmo de *Forward Propagation* (Figura 6), la función de pérdida, el algoritmo de *Back Propagation* y la actualización de los parámetros (Figura 7).

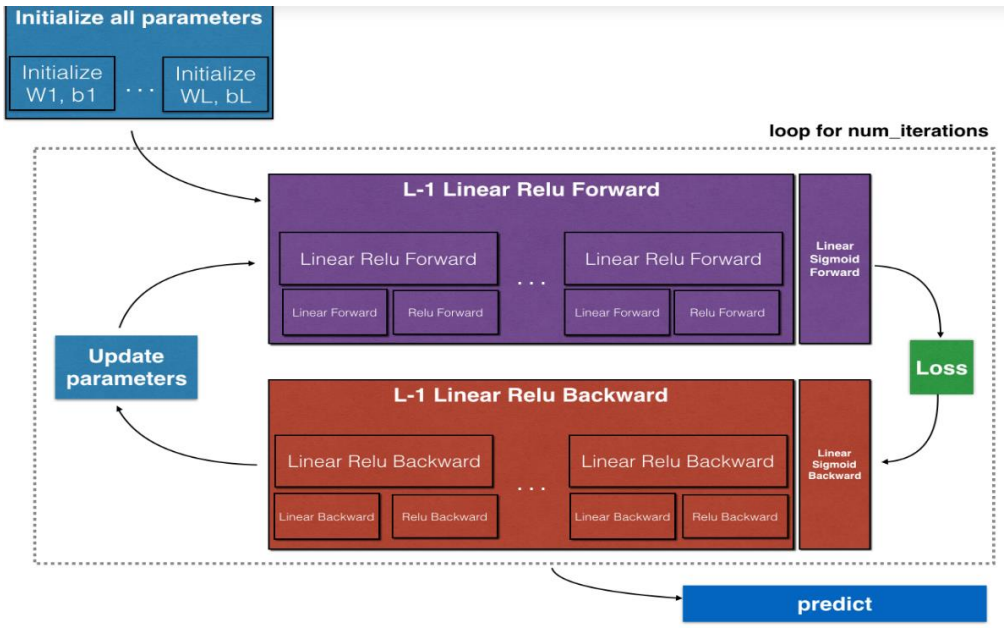


Figura 5. Modelo integral de una *Artificial Neural Network*

Fuente: Ng (2018).

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

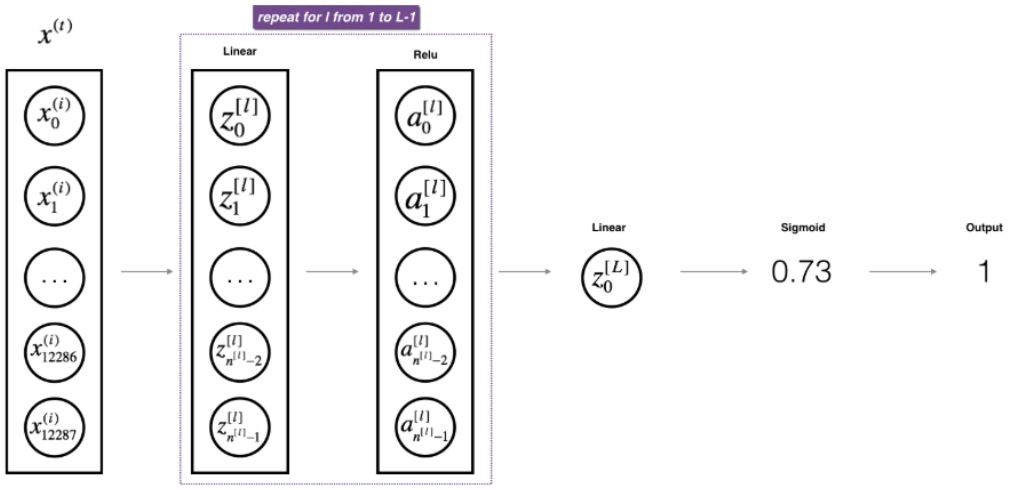


Figura 6. Algoritmo *Forward Propagation* en una *Artificial Neural Network*

Fuente: Ng (2018).

Donde matemáticamente se expresa de la siguiente forma:

$$z^{[1](i)} = W^{[1]}x^{(i)} + b^{[1]}$$

$$a^{[1](i)} = f(z^{[1](i)})$$

$$z^{[2](i)} = W^{[2]}x^{(i)} + b^{[2]}$$

$$a^{[L-1](i)} = sigmoid(z^{[L-1](i)})$$

$$y^{(i)}_{prediction} = \begin{cases} 1\,, si\ a^{[L-1](i)} > 0.5 \\ 0 \end{cases}$$

Donde:

L: *Número de capas en la red neuronal.*

$x^{(i)}$: *Iésimo ejemplo de entrenamiento.*

$W^{[l]}$: *Matriz de pesos en la capa* l.

$b^{[l]}$: *Bias en la capa* l, *variable que apoya en generalización del modelo.*

$a^{[l]}$: *Predicción en la capa* l.

$z^{[l]}$: *Resultado de la función lineal en la capa* l.

f: *Función de activación*

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

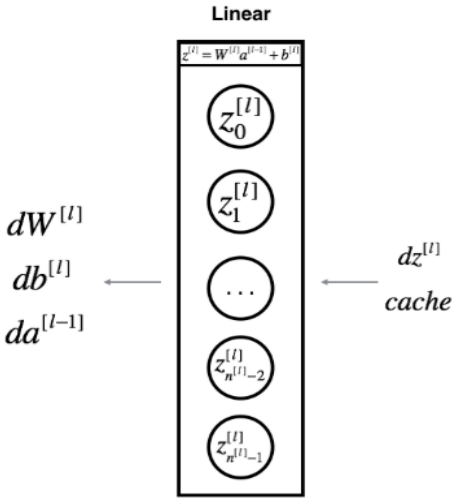


Figura 7. Algoritmo *Backward Propagation* en una *Artificial Neural Network*

Fuente: Ng (2018).

La expresión matemática de *Backward Propagation* se expresa a través de la siguiente forma:

$$dW^{[l]} = \frac{\partial L}{\partial W^{[l]}} = \frac{1}{m} dZ^{[l]} A^{[l-1]T}$$

$$db^{[l]} = \frac{\partial L}{\partial b^{[l]}} = \frac{1}{m} \sum_{i=1}^{m} dZ^{[l](i)}$$

$$dA^{[l-1]} = \frac{\partial L}{\partial W A^{[l-1]}} = W^{[l]T} dZ^{[l]}$$

Donde:

l: $Número\ de\ capa\ de\ la\ red\ neuronal$

$W^{[l]}$: $Matriz\ de\ pesos\ en\ la\ capa\ l.$

$b^{[l]}$: $Bias\ en\ la\ capa\ l.$

$A^{[l]}$: $Predicción\ en\ la\ capa\ l.$

$Z^{[l]}$: $Resultado\ de\ la\ función\ lineal\ en\ la\ capa\ l.$

L: $Función\ error.$

m: $Número\ de\ ejemplos\ de\ entrenamiento\ en\ el\ dataset.$

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

1.1.9.4 Decision Tree

Según (Shalev & Ben, 2014) definen *Decision Tree (DT)* como un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión. El objetivo es crear un modelo que permita predecir el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas de las características de los datos.

Los criterios de decisión de un DT se dan por la siguiente expresión:

$$\Delta I(t) = I(t) - \frac{N_{tS}}{N_t} I(t_S) - \frac{N_{tN}}{N_t}(t_N)$$

Donde:

t: $Representa\ un\ nodo\ en\ el\ árbol$
N_t: $Número\ total\ de\ muestras\ en\ el\ nodeo\ padre\ t.$
N_{tS}: $Número\ total\ de\ muestras\ enviados\ al\ nodo\ SI.$
N_{tN}: $Número\ total\ de\ muestras\ enviados\ al\ nodo\ No.$

1.1.9.5 k-Nearest Neighbors (kNN)

Según Ng (2018) menciona que *kNN* es uno de los muchos algoritmos de aprendizaje supervisado utilizados en el campo de la minería de datos y el aprendizaje automático, es un clasificador donde el aprendizaje se basa en cuán similar es un dato a otro. El entrenamiento está formado por vectores de *n* dimensiones.

Para calcular la distancia entre dos puntos (x, x'), la nueva muestra y todos los demás datos que tiene en su conjunto de datos existen varias formas de obtener este valor, donde la distancia euclidiana ya que es uno de los más usados, y está determinado por la siguiente expresión:

$$\rho(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

Donde:

ρ: $función\ de\ distancia\ euclidiana.$

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

$x_i: i-ésimo\ ejemplo\ de\ entrenamiento.$

$x'_i: i-ésimo\ ejemplo\ de\ predicción.$

$n: cantidad\ de\ atributos$

1.1.9.6 Random Forest (RF)

Breiman (2001) propuso el algoritmo de bosque aleatorio como un método de clasificación y regresión de propósito general. Los bosques aleatorios son una combinación de predictores de árboles, de modo que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles en el bosque. Este autor formalmente define un bosque aleatorio como un clasificador que consiste en una colección de clasificadores estructurados en árbol:

$$h(x, \Theta_k)$$

Donde:

$h: Clasificador\ estructurado\ en\ árbol$

$k: k-ésimo\ árbol$

$x: Vector\ de\ atributos\ de\ entrada$

$\Theta_k: Vectores\ aleatorios\ independientes\ del\ árbol\ k.$

Ayma (2019) indica que la salida final, de un bosque aleatorio, corresponde a la clase de moda entre los árboles, cada árbol del bosque es diferente respecto a los atributos y conjunto de entrenamiento, en donde la selecciona aleatoria de atributos de un espacio, $X\ \epsilon\ R^l$, se crean t árboles de decisión con diferentes espacios de atributos $X_s \epsilon R^{l_s}, R^{l_s} \subseteq R^l$. Así mismo indica que la selección aleatoria de muestras de entrenamiento, X_e, consiste en dividir X_e en T subconjuntos X_t, para entrenar cada árbol de decisión t con un subconjunto de entrenamiento X_t.

En los árboles estándar, cada nodo se divide utilizando la mejor división entre todas las variables. En un bosque aleatorio, cada nodo se divide utilizando el mejor entre un subconjunto de predictores elegidos aleatoriamente en ese nodo. Esta estrategia se desempeña muy bien en comparación con muchos otros clasificadores, incluido el análisis discriminante, las máquinas de vectores de soporte y las redes neuronales, y es robusta contra el sobreajuste (Liaw & Wiener, 2002).

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

1.1.10 Proceso de entrenamiento de modelos de *Machine Learning*

Raschka & Vahid (2017) describe el proceso de entrenamiento de los modelos de *Machine Learning* como se muestra en la Figura 8, el proceso se origina con el pre procesamiento de datos, seguidamente se realiza la configuración del *dataset* en los modelos de *Machine Learning* seleccionados, posteriormente se establece los hiperparámetros y las métricas para obtener los modelos finales que puedan aprender del *dataset*.

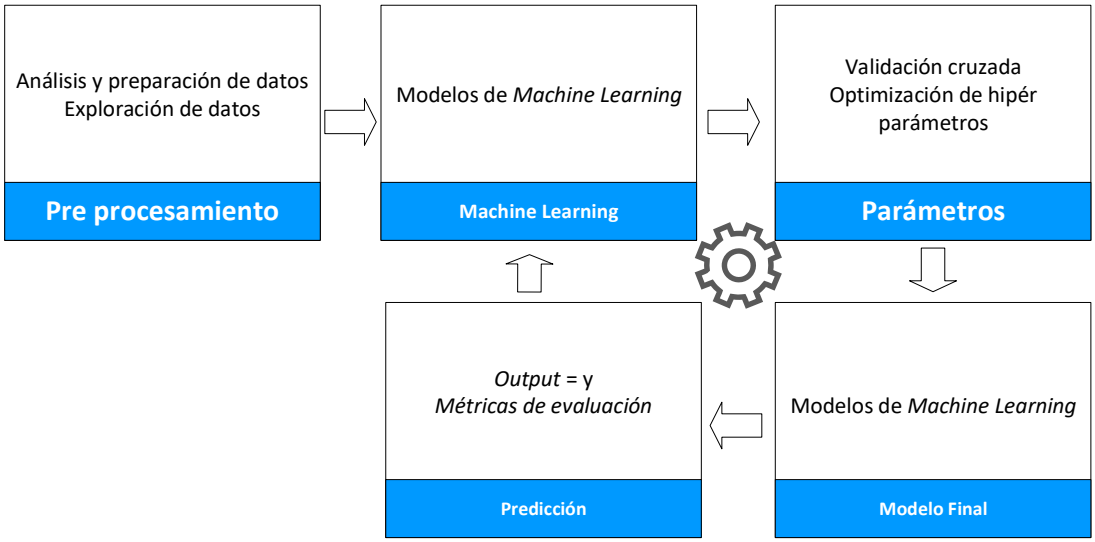


Figura 8. Entrenamiento de modelos de *Machine Learning*

Fuente: Adaptado de Raschka & Vahid (2017)

1.1.11 Codificación *One Hot*

Brownlee (2017) sostiene que la codificación *One Hot* es un proceso mediante el cual las variables categóricas se convierten en una forma que podría proporcionarse a los algoritmos de *Machine Learning* para hacer un mejor trabajo en la predicción. Choong & Lee (2017) describen que este método produce un vector con una longitud igual al número de categorías en el conjunto de datos. Si un punto de datos pertenece a la categoría *i-ésima*, a los componentes de este vector se les asigna el valor 0, excepto el componente *i-ésimo*, al que se le asigna un valor de uno, realizando un seguimiento de las categorías de una manera numéricamente significativa.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

1.1.12 Métricas de evaluación del rendimiento de un modelo

Kelleher et al., (2015) sostienen que una vez que se haya construido un modelo, la pregunta más importante que surge es qué tan bueno es el modelo. Por lo tanto, evaluar su modelo es la tarea más importante (Albon, 2018; Hossin & M.N, 2015; Mueller & Guido, 2016). Por lo que hace énfasis en la matriz de confusión (Tabla 1), que es una tabla que se usa para describir el rendimiento de un modelo de clasificación en un conjunto de datos de prueba para los que se conocen los valores verdaderos (Pina, 2018).

Tabla 1

Matriz de confusión

		Clase predecida	
		Clase = Si	Clase = No
Clase actual	Clase = Si	Verdadero positivo (TP)	Falso negativo (FN)
	Clase = No	Falso positivo (FP)	Verdadero negativo (TN)

Fuente: Pina (2018)

Para comprender las métricas: *Accuracy, Precision, Recall y F1 Score*, es importante comprender los parámetros *TP, FN, FP* y *TN*. Verdadero positivo y verdadero negativo son las observaciones que se predicen correctamente y, por lo tanto, se muestran en verde. Se quiere minimizar los falsos positivos y falsos negativos para que se muestren en color rojo. Estos términos son un poco confusos, y son explicados a continuación:

Positivos verdaderos (*TP*): estos son los valores positivos pronosticados correctamente, lo que significa que el valor de la clase real es sí y el valor de la clase pronosticada también es sí. Por ejemplo, si el valor de clase real indica que este pasajero sobrevivió y la clase pronosticada le dice lo mismo.

Verdaderos negativos (*TN*): estos son los valores negativos predichos correctamente, lo que significa que el valor de la clase real es no y el valor de la clase pronosticada también es no. Por ejemplo, si la clase real dice que este pasajero no sobrevivió y la

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

clase prevista le dice lo mismo. Falsos positivos y falsos negativos, estos valores ocurren cuando su clase real contradice con la clase predicha.

Positivos falsos (*FP*): cuando la clase real es no y la clase predicha es sí. Por ejemplo, si la clase real dice que este pasajero no sobrevivió, pero la clase predicha le dice que este pasajero sobrevivirá.

Falsos negativos (*FN*): cuando la clase real es sí, pero la clase predicha en no. Por ejemplo, si el valor real de la clase indica que este pasajero sobrevivió y la clase pronosticada le dice que el pasajero no sobrevivirá.

Accuracy es la medida de rendimiento más intuitiva y es simplemente una relación entre la observación predicha correctamente y el total de observaciones. Si se tiene una alta precisión, entonces nuestro modelo es el mejor, siempre y cuando la precisión es una gran medida, pero solo cuando tiene conjuntos de datos simétricos donde los valores de falsos positivos y falsos negativos son casi iguales.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision es la relación entre las observaciones positivas predichas correctamente y el total de observaciones positivas predichas.

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

Recall es la proporción de observaciones positivas pronosticadas correctamente en el conjunto de observaciones positivas pronosticadas (correcta o incorrectamente).

$$Recall = \frac{TP}{TP + TN}$$

F1 Score es la media armónica de *Precision* y *Recall*. Por lo tanto, este puntaje tiene en cuenta tanto los falsos positivos como los falsos negativos. Intuitivamente, no es tan fácil de entender como la *Accuracy*, pero *F1 Score* suele ser más útil que la *Accuracy*, especialmente si tiene una distribución de clase desigual. *Accuracy* funciona mejor si los falsos positivos y los falsos negativos tienen un costo similar. Si el costo de los falsos positivos y los falsos negativos son muy diferentes, es mejor tener en cuenta tanto la *Precision* como *Recall*.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

$$F1\ Score = \frac{2*(Recall*Precision)}{Recall+Precision}$$

AUC ROC

Narkhede (2018) sostiene que *AUC ROC* (*Area Under The Curve ROC*) es una de las métricas de evaluación más importantes para verificar el rendimiento de cualquier modelo de clasificación, en especial clasificación binaria, también conocida como área bajo la curva *ROC*. Por su parte, Véliz (2016) menciona que la curva *ROC* es un gráfico que muestra el rendimiento de un modelo de clasificación, basado en dos variables: la sensibilidad (*sensitivity*) y la especificidad (*specificity*).

Sensitivity: es una medida de la capacidad de acierto de un evento y se define como el número de categorías positivas bien predichas dividido por el total de categorías positivas.

Specificity: es una medida de la capacidad de acierto del evento complementario al anterior. Se define como el número de categorías falsas bien predichas dividido por el total de categorías falsas.

En la Figura 9 se muestra la curva *ROC* correspondiente a diferentes puntos de corte de un modelo, si el modelo es perfecto, hay una región en la que cualquier punto de corte tiene sensibilidad y especificidad iguales a 1 y la curva tiene solo el punto (0,1). Por el contrario, si el modelo no ayuda en la clasificación, la sensibilidad es igual a la tasa de falsos positivos y la curva es la diagonal que va desde el punto (0,0) a (1,1).

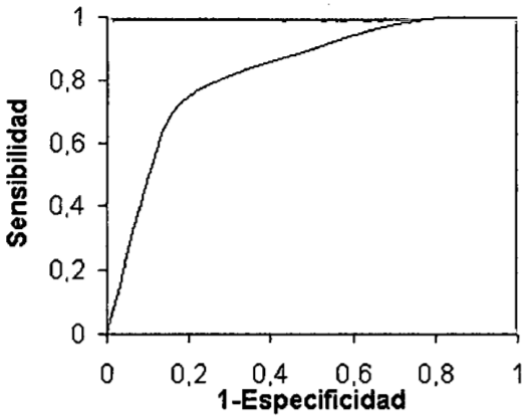


Figura 9. Curva *ROC*

Fuente: Véliz (2016)

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Por tanto, el modelo tiene mejor desempeño si la curva *ROC* correspondiente se aleja más de la diagonal principal. La Tabla 2 presenta la interpretación del área bajo la curva ROC (Véliz, 2016).

Tabla 2

Interpretación AUC

Área bajo la curva ROC	**Poder Discriminante**
Área ROC = 0.5	Nulo (como lanzar una moneda)
0.7<= Área ROC < 0.8	Aceptable
0.8<= Área ROC < 0.9	Excelente
Área ROC >=0.9	Excepcionalmente buena

Fuente: Véliz (2016).

1.1.13 *Scikit Learn*

Scikit Learn es un módulo de Python que integra una amplia gama de algoritmos de *Machine Learning* de última generación para problemas supervisados y no supervisados de mediana escala (Nelli, 2018). Pedregosa et al. (2011) enfatiza la facilidad de uso, el rendimiento, la documentación y la coherencia de la API. *Scikit Learn* tiene dependencias mínimas y se distribuye bajo la licencia *Berkeley Software Distribution* simplificada, lo que fomenta su uso en entornos académicos y comerciales (Developers, 2016).

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

1.2 Antecedentes

A continuación, se presentan la revisión de la literatura, entre ellos diversas tesis y artículos científicos publicados en el ámbito global, nacional y local sobre *Machine Learning*, riesgo crediticio, otorgamiento de créditos y entidades financieras.

Tesén (2017) estudió la eficacia de los modelos de *Machine Learning* evaluando el riesgo crediticio de las personas naturales en una institución financiera de Chiclayo, comprobó que los modelos de *Machine Learning* evalúan eficazmente el riesgo crediticio de personas naturales comparados al modelo clásico de *CreditScoring*, basado en una regresión lineal. Además, realizó pruebas utilizando diferentes modelos, como resultado obtuvo que el mejor modelo, que utiliza *Machine Learning*, es el de Redes Neuronales con un 81.10% en la etapa de validación.

Kalayci et al. (2018) realizaron un estudio de análisis de riesgo crediticio en PYMES usando modelos *Machine Learning*, evaluando la información histórica del cliente y el comportamiento de pago en un periodo de seis meses, para luego predecir la puntualidad de pago en los siguientes seis meses, llegando a la conclusión que los modelos de *Machine Learning* tienen gran importancia en la predicción de puntualidad de pago de los créditos.

Addo et al. (2018) aplicaron modelos de *Machine Learning* y *Deep Learning* en el análisis de riesgo crediticio, para lo cual construye clasificadores binarios basado en datos reales para predecir la probabilidad del pago de un crédito, en la que hallaron diez características principales para realizar el modelamiento, llegando a la conclusión que los modelos basados en árboles son más estables que los modelos basados en *ANN* de múltiples capas.

Turkson et al. (2016) aplicaron un enfoque de *Machine Learning* para realizar la predicción de solvencia de créditos bancarios, en la cual utilizaron datos de créditos reales de un determinado banco para el entrenamiento y elección del mejor modelo de *Machine Learning*, llegando a la conclusión que los modelos *Nearest Centroid* y *Gaussian Naive Bayes* son los mejores obteniendo un 80% de precisión en determinar si un cliente incumplirá o no el pago de su crédito.

Kruppa et al. (2013) estudiaron el riesgo de créditos de producto consumo, buscando estimar la probabilidad individual de pago utilizando modelos de *Machine Learning* tales como *Ramdom Forest*, *k-nearest neighbors* y *bagget k-nearest neighbors* llegando a la conclusión que *Random Forest* supero a los demás modelos.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Altman & Saunders (1998) realizaron la medición del riesgo crediticio con datos de los 20 últimos años centrándose en la evolución de la literatura sobre la medición del riesgo de crédito de los préstamos individuales y las carteras de préstamos de los últimos 20 años y a su vez llegaron a plantear un nuevo enfoque basado en un marco de riesgo de mortalidad para medir el riesgo y los rendimientos de préstamos y bonos.

Gyorfi et al. (2012) investigaron modelos basados en *Machine Learning* diseñando estrategias de inversión para los mercados financieros, generando información sobre esta aplicación orientada a la matemática, ciencia de la computación, finanzas y otros; donde los autores demostraron que su modelo ofrece cierta promesa en el análisis de las estructuras de riesgo-rendimiento de las carteras de instrumentos de deuda expuestos a riesgo de crédito.

Saju & Chacko (2017) investigaron si las actividades de microfinanzas dirigidas a los consumidores de la pirámide son sostenibles; el estudio sigue una metodología mixta; las opiniones de los gerentes sobre la sostenibilidad de los programas se evaluaron al analizar sus respuestas en las áreas de desvío de fondos, costos operativos, tasas de interés y tasa de rendimiento de los préstamos a través de entrevistas semiestructuradas; los hallazgos de esta investigación se complementan con el trabajo existente para presentar una comprensión integral de este tema al investigar el aspecto de sostenibilidad de estos programas desde la dimensión de clientes y prestamistas.

Jarrow (2009) investigó a diferentes modelos de riesgo de crédito, incluyendo temas de modelos estructurales y de forma reducida, información incompleta, derivados de crédito y contagio predeterminado. Se argumenta que los modelos de forma reducida y no los modelos estructurales son apropiados para la fijación de precios y la cobertura de valores de riesgo crediticio. Se discuten las direcciones para futuras investigaciones, llegando a aportar literatura sobre los diferentes modelos existentes.

Sadatrasoul et al. (2013) centraron su investigación en la revisión de literatura académica y sistemática e incluye todas las revistas en la base de datos de revistas en línea de *Science Direct*. Los artículos se clasifican de acuerdo al puntaje crediticio de empresas, individuales y pequeñas y medianas (PyME), en la cual se aplica técnicas de minería de datos. Los métodos de selección de variables también se investigan por separado dado que las mismas son importante en el problema de calificación crediticia; los hallazgos de la revisión revelan que las técnicas de extracción de datos se aplican principalmente al

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

puntaje crediticio individual y existen algunas investigaciones sobre la calificación crediticia de empresas y PyME.

Valencia (2017) elaboró un modelo de *Scoring* para el otorgamiento de crédito en las pequeñas y medianas empresas en la cual determina las variables que influyen en el otorgamiento de créditos para luego hacer uso de herramientas estadísticas y econométricas para la construcción de su modelo.

Kourou et al. (2015) investigaron la aplicación de *Machine Learning* en el pronóstico y predicción del cáncer, la cual se han convertido en una necesidad, ya que puede facilitar el manejo clínico posterior de los pacientes, donde en la investigación realizó la investigación de la aplicación de *Machine Learning* en el diagnóstico del cáncer, aplicando una diversidad de técnicas redes neuronales artificiales, las redes bayesianas, las máquinas de vectores de soporte y los árboles de decisión, teniendo como resultado una herramienta prometedora para la inferencia en el dominio del cáncer.

Tack (2018) realizó la aplicación de *Machine Learning* en la fisioterapia musculoesquelética, en las cuales se centra tanto en el aprendizaje supervisado y no supervisado en la medicina musculoesquelética a través de imágenes de diagnóstico, datos de medición de pacientes y apoyo a decisiones clínicas, llegando a la siguiente conclusión: la alfabetización de datos debe ser un componente de los planes de desarrollo profesional para ayudar a los fisioterapeutas en la aplicación de *Machine Learning* y la preparación de sistemas de tecnología de la información para utilizar estas técnicas.

Khashman (2010) desarrolló un sistema de evaluación de riesgo crediticio que utiliza modelos de redes neuronales supervisadas basadas en el algoritmo de aprendizaje de propagación hacia atrás, donde se investigó nueve esquemas de aprendizaje con diferentes proporciones de datos de capacitación a validación, y se ha proporcionado una comparación entre los resultados de su implementación.

Huang et al. (2015) exploraron la tendencia del *Machine Learning* extremo y sus aplicaciones, en la cual investiga estado actual de la investigación teórica y los avances prácticos sobre este tema, lo cual es aplicados a una variedad de dominios, como ingeniería biomédica, visión computacional, identificación de sistemas y control y robótica, generando mayor literatura sobre *Extreme Machine Learning*.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Doulah & Alam (2018) compararon ocho algoritmos de clasificación de *Machine Learning*, como los árboles de decisión, el bosque aleatorio, la red neuronal artificial, la máquina de vectores de apoyo, el análisis discriminante lineal, los vecinos más cercanos a k, la regresión logística y los ingenuos Bayes en datos ecológicos. Teniendo como objetivo comparar diferentes algoritmos de clasificación de aprendizaje automático en conjuntos de datos ecológicos. En este análisis hemos comprobado la prueba de precisión entre los algoritmos, donde llegaron a la a la conclusión de que el Análisis Discriminante Lineal y los vecinos más cercanos a k son los mejores métodos entre todos los demás algoritmos.

Buczak & Guven (2016) describieron una encuesta enfocada en los métodos de *Machine Learning* (*ML*) y de *Data Minng* (*DM*) para el análisis cibernético para apoyar la detección de intrusos. Debido a que los datos son tan importantes en los enfoques de *ML / DM*, describió conjuntos de datos cibernéticos bien conocidos utilizados en *ML / DM*. Abordando la complejidad de los algoritmos *ML / DM*.

Rosten & Drummond (2006) demostraron que el *Machine Learning* se puede usar para derivar un detector de características que puede procesar completamente el video en vivo usando menos del 7% del tiempo de procesamiento disponible, concluyen que, el detector investigado supera significativamente a los detectores de características existentes de acuerdo con este criterio.

Flores & Ramon (2014) exploran las estrategias cooperativas enfocados en la probabilidad de incumplimiento, para lo cual utilizan clasificadores estadísticos y *Machine Learning*, donde evalúan el rendimiento de diferentes modelos, obteniendo resultados que indican que las estrategias ajustadas por correlación son técnicas prometedoras para la administración de cartera de créditos de bajo incumplimiento, y las mismas proporcionan estimaciones de riesgo crediticito precisas y bien calibradas.

Arango & Restrepo (2017) diseñan un modelo de scoring enfocado en el otorgamiento de créditos de producto consumo en una compañía de Colombia, donde los autores utilizan los modelos análisis discriminante, modelo probabilístico, modelo logístico y las redes neuronales artificiales, así mismo los datos de la misma compañía son utilizados para realizar el entrenamiento de los modelos, obteniendo que regresión logística permite predecir el comportamiento en función de las variables analizadas.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Identificación del problema

La inteligencia artificial (IA) está cambiando drásticamente el entorno empresarial y social, así como en el ámbito de las finanzas y microfinanzas, en el cual la IA puede realizar tareas de forma eficiente y eficaz en comparación a los seres humanos, permitiendo tomar decisiones más asertivas a éstos usando basado en datos históricos (Rouhiainen, 2019), pero este es un gran desafío que conlleva retos complicados que involucra a personal especializado para la selección y recolección de estos datos (Zhao, 2018); en el ámbito empresarial se tiene a *Machine Learning* como uno de los campos más utilizados de la IA para el análisis financiero en el sector urbano o rural (Wang & Tao, 2008).

En la actualidad, en América Latina, las actividades económicas del sector rural impulsan la economía de un país, sin embargo, este sector de la población no cuenta con un apoyo adecuado que promueva su desarrollo y crecimiento, que podría darse con el acceso a fuentes de financiamiento formales que permitan impulsar sus actividades económicas a nivel microfinanciero. En el Perú las entidades encargadas de otorgar estos tipos de servicios de financiamiento al sector rural lo realizan a través de los microcréditos (Delfiner et al., 2006), las cuales enfrentan el problema de la existencia de un nivel de riesgo, que se traduce en la morosidad por parte de los clientes hasta un punto de incobrabilidad, generando una gran incertidumbre al momento de otorgar un microcrédito (Tesén, 2017).

La entidad microfinanciera analizada no es ajena a esta problemática dado que actualmente presenta dificultades en cuanto al otorgamiento de microcréditos, pese a

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

tener personal especializado denominado asesores de negocios, se observa que la entidad tiene clientes que presentan retrasos en el pago de sus cuotas, en algunos casos se enfrenta al incumplimiento total del pago, situación en el cual la entidad opta por una serie de medidas que van desde las vías administrativas hasta las vías judiciales; esta problemática se expresa a través del índice de morosidad reportado ante su ente supervisor.

En estos últimos años la entidad microfinanciera presentó un índice de morosidad muy variable, en el año 2015 alcanzó un 6.5%, en el año 2016 alcanzó 6.1% y en el año 2017 alcanzó un 6.3%; esta variabilidad demuestra una deficiencia en el nivel de asertividad al momento del otorgamiento de microcréditos y se puede observar la necesidad de proponer un modelo predictivo basado *Machine Learning* para mejorar el nivel de asertividad en el otorgamiento de microcréditos en la entidad.

2.2 Enunciados del problema

Considerando la situación descrita, en la presente investigación se realizó la siguiente formulación del problema: ¿Cuál de los modelos de *Machine Learning* permite mejorar el nivel de asertividad en el otorgamiento de microcréditos?

2.3 Justificación

La toma de decisión de otorgar un crédito se torna en una tarea compleja de realizar por lo que muchas empresas optan por construir modelos que permitan identificar si un determinado microcrédito es otorgado o no (Tesén, 2017), pero cada entidad tiene particularidades tales como sus clientes objetivos, metodología de trabajo y productos de créditos, motivo por el cual cada una de estas requiere realizar una investigación independiente para plantear un modelo adecuado que le permita mejorar el nivel de asertividad en el otorgamiento de microcréditos.

Desde el punto de vista teórico, se plantea un modelo que puede servir como referencia para el otorgamiento de microcréditos haciendo uso de la inteligencia artificial a través de modelos de *Machine Learning*, que permitan mejorar el nivel de asertividad en el otorgamiento de un microcrédito en el sector rural, lugar donde la información histórica de la persona es nula o faltante y requieren de nuevas herramientas para mejorar el nivel de asertividad basado en casos históricos de otros clientes para predecir el pago de crédito de estas personas (Wendel & Harvey, 2006).

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Desde el punto de vista metodológico en la presente investigación se propone un proceso de para determinar el modelo más asertivo en el otorgamiento de microcréditos permitiendo ser referenciado en investigaciones futuras relacionadas a la predicción de otorgamiento de microcréditos con *Machine Learning*.

Desde un punto de vista práctico, la investigación busca la aplicabilidad del modelo en la entidad microfinanciera con el fin de reducir sus ratios de mora que están relacionados directamente al nivel de asertividad del otorgamiento de créditos y permitir de esta forma a la población del sector rural acceder a una fuente de financiación formal.

2.4 Objetivos

2.4.1 Objetivo General

Determinar el mejor modelo de *Machine Learning* que permite mejorar el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

2.4.2 Objetivos Específicos

- Determinar las variables para los modelos de *Machine Learning*.
- Determinar el nivel de asertividad de los modelos de *Machine Learning* a través de las variables determinadas.

2.5 Hipótesis

2.5.1 Hipótesis general

Artificial Neural Network es el mejor modelo que nos permite mejorar el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

2.5.2 Hipótesis especificas

- El modelo Regresión Logística mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.
- El modelo *Random Forest* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.
- El modelo *Support Vector Machine* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

- El modelo *Artificial Neural Networks* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.
- El modelo *Decision Tree* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.
- El modelo *k-Nearest Neighbors* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio

La presente investigación se realizó en la ciudad de Puno, ubicada 3,820 metros sobre el nivel del mar, tiene una población de 1’172,697 habitantes, con el 46.22% en el sector rural (INEI, 2019). La actividad económica principal el sector agropecuario, siendo este el foco de la entidad microfinanciera con la que se realizó la presente investigación.

3.2 Población

Se ha considerado la entidad microfinanciera con mayor participación en el mercado, cuya la población estuvo conformada por 15,015 con al menos un microcrédito del producto agropecuario en el año 2017.

3.3 Muestra

La selección de la muestra ha sido de tipo no probabilístico, utilizando el muestreo por conveniencia (Sampieri et al., 2014), este tipo de muestreo se caracteriza por obtener muestras accesibles representativas, por lo que, se consideró como muestra a toda la población, conformada por los 15,015 clientes de la entidad, esto para obtener un mejor entrenamiento de los modelos de *Machine Learning*.

3.4 Método de investigación

La investigación es no experimental de tipo transversal, su propósito es describir variables y analizar su incidencia e interrelación en un momento dado (Sampieri et al., 2014), en este caso en el periodo de un año, de marzo de 2017 a marzo de 2018.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Se ha empleado un diseño de investigación descriptiva comparativa, se han establecido variables y propiedades de un determinado fenómeno para ser analizadas y comparadas. En este estudio el objetivo es conocer el mejor modelo de *Machine Learning* que permita mejorar el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero, para lo cual se ha comparado el nivel de asertividad en el otorgamiento de créditos de los modelos más usados de *Machine Learning*.

3.5 Descripción detallada de métodos por objetivos específicos

El proceso para determinar el modelo más asertivo en el otorgamiento de microcréditos propuesto en la presente investigación es presentado en la Figura 10, este proceso se divide en tres hitos, el hito de especificación, en el que se realiza la revisión del proceso de otorgamiento de microcréditos y la revisión de investigaciones que ayuden a determinar las variables que influyen en el otorgamiento de microcréditos; el hito de implementación, focalizado en el entrenamiento de los modelos seleccionados de *Machine Learning* a partir del *dataset* pre procesado; finalmente, en el hito de evaluación, se realiza la evaluación de los modelos a través de métricas para seleccionar el modelo más asertivo.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

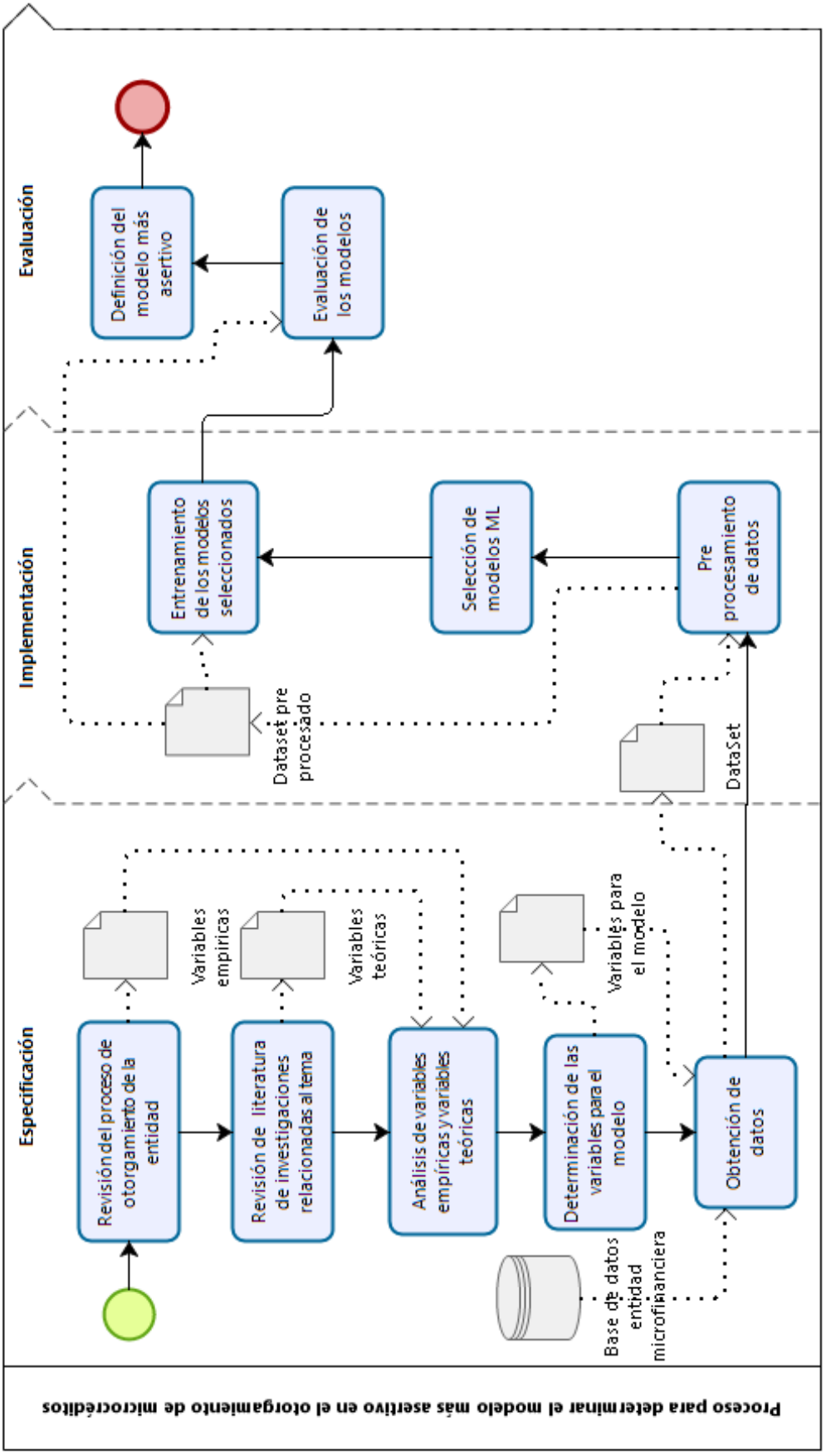


Figura 10. Proceso para determinar el modelo más asertivo en el otorgamiento de microcréditos

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

3.5.1 Especificación

Tiene como objetivo la determinación de las variables para el modelo expresadas en un *dataset*, para el cual se realiza la revisión del proceso de otorgamiento de microcrédito de la entidad, a través del mapeo de procesos con el software *Bizagi* (*Bizagi*, 2019), que explora los procesos de captación, evaluación y aprobación, y permite identificar las variables empíricas de este proceso; además se realizó la revisión de la literatura para obtener las variables teóricas, para lo cual se exploraron investigaciones relacionadas al otorgamiento y análisis de riesgo crediticio; finalmente se realizó el análisis de las variables empíricas y teóricas a través del cruce de variables.

3.5.2 Implementación

Tiene como objetivo obtener el entrenamiento de los modelos de *Machine Learning*, a partir del *dataset* obtenido en el hito anterior; para luego realizar el pre procesamiento de los datos como se observa en la Figura 10; en la que realiza la limpieza y el análisis de la distribución de los mismos, se aplica la técnica de codificación *One Hot* para obtener como resultado el *dataset* pre procesado para el entrenamiento y evaluación de los modelos.

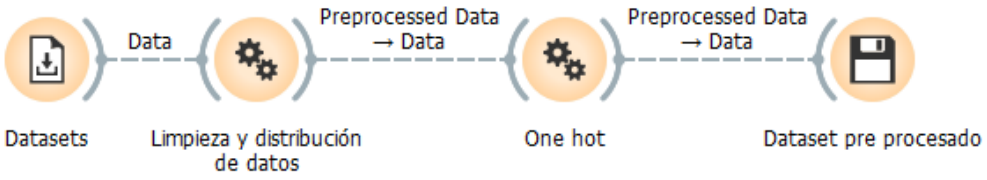


Figura 11. Pre procesamiento de datos

Para el entrenamiento de los modelos se selecciona un total de seis modelos de *Machine Learning* de aprendizaje supervisado, los cuales fueron: Regresión Logística (RL), *Random Forest (RF)*, *Support Vector Machine (SVM)*, *Artificial Neural Network (ANN)*, *Decision Tree (dTree)*, *k-Nearest Neighbors (kNN)*, *Linear Discriminant Analysis (LDA) y Multinomial Regresion (MR)*, que son frecuentemente utilizados en investigaciones relacionadas, como se observa en la Tabla 3.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 3

Modelos de Machine Learning en investigaciones relacionadas

Referencia	RL	*RF*	*SVM*	*ANN*	*dTree*	*NBC*	*kNN*	*LDA*	*MR*
Chiranjit & Andreas (2017)		x	x	x		x	x		
Flores & Ramon (2014)	x		x	x			x	x	
Addo et al. (2018)	x	x							x
Turkson et al. (2016)	x		x	x					
Kruppa et al. (2013)	x						x		
Kalayci et al. (2018)		x	x	x					
Arango & Restrepo (2017)	x	x		x					
Arango (2017)				x					
Millán & Caicedo (2018)	x			x	x				

Finalmente, el entrenamiento de los modelos de *Machine Learning* (Figura 12), se realiza usando el *dataset* pre procesado para el entrenamiento de cada uno de los modelos de *Machine Learning*, finalmente se realiza la prueba de cada modelo y se determina sus puntajes de acuerdo a sus métricas definidas en la investigación.

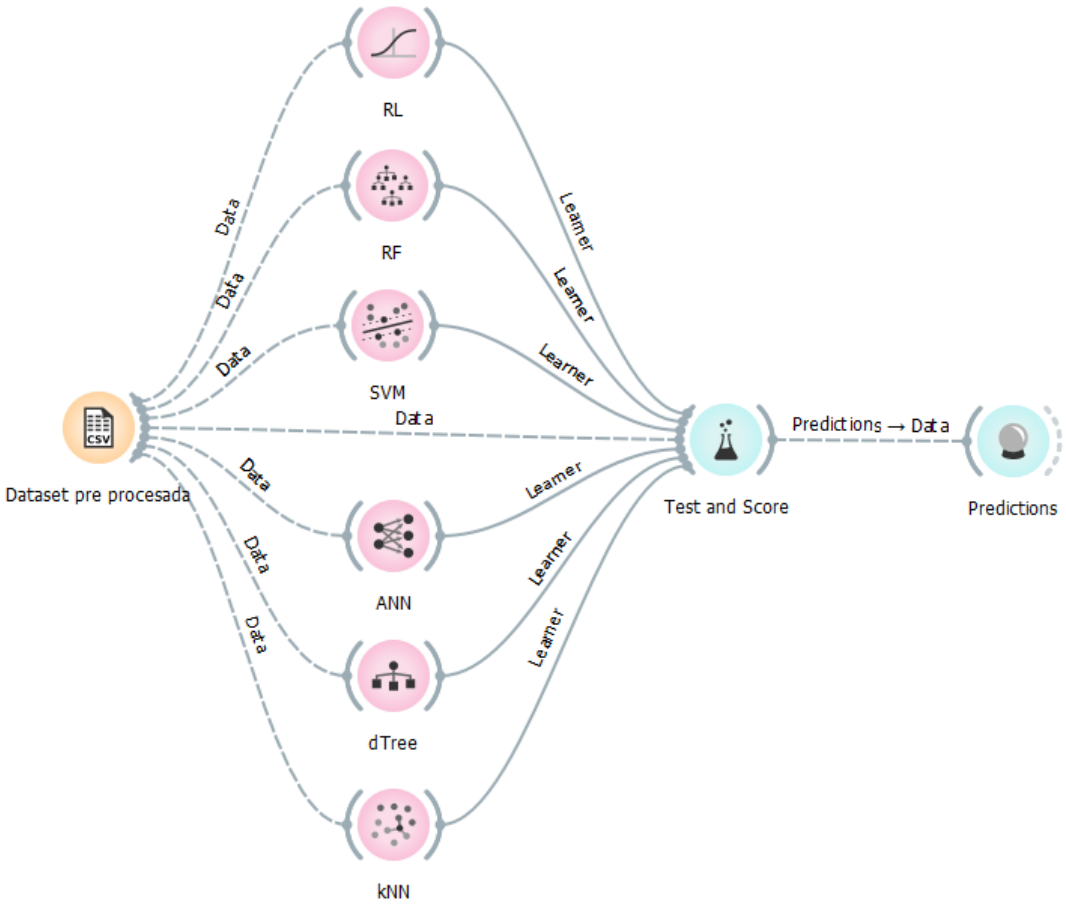


Figura 12. Entrenamiento de modelos de *Machine Learning* de la investigación

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

3.5.3 Evaluación

Tiene como objetivo la determinación del modelo más asertivo a través de la evaluación de los modelos de *Machine Leaning* mediante las métricas *Accuracy (Acc)*, *Precision (Pre)*, *Recall (Rec)*, *F1 Score (F1)* y *AUC-ROC (AUC)* que son los frecuentemente utilizados para la evaluación de modelos de *Machine Learning* como se observa en la Tabla 4.

Tabla 4

Métricas usadas para la evaluación de modelos de Machine Learning en investigaciones relacionadas

Referencia	*Acc*	*Pre*	*Rec*	*F1*	*AUC*
Chiranjit & Andreas (2017)	x	x	x	x	
Flores & Ramon (2014)	x				x
Addo et al. (2018)					x
Turkson et al. (2016)	x	x	x	x	
Kruppa et al. (2013)					x
Millán & Caicedo (2018)					x

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

En este capítulo se presentan los resultados obtenidos del desarrollo de los objetivos de la investigación y su respectiva, en la sección 4.1. se presenta la determinación de las variables consideradas para el modelo y en la sección 4.2 se presenta el resultado del entrenamiento y evaluación de los modelos para la selección del modelo más asertivo a través de métricas.

4.1 Resultado conforme al primer objetivo específico

4.1.1 Revisión del proceso de otorgamiento de la entidad

El proceso de otorgamiento de la entidad microfinanciera se divide en 4 procesos como se muestra en la Figura 13, captación, evaluación, aprobación y desembolso; la cual se asemeja al modelo expuesto por Morales & Morales (2014).

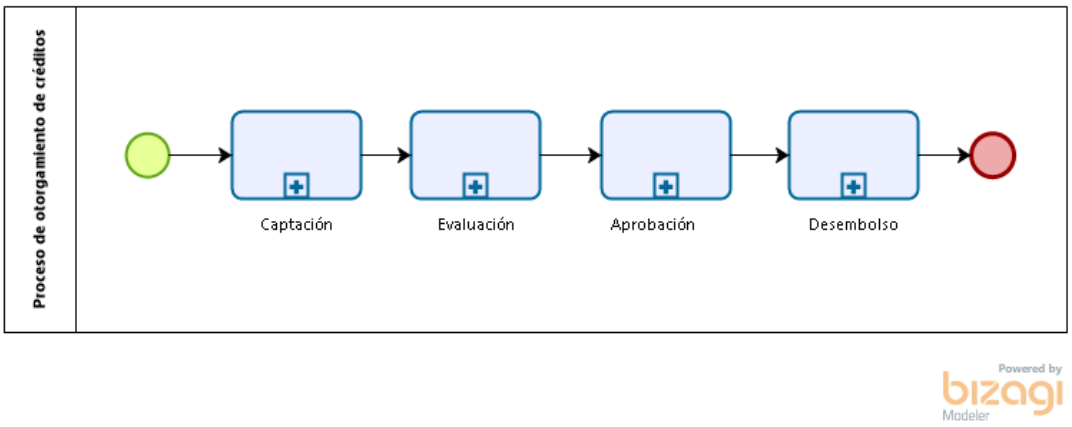


Figura 13. Proceso de otorgamiento de crédito de la entidad

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

El proceso captación se divide en dos tipos (Figura 14): la captación activa, donde el asesor de negocios visita a clientes en sus negocios o donde realizan sus actividades económicas generalmente ubicadas en las zonas rurales del Perú (Anexo 1); la captación pasiva, que se da cuanto el cliente se apersona a las oficinas buscando un crédito para un determinado fin, el cual es atendido por plataforma, donde es informado sobre los productos de créditos que ofrece la entidad microfinanciera (Anexo 2).

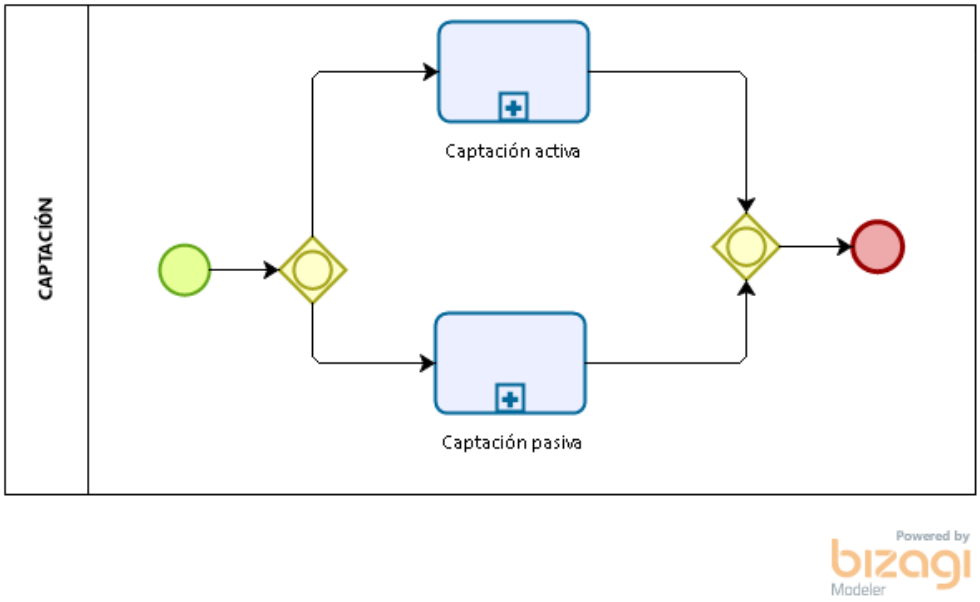


Figura 14. Proceso de captación

En ambos casos se le ofrece al cliente los productos relacionados a su actividad económica con distintos fines, si el cliente se interesa por el producto de crédito que se le está ofreciendo, el asesor de negocios recolecta más información del cliente y realiza una evaluación rápida basado en su conocimiento empírico y la información histórica que tenga a disposición si el cliente pasa este filtro, se registra o actualiza su información para luego solicitar el crédito en el sistema financiero de la entidad. Este registro o actualización contempla los datos del cliente; como número de documento, fecha de nacimiento, actividad económica, grado de instrucción entre otros; datos de la solicitud del crédito como monto, tasa, plazo, destino del crédito entre otros; y datos de la garantía que el cliente otorgará para garantizar el microcrédito en caso que sea necesario, teniendo como resultado una solicitud de crédito.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

En el proceso de evaluación (Anexo 3) el asesor de negocios se encarga de elaborar una hoja de ruta en el sistema para reunir y verificar la información cuantitativa y cualitativa del cliente *in situ*, plasmado en un expediente que es evaluado por un asesor de negocios que posteriormente lo eleva ante un comité de créditos; conformado por cinco asesores de negocios como mínimo, los cuales deben ser de diferentes niveles, estos niveles son determinados por el tiempo de experiencia y la magnitud de la cartera de créditos que administre; en donde el asesor de negocios expone la solicitud del microcrédito, y el comité tiene la potestad de observar, aprobar o rechazar la solicitud de crédito de acuerdo a su conocimiento empírico.

En el proceso de aprobación (Anexo 4) la solicitud de microcrédito pasa por diferentes niveles de aprobación de acuerdo a las condiciones del mismo para determinar si debe ser otorgado o no. En el proceso de desembolso se realiza el registro de firmas en los contratos que indican las condiciones del microcrédito para concretarse el desembolso del monto acordado.

Finalmente, como resultado de la revisión del proceso de otorgamiento microcrédito de la entidad, se logró identificar un total de 34 variables empíricas que se observan en la Tabla 5.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Tabla 5

Variables empíricas

Nro.	Variables	Descripción
1	Capital de desembolso	Monto que el cliente solicita en su crédito
2	Monto de la cuota	Monto de la cuota de acuerdo periodo
3	Destino del crédito	Destino del crédito donde indica en que se utilizará el capital desembolsado
4	Número de cuotas	Número de cuotas del crédito
5	Oficina	Oficina de la entidad donde solicitan el crédito.
6	Interés Pactado	Interés pactado con el cliente.
7	Tipo de solicitud de crédito	Indica si la solicitud es un otorgamiento nuevo, ampliación, reprogramación o refinanciamiento.
8	Clasificación del cliente	Clasificación interna del cliente de acuerdo a la entidad
9	Tipo de cliente	Tipo de cliente en la cual puede ser nuevo o recurrente
10	Tipo de crédito	Indica si el crédito es nuevo o es una ampliación o refinanciamiento
11	Edad	Edad del cliente.
12	Ubigeo	Indica el Ubigeo de nacimiento del cliente
13	Tipo Vivienda	Indica el tipo de vivienda del cliente que puede ser rustica, material noble u otros.
14	Estado Civil	Estado civil del cliente
15	Sexo	Indica el género del cliente
16	Nivel Instrucción	Nivel de instrucción del cliente que pueden ser sin estudios, primaria, secundaria o superior.
17	Demanda	Indica si el cliente tenia demanda por alimentos
18	Ocupación	Ocupación del cliente
19	Profesión	Profesión del cliente en caso de que el cliente tenga nivel de instrucción superior
20	Cuentas de ahorro	Indica si el cliente cuenta con cuentas de ahorro vigentes
21	Sector Económico	Sector económico al que pertenece de acuerdo a su actividad primaria que realiza el cliente.
22	Actividad Primaria	Actividad primaria del cliente
23	Tipo Actividad Primaria	Tipo de Actividad primaria
24	Años en actividad primaria	Número de años en la actividad primaria
25	Actividad secundaria	Actividad primaria del cliente
26	Tipo de actividad secundaria	Tipo de Actividad secundaria
27	Años en actividad secundaria	Número de años en la actividad secundaria
28	Saldo	Monto total de deudas al momento del otorgamiento
29	Plazo	Indica el plazo del crédito en días
30	Garantía	Indica si el cliente tiene garantía o no
31	Número de dependientes	Indica el número de dependientes que el cliente tiene
32	Atraso promedio	Indica el número de días que el cliente se atrasó en pago de su cuota.
33	Número de condonaciones	Indica el número de condonaciones que cliente solicito en su créditos
34	Número de créditos	Indica el número de créditos vigentes que tiene el cliente

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

4.1.2 Revisión de literatura sobre otorgamiento de crédito

En la literatura, los autores trabajan con un promedio de 13 variables para el otorgamiento de crédito y *Machine Learning*, estas sirven para realizar el entrenamiento de sus modelos. La tabla 6 resume las variables utilizadas por autor, considerando los trabajos de Ochoa et al. (2010), Arango & Restrepo (2017), Arango (2017), Millán & Caicedo (2018), Turkson et al. (2016), Kruppa et al. (2013) y Valencia (2017), observando que dichos autores trabajan con un promedio de 13 variables para el entrenamiento de sus modelos.

Además, se logró identificar un total de 34 variables teóricas involucradas en el proceso en las cuales resaltan Monto, Edad, Sexo, Estado civil, Plazo, Garantía, Nivel Educativo, Ingreso total, Antigüedad laboral, Dependientes, Tipo de vivienda, Reestructurado, Ocupación, Egreso total, Estrato socioeconómico y Tasa de interés.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 6

Variables teóricas

	Ochoa et al. (2010)	Arango & Restrepo (2017)	Arango (2017)	Millán & Caicedo (2018)	Turkson et al. (2016)	Kruppa et al. (2013)	Valencia (2017)
Monto	x	x	x	x		x	x
Edad	x	x	x	x	x	x	
Sexo	x	x	x	x	x	x	
Estado civil	x	x	x	x	x		
Plazo	x	x	x	x			x
Garantía	x	x	x	x			
Nivel Educativo	x	x	x		x		
Ingreso total	x	x	x	x			
Antigüedad laboral	x	x	x			x	
Dependientes	x	x	x	x			
Tipo de vivienda	x	x	x	x			
Reestructurado	x	x	x				
Ocupación	x	x		x			
Egreso total		x	x	x			
Estrato socioeconómico	x	x	x				
Tasa de interés		x	x	x			
Oficina	x	x					
Categoría	x	x					
Tipo de contrato	x	x					
Antigüedad en la institución	x	x					
Forma de pago	x	x					
Activo			x				x
Cubrimiento de interés			x				x
Capacidad de pago	x						
Fecha de otorgamiento						x	
Ubigeo						x	
Ventas netas							x
Días de mora		x					
Ahorro		x					
Saldo		x					
Número de créditos		x					
Cuota			x				
Histórico central riesgo			x				
Vivienda			x				

La descripción de cada una de estas variables teóricas identificadas se observa en la Tabla 7.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Tabla 7

Descripción de las variables teóricas

Variable	Descripción
Monto	Indica el monto de desembolso
Edad	Edad del cliente al momento del otorgamiento
Sexo	Genero del cliente
Estado civil	Estado civil del cliente según SBS
Plazo	Plazo que tiene el crédito
Garantía	Garantía del cliente
Nivel Educativo	Nivel educativo del cliente
Ingreso total	La suma de sus ingresos mensuales
Antigüedad laboral	Número de años de experiencia laboral
Dependientes	Número de dependientes tales como hijos, padres de la tercera edad entre otros
Tipo de vivienda	Tipo de vivienda del cliente tal como material noble, adobe entre otros
Reestructurado	Indica si el crédito fue refinanciado, ampliado o condonado
Ocupación	Ocupación del cliente al momento de realizar el otorgamiento de crédito
Egreso total	Gastos totales del cliente, así como los deberes del mismo
Estrato socioeconómico	Nivel económico de acuerdo a su clasificación
Tasa de interés	Tasa de interés del crédito otorgado
Oficina	Oficina donde se otorgó el crédito
Categoría	Categoría del cliente de acuerdo a la entidad.
Tipo de contrato	Tipo de contrato del cliente tal como a tiempo completo, tiempo parcial, nombrado entre otros
Antigüedad en la institución	Indica la antigüedad del cliente en la entidad
Forma de pago	La forma de pago indica el periodo del crédito por ejemplo mensual, trimestral, semestral, etc.
Activo	Total de activos del cliente
Cubrimiento de interés	Indica el número de veces que el cliente puede pagar el interés
Capacidad de pago	Evalúa su capacidad de pago mensual
Fecha de otorgamiento	Fecha que realiza el desembolso del crédito
Ubigeo	Ubicación geográfica expresado a través de Ubigeo (departamento, provincia y distrito)
Ventas netas	Total de ventas realizadas en los últimos tres meses
Días de mora	Cantidad promedio de días de atraso en su crédito
Ahorro	Indica la cantidad de cuentas de ahorro del cliente.
Saldo	La suma del total de sus créditos directos e indirectos
Número de créditos	Número total de créditos directos vigentes
Cuota	Monto de la cuota en un determinado periodo
Histórico central riesgo	Atraso promedio según la entidad supervisora
Vivienda	Indica el costo de la vivienda

4.1.3 Análisis de variables empíricas y variables teóricas

En este proceso de análisis se realizó el cruce de variables empíricas, identificadas en la revisión del proceso de otorgamiento de microcréditos de la entidad, y las variables teóricas, identificadas de la revisión de las investigaciones relacionadas, de este cruce

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

se obtuvo el resultado que se muestra en la Tabla 8, en el que se identifica la coincidencia de 25 variables empíricas con 22 variables teóricas, vale decir que estas 25 variables empíricas son respaldadas a través de las variables teóricas aplicadas en las finanzas.

Las variables teóricas identificadas son el Nivel educativo, Antigüedad Laboral y Cuota tienen más de una coincidencia con respecto a las variables empíricas, la variable Nivel educativo coincide con las variables Nivel de instrucción y Profesión que están directamente relacionadas, por tanto, se considera estas dos últimas para expresar mejor el Nivel educativo del cliente; la variable Antigüedad laboral coincide con las variables Años en la actividad primaria y Años en la actividad secundaria, lo que refleja que microfinanzas los clientes no solo se dedican a una determinada actividad, si no que realizan frecuentemente dos o más actividades, y la variable Cuota coincide con las variables Monto de cuota y Número de cuotas observando que en las variables empíricas expresan tanto la cantidad de cuotas, así como el monto de la misma.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

TESIS EPG UNA - PUNO
Universidad Nacional del Altiplano

Tabla 8:

Cruce de variables empíricas con variables teóricas

		Variables teóricas																																
		Monto	Edad	Sexo	Estado civil	Plazo	Garantía	Nivel Educativo	Ingreso total	Antigüedad laboral	Dependientes	Tipo de vivienda	Reestructurado	Ocupación	Egreso total	Estrato socioeconómico	Tasa de interés	Oficina	Categoría	Tipo de contrato	Antigüedad en la institución	Forma de pago	Activo	Cubrimiento de interés	Capacidad de pago	Fecha de otorgamiento	Ubigeo	Ventas netas	Días de mora	Ahorro	Saldo	Número de créditos	Cuota	Histórico central riesgo	Vivienda
Variables empíricas	Capital de desembolso	X																																
	Monto de la cuota																																X	
	Destino del crédito																																	
	Número de cuotas																																X	
	Oficina																	X																
	Interés																X																	
	Pactado																																	
	Tipo de solicitud de crédito												X																					
	Clasificación del cliente																																	X
	Tipo de cliente																		X															
	Tipo de crédito																																	
	Edad		X																															
	Ubigeo																										X							
	Tipo Vivienda											X																						
	Estado Civil				X																													
	Sexo			X																														
	Nivel Instrucción							X																										
	Demanda																																	
	Ocupación													X																				

Profesión			X						
Cuentas de ahorro							X		
Sector Económico									
Actividad Primaria									
Tipo Actividad Primaria									
Años en actividad primaria				X					
Actividad secundaria									
Tipo de actividad secundaria									
Años en actividad secundaria				X					
Saldo								X	
Plazo	X								
Garantía		X							
Número de dependientes					X				
Atraso promedio						X			
Número de condona-ciones									
Número de créditos									X

En este proceso de análisis se identificó un total de 9 variables empíricas, que se observa en la Tabla 9, las cuales no tuvieron coincidencia con las variables teóricas, sin embargo, en la presente investigación se consideró estas variables debido a que éstas son relacionadas a las microfinanzas a diferencia de las variables teóricas relacionadas a las finanzas en general.

Tabla 9

Variables empíricas sin coincidencia con las variables teóricas

Nro.	Variable empírica
1	Destino del crédito
2	Tipo de crédito
3	Demanda
4	Sector Económico
5	Actividad Primaria
6	Tipo Actividad Primaria
7	Actividad secundaria
8	Tipo de actividad secundaria
9	Número de condonaciones

De la Tabla 9, en el caso de las variables Sector Económico, Actividad Primaria, Tipo de Actividad Primaria, Actividad Secundaria y Tipo de Actividad secundaria van relacionados a las actividades económicas que realizan las personas del sector rural.

La variable Destino de crédito indica en que se invertirá el monto desembolsado, como puede ser capital de trabajo, compra de activos, libre disponibilidad entre otro; así mismo la variable Tipo de crédito indica si el crédito es nuevo, ampliación o refinanciamiento que apoya en la identificación del cliente; la variable Demanda indica el cumplimiento de sus obligaciones del cliente ante su familia permitiendo conocer el nivel de responsabilidad en razón a sus obligaciones que permite perfilar al cliente; y finalmente la variable Condonaciones indica si el cliente requirió realizar una reducción de su crédito a razón de mora para que el mismo sea pagado permitiendo conocer el antecedente del cliente. Estas nueve variables son consideradas como variables significativas para el entrenamiento de los modelos en otorgamiento de microcréditos.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

4.1.4 Determinación de las variables para el modelo

En el proceso de determinación de las variables independientes se utilizaron las 25 variables del cruce de las variables teóricas y empíricas, las cuales son respaldadas teóricamente a través de las finanzas, adicionando las 9 variables significativas descritas en la sección anterior, haciendo un total de 34 variables independientes para entrenamiento y validación del modelo como se observa en la Tabla 10.

En este proceso también se determinó la variable dependiente, que es lo que se busca predecir con el modelo siendo ésta la predicción del otorgamiento de un microcrédito.

Tabla 10

Variables de entrada del modelo

Variable	Variable
Capital de desembolso	Años en actividad primaria
Monto de la cuota	Años en actividad secundaria
Número de cuotas	Saldo
Oficina	Plazo
Interés Pactado	Garantía
Tipo de solicitud de crédito	Número de dependientes
Clasificación del cliente	Atraso promedio
Tipo de cliente	Número de créditos
Edad	Destino del crédito
Ubigeo	Tipo de crédito
Tipo Vivienda	Demanda
Estado Civil	Sector Económico
Sexo	Actividad Primaria
Nivel Instrucción	Tipo Actividad Primaria
Ocupación	Actividad secundaria
Profesión	Tipo de actividad secundaria
Cuentas de ahorro	Número de condonaciones

En el proceso Obtención de datos se realiza la extracción de los datos de la base de datos de la entidad microfinanciera en función a las 34 variables independientes y la variable dependiente, cuyas etiquetas fueron definidas en razón del cumplimiento de cada microcrédito otorgado a los clientes.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

4.1.5 Discusión

En la presente investigación se identificó un total de 34 variables independientes y una variable dependiente, las variables independientes se obtuvieron a partir de la revisión del proceso de otorgamiento de crédito y la revisión de la literatura existente, obteniendo un total de 25 variables a partir del cruce de las mismas las cuales reflejan el sustento teórico a las variables empíricas identificadas en el proceso que respaldan su validez en el aspecto financiero, a éstas se adicionan 9 variables significativas en el ámbito microfinanciero, en concordancia con Arango (2017), quién menciona, que las variables independientes deben ser aquellas que guarden relación con la variable dependiente, para no distorsionar el modelo. Sin embargo, una de las limitaciones identificadas en la revisión de la literatura es que los modelos están aplicados a un ámbito financiero que se asemeja al microfinanciero teniendo algunas particularidades.

En la investigación se resalta una actividad considerable con respecto a la integración de diferentes variables independientes en relación al campo de finanzas y microfinanzas garantizando la incorporación de características relevantes para el modelo. Además, el proceso de obtención de datos se completaron los datos, tal como sugiere (Zhou, 2018), quien destaca que en muchas tareas es difícil obtener información de supervisión sólida, como etiquetas de verdad completas y correctas, debido al alto costo del proceso de etiquetado de datos y sus características.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

4.2 Resultado conforme al segundo objetivo especifico

Para el entrenamiento de los de modelos se inició con la preparación de datos de las variables identificadas en el punto 4.1, seguidamente se procedio a entrenar los modelos Regresión Logística, *Random Forest, Support Vector Machine, Artificial Neural Networks, Decision Tree y k-Nearest Neighbors.*

4.2.1 Pre procesamiento de datos

Una vez definidas las variables independientes de entrada *X*, y la variable dependiente de salida *y*, que indica si un determinado microcrédito debe ser otorgado o no; se procedió a realizar la extracción de datos de la base de datos de la entidad teniendo un total de 17,454 registros de créditos de los 15,015 clientes correspondientes al periodo de marzo del 2017 a marzo de 2018, los cuales fueron exportados en formato CVS, definiendo el nombre de las variables como se observa en la Tabla 11.

Tabla 11

Nombre de variables para el entrenamiento

Variable	Nombre de la variable	Uso en el modelo	Tipo de variable
Capital de desembolso	nCapitalDesembolso	Entrada	Numérica
Edad	dAnhoNac, dMesNac	Entrada	Numérica
Sexo	idSexo	Entrada	Numérica
Estado Civil	idEstadoCivil	Entrada	Categórica
Plazo	nPlazo	Entrada	Numérica
Garantía	lGarantia	Entrada	Categórica
Nivel Instrucción	idNivelInstruccion	Entrada	Categórica
Años en actividad primaria	nActPriAnios	Entrada	Numérica
Número de dependientes	nNumPerDepend	Entrada	Numérica
Tipo Vivienda	idTipoVivienda	Entrada	Categórica
Tipo de solicitud de crédito	idOperacion	Entrada	Categórica
Ocupación	idOcupacion	Entrada	Categórica
Interés Pactado	nInteresPactado	Entrada	Numérica
Oficina	idEstablecimiento	Entrada	Categórica
Clasificación del cliente	idTipoCliente	Entrada	Categórica
Ubigeo	idUbigeo	Entrada	Categórica
Atraso promedio	nAtrasoPromedio	Entrada	Numérica
Cuentas de ahorro	nAhorro	Entrada	Numérica
Saldo	nTotalDeudas	Entrada	Numérica
Número de créditos	nCreditosVigentes	Entrada	Numérica
Destino del crédito	idDestino	Entrada	Categórica
Monto cuota	nMontoCuota	Entrada	Numérica
Número de cuotas	nCuotas, nMontoCuota	Entrada	Numérica

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Tipo de cliente	IdClasificaDeudor	Entrada	Categórica
Tipo de crédito	idTipoCuentaCredito	Entrada	Categórica
Demanda	nDemanda	Entrada	Numérica
Profesión	idProfesion	Entrada	Categórica
Sector Económico	idSectorEcon	Entrada	Categórica
Actividad Primaria	idActividadInternaPri	Entrada	Categórica
Tipo Actividad Primaria	idTipoActividadPri	Entrada	Categórica
Actividad secundaria	idActividadInternaSec	Entrada	Categórica
Tipo de actividad secundaria	idTipoActividadSec	Entrada	Categórica
Años en actividad secundaria	nActSecAnios	Entrada	Numérica
Número de condonaciones	nCondonaciones	Entrada	Numérica
Crédito otorgado	lOtorgarCredito	Salida	Categórica

En este proceso de pre procesamiento se consideró a los datos faltantes o nulos que afectan en el entrenamiento de los modelos y se realizó la evaluación y gestión de los datos faltantes, los cuales pudieron ser consecuencia de error humano que no ingresaron o no fueron declarados, para lo cual se evaluó la acumulación de los valores faltantes identificados los mismos como valores nulos, teniendo como resultado la Figura 15.

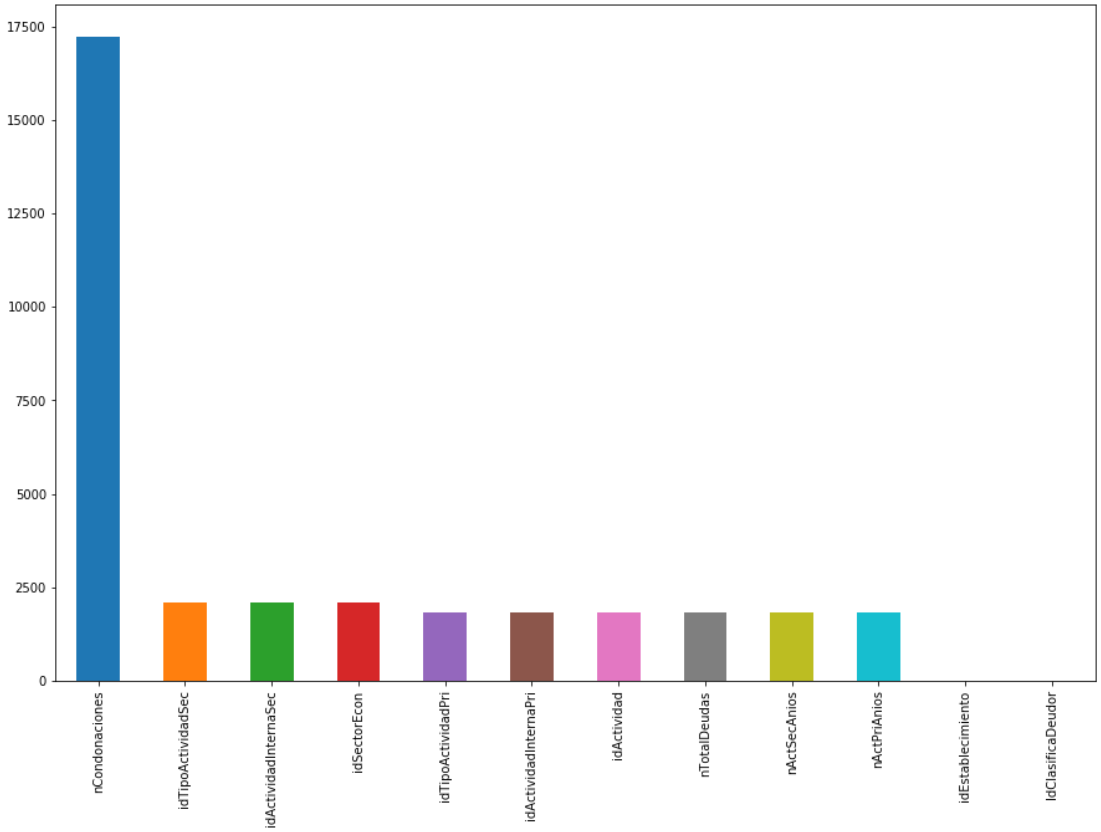


Figura 15. Datos faltantes

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

La Figura 15 presenta valores nulos que son considerados como datos faltantes detallados en la Tabla 12.

Tabla 12

Variables con valores faltantes

Variables	Cantidad de datos faltantes	Porcentaje
nCondonaciones	17,209	99.76%
idTipoActividadSec	2,102	12.18%
idActividadInternaSec	2,102	12.18%
idSectorEcon	2,101	12.18%
idTipoActividadPri	1,839	10.66%
nTotalDeudas	1,838	10.65%
nActSecAnios	1,838	10.65%
nActPriAnios	1,838	10.65%
idActividadInternaPri	1,838	10.65%
idActividad	1,838	10.65%
idTipEvalCred	1,838	10.65%
idEstablecimiento	2	0.01%
IdClasificaDeudor	1	0.01%

Fuente: Elaboración propia

- ***nCondonaciones***: indica el número de condonaciones que el cliente tuvo durante el pago de su crédito y la cantidad de datos faltantes implica el 99.60%, lo cual es significativo pero los datos faltantes del mismo indican que no existió condonación alguna durante el transcurso de su crédito, por lo tanto, los datos faltantes de esta variable son completados con el valor cero.
- ***idTipoActividadSec***: clasifica el tipo de actividad secundaria que realiza el cliente donde la clasificación se registra de acuerdo Superintendencia de Banca y Seguros, que es el ente regulador de la entidad, a su vez se identifica que la cantidad de datos faltantes implica el 12.04% lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.
- ***idActividadInternaSec***: clasifica el tipo de actividad que el cliente realiza de forma secundariamente, de acuerdo a la entidad, también se ha identificado un 12.04% de datos faltantes, que no representa un valor significativo, por ende, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

- ***idSectorEcon***, identifica el sector económico en el que clasificaron al cliente al momento de su registro, y se identifica un 12.04% de datos, lo cual no es significativo, por ende, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.
- ***idTipoActividadPri***, identifica el tipo de actividad primaria que realiza el cliente, la clasificación se registra de acuerdo a la Superintendencia de Banca y Seguros que es el ente regulador de la entidad, a su vez se identifica que la cantidad de datos faltantes implica el 10.54% lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.
- ***nTotalDeudas***, indica la suma de deudas con otras entidades financieras, considerando los créditos directos e indirectos, donde se identifica el 10.53% de datos faltantes, lo cual no es significativo, por lo tanto, los datos faltantes se completaron con el valor promedio de la variable al ser una variable de tipo numérica.
- ***nActSecAnios***, indica el número de años que el cliente se viene dedicando a su actividad secundaria, donde se identifica el 10.53% de datos faltantes, lo cual no es significativo, por lo tanto, los datos faltantes se completaron con el valor promedio de la variable al ser una variable de tipo numérica.
- ***nActPriAnios***, indica el número de años que el cliente se viene dedicando a su actividad primaria, donde se identifica el 10.53% de datos faltantes, lo cual no es significativo, por lo tanto, los datos faltantes se completaron con el valor promedio de la variable al ser una variable de tipo numérica.
- ***idActividadInternaPri***, clasifica la actividad interna primaria a la que el cliente se dedica, registrada por la entidad, a su vez se identifica que la cantidad de datos faltantes implica el 10.53% lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.
- ***idActividad***, clasifica la actividad económica a la que se dedica el cliente, a su vez se identifica que la cantidad de datos faltantes implica el 10.53% lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

- ***idTipEvalCred***, clasifica el tipo de evaluación que se le dio al cliente al momento de realizar el otorgamiento del crédito, a su vez se identifica que la cantidad de datos faltantes implica el 10.53% lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.
- ***idEstablecimiento***, clasifica la oficina donde se le otorgó el crédito, a su vez se identifica que la cantidad de datos faltantes implica el 0.01% lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.
- ***IdClasificaDeudor***, indica la clasificación que tenía el cliente al momento de realizar el desembolso del crédito, a su vez se identifica que la cantidad de datos faltantes implica el 10.65%, lo cual no es significativo, por tanto, los datos faltantes se completaron con la moda de la variable al ser un tipo de dato categórico.

Así mismo, como parte de este proceso de preparación se realiza la exploración de datos, para lo cual se analizaron 16 variables de tipo numérico en total, presentados en la Tabla 13; este análisis considera si es necesario aplicar la normalización de los mismo para obtener datos simétricos.

Tabla 13

Variables tipo numérico consideradas para el modelo

Variables procesamiento	Tipo de variable
nCapitalDesembolso	Numérico
dAnhoNac	Numérico
dMesNac	Numérico
nPlazo	Numérico
nActPriAnios	Numérico
nNumPerDepend	Numérico
nInteresPactado	Numérico
nAtrasoPromedio	Numérico
nAhorro	Numérico
nTotalDeudas	Numérico
nCreditosVigentes	Numérico
nMontoCuota	Numérico
nCuotas	Numérico
nDemanda	Numérico
nActSecAnios	Numérico
nCondonaciones	Numérico

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

De la Tabla 13, las variables *nAnhoNac* y *nMesNac*, se consideraron como datos de tipo categóricos, debido a que *nMesNac* indica el mes de nacimiento del cliente que puede ser en el mes de enero, febrero, marzo, abril, mayo, junio, julio, agosto, setiembre, octubre, noviembre y diciembre, en cuanto a la variable *nAnhoNac* indica el año de nacimiento del cliente, por ende, se tiene solo 14 variables de tipo numérico; por tanto se realizó el análisis de las 14 variables según se observa en la Tabla 14, en la cual se consideran los campos de contador *(count)*, promedio *(mean)*, desviación estándar (*std*), mínimo *(min)*, 25%, 50%, 75% y el máximo *(max)*.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 14

Descripción de variables de tipo numérico

	nCapitalDesembolso	nMontoCuota	nPlazo	nActPriAnios	nActSecAnios	nNumPerDepend	nInteresPactado	nAtrasoPromedio	nAhorro	nTotalDeudas	nCreditosVigentes	nCuotas	nCondonaciones	nDemanda
count	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454
mean	4,914.70	1,046.68	408.76	10.75	4.04	0.86	1,514.71	1.39	0.09	2,120.21	1.04	9.42	0.02	0.00
std	4,638.45	1,596.82	157.52	7.87	6.02	1.23	1,843.19	7.19	0.45	4,648.92	0.38	6.48	0.27	0.00
min	300.00	0.00	30.00	0.00	0.00	0.00	14.49	0.00	0.00	0.00	0.00	1.00	0.00	0.00
25%	2,000.00	214.40	360.00	5.00	0.00	0.00	456.54	0.00	0.00	0.00	1.00	4.00	0.00	0.00
50%	3,001.00	500.00	365.00	10.00	0.00	0.00	871.57	0.00	0.00	0.00	1.00	9.00	0.00	0.00
75%	6,000.00	1,170.80	456.00	15.00	6.00	2.00	1,818.54	1.00	0.00	2,328.18	1.00	12.00	0.00	0.00
max	60,000.00	26,809.50	3,240.00	100.00	100.00	21.00	29,831.95	259.00	20.00	104,051.75	3.00	60.00	10.00	0.00

TESIS EPG UNA - PUNO
Universidad Nacional del Altiplano

Según la Tabla 14 en la variable *nDemanda*, se puede observar que tanto los valores mínimo y máximo son 0, lo cual implica que no posee ningún valor, por lo tanto, se elimina del *dataset* al no tener valor en para el entrenamiento de los modelos de *Machine Leaning*.

Las variables con mayor valor en desviación estándar son: *nCapitalDesembolso*, *nTotalDeudas* y *nMontoCuota*, lo cual implica que son asimétricos, por ende, fue necesario verificar la distribución de cada una de estas variables a fin de corregir su simetría.

nCapitalDesembolso, indica el monto del préstamo que se otorgó al cliente en una determinada fecha, como se puede observar en la Tabla 4, tiene el valor máximo de 60,000.00 y mínimo de 300.00 nuevos soles y con una desviación estándar muy elevada de 4,914.00, lo cual indica de los datos de esta variable se encuentran dispersos, en Figura 16 se tiene la distribución de la variable *nCapitalDesembolso*, la cual se encuentra sesgada positivamente, obteniendo un coeficiente de asimetría de 2.37.

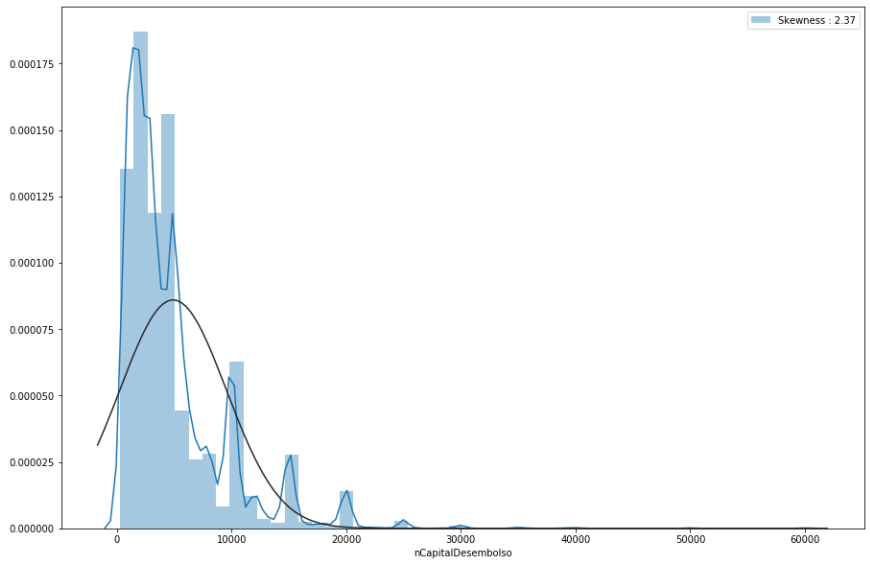


Figura 16. Distribución de variable Capital de desembolso *(*nCapitalDesembolso*)*

Por lo tanto, utilizaremos la función *log1p* de la librería *numpy*, que fue aplicado para transformar variable, teniendo el resultado mostrado en la Figura 17, donde se observa un coeficiente de asimétrica -0.10, el cual está más cerca a cero, por tanto, es cercano a ser simétrico.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

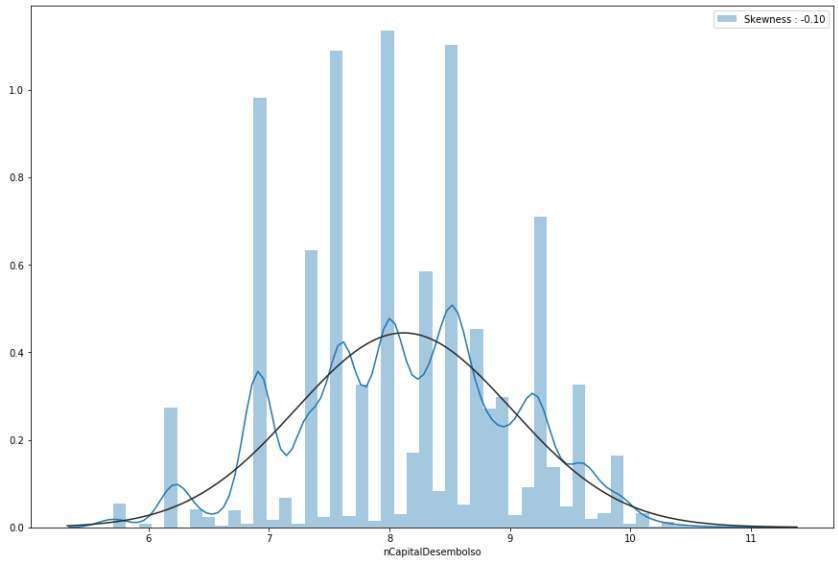


Figura 17. Distribución de variable Capital de desembolso transformada

***nTotalDeudas*:** Representa el monto de créditos directos e indirectos que el cliente tenía al momento de otorgar el crédito, el rango está dentro de 0.00 y 104,051.75, a su vez la desviación estándar es 4,648.92, los que indica que los datos se encuentran dispersos, y de acuerdo a la Figura 18 se tiene un coeficiente de asimetría de 5.16, que implica que la distribución se encuentra sesgada positivamente.

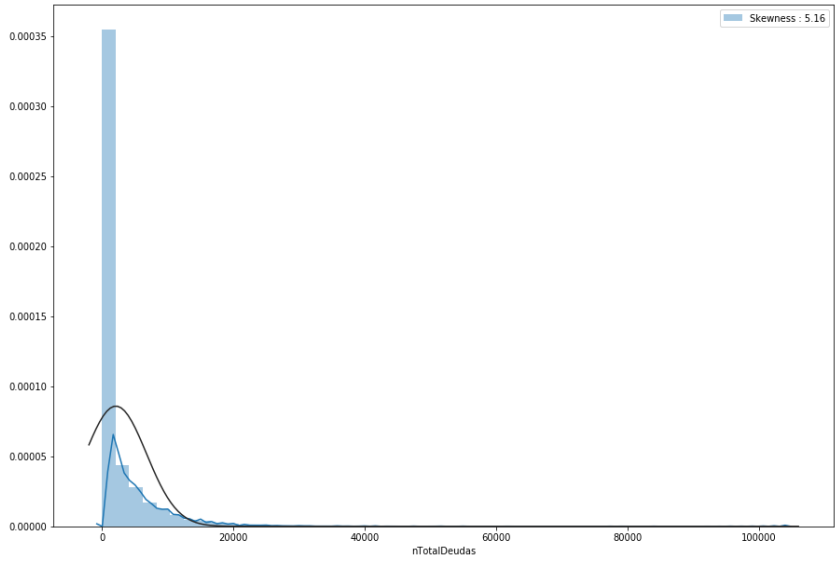


Figura 18. Distribución de variable Saldo (*nTotalDeudas*)

Se utilizará la función *log1p* para transformar los datos de la variable *nTotalDeudas*, obteniendo como resultado un coeficiente de asimétrica 0.46 de acuerdo a la Figura 19, el cual está más cerca a cero, por tanto, es cercano a ser simétrico.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

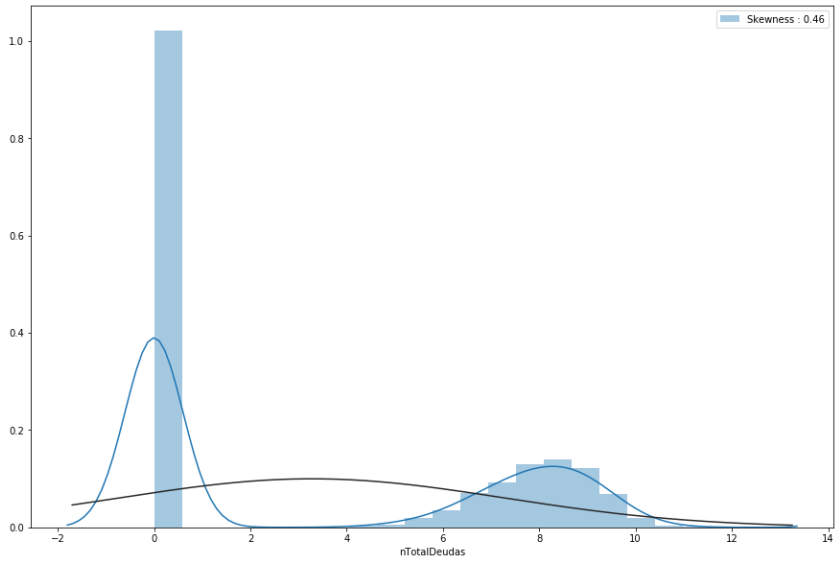


Figura 19. Distribución de variable Saldo transformada

***nMontoCuota*:** Representa el monto de cuota que el cliente debería pagar en un determinado tiempo determinado al momento del otorgamiento del crédito, el rango se encuentra entre 0.00 y 26,809.50, a su vez la desviación estándar es 1,596.82 y con un coeficiente de asimetría de 4.19 y una distribución sesgada positivamente como se muestra en la Figura 20.

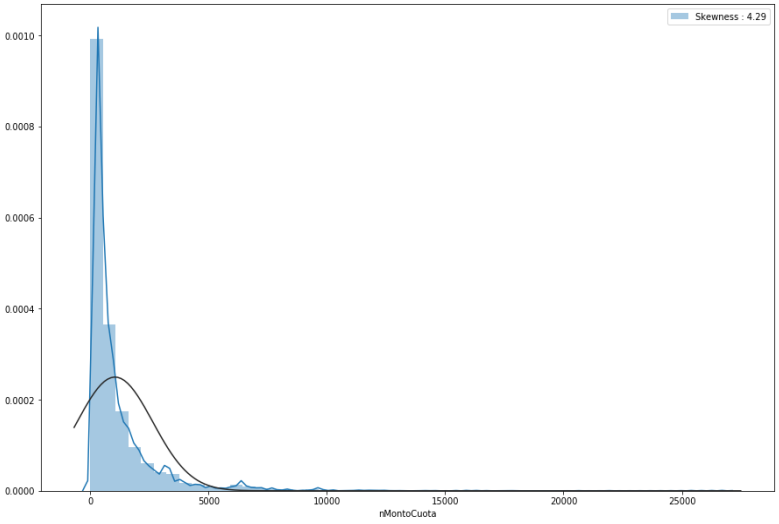


Figura 20. Distribución de variable Cuota (*nMontoCuota*)

Se utilizó la función *log1p* para transformar los datos de la variable *nMontoCuota*, obteniendo como resultado un coeficiente de asimétrica 0.46 de acuerdo a la Figura 21, con esto se reduce el coeficiente de asimetría.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

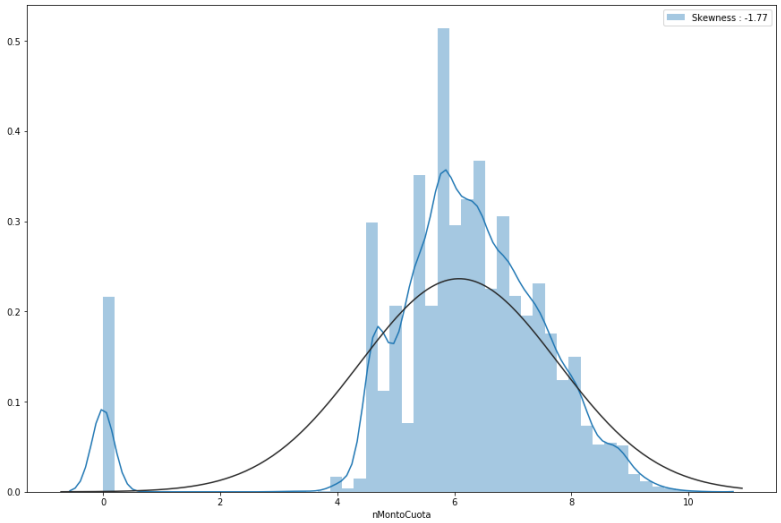


Figura 21. Distribución de variable Cuota transformada

Así mismo se verifica las variables *nPlazo, nAtrasoPromedio, nActPriAnios, nCuotas, nActSecAnios y nInteresPactado*, las cuales tiene un coeficiente de asimetría alto como se muestra en la Figura 22. Las mismas fueron transformadas con la función *Log1p* para reducir el coeficiente de asimetría, teniendo como resultado la Figura 23, en la cual los mismo redujeron el coeficiente de asimetría, de forma que se aproxima a cero.

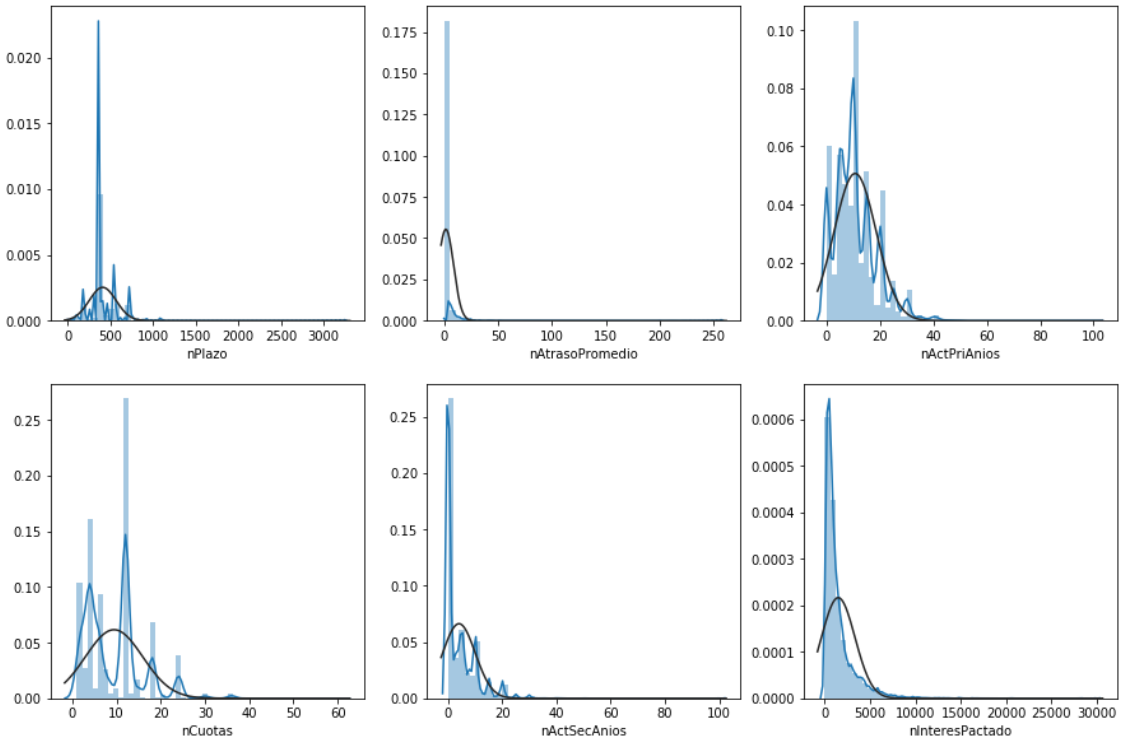


Figura 22. Distribución de variables faltantes

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

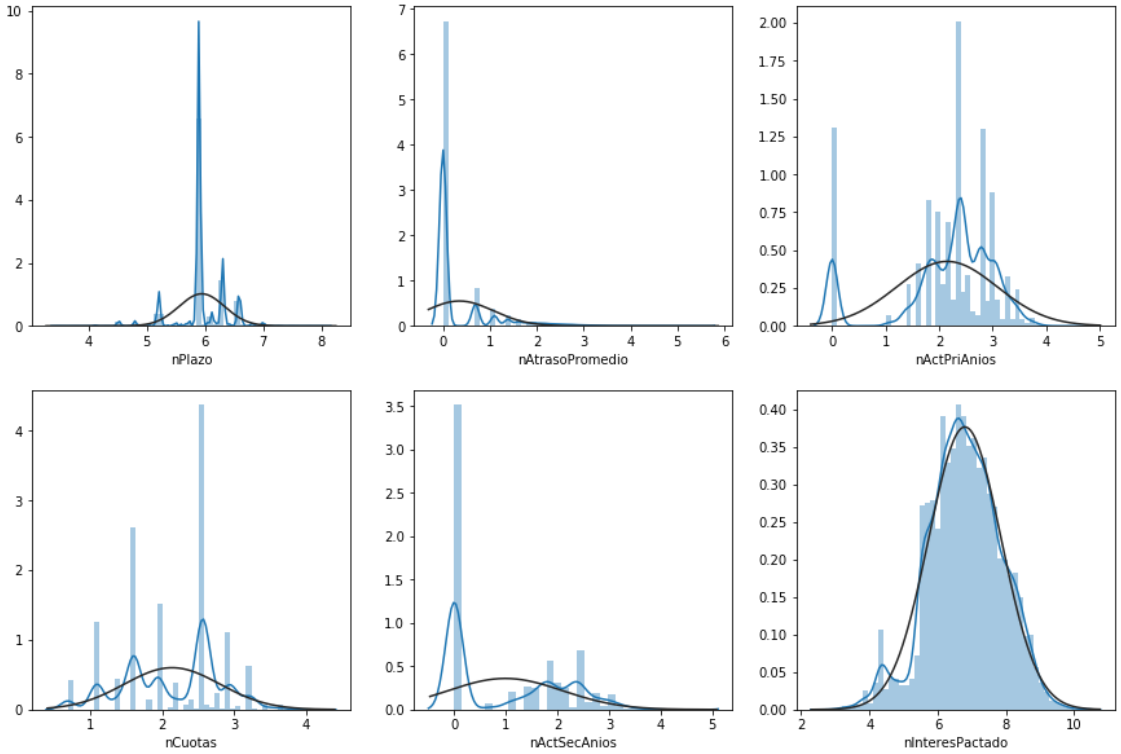


Figura 23. Distribución de variables faltantes transformados

Obteniendo los resultados como se muestra en la Tabla 15, en la cual se observa el resumen de los datos pre procesados y que son exportados para el entrenamiento de los modelos de *Machine Learning*.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 15

Resumen de datos pre procesados para entrenamiento de los modelos

	nCapitalDesembolso	nMontoCuota	nPlazo	nActPriAnios	nActSecAnios	nNumPerDepend	nInteresPactado	nAtrasoPromedio	nAhorro	nTotalDeudas	nCreditosVigentes	nCuotas	nCondonaciones	lOtorgarCredito
count	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454	17,454
mean	8.12	6.08	5.94	2.15	0.99	0.86	6.80	0.35	0.09	3.28	1.04	2.14	0.02	0.77
std	0.90	1.69	0.39	0.94	1.11	1.23	1.06	0.73	0.45	4.00	0.38	0.66	0.27	0.42
min	5.71	0.00	3.43	0.00	0.00	0.00	2.74	0.00	0.00	0.00	0.00	0.69	0.00	0.00
25%	7.60	5.37	5.89	1.79	0.00	0.00	6.13	0.00	0.00	0.00	1.00	1.61	0.00	1.00
50%	8.01	6.22	5.90	2.40	0.00	0.00	6.77	0.00	0.00	0.00	1.00	2.30	0.00	1.00
75%	8.70	7.07	6.12	2.77	1.95	2.00	7.51	0.69	0.00	7.75	1.00	2.56	0.00	1.00
max	11.00	10.20	8.08	4.62	4.62	21.00	10.30	5.56	20.00	11.55	3.00	4.11	10.00	1.00

TESIS EPG UNA - PUNO
Universidad Nacional del Altiplano

Finalmente, como parte del proceso de pre procesamiento de los datos se aplicó la técnica de codificación *One Hot* a las variables de tipo categórico, obteniendo como resultado un *dataset* pre procesado para el uso en el entrenamiento y validación de los modelos.

4.2.2 Entrenamiento de modelos

En este proceso se realizó entrenamiento de los modelos seleccionados en punto 3.5.2, los cuales fueron sometidos a la misma situación para luego ser evaluados y comparados.

4.2.3 Regresión Logística

En el entrenamiento de algoritmo Regresión Logística se utilizó la librería *Scikit Learn*, específicamente la función *SGDClassifier*, obteniendo el rendimiento del algoritmo a través de la Curva *ROC* como se muestra en la Figura 24, obteniendo la medida bajo el área de la curva *ROC* (*AUC*) de 0.8607, lo que indica que el algoritmo tiene un rendimiento óptimo, dado que se encuentra encima de la línea de no discriminación.

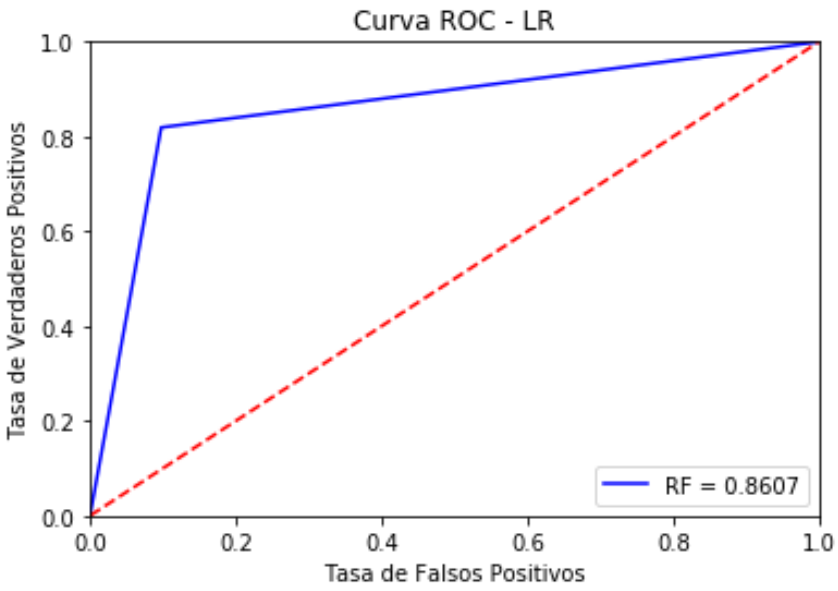


Figura 24. Curva *ROC* para Regresión Logística

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

A su vez se determina las siguientes métricas, que se puede observar en la Tabla 16, donde se obtiene en entrenamiento una exactitud (*accuracy*) 0.8440, mientras que en pruebas se obtiene una exactitud de 0.8390, con una precisión (*precision*) de 0.9638, una exhaustividad (*recall*) de 0.8191 y F1 de 0.8856, por lo que se determina que el algoritmo de *Machine Learning* Regresión Logística, tiene un óptimo desempeño.

Tabla 16:

Métricas resultantes de entrenamiento de Regresión Logística

Train Accuracy	***Test Accuracy***	***Precision***	***Recall***	***F1***
0.8440	0.8390	0.9638	0.8191	0.8856

4.2.4 *Random Forest*

Breiman (2001) define un bosque aleatorio como un clasificador que consiste en una colección de clasificadores estructurados en árbol, en el cual se utilizó la librería *Scikit Learn*, específicamente la función *RandomForestClassifier*, obteniendo el rendimiento del algoritmo como se muestra en la Figura 25, obteniendo la medida bajo el área de la curva *ROC (AUC)* de 0.6635, lo que indica que el algoritmo tiene un rendimiento óptimo, dado que se encuentra encima de la línea de no discriminación.

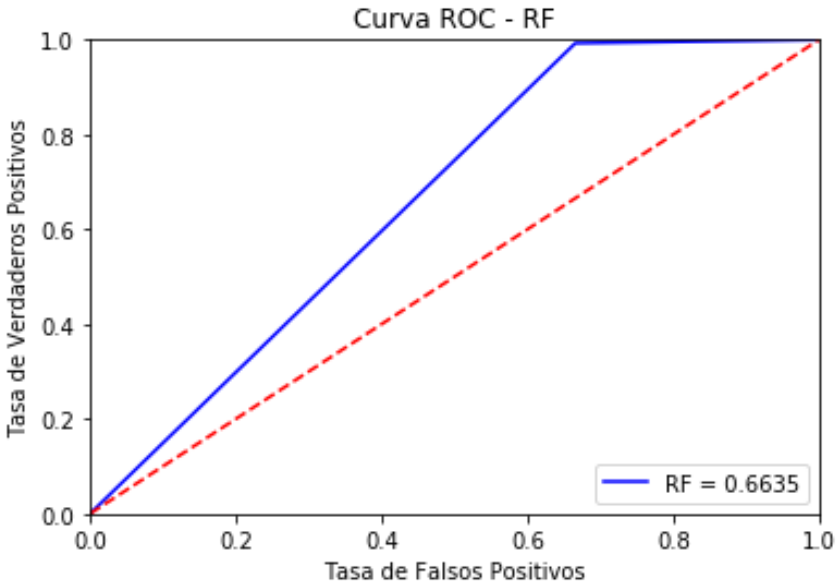


Figura 25. Curva *ROC* para *Random Forest*

A su vez se determina las siguientes métricas, que se puede observar en la Tabla 17, donde se obtiene en entrenamiento una exactitud (*accuracy*) 0.8494, mientras que en pruebas se obtiene una exactitud de 0.8351, con una precisión (*precision*) de 0.8257,

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

una exhaustividad (*recall*) de 0.9928 y *F1* de 0.9016, por lo que se determina que el algoritmo de *Machine Learning Random Forest*, tiene un óptimo desempeño.

Tabla 17

Métricas resultantes de entrenamiento de Random Forest

Train Accuracy	*Test Accuracy*	*Precision*	*Recall*	*F1*
0.8494	0.8351	0.8257	0.9928	0.9016

4.2.5 *Support Vector Machine*

Según Liang et al. (2016) indican que la idea principal de este algoritmo es determinar el hiperplano de separación que maximiza el margen entre dos clases de datos de entrenamiento, para lo cual se utilizó la librería *Scikit Learn*, específicamente la función *LinearSVC*, obteniendo el rendimiento del algoritmo como se muestra en la Figura 26, obteniendo la medida bajo el área de la curva *ROC (AUC)* de 0.8444, lo que indica que el algoritmo tiene un rendimiento óptimo, dado que se encuentra encima de la línea de no discriminación.

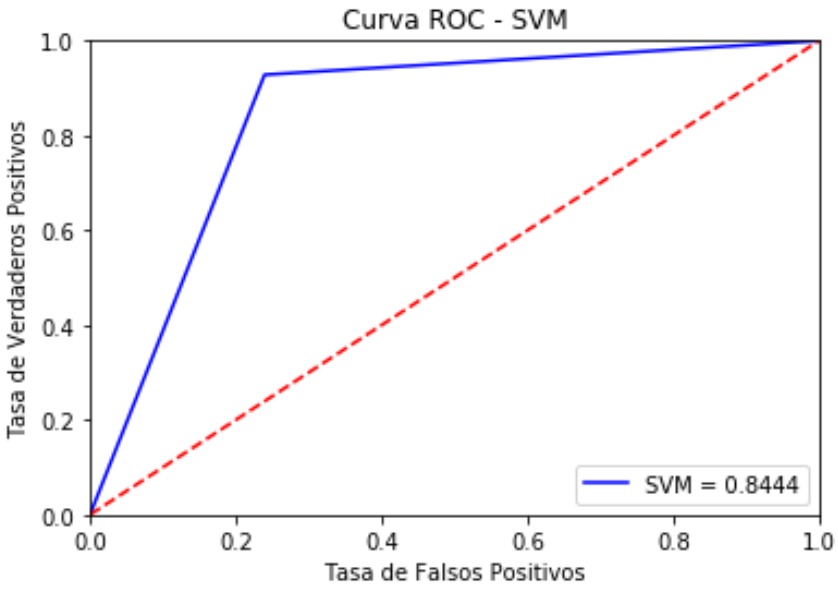


Figura 26: Curva *ROC* para *Support Vector Machine*

A su vez se determinan las siguientes métricas, la cuales se observan en la Tabla 18, donde se obtiene en entrenamiento una exactitud (*accuracy*) 0.8895, mientras que en pruebas se obtiene una exactitud de 0.8881, con una precisión (*precision*) de 0.9249, una exhaustividad (*recall*) de 0.9282 y *F1* de 0.9266, por lo que se determina que el algoritmo de *Machine Learning Support Vector Machine*, tiene un óptimo desempeño.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 18

Métricas resultantes de entrenamiento de Support Vector Machine

Train Accuracy	Test Accuracy	Precision	Recall	F1
0.8895	0.8881	0.9249	0.9282	0.9266

4.2.6 *Artificial Neural Networks*

Según Shanmuganathan (2016) es un modelo computacional inspirado biológicamente, consiste en elementos de procesamiento y conexiones entre ellos con coeficientes unidos a las conexiones simulando las redes neuronales cerebrales, en esta investigación se utilizó la librería *Keras* una de las más usadas en los medios académicos, específicamente las función *Sequential* y *Dense*, obteniendo el rendimiento del algoritmo que se muestra en la Figura 27, obteniendo la medida bajo el área de la curva *ROC (AUC)* de 0.9372, lo que indica que el algoritmo tiene un rendimiento óptimo, dado que se encuentra encima de la línea de no discriminación.

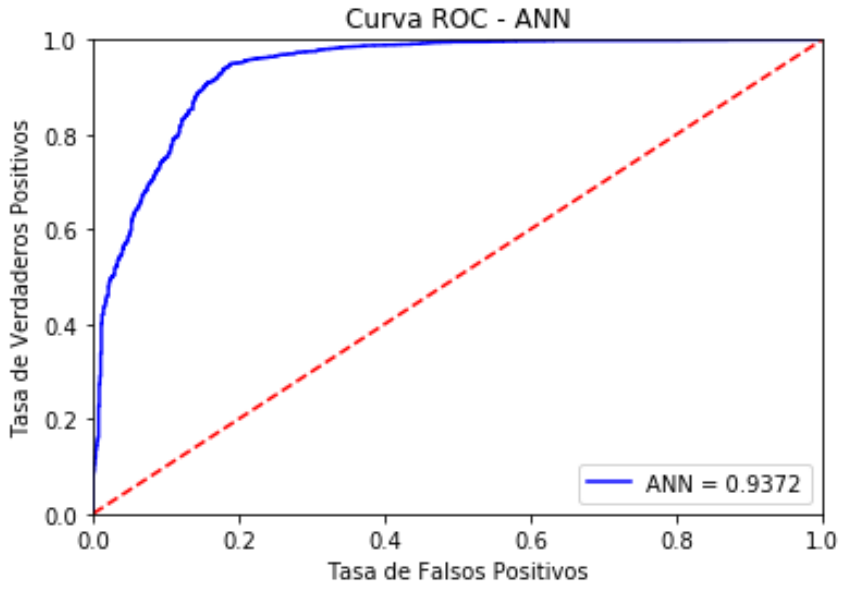


Figura 27. Curva *ROC* para *Artificial Neural Networks*

A su vez se determina las siguientes métricas, que se puede observar en la Tabla 19, donde se obtiene en entrenamiento una exactitud (*accuracy*) 0.9317, mientras que en pruebas se obtiene una exactitud de 0.9119, con una precisión (*precision*) de 0.9184, una exhaustividad (*recall*) de 0.9705 y *F1* de 0.9437, por lo que se determina que el algoritmo de *Machine Learning Artificial Neural Networks*, tiene un óptimo desempeño.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 19

Métricas resultantes de entrenamiento de Artificial Neural Network

Train Accuracy	Test Accuracy	Precision	Recall	F1
0.9317	0.9119	0.9184	0.9705	0.9437

4.2.7 *Decision Tree*

Luego del entrenamiento del modelo se obtiene el rendimiento del algoritmo como se muestra en la Figura 28, donde se obtiene la medida bajo el área de la curva *ROC (AUC)* de 0.8880, lo que indica que el algoritmo tiene un rendimiento óptimo, dado que se encuentra encima de la línea de no discriminación.

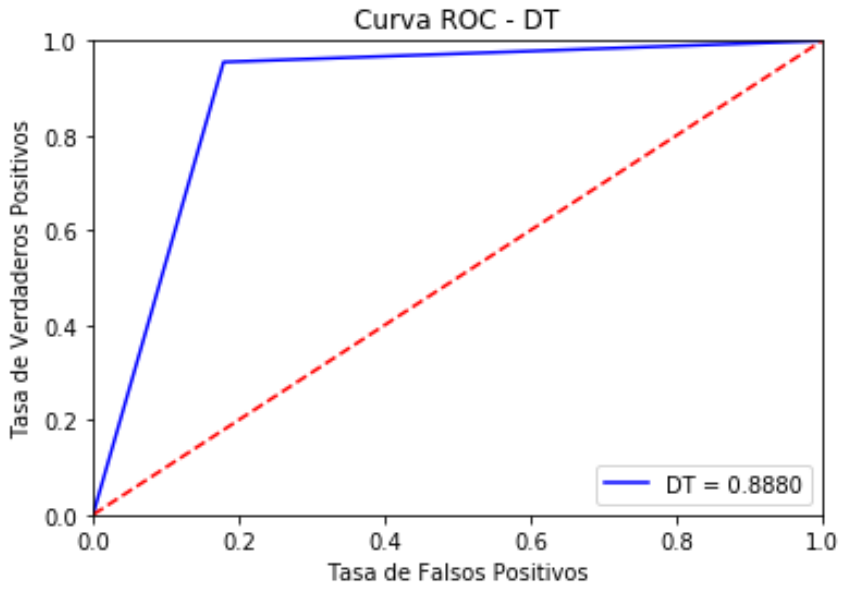


Figura 28. Curva *ROC* para *Decision Tree*

A su vez se determina las siguientes métricas, que se puede observar en la Tabla 20, donde se obtiene en entrenamiento una exactitud (*accuracy*) 0.9665, mientras que en pruebas se obtiene una exactitud de 0.9230, con una precisión (*precision*) de 0.9443, una exhaustividad (*recall*) de 0.9551 y *F1* de 0.9496, por lo que se determina que el algoritmo de *Machine Learning Decision Tree*, tiene un óptimo desempeño.

Tabla 20

Métricas resultantes de entrenamiento de Decision Tree

Train Accuracy	Test Accuracy	Precision	Recall	F1
0.9665	0.9230	0.9443	0.9551	0.9496

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

4.2.8 *k-Nearest Neighbors*

Con este modelo se obtiene el rendimiento del algoritmo mostrado en la Figura 29, obteniendo la medida bajo el área de la curva *ROC (AUC)* de 0.6598, lo que indica que el algoritmo tiene un rendimiento óptimo, dado que se encuentra encima de la línea de no discriminación.

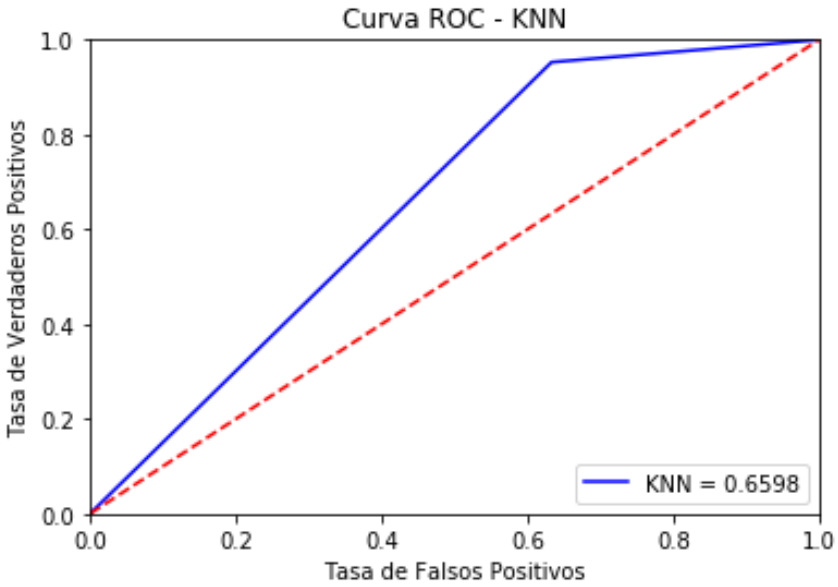


Figura 29. Curva *ROC* para *k-Nearest Neighbors*

A su vez se determina las siguientes métricas, que se puede observar en la Tabla 21, donde se obtiene en entrenamiento una exactitud (*Accuracy*) 0.8831, mientras que en pruebas se obtiene una exactitud de 0.8124, con una precisión (*Precision*) de 0.8270, una exhaustividad (*recall*) de 0.9527 y *F1* de 0.8854, por lo que se determina que el algoritmo de *Machine Learning k-Nearest Neighbors*, tiene un óptimo desempeño.

Tabla 21

Métricas resultantes de entrenamiento de k-Nearest Neighbors

Train Accuracy	*Test Accuracy*	*Precision*	*Recall*	*F1*
0.8831	0.8124	0.8270	0.9527	0.8854

4.2.9 Comparación de modelos de *Machine Learning*

Luego de realizado el entrenamiento, se realizó la evaluación de los modelos de *Machine Learning*, obteniendo los resultados como se observa en la Tabla 22 y en la Figura 30, en la cual *DT* se obtiene una mejor puntuación en *Accuracy* seguido de *ANN*, que refleja una medición global del otorgamiento de microcréditos, dado que

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

esta métrica depende del balanceo de los casos positivos y negativos en nuestro *dataset*.

Tabla 22

Métricas resultantes de los Modelos

	Regresión Logistica (RL)	***Random Forest (RF)***	***Support Vector Machine (SVM)***	***Artificial Neural Networks (ANN)***	***Decision Tree (DT)***	***k-Nearest Neighbors (kNN)***
Accuracy	83.90 %	83.51 %	88.81 %	91.19 %	92.30 %	81.24 %
Precision	96.38 %	82.57 %	92.49 %	91.84 %	94.43 %	82.70 %
Recall	81.91 %	99.28 %	92.82 %	97.05 %	95.51 %	95.27 %
F1	88.56 %	90.16 %	92.66 %	94.37 %	94.96 %	88.54 %
ROC	86.07 %	66.35 %	84.44 %	93.72 %	88.80 %	65.98 %

Si se analiza *Precision*, los modelos *LR* y *DT* obtienen el mejor puntaje que indica la ratio de aciertos en casos positivos en el otorgamiento de microcréditos, pues esta métrica se enfoca en los casos positivos y no toma en cuenta los aciertos negativos. Aplicando *Recall*, el modelo con mejor puntuación es *RF*, seguido de *ANN*, que indica el ratio de microcréditos identificados como otorgados del total de los que se deberían otorgar, esta métrica se enfoca también en los casos positivos. En *F1 Score* el primer lugar lo ocupa *DT*, seguido de *ANN*, que indica la precisión de los microcréditos otorgados, esta métrica involucra el equilibrio entre las métricas *Precision* y *Recall*, se centra en los casos positivos. Finalmente, en *AUC ROC* el modelo con mejor puntuación en ANN, seguido de DT, que indica el ratio de créditos a ser otorgados y no otorgados correctamente, esta métrica se enfoca en los casos positivos como los negativos. Cabe señalar que los modelos fueron sometidos al mismo escenario para luego ser evaluados y comparados.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

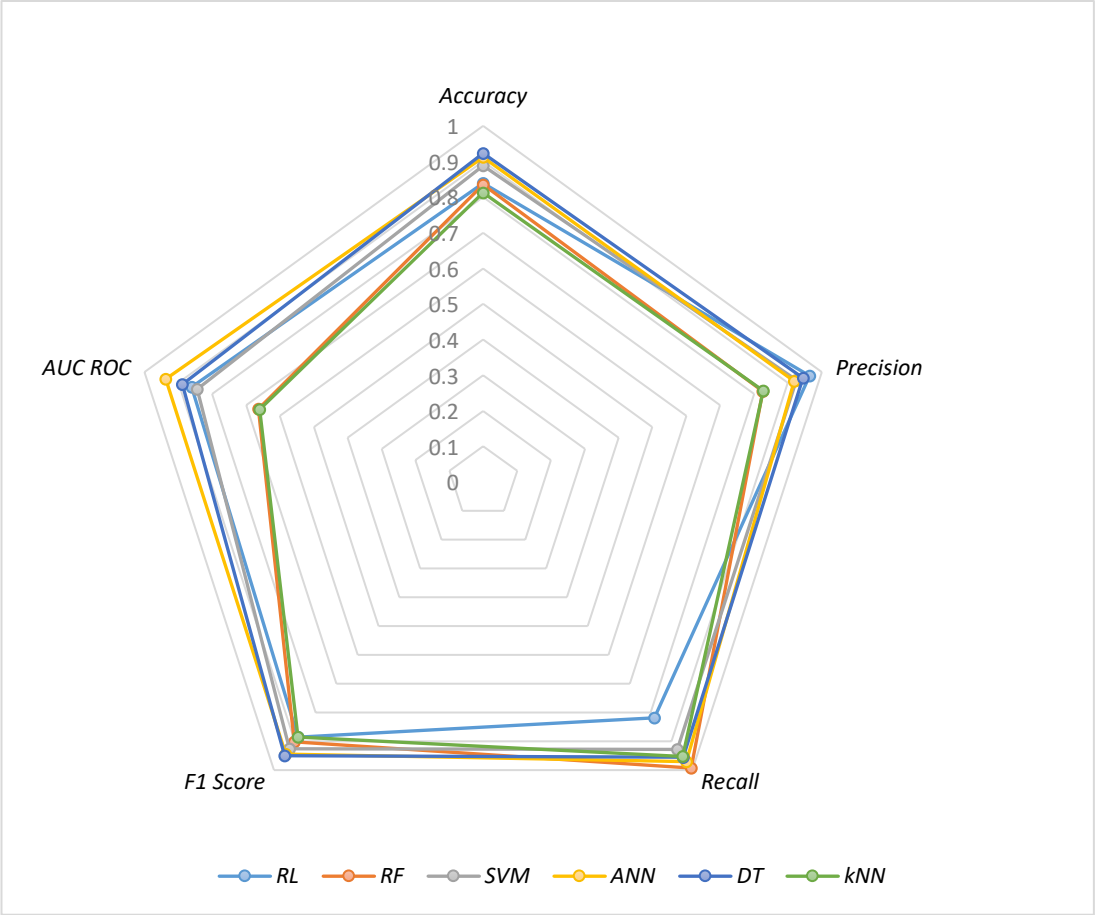


Figura 30. Comparación de los modelos de *Machine Learning*

Para medir el *Nivel de asertividad* de los modelos se han comparado las métricas estudiadas, como se muestra en la Figura 30, donde se puede observar que los modelos *ANN*, *SVM* y *RL* tienen los mejores resultados en relación a las cinco métricas, lo que hace difícil la selección del modelo más asertivo para el otorgamiento de microcréditos. Evaluando las métricas seleccionadas, se sabe que la métrica *Accuracy* mide de forma global a cada modelo, mientras que las métricas *Precision, Recall* y *F1 Score* se centran en los aciertos positivos, sin embargo, se deben considerar también los aciertos negativos, finalmente AUC ROC considera tanto los aciertos positivos y negativos, lo que permite medir de forma más apropiada el Nivel de asertividad en el otorgamiento de microcréditos, reduciendo el nivel de riesgo. Esta selección coincide con la encontrada por autores como (Kruppa et al., 2013), (Addo et al., 2018), (Hossin & M.N, 2015), (Flores & Ramon, 2014) y (Millán & Caicedo, 2018).

Considerando los resultados de la métrica *AUC ROC*, el rendimiento de los modelos se obtiene a través de la curva *ROC*, mostrada en las Figuras 24, 25, 26, 27, 28 y 29, evaluando su rendimiento según la línea de no discriminación. Como se aprecia en las

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

figuras, el rendimiento de todos los modelos se considera óptimo, pues se encuentran por encima de la línea de no discriminación, esto se ve en los modelos de RL (Figura 24); *RF*, (Figura 25); *SVM* (Figura 26); *ANN* (Figura 27); *DT* (Figura 28) y *kNN* (Figura 29). Como se observa, el modelo más asertivo en el otorgamiento de microcréditos es el *Artificial Neural Networks (ANN)*, con un nivel de asertividad del 93.72 %, el cual determina que créditos deben ser otorgados o no otorgados.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

4.2.10 Discusión

Para la determinación del modelo más asertivo se seleccionaron los modelos de *Machine Learning* más relevantes en la predicción de otorgamiento de crédito en concordancia con Turkson et al. (2016)**,** que menciona que los modelos de *Machine Learning* aplicados a las finanzas tiene un nivel de predicción relativamente bueno en el otorgamiento de crédito, sin embargo una limitación identificada es la falta de modelos *Machine Leaning* en el campo de las microfinanzas, punto que es corroborado en la presente investigación al obtener un buen nivel de asertividad en cada uno de los modelos entrenados debido al uso del proceso propuesto en la presente investigación que da mayor énfasis en la determinación de las variables.

Así mismo Turkson et al. (2016) y Chakraborty & Joseph (2017) utilizan las *Acurracy*, *Presicion, Recall, F-1 Score* que les resultaron adecuados para evaluar modelos de *Machine Learning*, así mismo Flores & Ramon (2014), Addo et al. (2018), Kruppa et al. (2013), Millán & Caicedo (2018) utilizan la métrica *AUC ROC* para medir sus modelos, resultándoles adecuados para los mismos; en la presente investigación se utilizó todas las métricas utilizadas por estos autores, dado que cada métrica tiene una formula distinta para medir el desempeño de cada modelo, resultando la métrica AUC ROC como la métrica que dio mejores resultados para obtener el Nivel de asertividad, debido a su naturaleza que mide tanto los aciertos positivos y aciertos negativos.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

4.3 Prueba de hipótesis

Para la prueba de hipótesis con respecto modelo predictivo de análisis de riesgo crediticio usando *Machine Learning* en una entidad del sector microfinanciero, se ha utilizado como método de prueba de hipótesis la denominada prueba t-student, donde se ha utilizado el valor de $\mu_0 = 76.06$, que fue determinado a través de los casos de aciertos en la entidad microfinanciera que implicaría el nivel de asertividad humano, así mismo en el proceso que se ha realizado utilizando el software SPSS, para lo cual se plantearon las siguientes hipótesis estadísticas:

CON RESPECTO AL MODELO REGRESIÓN LOGISTICA

H_O: El modelo Regresión Logística no mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

H_a: El modelo Regresión Logística mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

$$H_O: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Tabla 23

Prueba de muestras única con respecto a Regresión Logística.

Prueba de muestra única

	Valor de prueba = 76.06					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de asertividad	4.501	4	.011	11.30400	4.3313	18.2767

En la Tabla 23, se muestra que $t_c = 4.501$, así mismo se conoce que $t_t = 2.1318$, por lo que $t_c > t_t$, entonces se rechaza la hipótesis nula y se acepta la hipótesis alterna, lo que afirma que el modelo de Regresión Logística mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

CON RESPECTO AL MODELO *RANDOM FOREST*

H_O: El modelo *Random Forest* no mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

H_a: El modelo *Random Forest* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

$$H_O: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Tabla 24

Prueba de muestras única con respecto al modelo Random Forest.

Prueba de muestra única

	Valor de prueba = 76.06					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de asertividad	1,538	4	,199	8,31400	-6,6959	23,3239

En la Tabla 24, se muestra que $t_c = 1.538$, así mismo se conoce que $t_t = 2.1318$, por lo que $t_c \leq t_t$, entonces se acepta la hipótesis nula y se rechaza la hipótesis alterna, lo que implica que el modelo de *Random Forest* no mejorar el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

CON RESPECTO AL MODELO *SUPPORT VECTOR MACHINE*

H_O: El modelo *Support Vector Machine* no mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

H_a: El modelo *Support Vector Machine* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

$$H_O: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

Tabla 25

Prueba de muestras única con respecto al modelo Support Vector Machine.

Prueba de muestra única						
	Valor de prueba = 76.06					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de asertividad	8,692	4	,001	14,18400	9,6532	18,7148

En la Tabla 25, se muestra que $t_c = 8.692$, así mismo se conoce que $t_t = 2.1318$, por lo que $t_c > t_t$, entonces se rechaza la hipótesis nula y se acepta la hipótesis alterna, lo que afirma que el modelo de *Support Vector Machine* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

CON RESPECTO AL MODELO *ARTIFICIAL NEURAL NETWORKS*

H_O: El modelo *Artificial Neural Networks* no mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.
H_a: El modelo *Artificial Neural Networks* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

$$H_O: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Tabla 26

Prueba de muestras única con respecto al modelo Artificial Neural Networks.

Prueba de muestra única						
	Valor de prueba = 76.06					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de asertividad	16,986	4	,000	17,57400	14,7014	20,4466

En la Tabla 26, se muestra que $t_c = 16.986$, así mismo se conoce que $t_t = 2.1318$, por lo que $t_c > t_t$, entonces se rechaza la hipótesis nula y se acepta la hipótesis alterna, lo que afirma que el modelo de *Artificial Neural Networks* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

CON RESPECTO AL MODELO *DECISION TREE*

H_O: El modelo *Decision Tree* no mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

H_a: El modelo *Decision Tree* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

$$H_O: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Tabla 27

Prueba de muestras única con respecto al modelo Decision Tree.

Prueba de muestra única

	Valor de prueba = 76.06					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de asertividad	13,968	4	,000	17,14000	13,7330	20,5470

En la Tabla 27, se muestra que $t_c = 13.968$, así mismo se conoce que $t_t = 2.1318$, por lo que $t_c > t_t$, entonces se rechaza la hipótesis nula y se acepta la hipótesis alterna, lo que afirma que el modelo de *Decision Tree* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

CON RESPECTO AL MODELO *K-NEAREST NEIGHBORS*

H_O: El modelo *k-Nearest Neighbors* no mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

H_a: El modelo *k-Nearest Neighbors* mejora el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

$$H_O: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Tabla 28

Prueba de muestras única con respecto al modelo k-Nearest Neighbors.

Prueba de muestra única

	Valor de prueba = 76.06					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Nivel de asertividad	1,374	4	,241	6,68600	-6,8275	20,1995

En la Tabla 28, se muestra que $t_c = 1.374$, así mismo se conoce que $t_t = 2.1318$, por lo que $t_c \leq t_t$, entonces se acepta la hipótesis nula y se rechaza la hipótesis alterna, lo que implica que el modelo de *k-Nearest Neighbors* no mejorar el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

Finalmente, de las Tablas 23 - 28, se puede observar que *Artificial Neural Network* es el mejor modelo que nos permite mejorar el nivel de asertividad en el otorgamiento de microcréditos en una entidad del sector microfinanciero.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

CONCLUSIONES

Se determinó que el modelo más asertivo para el otorgamiento de microcréditos en la entidad microfinanciera siendo éste *Artificial Neural Network* en comparación a los modelos Regresión Logística, *Random Forest*, *Support Vector Machine*, *Decision Tree* y *k-Nearest Neighbor* usando el proceso propuesto en la presente investigación, lo que permitió reducir el nivel de riesgo en el otorgamiento de microcréditos.

La exploración del proceso de otorgamiento de la entidad microfinanciera y la revisión de investigaciones relacionadas al tema permitió determinar un total de 34 variables significativas que se observa en la Tabla 10, las cuales fueron utilizadas para el entrenamiento de los modelos de *Machine Learning*, garantizando que las mismas guarden relación al otorgamiento de un microcrédito.

Se utilizaron modelos de *Machine Learning* aplicados a riesgo crediticio, los cuales fueron seleccionados, entrenados y evaluados, usando diversas librerías. Se ha calculado el Nivel de asertividad, a través de diferentes métricas, como *Accuracy*, *Precision*, *Recall*, *F1 Score* y *AUC ROC*, encontrándose el modelo con mejor desempeño en la evaluación de otorgamiento de crédito, que ha sido *Artificial Neural Network (ANN)*, el cual ha obtenido un Nivel de asertividad de 93.72 %, sobre los modelos Regresión Logística (86.07 %), *Random Forest* (66.35 %), *Support Vector Machine* (84.44 %), *Decision Tree* (88.80 %) y *k-Nearest Neighbor* (65.98 %).

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

RECOMENDACIONES

Se recomienda el uso del proceso propuesto en la presente investigación para determinar el modelo más asertivo en la predicción de otorgamiento de microcréditos de entidades dedicadas al mismo rubro.

Se recomienda que el personal especializado haga uso del modelo más asertivo *(ANN)*, como herramienta de ayuda a la toma de decisiones, al momento de otorgar un microcrédito en la entidad microfinanciera. Así mismo, para los desarrolladores, se sugiere que el proceso propuesto sea probado en otros segmentos financieros u otras regiones del ámbito rural.

Se recomienda el uso de la técnica de codificación *One Hot*, que permitió asegurar un mejor comportamiento en el proceso de entrenamiento de los diferentes modelos de *Machine Learning*, obteniéndose una información de supervisión sólida, con etiquetas y 31 variables de entrada completas y correctas, de otorgar un crédito o no.

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

BIBLIOGRAFÍA

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. In *SSRN*. doi: 10.2139/ssrn.3155047

Albon, C. (2018). *Machine Learning with Python Cookbook*. Sebastopol, Estados Unidos: O'Reilly Media, Inc.

Altman, E., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, *21*.

Arango, L., & Restrepo, D. (2017). *Diseño de un modelo de scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento colombiana*, 86. Recuperado de: https://repository.eafit.edu.co/xmlui/bitstream/handle/10784/12434/Laura_Arango Duque_Daniel_RestrepoBaena_2017.pdf?sequence=2&isAllowed=y

Arango, S. (2017). *Modelo de puntaje para el otorgamiento de créditos hipotecarios basados en el comportamiento de precio de la vivienda en Medellín*. Universidad EIA, Medellin, Colombia.

Ayma, V. (2019). *Métodos Basados en Búsqueda*. Machine Learning Seminar, Seminario llevado por la Universidad Nacional de San Antonio Abad del Cusco en la Ciudad de Cuzco, Perú.

Baştanlar, Y., & Özuysal, M. (2014). *Introduction to machine learning Methods in Molecular Biology*. Dordrecht, London: Humana Press.

Berger, M., Yonas, A. B., & Lloreda, M. (2003). *The Second Story Wholesale Microfinance in Latin America*. Washington, D. C. Inter-American Development Bank

Bizagi. (2019). *Bizagi Modeler*. Matrid, España. Recuperado de https://www.bizagi.com/es/plataforma/modeler

Breiman, L. (2001). Randomforest. *Machine Learning*, (45), 5-32. doi: 10.1017/CBO9781107415324.004

Brownlee, J. (2017). *How to One Hot Encode Sequence Data in Python*. Australia: Machine Learning Mastery Pty. Ltd. Recuperado de https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

python/

Buczak, A., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, (18), 1153-1176.

Chakraborty, C., & Joseph, A. (2017). Machine Learning at Central Banks. *SSRN Electronic Journal*, (674), 13-85. doi: 10.2139/ssrn.3031796

Chavez, E. (2017). *Propuesta de mejora del proceso de créditos y cobranzas para optimizar la liquidez en la empresa hellmann worldwide logistics S.A.C., 2017*. Universidad San Ignacio de Loyola Lima, Perú.

Choong, A., & Lee, N. (2017). Evaluation of Convolutionary Neural Networks Modeling of DNA Sequences using Ordinal versus one-hot Encoding Method. BioRxiv. doi: 10.1101/186965

Superintendencia de Banca y Seguros - SBS (2008). *Resolución SBS N° 11356 - 3028*. Lima - Perú Recuperado de: https://www.sbs.gob.pe/Portals/0/jer/pfrpv_normatividad/20160719_Res-11356-2008.pdf

Caja Rural de Ahorro y Credito los Andes - CRACLASA (2019). *Crédito Agropecuario | Caja Rural de Crédito y Ahorro los Andes*. Puno, Perú. Recuperado de http://cajarurallosandes.com/web/category/creditos/credito-pecuario/

Instituto Nacional del Estadistica e Informática - INEI (2019). *Población y vivienda: Magnitud y Crecimiento Poblacional*. Lima, Perú. Recuperado de https://www.inei.gob.pe/estadisticas/indice-tematico/poblacion-y-vivienda/

Deisenroth, M., Faisal, A., & Soon, C. (2019). *Mathematics for Machine Learning*. Cambridge, United Kingdom: Cambridge University Press.

Delfiner, M., Pailhe, C., & Peron, S. (2006). Microfinanzas: Un análisis de experiencias y alternativas de regulacion. *Mpra*, *497*(15211), 1–45. Recuperado de http://mpra.ub.uni-muenchen.de/ 497/

Developers, S.-L. (2016). *Developer's Guide*. Recuperado de https://scikit-learn.org/stable/developers/index.html

Doulah, S., & Alam, A. (2018). *Ecological Data Analysis Based on Machine Learning Algorithms*. Bangladés Hajee Mohammad Danesh Science and Technology

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

University

Flores, R., & Ramon, J. (2014). Modelling credit risk with scarce default data: on the suitability of cooperative bootstrapped strategies for small low-default portfolios. *Journal of the Operational Research Society*, (65), 416 - 434

Goodfellow, I., Bengio, Y., & Courville Aaron. (2019). *Deep Learning. In Data Science*. Cabridge, Estados Unidos: MIT Press, doi:10.1016/b978-0-12-814761-0.00010-1

Guajardo, J. (1991). *Estrategias y técnicas para optimizar el crédito y la cobranza*. Universidad Autonoma de Nuevo León. Monterrey, Mexico.

Gyorfi, L., Ottucsak, G., & Walk, H. (2012). *Machine Learning for Financial Engineering*. Covent Garden, London: Imperial College Press.

Hossin, M., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, (5), 1–11. doi: 10.5121/ijdkp.2015.5201

Huang, G., Huang, G.-B., Song, S., & You, K. (2015). Trends in extreme learning machines: A review. *Neural Networks*, (61), 32-48.

Jarrow, R. (2009). Credit Risk Models. *Annual Review of Financial Economics*, (1), 37 - 68. doi: 10.1146/annurev.financial.050808.114300

Jiménez, N. (2016). *La gestión de la calidad crediticia como alternativa de solución a los problemas de morosidad de la cartera de la micro y pequeña empresa y su efecto en los resultados económicos y financieros de la Caja Municipal de Ahorro y Crédito del Santa - años 2014*. Universidad Católica los Ángeles de Chimbote, Chimbote - Perú.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. In *Science*, (349), 255- 260. doi: 10.1126/science.aaa8415

Kalayci, S., Kamasak, M., & Arslan, S. (2018). Credit risk analysis using machine learning algorithms. *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018*, 1-4, doi: 10.1109/SIU.2018.8404353

Kelleher, J. D., Namee, B. Mac, & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. London, England: MIT Press.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications,*

Repositorio Institucional UNA-PUNO
No olvide citar esta tesis

(37), 6233-6239. doi: 10.1016/j.eswa.2010.02.101

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Informatica (Ljubljana)*, (31), 249- 268.

Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and predictio. *Computational and Structural Biotechnology Journal*, (13), 8-17.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, (40), 5125-5131. doi: 10.1016/j.eswa.2013.03.019

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016). Big data application in education: Dropout prediction in edx MOOCs. *Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016*, 440–443. doi: 10.1109/BigMM.2016.70

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *Forest*, (2), 18 - 22.

Lizarzaburu, E. (2014). Sistema financiero peruano: área de tesorería. *Strategy & Management Business Review*, (5), 33-70.

Luger, G. F. (2013). *Livro Inteligência Artificial - 6ª Edição* São Paulo, Brazil: Pearson Education do Brasil Ltda.

Millán, J., & Caicedo, E. (2018). Modelos para otorgamiento y seguimiento en la gestión de riesgo de crédito. *Revista de Métodos Cuantitativos Para La Economía y La Empresa*, (25), 23 - 41.

Mitchell, T. M. (1997). *Machine Learning.* Redmond, Estados Unidos: McGraw-Hill.

Morales, J., & Morales, A. (2014). *Crédito y cobranza*. Distrito Federal, Mexico: Grupo Editorial Patria.

Mueller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python*. Estados Unidos: O'Reilly Media, Inc.

Narkhede, S. (2018). *Understanding AUC - ROC Curve*. Recuperado de https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nelli, F. (2018). *Python Data Analytics: With Pandas, NumPy, and Matplotlib*. Apress. doi: 10.1007/978-1-4842-3913-1

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Ng, A. (2018). *Machine Learning Yeraning*. Deeplearning.ai Project. Recuperado de https://www.deeplearning.ai/project/

Nguyen, L. (2016). Tutorial on Support Vector Machine. *Some Novel Algorithms for Global Optimization and Relevant Subjects*, (6).

Ochoa, J., Galeano, W., & Agudelo, L. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de Coyuntura Económica*, *16*, 191–222.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, (12), 2825 - 2830.

Pina, K. (2018). *Matriz de confusión*. Recuperado de https://koldopina.com/matriz-de-confusion/

Raschka, S., & Vahid, M. (2017). *Python Machine Learning Second Edition*. In *Packt Publishing* (2). United Kingdom: Cambridge University Press. doi: 10.1017/CBO9781107415324.004

Rosten, E., & Drummond, T. (2006). *Machine learning for high-speed corner detection*. Computer Vision - ECCV 2016, (3951), 430 - 443. doi: 10.1007/11744023_34

Rouhiainen, L. (2019). *Inteligencia Artificial para empresas*. Recuperado de https://libro.ai/wp-content/uploads/2019/03/Informe_AI_2019_Act.pdf

Sadatrasoul, S. M., Gholamian, M., Siami, M., & Hajimohammadi, Z. (2013). Credit scoring in banks and financial institutions via data mining techniques: A literature review. *Journal of AI and Data Mining Journal of AI and Data Mining*, (1), 119 - 129. doi: 10.22044/jadm.2013.124

Saju, J., & Chacko, J. (2017). Sustainable development of microfinance customers: An empirical investigation based on India. *Journal of Enterprise Information Management*, (30), 1 - 30. doi: 10.1108/EL-01-2017-0019

Sampieri, R., Collado, C., & Baptista, M. (2014). *Metodología de la investigación Sexta Edición*. Mexico: Mc Graw Hill Education.

Shalev, S., & Ben, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, Estados Unidos: Cambridge University Press.

Shanmuganathan, S. (2016). *Artificial Neural Network Modelling: An Introduction*.

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Switzerland: Sprinter International Publishing.

Tack, C. (2018). Artificial intelligence and machine learning applications in musculoskeletal physiotherapy. *Musculoskeletal Science and Practice*, (39), 164 - 169. doi: 10.1016/j.msksp.2018.11.012

Tesén, A. (2017). *Eficacia de los modelos de aprendizaje de maquina para evaluar el riesgo crediticio de personas naturales en una institución financiera de Chiclayo*. Universidad Nacional de Santa, Chimbote - Perú.

Turkson, R. E., Baagyere, E. Y., & Wenya, G. E. (2016). A machine learning approach for predicting bank credit worthiness. *2016 3rd International Conference on Artificial Intelligence and Pattern Recognition, AIPR 2016*, 1 - 7. doi: 10.1109/ICAIPR.2016.7585216

Valencia, A. (2017). *Modelo Scoring para el otorgamiento de crédito de las pymes*. Universidad EAFIT, Medellin, Colombia.

Véliz, C. (2016). *Análisis Multivariantes. Métodos estadísticos multivariantes para la Investigación*. Mexico: Cengage Learning Editores.

Wang, J., & Tao, Q. (2008). Machine learning: The state of the art. *IEEE Intelligent Systems*, *23*(6), 49–55. doi: 10.1109/MIS.2008.107

Wendel, C., & Harvey, M. (2006). SME credit scoring : key initiatives, opportunities, and issues. *Financial Sector Vice Presidency*, (10), 1 - 6.

Zhao, H. (2018). Inteligencia artificial para el bien en el mundo. *ITU News Magazine* (1), 2- 5.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, (5), 44-53. doi: 10.1093/nsr/nwx106

Repositorio Institucional UNA-PUNO

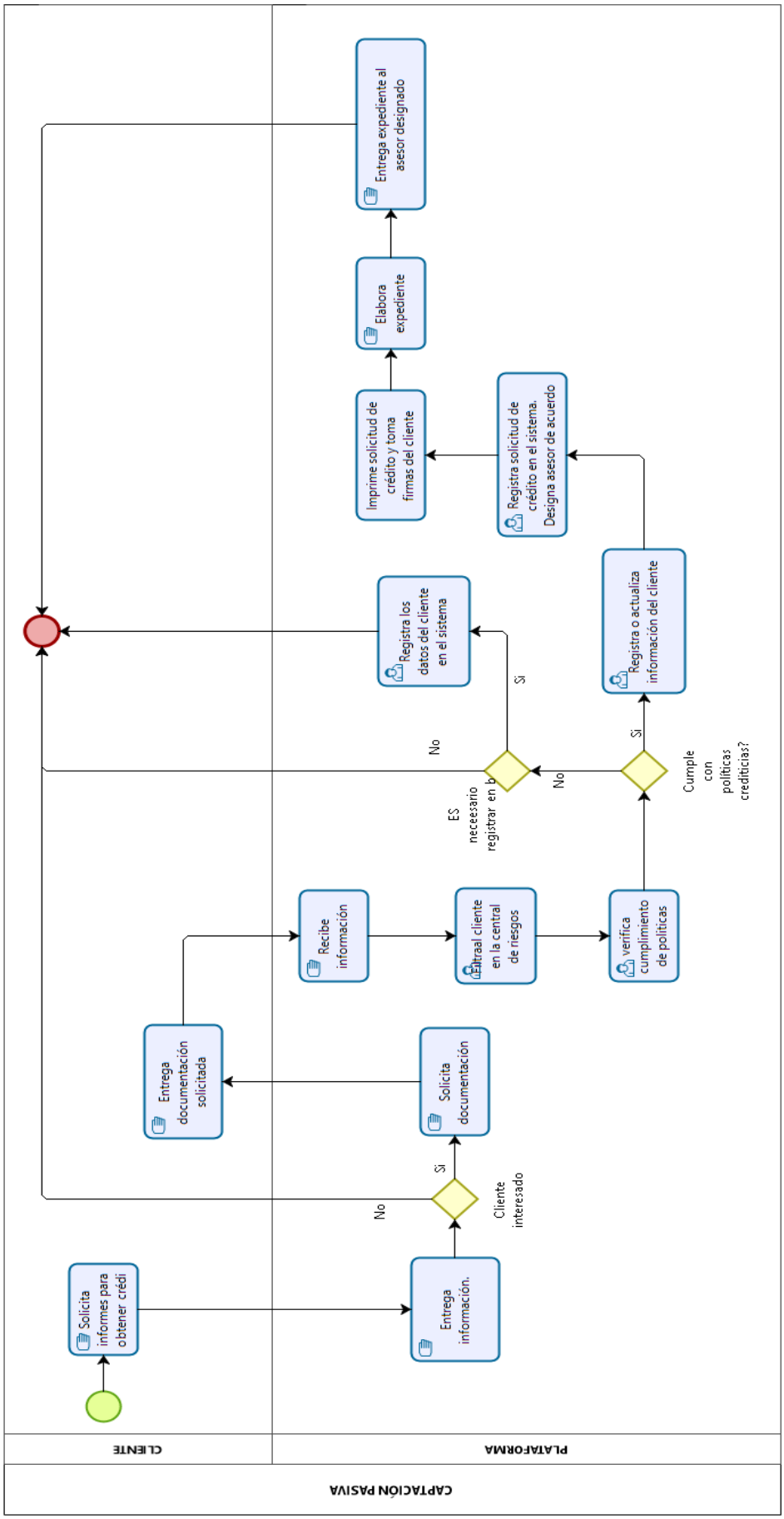
No olvide citar esta tesis

ANEXOS

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

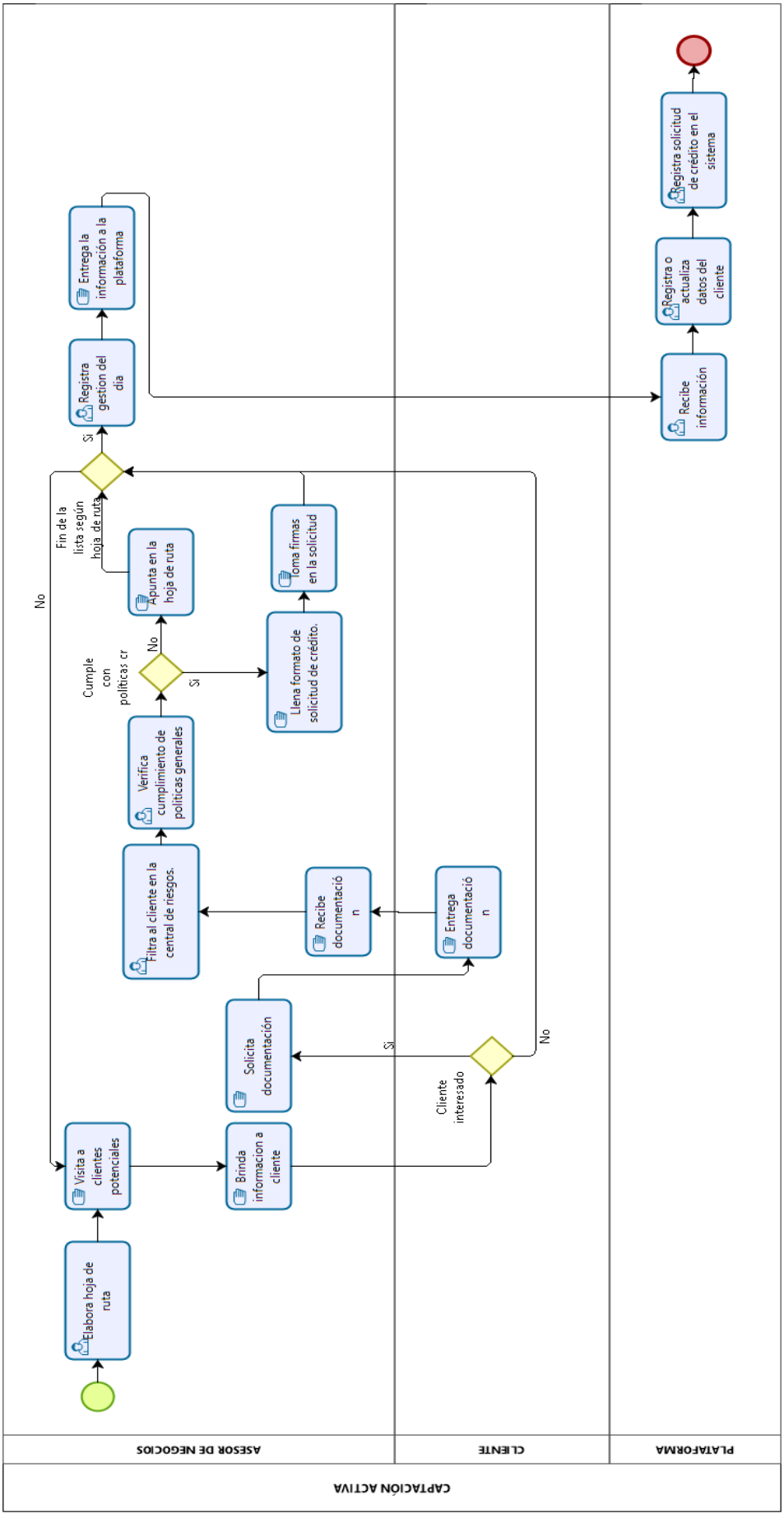
Anexo 1: Proceso de captación pasiva



Repositorio Institucional UNA-PUNO

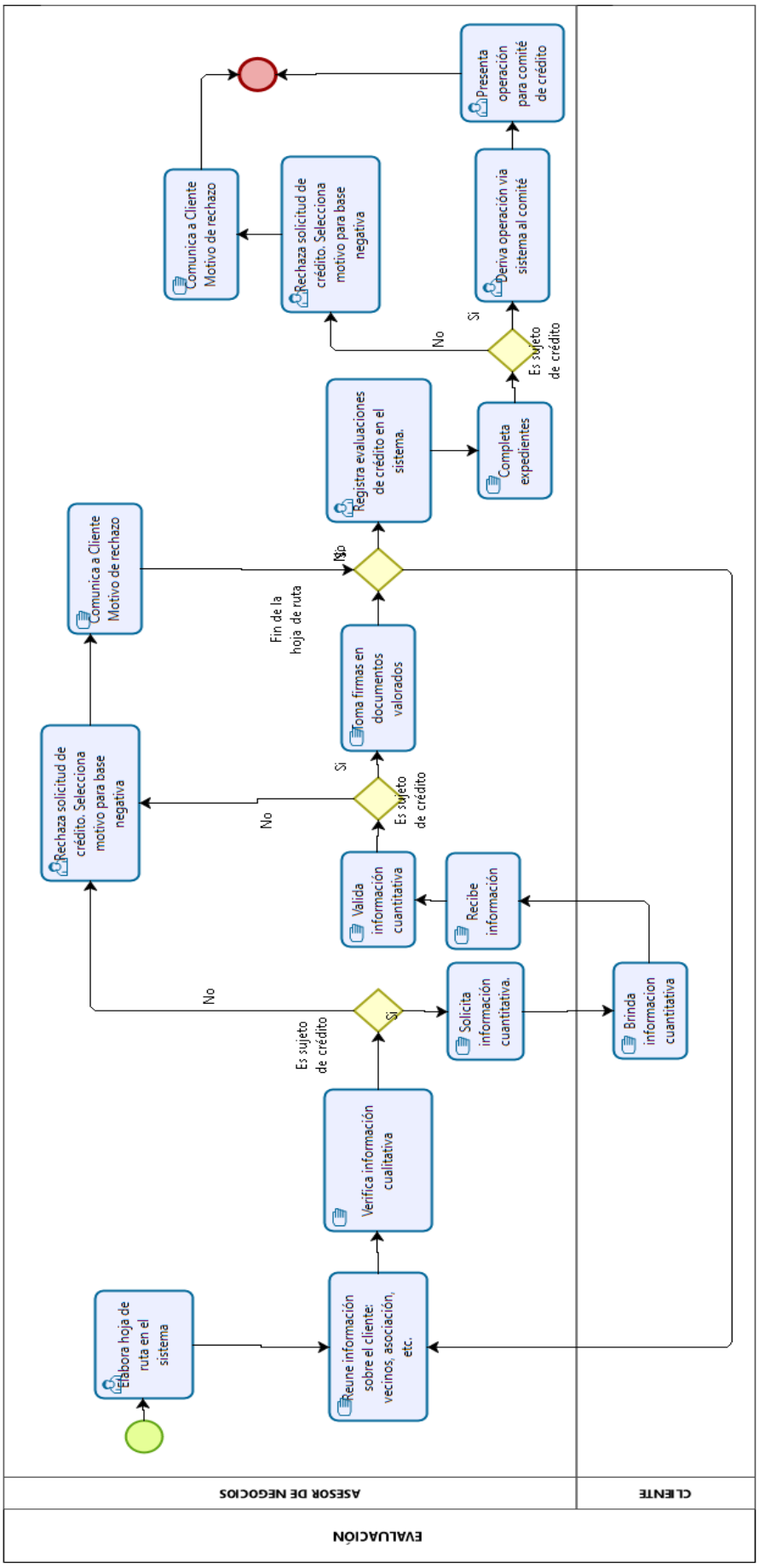
No olvide citar esta tesis

Anexo 2: Proceso de captación activa



TESIS EPG UNA - PUNO

Universidad Nacional del Altiplano



Anexo 3: Proceso de evaluación de créditos

Repositorio Institucional UNA-PUNO

No olvide citar esta tesis

Anexo 4: Proceso de aprobación de créditos